# Mixed Model Methods for Genetic Analysis

Classnotes for AnSc 8141
(3 credits)

Yang Da
Department of Animal Science
University of Minnesota
Fall 2023

# CHAPTER 1: THEORETICAL FOUNDATION

The theoretical foundation of mixed model methods for genetic analysis including the following:

1) Partition of phenotypic values and variance,
2) Partition of genotypic values and variance,
3) Additive and dominance relationships and coancestry,
4) Covariance between relatives.

The partition of phenotypic values and variance provides the basic statistical model for genetic analysis.

The partition of genotypic values and variance provides a quantitative genetics approach to test and detect genetic effects, to estimate heritability using pedigree and genomic information, to define genomic relationships, and to implement genomic prediction.

Additive and dominance relationships among individuals typically are required for mixed models assuming additive and dominance effects, and are the theoretical and computational components of covariance between relatives. Relationships among individuals can be pedigree or genomic relationships. Additive models are most widely used models.

Covariance between relatives is required for mixed models and combines the results of relationship between individuals and the partition of the genetic variance.

The concepts of Hardy-Weinberg equilibrium (HWE), Hardy-Weinberg disequilibrium (HWD), linkage equilibrium (LE), linkage disequilibrium (LD), identify by descent (IBD) and identify by state (IBS) are often involved, and such concepts are described in Appendices.

## 1.1 Partition of Phenotypic Values and Variance

The phenotypic value of a trait or phenotype for an individual is assumed to be the sum of a genotypic value of the individual and a random residual that is not explained by the individual's genotype. Let:

y = the phenotypic value of a trait or phenotype for an individual,
g = the genetic value of a trait or phenotype for the individual,
e = the random residual of a trait or phenotype for the individual,
$\mathrm{Var}(y) = \sigma_y^2$ = variance of y in the population = phenotypic variance,
$\mathrm{Var}(g) = \sigma_g^2$ = variance of g in the population = genetic variance,
$\mathrm{Var}(e) = \sigma_e^2$ = variance of e in the population = residual variance.

Then, the phenotypic values and variance are partitioned as:

$$y = g + e \qquad [1.1.1]$$
$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2 \qquad [1.1.2]$$

The partition of the phenotypic variance by Equation [1.1.2] assumes no covariance between the genotypic values and the residuals in the population.

## 1.2 Single-locus Partition of Genotypic Values and Variance

The partion of the genotypic value (g) in Equation [1.1.1] and the partition of the genetic variance ( $\sigma_g^2$ ) in Equation [1.1.2] is the foundation of quantitative genetics (QG), and these partitions are a process of 'genetic partition', which provides a QG approach to model genetic effects that could include various types of genetic effects. Genetic partition originated from Fisher (1918). Future models based on the genetic partion for testing genetic effects and for genomic estimation and prediction will be referred to as the 'QG models'. This section summarizes the main results of the genetic partion assuming additive and dominance effects. Derivation details are provided in Appendix 1.

The main objective of this section is to define additive effect, additive value, additive variance, dominance effect, dominance value (dominance deviation), dominance variance, and genotypic variance.

**Partiton of genetic values**

A bi-allelic locus with HWE is assumed to affect the quantitative trait, with alleles $A_1$ and $A_2$, and allele frequencies $p(A_1) = p$, and $p(A_2) = q$.

**Table 1.2.1** Calculation of population mean and average effect ($N = N_{11} + N_{12} + N_{22}$)

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Number of individuals | $N_{11}$ | $N_{12}$ | $N_{22}$ |
| Genotypic frequency: general expression | $P_{11} = N_{11}/N$ | $P_{12} = N_{12}/N$ | $P_{22} = N_{22}/N$ |
| Genotypic frequency under HWE | $p^2$ | $2pq$ | $q^2$ |
| Number of $A_1$ | 2 | 1 | 0 |
| Number of $A_2$ | 0 | 1 | 2 |
| Genotypic value | $g_{11}$ | $g_{12}$ | $g_{22}$ |

From the aboe table, the population mean ($\mu$), allelic mean ( $\mu_i$, $i = 1, 2$ ), average effect or allelic effect ( $a_2$ ), and the additive effect or average effect of gene substitution ($\mu$) are:

$$\mu = p^2 g_{11} + 2pq\, g_{12} + q^2 g_{22} \quad [1.2.1]$$

$$\mu_1 = pg_{11} + qg_{12} \quad [1.2.2]$$

$$\mu_2 = pg_{12} + qg_{22} \quad [1.2.3]$$

$$a_1 = \mu_1 - \mu = q[p(g_{11} - g_{12}) + q(g_{12} - g_{22})] \quad [1.2.4]$$

$$a_2 = \mu_2 - \mu = -p[p(g_{11} - g_{12}) + q(g_{12} - g_{22})] \quad [1.2.5]$$

$$\alpha = a_1 - a_2 = \mu_1 - \mu_2 = p(g_{11} - g_{12}) + q(g_{12} - g_{22}) \quad [1.2.6]$$

$$= p\,g_{11} + (q - p)\,g_{12} - q\,g_{22} \quad [1.2.7]$$

$$= \begin{bmatrix} p & q-p & -q \end{bmatrix}\begin{bmatrix} g_{11} & g_{12} & g_{22} \end{bmatrix}' = \mathbf{c}_a'\mathbf{g} \quad [1.2.8]$$

where

$$\mathbf{c}_a' = \begin{bmatrix} p & q-p & -q \end{bmatrix} = \text{row vector of additive contrast coefficients} \quad [1.2.9]$$

$$\mathbf{g} = \begin{bmatrix} g_{11} & g_{12} & g_{22} \end{bmatrix}' = \text{column vector of genotypic values} \quad [1.2.10]$$

Equations [1.2.7]-[1.2.10] show that additive effect (α) is a contrast of the three genotypic values, because the coefficients of the three genotypic values add to '0'. Therefore, Equation [1.2.7] is the 'additive contrast' of the three genotypic values and provides a method for testing the statistical significance of the additive effect by testing the significance of the additive contrast.

From Equations [1.2.4]-[1.2.6],

$$a_1 = \mu_1 - \mu = q\alpha \quad [1.2.11]$$

$$a_2 = \mu_2 - \mu = -p\alpha \quad [1.2.12]$$

The additive value (breeding value) of each genotype ($a_{ij}, i, j = 1, 2$) is defined as sum of the allelic effects of the genotype:

$$a_{11} = 2a_1 = 2q\alpha \quad \text{for } A_1A_1 \quad [1.2.13]$$

$$a_{12} = a_1 + a_2 = (q - p)\alpha \quad \text{for } A_1A_2 \quad [1.2.14]$$

$$a_{22} = 2a_2 = -2p\alpha \quad \text{for } A_1A_2 \quad [1.2.15]$$

Equations [1.2.13]-[1.2.15] are the theoretical foundation of the QG model for defining genomic additive relationships, genomic estimation of additive heritability, and genomic prediction of additive values.

Dominance value or dominance deviation ($d_{ij}, i, j = 1, 2$) is the deviation of a genotypic value from its mean and additive value:

$$d_{11} = g_{11} - \mu - 2a_1 = -2q^2[g_{12} - \tfrac{1}{2}(g_{11} + g_{22})] \quad \text{for } A_1A_1 \quad [1.2.16]$$

$$d_{12} = g_{12} - \mu - (a_1 + a_2) = 2pq[g_{12} - \tfrac{1}{2}(g_{11} + g_{22})] \quad \text{for } A_1A_2 \quad [1.2.17]$$

$$d_{22} = g_{22} - \mu - 2a_2 = -2p^2[g_{12} - \tfrac{1}{2}(g_{11} + g_{22})] \quad \text{for } A_2A_2 \quad [1.2.18]$$

Dominance effect is defined as the difference between the dominance value of the heterozygous genotype ($A_1A_2$) and the average of the dominance values of the homozygous

genotypes ($A_1A_1$ and $A_2A_2$):

$$\delta = d_{12} - \tfrac{1}{2}(d_{11} + d_{22}) = g_{12} - \tfrac{1}{2}(g_{11} + g_{22}) \quad [1.2.19]$$

$$= \begin{bmatrix} -\tfrac{1}{2} & 1 & -\tfrac{1}{2} \end{bmatrix} \begin{bmatrix} g_{11} & g_{12} & g_{22} \end{bmatrix}' = \mathbf{c}_d' \mathbf{g} \quad [1.2.20]$$

where

$$\mathbf{c}_d' = \begin{bmatrix} -\tfrac{1}{2} & 1 & -\tfrac{1}{2} \end{bmatrix} = \text{row vector of contrast coefficients} \quad [1.2.21]$$

Equations [1.2.19]-[1.2.21] show that dominance effect (δ) is a contrast of the three genotypic values, because the coefficients of the three genotypic values add to '0'. Therefore, Equation [1.2.19] or [1.2.20] is the 'dominance contrast' of the three genotypic values and provides a method for testing the statistical significance of the dominance effect by testing the significance of the dominance contrast.

From Equations [1.2.16]-[1.2.19],

$$d_{11} = -2q^2 \delta \quad \text{for } A_1A_1 \quad [1.2.22]$$

$$d_{12} = 2pq \, \delta \quad \text{for } A_1A_2 \quad [1.2.23]$$

$$d_{22} = -2p^2 \delta \quad \text{for } A_2A_2 \quad [1.2.24]$$

Equations [1.2.22]-[1.2.24] are the theoretical foundation of the QG model for defining genomic dominance relationships, genomic estimation of dominance heritability, and genomic prediction of dominance values.

With the additive values of Equations [1.2.13]-[1.2.15] and the dominance values of Equations [1.2.22]-[1.2.24], the genotypic values of Equation [1.2.10] are partitioned as:

$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} + \begin{bmatrix} a_{11} \\ a_{12} \\ a_{22} \end{bmatrix} + \begin{bmatrix} d_{11} \\ d_{12} \\ d_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 2q \\ q-p \\ -2p \end{bmatrix} \alpha + \begin{bmatrix} -2q^2 \\ 2pq \\ -2p^2 \end{bmatrix} \delta \quad [1.2.25]$$

or,

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{w}_\alpha \alpha + \mathbf{w}_\delta \delta = \mathbf{1}\mu + \mathbf{a} + \mathbf{d} \quad [1.2.26]$$

where $\mathbf{1} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}'$ = 3×1 column vector of 1's , and

$$\mathbf{w}_\alpha = \begin{bmatrix} 2q \\ q-p \\ -2p \end{bmatrix} = \text{model matrix of } \alpha \quad [1.2.27]$$

$$\mathbf{w}_\delta = \begin{bmatrix} -2q^2 \\ 2pq \\ -2p^2 \end{bmatrix} = \text{model matrix of } \delta \quad [1.2.28]$$

$$\mathbf{a} = \mathbf{w}_\alpha \alpha = \text{additive values} \quad [1.2.29]$$

$$\mathbf{d} = \mathbf{w}_\delta \delta \text{ = dominance values} \quad [1.2.30]$$

Equations [1.2.25]-[1.2.30] are the theoretical foundation of the QG model with additive and dominance effects for genomic estimation and prediction.

**Partiton of genetic variance**

Additive variance ($\sigma_a^2$) is the variance of additive values defined by Equations [1.2.13]-[1.2.15], dominance variance ($\sigma_d^2$) is the variance of dominance values defined by Equations [1.2.22]-[1.2.24], and genotypic variance ($\sigma_g^2$) is the variance of genotypic values defined in Table 1.2.1. These variances are:

$$\sigma_a^2 = 2pq\alpha^2 \quad [1.2.31]$$

$$\sigma_d^2 = 4p^2q^2\delta^2 \quad [1.2.32]$$

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 \quad [1.2.33]$$

With the partition of genotypic values into additive and dominance values, and the partition of genotypic variance into additive and dominance variances, the phenotypc values and variances can be expressed as:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{22} \end{bmatrix} = \begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ g_{22} \end{bmatrix}$$
$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 2q \\ q-p \\ -2p \end{bmatrix} \alpha + \begin{bmatrix} -2q^2 \\ 2pq \\ -2p^2 \end{bmatrix} \delta + \begin{bmatrix} e_{11} \\ e_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} a_{11} \\ a_{12} \\ a_{22} \end{bmatrix} + \begin{bmatrix} d_{11} \\ d_{12} \\ d_{22} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ g_{22} \end{bmatrix} \quad [1.2.34]$$

or,

$$\mathbf{y} = \mathbf{g} + \mathbf{e} = \mathbf{1}\mu + \mathbf{w}_\alpha \alpha + \mathbf{w}_\delta \delta + \mathbf{e} = \mathbf{1}\mu + \mathbf{a} + \mathbf{d} + \mathbf{e} \quad [1.2.35]$$

With the partition of genotypic variance into additive and dominance variances, the phenotypc variances can be expressed as:

$$\sigma_y^2 = \text{var}(y) = \text{var}(G) + \text{var}(e) = \sigma_a^2 + \sigma_d^2 + \sigma_e^2 \quad [1.2.36]$$

The partion of phenotypic variance by Equation [1.2.36] provides variance components for mixed models and for variance-covariance matrices among individuals.

## 1.3 Genetic Values and Variances of Multiple Loci

The genetic partition leading to Equation [1.2.36] assumes a single bi-allelic locus. For multiple loci such as the genome-wide single nucleotide polymorphism (SNP) markers, the total genetic value is assumed to be the sum of the genetic values of all loci and the total genetic variance is assumed to be the sum of the genetic variances of all loci (VanRaden, 2008). Based on these assumptions, the genetic values and variances of multiple loci are:

$$\mathbf{a} = \sum_{i=1}^{m} \mathbf{a}_i \text{ = total additive value of m loci} \quad [1.3.1]$$

$$\mathbf{d} = \sum_{i=1}^{m} \mathbf{d}_i \text{ = total dominance value of m loci} \quad [1.3.2]$$

$$\mathbf{g} = \sum_{i=1}^{m} \mathbf{g}_i \text{ = total genetic value of m loci} \quad [1.3.3]$$

$$\sigma_a^2 = \sum_{i=1}^{m} \sigma_{ai}^2 \text{ = total additive variance of m loci} \quad [1.3.4]$$

$$\sigma_d^2 = \sum_{i=1}^{m} \sigma_{di}^2 \text{ = total dominance variance of m loci} \quad [1.3.5]$$

$$\sigma_g^2 = \sum_{i=1}^{m} \sigma_{ai}^2 + \sum_{i=1}^{m} \sigma_{di}^2 = \sum_{i=1}^{m} (\sigma_{ai}^2 + \sigma_{di}^2) \text{ = total genetic variance of m loci} \quad [1.3.6]$$

These results have been used in genomic prediction of genetic values and genomic estimation of heritabilities using genome-wide SNP markers.

## 1.4 Genetic Partion for Multiallelic Loci

The genetic partition of a multiallelic locus has the same general expressions of the genetic values and variances as those for a bi-allelic locus (Da, 2015):

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{w}_{\alpha h}\boldsymbol{\alpha} + \mathbf{w}_{\delta h}\boldsymbol{\delta} = \mathbf{1}\mu + \mathbf{a} + \mathbf{d} \quad [1.4.1]$$

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 \quad [1.4.2]$$

The main differences between multiallelic and biallelic loci are in the calculations of the $\mathbf{w}_{\alpha h}$ and $\mathbf{w}_{\delta h}$ of Equation [1.4.1], which are more tedious than the calculations of the $\mathbf{w}_{\alpha h}$ and $\mathbf{w}_{\delta h}$ matrices of Equation [1.2.26]. The $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ vectors each is a $(h-1)\times 1$ column vector rather than a single parameter in Equation [1.2.26], where h = number of alleles. Although Equation [1.4.2] used the same notations as Equation [1.2.32], the calculations of the additive and dominance variances in Equations [1.4.2] are different from the calculations of additive and dominance variances of Equations [1.2.31] and [1.2.32].

## 1.5 Genetic Partion of Two-locus Genotypic Values

For a two-locus genotype, Locus 1 and Locus 2, interactions between these two loci may exist, and these interactions can be partitioned into additive × additive (A×A), additive × dominance (A×D) and dominance × dominance (A×D) epistasis values (Kempthorne, 1954; Cockerham 1954;

Henderson, 1985). The two-locus genotypic values and variance in a population of n individuals are partitioned as:

$$\begin{aligned}\mathbf{g} &= \mathbf{1}\mu + \mathbf{w}_{\alpha}\alpha + \mathbf{w}_{\delta}\delta + \mathbf{w}_{\alpha\alpha}(\alpha\alpha) + \mathbf{w}_{\alpha\delta}(\alpha\delta) + \mathbf{w}_{\delta\delta}(\delta\delta) \\ &= \mathbf{1}\mu + \mathbf{a} + \mathbf{d} + \mathbf{aa} + \mathbf{ad} + \mathbf{dd}\end{aligned} \quad [1.5.1]$$

$$\sigma^2_g = \sigma^2_a + \sigma^2_d + \sigma^2_{aa} + \sigma^2_{ad} + \sigma^2_{dd} \quad [1.5.2]$$

where $\alpha\alpha$ = A×A effect with the interpretation of allele × allele interaction, $\mathbf{w}_{\alpha\alpha}$ = model matrix of $\alpha\alpha$, $\alpha\delta$ = A×D effect with the interpretation of allele × genotype or genotype × allele interaction, $\mathbf{w}_{\alpha\delta}$ = model matrix of $\alpha\delta$, $\delta\delta$ = D×D effect with the interpretation of genotype × genotype interaction, $\mathbf{w}_{\delta\delta}$ = model matrix of $\delta\delta$, $\mathbf{aa}$ = A×A values, $\mathbf{ad}$ = A×D values, $\mathbf{dd}$ = D×D values, $\sigma^2_{aa}$ = A×A variance, $\sigma^2_{ad}$ = A×D variance, $\sigma^2_{dd}$ = D×D variance.

## 1.6 Covariance between Relatives

**Covariance between genetic values with additive and dominance values**

Mixed models typically use variance-covariance matrices among individuals based on genetic relationships among individuals. Let $y_j$ = the phenotypic value for individual j, $y_k$ = the phenotypic value for individual k. Then, based on the phenotypic model of Equation [1.1.1], the covariance between individuals j and k is:

$$\begin{aligned}\text{cov}(y_j, y_k) &= \text{cov}(g_j + e_j, g_k + e_k) \\ &= \text{cov}(g_j, g_k) + \text{cov}(g_j, e_k) + \text{cov}(e_j, g_k) + \text{cov}(e_j, e_k) \\ &= \text{cov}(g_j, g_k)\end{aligned} \quad [1.6.1]$$

Equation [1.6.1] assumes no covariance between genetic values and random residuals or between random residuals, i.e., + $\text{cov}(g_j, e_k) = \text{cov}(e_j, g_k) = \text{cov}(e_j, e_k) = 0$. The genetic covariance between two individuals (Cockerham, 1954; Henderson, 1985; Falconer and Mackay, 1996) is:

$$\text{cov}(g_j, g_k) = \sigma^2_a A_{jk} + \sigma^2_d D_{jk} \quad [1.6.2]$$

where $A_{jk}$ = numerator additive relationship between j and k, $D_{jk}$ = dominance relationship between j and j, Ignoring dominance, Equation [1.6.2] reduces to:

$$\text{cov}(g_j, g_k) = \sigma^2_a A_{jk} \quad [1.6.3]$$

Equation [1.6.3] is the additive model that is most widely used in genetic studies of quantitative traits.

## Pedigree additive relationship and coancestry

Additive relationship between individuals j and k ($A_{jk}$) based on pedigree information is calculated as twice the coefficient of coancestry coefficient ($f_{jk}$) (Falconer and Mackay, 1996):

$$A_{jk} = 2f_{jk} \quad [1.6.3]$$

The coefficient of coancestry or kinship coefficient between two individuals is the IBD probability of two randomly sampled alleles each from one individual:

$$f_{jk} = \text{conancestry coefficient} = \text{IBD probability between individuals j and k} \quad [1.6.4]$$

The inbreeding coefficient (f) of the child is the coancestry coefficient of parents j and k. Therefore,

$$f = f_{jk} = A_{jk}/2 \quad [1.6.5]$$

Note that

$$A_{jk} = 1 + f_{jk} \quad \text{if } j = k \quad [1.6.6]$$
$$= 0 \quad \text{if j and k are unrelated.} \quad [1.6.7]$$

Equations [1.6.6] and [1.6.7] set the [0, 2] boundary values of additive relationship. The upper limt is '2' for the additive relationship of an individual with oneself if the coancestry is 100%, and the lower limit is '0' for unrelated individuals. Additive relationship or coancestry is assumed to be '0' if any of the two individuals is unknown although some individuals without pedigree information could be related. The requirement of pedigree information is a weakness of pedigree relationship. In contrast, genomic relationships to be covered in later chapters do not require pedigree information and could compensate the weakness of pedigree relationships.

Coancestry coefficient and its corresponding additive relationship can be calculated using parental and parent-offspring information. Let the parents of individual j be denoted by 1 and 2, and the parents of individual k be denoted by 3 and 4. Then,

$$f_{jk} = (f_{13} + f_{14} + f_{23} + f_{24})/4 \quad [1.6.8]$$
$$= (f_{j3} + f_{j4})/2 \quad \text{if j is not younger than k} \quad [1.6.9]$$

$$A_{jk} = (A_{13} + A_{14} + A_{23} + A_{24})/4 \quad [1.6.10]$$
$$= (f_{j3} + f_{j4})/2 \quad \text{if j is not younger than k} \quad [1.6.11]$$

Based on Equations [1.6.8] and [1.6.9], the tabular method (Tier, 1990) is efficient for calculating coancestry coefficients for large and complex pedigrees.

**Dominance relationship**

Dominance relationship is the probability that the two genotypes of individual j and k are identical by descent. Let $A_1A_2$ = the genotype of individual j, and $A_3A_4$ = the genotype of individual k. Then, the dominance relationship between individuals j and k is:

$$D_{jk} = p(A_1A_2 = A_3A_4) \qquad [1.6.12]$$

If j equals to k, $D_{jk} = p(A_1A_2 = A_3A_4) = 1$. For $A_1A_2$ to be equal to $A_3A_4$, one of the following two events must be true:

$A_1 = A_3$ and $A_2 = A_4$, or $A_1 = A_4$ and $A_2 = A_3$.

Therefore,

$$\begin{aligned} D_{jk} &= p(A_1A_2 = A_3A_4) = p(A_1 = A_3 \text{ and } A_2 = A_4) + p(A_1 = A_4 \text{ and } A_2 = A_3) \\ &= p(A_1 = A_3)p(A_2 = A_4) + p(A_1 = A_4)p(A_2 = A_3) \quad \text{(assuming no inbreeding)} \\ &= f_{ik}f_{jl} + f_{il}f_{jk} \qquad [1.6.14] \\ &= \left[(2f_{ik})(2f_{jl}) + (2f_{il})(2f_{jk})\right]/4 \\ &= (A_{ik}A_{jl} + A_{il}A_{jk})/4 \qquad [1.6.15] \end{aligned}$$

Dominance relationship of an individual with oneself is 1, i.e., $D_{ii} = 1$.

**Example 1.6.1**: A pedigree for calculation of additive and dominance relationships:

| Individual | sire | dam |
|---|---|---|
| 1 | unknown | unknown |
| 2 | unknown | unknown |
| 3 | 1 | 2 |
| 4 | 1 | 2 |
| 5 | 1 | unknown |

Additive relationship between individuals 1 and 2 is zero, because individuals 1 and 2 do not have common parents, i.e., $A_{12} = 0$. Additive relationship of any of the four individuals with itself is equal to 1, i.e., $A_{ii} = 1 + f_i = 1$, knowing $f_i = 0$, for i = 1 to 5. Additive relationships among different individuals are:

$A_{12} = A_{25} =$ **0 = additive relationship between unrelated individuals**

$A_{13} = A_{14} = (A_{11} + A_{12})/2 = (1 + 0)/2 =$ **½ = parent-offspring additive relationship**

$A_{23} = A_{24} = (A_{22} + A_{12})/2 = (1 + 0)/2 = ½ =$ parent-offspring additive relationship

$A_{15} = (A_{11} + A_{1,unknown})/2 = (1 + 0)/2 = ½ =$ parent-offspring additive relationship

$A_{34} = (A_{11} + A_{12} + A_{22} + A_{12})/4 = (1 + 0 + 1 + 0)/4 =$ **½ = full-sib additive relationship**

$A_{35} = A_{45} = (A_{11} + A_{1,unknown} + A_{12} + A_{2,unknown})/4$

$= (1 + 0 + 0 + 0)/4 =$ **¼ = half-sib additive relationship**

Dominance relationships among different individuals are:

$D_{12} = D_{25} =$ **0 = dominance relationship between unrelated individuals**

$D_{13} = D_{14} = D_{23} = D_{24} = D_{15} =$ **0 = parent-offspring dominance relationship**

$D_{34} = (A_{11}A_{22} + A_{12}A_{12})/4 =$ ¼ **= full-sib dominance relationship**

$D_{35} = D_{45} = (A_{11}\, A_{2,unknown} + A_{12}\, A_{2,unknown})/4 =$ **0 = half-sib dominance relationship**

The above numerical results have some examples of typical pedigree relationships:

| | | |
|---|---|---|
| $A_{jk} = 0$ | for unrelated and unklnown individuals | [1.6.16] |
| $= 0.25$ | for half sibs | [1.6.17] |
| $= 0.5$ | for parent-offspring and full sibs | [1.6.18] |
| $f_{jk} = 0$ | for unrelated and unklnown individuals | [1.6.19] |
| $= 0.125$ | for half sibs | [1.6.20] |
| $= 0.25$ | for parent-offspring and full sibs | [1.6.21] |
| $D_{jk} = 0$ | for unrelated individuals, parent offspring and half sibs | [1.6.22] |
| $= 0.25$ | for full sibs | [1.6.23] |

These expected pedigree relationships provide important standards for interpreting genomic relationships.

## 1.7 Matrix Notations of Mathematical Expectation, Variance and Covariance Matrices

Let $\mathbf{X}_{n\times p}$ and $\mathbf{Y}_{n\times p}$ be random matrices, and let $\mathbf{A}_{s\times n}$ and $\mathbf{B}_{p\times q}$ be conformable matrices of constants. Then, the following formulations are useful results of mathematical expectations and variance-covariance matrices:

$$E(\mathbf{X} + \mathbf{Z}) = E(\mathbf{X}) + E(\mathbf{Z}) \quad [1.7.1]$$

$$E(\mathbf{By}) = \mathbf{B}E(\mathbf{y}) \quad [1.7.2]$$

$$E\left(\mathbf{A}_{s\times n}\mathbf{X}_{n\times p}\mathbf{B}_{p\times q}\right) = \mathbf{A}E(\mathbf{X})\mathbf{B} \quad [1.7.3]$$

$$\mathrm{cov}\left(\mathbf{x}_{n\times 1}, \mathbf{y}_{m\times 1}\right) = \mathbf{C}_{n\times m} \quad \text{(Mardia et al., 1979)} \quad [1.7.4]$$

or,

$$\mathrm{cov}\left(\mathbf{x}_{n\times 1}, \mathbf{y}_{m\times 1}'\right) = \mathbf{C}_{n\times m} \quad \text{(Searle et al., 1992)} \quad [1.7.5]$$

$$\mathrm{cov}\left(\mathbf{x}_{n\times 1} + \mathbf{z}_{n\times 1}, \mathbf{y}_{m\times 1}'\right) = \mathrm{cov}\left(\mathbf{x}_{n\times 1}, \mathbf{y}_{m\times 1}'\right) + \mathrm{cov}\left(\mathbf{z}_{n\times 1}, \mathbf{y}_{m\times 1}'\right) \quad [1.7.6]$$

$$\text{var}(c\mathbf{y}) = c^2\left[\text{var}(\mathbf{y})\right] \quad \text{with var}(\mathbf{y}) = \mathbf{V},\ c = \text{constant} \qquad [1.7.7]$$

$$\text{var}(\mathbf{By}) = \mathbf{B}\left[\text{var}(\mathbf{y})\right]\mathbf{B}' = \mathbf{BVB}' \quad \text{with var}(\mathbf{y}) = \mathbf{V} \qquad [1.7.8]$$

$$\text{var}(\mathbf{a}+\mathbf{d}) = \text{var}(\mathbf{a}) + \text{cov}(\mathbf{a},\mathbf{d}') + \text{cov}(\mathbf{d},\mathbf{a}') + \text{var}(\mathbf{d}) \qquad [1.7.9]$$

## 1.8 Variance-covariance Matrix of Genetic Values

**Vairiance-covarice matrices of additive and dominance values**

Covariance matrix of genetic values contains variances and covariances of genetic values for all individuals in the sample. Assuming n individuals, the genotypic values of Equation [1.2.26] after omitting the μ term is:

$$\mathbf{g} = \mathbf{a} + \mathbf{d} \qquad [1.8.1]$$

where $\mathbf{g} = (g_1, g_2, ..., g_n)'$ = n×1 vector of genetic values of n individuals, $\mathbf{a}$ = n×1 vector of additive effects, and $\mathbf{d}$ = n×1 vector of dominance effects. Assume no inbreeding and no covariance between additive and dominance effects, the variance-covariance matrix of $\mathbf{g}$ is

$$\mathbf{G} = \text{var}(\mathbf{g}) = \mathbf{G}_a + \mathbf{G}_d = \sigma_a^2\mathbf{A} + \sigma_d^2\mathbf{D} \qquad [1.8.2]$$

where

$\mathbf{G}$ = n×n variance-covariance matrix of genetic values

$\mathbf{G}_a = \sigma_a^2\mathbf{A}$ = n×n variance-covariance matrix of additive values [1.8.3]

$\mathbf{G}_d = \sigma_d^2\mathbf{D}$ = n×n variance-covariance matrix of dominance values [1.8.4]

$\mathbf{A}$ = n×n pedigree additive relationship matrix [1.8.5]

$\mathbf{D}$ = n×n pedigree dominance relationship matrix [1.8.6]

The mathematical expectation and variance-covariance matrix of $\mathbf{g}$, $\mathbf{a}$ and $\mathbf{d}$ are:

$$E\begin{pmatrix}\mathbf{a}\\ \mathbf{d}\end{pmatrix} = \mathbf{0} \qquad [1.8.7]$$

$$\text{var}\begin{pmatrix}\mathbf{g}\\ \mathbf{a}\\ \mathbf{d}\end{pmatrix} = \begin{pmatrix}\text{var}(\mathbf{g}) & \text{cov}(\mathbf{g},\mathbf{a}') & \text{cov}(\mathbf{g},\mathbf{d}')\\ \text{cov}(\mathbf{a},\mathbf{g}') & \text{var}(\mathbf{a}) & \text{cov}(\mathbf{a},\mathbf{d}')\\ \text{cov}(\mathbf{d},\mathbf{g}') & \text{cov}(\mathbf{d},\mathbf{a}') & \text{var}(\mathbf{d})\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \sigma_a^2\mathbf{A} & \sigma_d^2\mathbf{D}\\ \sigma_a^2\mathbf{A} & \sigma_a^2\mathbf{A} & \mathbf{0}\\ \sigma_d^2\mathbf{D} & \mathbf{0} & \sigma_d^2\mathbf{D}\end{pmatrix} \qquad [1.8.8]$$

**Covariance between relatives with additive, dominance and epistasis values**

Assuming no inbreeding, the general formula of covariance between the genotypic values of two relatives (Cockerham, 1954) is:

$$\text{cov}(g_j, g_k) = a_{jk}\sigma_a^2 + d_{jk}\sigma_d^2 + (a_{jk})^2\sigma_{aa}^2 + a_{jk}d_{jk}\sigma_{ad}^2 + (d_{jk})^2\sigma_{dd}^2 + ... \qquad [1.8.9]$$

where $\sigma^2_{aa}$ = variance of additive × additive interactions, $\sigma^2_{ad}$ = variance of additive × dominance interactions, $\sigma^2_{dd}$ = variance of dominance × dominance interactions. The variance-covariance matrix of the genetic values for q individuals (Henderson, 1985) is:

$$\text{var}(\mathbf{g}) = \sigma^2_a\mathbf{A} + \sigma^2_d\mathbf{D} + \sigma^2_{aa}\ \mathbf{A}\#\mathbf{A} + a_{jk}d_{jk}\sigma^2_{ad}\mathbf{A}\#\mathbf{D} + \sigma^2_{dd}\mathbf{D}\#\mathbf{D} + ... \qquad [1.8.10]$$

where # denotes the Hadamard product, which is element-wise multiplication. Let **C** = **A**#**B**. Then, the element ij in **C** ($c_{ij}$) is the product of the corresponding elements in **A** and **B**, i.e.,

$c_{ij} = (a_{ij})(b_{ij})$, with $a_{ij}$ = element ij in **A**, and $b_{ij}$ = element ij in **B**.

**Example 1.8.1**: Using the pedigree results of the two full sibs in Example 1.6.1, $a_{34}$ = 1/2, and $d_{34}$ = 1/4, the covariance between the genetic values of individuals 3 and 4 using Equation [1.8.9] is:

$$\text{cov}(g_3, g_4) = a_{34}\sigma^2_a + d_{34}\sigma^2_d + (a_{34})^2\sigma^2_{aa} + a_{34}d_{34}\sigma^2_{ad} + (d_{34})^2\sigma^2_{dd} + ...$$
$$\tfrac{1}{2}\sigma^2_a + \tfrac{1}{4}\sigma^2_d + (\tfrac{1}{2})^2\sigma^2_{aa} + (\tfrac{1}{2})(\tfrac{1}{4})\sigma^2_{ad} + (\tfrac{1}{4})^2\sigma^2_{dd} +$$

**Example 1.8.2**: The variance-covariance matrix of genetic values for individuals 1-4 in Example [1.6.1] using Equation [1.8.10] is:

$$\text{var}(g) = \text{var}\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \end{bmatrix} = \begin{bmatrix} \text{var}(g_1) & \text{cov}(g_1,g_2) & \text{cov}(g_1,g_3) & \text{cov}(g_1,g_4) \\ \text{cov}(g_2,g_1) & \text{var}(g_2) & \text{cov}(g_2,g_3) & \text{cov}(g_2,g_4) \\ \text{cov}(g_3,g_1) & \text{cov}(g_3,g_2) & \text{var}(g_3) & \text{cov}(g_3,g_4) \\ \text{cov}(g_4,g_1) & \text{cov}(g_4,g_2) & \text{cov}(g_4,g_3) & \text{var}(g_4) \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \tfrac{1}{2} & \tfrac{1}{2} \\ 0 & 1 & \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & 1 & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & \tfrac{1}{2} & 1 \end{bmatrix}\sigma^2_a + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tfrac{1}{4} \\ 0 & 0 & \tfrac{1}{4} & 1 \end{bmatrix}\sigma^2_d + \begin{bmatrix} 1 & 0 & \tfrac{1}{2} & \tfrac{1}{2} \\ 0 & 1 & \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & 1 & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & \tfrac{1}{2} & 1 \end{bmatrix}\#\begin{bmatrix} 1 & 0 & \tfrac{1}{2} & \tfrac{1}{2} \\ 0 & 1 & \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & 1 & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & \tfrac{1}{2} & 1 \end{bmatrix}\sigma^2_{aa}$$
$$+ \begin{bmatrix} 1 & 0 & \tfrac{1}{2} & \tfrac{1}{2} \\ 0 & 1 & \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & 1 & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} & \tfrac{1}{2} & 1 \end{bmatrix}\#\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tfrac{1}{4} \\ 0 & 0 & \tfrac{1}{4} & 1 \end{bmatrix}\sigma^2_{ad} + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tfrac{1}{4} \\ 0 & 0 & \tfrac{1}{4} & 1 \end{bmatrix}\#\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tfrac{1}{4} \\ 0 & 0 & \tfrac{1}{4} & 1 \end{bmatrix}\sigma^2_{dd} + ...$$

$$= \begin{bmatrix} 1 & 0 & ½ & ½ \\ 0 & 1 & ½ & ½ \\ ½ & ½ & 1 & ½ \\ ½ & ½ & ½ & 1 \end{bmatrix} \sigma_a^2 + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & ¼ \\ 0 & 0 & ¼ & 1 \end{bmatrix} \sigma_d^2 + \begin{bmatrix} 1 & 0 & ¼ & ¼ \\ 0 & 1 & ¼ & ¼ \\ ¼ & ¼ & 1 & ¼ \\ ¼ & ¼ & ¼ & 1 \end{bmatrix} \sigma_{aa}^2$$

$$+ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1/8 \\ 0 & 0 & 1/8 & 1 \end{bmatrix} \sigma_{ad}^2 + \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1/16 \\ 0 & 0 & 1/16 & 1 \end{bmatrix} \sigma_{dd}^2 + \ldots$$

## Implications of genetic relationships to practical applications

Examples 1.8.1 and 1.8.2 have the following implications to practical applications:

1) Individual relationships become smaller for higher-order interactions;
2) Confounding between interaction effects may exist, where confounding refers to the fact that two genetic effects have identical relationship matrices so that the two genetic effects cannot be separated;
3) Only full sibs have non-zero dominance related effects for practical purposes.

**Example 1.8.3**: Calculation of coancestry coefficients and additive relationships.

| ID | dam | sire |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 2 | 1 |
| 4 | 2 | 0 |
| 5 | 2 | 3 |
| 6 | 0 | 3 |

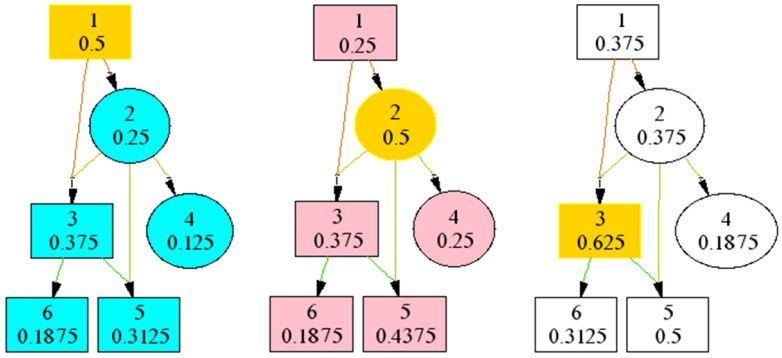



**Figure 1.8.1** Example of calculating coancestry coefficients. The pedigree in the left box has 6 individuals, with '0' denoting unknown parent. Each of the 3 pedigree figures displays coancestry coefficients between a selected individual (in gold color) and the other 5 individuals in the pedigree calculated using the Pedigraph program (Garbe and Da, 2008).

The additive relationship matrix for the six individuals is:

$$\mathbf{A}=\begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} & A_{15} & A_{16} \\ A_{21} & A_{22} & A_{23} & A_{24} & A_{25} & A_{26} \\ A_{31} & A_{32} & A_{33} & A_{34} & A_{35} & A_{36} \\ A_{41} & A_{42} & A_{43} & A_{44} & A_{45} & A_{46} \\ A_{51} & A_{52} & A_{53} & A_{54} & A_{55} & A_{56} \\ A_{61} & A_{62} & A_{63} & A_{64} & A_{65} & A_{66} \end{bmatrix}=\begin{bmatrix} 1 & \frac{1}{2} & \frac{3}{4} & \frac{1}{4} & \frac{5}{8} & \frac{3}{8} \\ & 1 & \frac{3}{4} & \frac{1}{2} & \frac{7}{8} & \frac{3}{8} \\ & & \frac{5}{4} & \frac{3}{8} & 1 & \frac{5}{8} \\ & & & 1 & \frac{7}{16} & \frac{3}{16} \\ & & & & \frac{11}{8} & \frac{1}{2} \\ & & & & & 1 \end{bmatrix}=\begin{bmatrix} 1 & 0.5 & 0.75 & 0.25 & 0.625 & 0.375 \\ & 1 & 0.75 & 0.5 & 0.875 & 0.375 \\ & & 1.25 & 0.375 & 1 & 0.625 \\ & & & 1 & 0.4375 & 0.1875 \\ & & & & 1.375 & 0.5 \\ & & & & & 1 \end{bmatrix}$$

Diving the above matrix by 2 yields the coancestry coefficients in Figure 1.8.1. Alternatively, multiply the coancestry coefficients in Figure 1.8.1 yields the corresponding additive relationships.

## 1.9 Fixed, Random, and Mixed Models

**Fixed model**: 'Effects' in a fixed model are 'fixed effects' with constant mean and zero variance.

$$y_{ij}=\mu+G_i+e_{ij}, \quad i=1,...n; j=1,.., n_i$$
$$E\left(y_{ij}\right)=E\left(\mu+G_i\right)=\mu+G_i$$
$$\text{var}\left(\mu+G_i\right)=0$$
$$\text{var}\left(y_{ij}\right)=\text{var}\left(e_{ij}\right)=\sigma_e^2$$

where $y_{ij}$ = observation j on individuals with fixed effect of blood group i, $\mu$ = the common mean of all observations treated as fixed effect, $G_i$ = fixed effect of blood group i, $e_{ij}$ = random residual of observation j on individuals with blood group i with zero mean and variance of $\sigma_e^2$, n = number of individuals, and $n_i$ = number of observations for individual i. In matrix notations, the fixed model is typically written as:

$$\mathbf{y}=\mathbf{Xb}+\mathbf{e} \qquad [1.9.1]$$

where $\mathbf{y}$ = N×1 column vector of observations, $\mathbf{b}$ = c×1 column vector of fixed effects, $\mathbf{X}$ = N×c model matrix, $\mathbf{e}$ = N×1 column vector of random residuals, c = number of levels of the fixed effects, and N = number of observations.

$$E(\mathbf{y})=\mathbf{Xb} \qquad [1.9.2]$$
$$\text{var}(\mathbf{y})=\text{var}(\mathbf{e})=\mathbf{R}=\sigma_e^2\mathbf{I} \qquad [1.9.3]$$

**Random model**: 'Effects' in the model are 'random effects' typically assumed to have zero mean and non-zero variances.

$$y_{ij} = u_i + e_{ij}, \qquad i = 1,...n; j = 1,.., n_i$$

where $y_{ij}$ = observation j of individual i, $u_i$ = 'random effect' of individual i with zero mean and variance of $\sigma_u^2$, $e_{ij}$ = random variable with zero mean and variance of $\sigma_e^2$. Typical assumptions are:

$$E\left(y_{ij}\right) = 0$$
$$\text{var}\left(y_{ij}\right) = \text{var}\left(u_i + e_{ij}\right) = \sigma_u^2 + \sigma_e^2$$

In matrix notations, the random model for n individuals can be written as:

$$\mathbf{y} = \mathbf{Zu} + \mathbf{e} \qquad [1.9.4]$$

where $\mathbf{u}$ = n×1 vector of random effects, and $\mathbf{Z}$ = N×n model matrix of $\mathbf{u}$. Assuming $\mathbf{u}$ is a vector of additive genetic effects, typical assumptions are:

$$E(\mathbf{y}) = \mathbf{0} \qquad [1.9.5]$$
$$\text{var}(\mathbf{u}) = \mathbf{G} \qquad [1.9.6]$$
$$\text{var}(\mathbf{e}) = \mathbf{R} = \sigma_e^2\mathbf{I} \qquad [1.9.7]$$
$$\begin{aligned}\text{var}(\mathbf{y}) = \mathbf{V} &= \text{var}(\mathbf{Zu} + \mathbf{e}) = \text{var}(\mathbf{Zu}) + \text{var}(\mathbf{e}) \\ &= \mathbf{Z}[\text{var}(\mathbf{u})]\mathbf{Z}' + \mathbf{R} \\ &= \mathbf{ZGZ}' + \mathbf{R} \qquad [1.9.8] \\ &= \mathbf{ZGZ}' + \sigma_e^2\mathbf{I} \qquad [1.9.9]\end{aligned}$$

where $\mathbf{V}$ = N×N matrix, $\mathbf{I}$ = N×N identity matrix. The above formula includes the assumption that $\mathbf{u}$ and $\mathbf{e}$ are uncorrelated.

**Mixed model**: A statistical model with both 'fixed' and 'random' effects is termed as mixed model and is typically written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad [1.9.10]$$
$$E(\mathbf{y}) = \mathbf{Xb}$$
$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} = \mathbf{ZGZ}' + \sigma_e^2\mathbf{I}, \text{ same as Equations 1.9.8 and 1.9.9.}$$

## Rule for defining 'fixed' or 'random' effects

No universally accepted rule is available for defining fixed or random effects. The following rules can be used:

- Effects with small number of levels defined as fixed effects
- Effects with large number of levels defined as random effects

## Comparison of 'fixed' and 'random' effects

- Fixed effects consume degrees of freedom (d.f.)
- Number of fixed effects cannot exceed the number of residual d.f.
  (residual d.f. = N − rank of $\mathbf{X}$ = N − r)
- Residual d.f. needs to be sufficiently large
- Fixed effects typically are subjected to significance tests

- Random effects typically are not subjected to significance tests
- Random effects do not consume degrees of freedom (d.f.)
- Number of random effects may exceed the number of observations

In genomic prediction, treating SNP markers as random effects is the only workable option because the number of markers is large.

**Exercises**

Define the dimensions for the following matrices: $\mathbf{Xb}$, $\mathbf{Zu}$, $\mathbf{R}$, $\mathbf{GZ}'$ and $\mathbf{ZGZ}'$.

# CHAPTER 2: BEST LINEAR UNBIASED PREDICTION (BLUP)

Best linear unbiased prediction (BLUP) from mixed models estimates the unknown means of phenotypic values using the generalized least squares estimator and then predicts genetic values using the phenotypic values after removing the phenotypic means (Henderson, 1984). BLUP has been the standard method for genetic evaluation using pedigree information, and its genomic version termed as GBLUP (genomic BLUP) that replaces pedigree relationships with genomic relationships has become a standard method for genomic evaluation.

## 2.1 Mixed Models for BLUP

The description of a mixed model generally requires two items: the mixed model, and the variance-covariance matrix of the phenotypic observations under the given mixed model. The general notations for the mixed model and its variance-covariance matrices are:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \quad [2.1.1]$$

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R} \quad [2.1.2]$$

where

$\mathbf{y}$ = N × 1 vector of phenotypic observations
$\mathbf{b}$ = c × 1 vector of fixed effects
$\mathbf{X}$ = N × c incidence matrix for fixed effects
$\mathbf{u}$ = n × 1 vector of random genetic values
$\mathbf{Z}$ = N × n incidence matrix to allocate phenotypic observations to individuals
= identify matrix if N = n
$\mathbf{e}$ = N × 1 vector of random residuals
N = number of phenotypic observations
c = number of levels of fixed effects = number of columns in $\mathbf{X}$
n = number of animals

$$\mathbf{G} = \text{var}(\mathbf{u}) \quad [2.1.3]$$

$$\mathbf{R} = \text{var}(\mathbf{e}) \quad [2.1.4]$$

For all single-trait mixed models, the assumption for $\mathbf{R}$ is:

$$\mathbf{R} = \sigma_e^2 \mathbf{I}_N \quad [2.1.5]$$

where $\sigma_e^2$ = residual variance, and $\mathbf{I}_N$ = N × N identity matrix.

With the general expressions of Equations [2.1.1] and [2.1.2], any or all types of the genetic values described in Chapter 1 and many other variations can be included in the mixed model by modeling the $\mathbf{u}$ and $\mathbf{G}$ matrices. Some of the variations including multiple traits and genomic prediction will be described in other chapters.

**Example 2.1.1: additive values**

the mixed model with additive values can be descrived as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \quad [2.1.6]$$

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R} = \sigma_a^2\mathbf{ZAZ}' + \sigma_e^2\mathbf{I}_N \quad [2.1.7]$$

where $\mathbf{a}$ = n × 1 column vector of additive values of individuals, $\sigma_a^2$ = additive variance, $\mathbf{A}$ = additive relationship matrix, and

$$\mathbf{G} = \text{var}(\mathbf{a}) = \sigma_a^2\mathbf{A} \quad [2.1.8]$$

The additive model of Equations [2.1.6] and [2.1.7] has been used most widely compared to more complex models.

**Example 2.1.2: additive and dominance values**

The mixed model with additive and dominance values requires new definitions of $\mathbf{Z}$, $\mathbf{u}$ and $\mathbf{G}$ matrices:

$$\mathbf{Z}_2 = \begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix} \quad [2.1.9]$$

$$\mathbf{u} = \begin{bmatrix}\mathbf{a} \\ \mathbf{d}\end{bmatrix} \quad [2.1.10]$$

$$\mathbf{G} = \text{var}\begin{bmatrix}\mathbf{a} \\ \mathbf{d}\end{bmatrix} = \begin{bmatrix}\text{var}(\mathbf{a}) & \mathbf{0} \\ \mathbf{0} & \text{var}(\mathbf{d})\end{bmatrix} = \begin{pmatrix}\sigma_a^2\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \sigma_d^2\mathbf{D}\end{pmatrix} \quad [2.1.11]$$

where $\mathbf{d}$ = n × 1 column vector of dominance values of individuals, $\sigma_d^2$ = dominance variance, $\mathbf{D}$ = dominance relationship matrix. Equation [2.1.11] assumes independent additive and dominance values. The mixed model with additive and dominance values can be descrived as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_2\mathbf{u} + \mathbf{e} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e} \quad [2.1.12]$$

$$\begin{aligned}\mathbf{V} &= \text{var}(\mathbf{Z}_2\mathbf{u}) + \text{var}(\mathbf{e}) = \mathbf{Z}_2\,\text{var}(\mathbf{u})\mathbf{Z}_2' + \mathbf{R} \\ &= \sigma_a^2\mathbf{ZAZ}' + \sigma_d^2\mathbf{ZDZ}' + \sigma_e^2\mathbf{I}\end{aligned} \quad [2.1.13]$$

The mixed model of Equations [2.1.12] and [2.1.13] can be considered as a special case of the mixed models with multiple types of genetic effects because the approach of model writing for mixed models with multiple types of genetic effects is similar to the approach Equations [2.1.9]-[2.1.13].

## 2.2 The Conditional Expectation (CE) Method of BLUP

The conditional expectation of **u** given **y** is:

$$E(\mathbf{u}/\mathbf{y}) = E(\mathbf{u}) + \text{cov}(\mathbf{u},\mathbf{y}')[\text{var}(\mathbf{y})]^{-1}(\mathbf{y} - \mathbf{Xb}) = \text{cov}(\mathbf{u},\mathbf{y}')[\text{var}(\mathbf{y})]^{-1}(\mathbf{y} - \mathbf{Xb}) \quad [2.2.1]$$

Then, the BLUP of **u** and **e** in Equation [2.1.1] can be obtained from Equation [2.2.1] by replacing **b** with its generalized least squats (GLS) estimator or best linear unbiased estimator (BLUE):

$$\hat{\mathbf{u}} = E(\mathbf{u}/\mathbf{y},\hat{\mathbf{b}}) = \text{cov}(\mathbf{u},\mathbf{y}')[\text{var}(\mathbf{y})]^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad [2.2.2]$$

$$= \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{GZ}'\mathbf{Py} \quad [2.2.3]$$

$$\hat{\mathbf{e}} = E(\mathbf{e}/\mathbf{y},\hat{\mathbf{b}}) = \text{cov}(\mathbf{e},\mathbf{y}')[\text{var}(\mathbf{y})]^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad [2.2.4]$$

$$= \mathbf{RV}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{RPy} \quad [2.2.5]$$

$$= \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}} \quad [2.2.6]$$

where

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \text{ = GLS estimator or (BLUE) of } \mathbf{b} \quad [2.2.7]$$

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \quad [2.2.8]$$

Equations [2.2.2] and [2.2.3] are the conditional expectation of **u** given **y**, and Equations [2.2.4] and [2.2.5] are the conditional expectation of **e** given **y**, except that **b** in the conditional expectation is replaced by its GLS estimator or BLUE. The BLUP of **u** given by Equations [2.2.2]- and [2.2.3] will be referred to as the "conditional expectation (CE) method" of BLUP. Equation [2.2.5] shows that $\hat{\mathbf{e}}$ is the BLUP of **e** but in practice $\hat{\mathbf{e}}$ is commonly calculated using Equation [2.2.8].

**Exercises**

1) Using the data in any of the SAS program, verify:
   $[\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{X} = \mathbf{X}$ (Searle et al., 1992; Harville, 1997)
2) Prove:
   $\mathbf{PX} = [\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{X} = \mathbf{0}$

## 2.3 The MME Method of BLUP

Mixed model equations (MME) for BLUP, discovered by Henderson, remove the need of $\mathbf{V}^{-1}$ and instead use $\mathbf{G}^{-1}$. For the additive model of Equations [2.1.6] and [2.1.7], $\mathbf{G}^{-1} = \mathbf{A}^{-1}/\sigma_a^2$. Henderson discovered an easy way to construct $\mathbf{A}^{-1}$ without actually inverting **A** (Henderson 1975). BLUP based on MME has been the standard procedure for the implementation of genetic evaluations using BLUP of additive models. The general formula of MME for the mixed model of Equations [2.1.1] and [2.1.2] is:

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \quad [2.3.1]$$

The BLUP of **u** is obtained by solving the MME. Equation [2.3.1] can be expressed as:

$$\mathbf{Cs} = \mathbf{r} \quad [2.3.2]$$

where

$$\mathbf{C} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} = \text{coefficient matrix of MME} \quad [2.3.3]$$

$$\mathbf{s} = \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \text{solution vector of MME} \quad [2.3.4]$$

$$\mathbf{r} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} = \text{right-hand-side of MME} \quad [2.3.5]$$

Then, the solution vector of Equation [2.3.4] is obtained as:

$$\mathbf{s} = \mathbf{C}^{-}\mathbf{r} \quad [2.3.6]$$

where $\mathbf{C}^{-}$ = a generalized inverse of **C**. The BLUP using the MME of Equation [2.3.1] will be referred to as the "MME method" of BLUP.

The CE and MME methods have identical results and offer alternative computing strategies for calculating BLUP.

## 2.4 Examples of CE and MME Methods of BLUP

The purpose of these examples is to show the CE formula of BLUP and the MME formula under two models, the mixed model with additive values only, and the mixed model with both additive and dominance values.

### Example 2.4.1: additive values

For the mixed model with additive values only, $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$ (Equation [2.1.5]) and $\mathbf{V} = \sigma_a^2\mathbf{ZAZ}' + \sigma_e^2\mathbf{I}$ (Equation [2.1.6]), the BLUP of additive values by the CE method is:

$$\hat{\mathbf{a}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_a^2\mathbf{AZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad [2.4.1]$$

Rewrite the MME of Equation [2.3.1] as:

$$\mathbf{X'R^{-1}X\hat{b}} + \mathbf{X'R^{-1}Z\hat{a}} = \mathbf{X'R^{-1}y} \quad [2.4.2]$$

$$\mathbf{Z'R^{-1}X\hat{b}} + (\mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}})\mathbf{\hat{a}} = \mathbf{Z'R^{-1}y} \quad [2.4.3]$$

Noting $\mathbf{R} = \sigma_e^2\mathbf{I}_N$ (Equation [2.1.5]) and $\mathbf{R}^{-1} = (1/\sigma_e^2)\mathbf{I}_N$, Equations [2.4.2] and [2.4.3] are:

$$\mathbf{X'X\hat{b}} / \sigma_e^2 + \mathbf{X'Z\hat{a}} / \sigma_e^2 = \mathbf{X'y} / \sigma_e^2 \quad [2.4.4]$$

$$\mathbf{Z'X\hat{b}} / \sigma_e^2 + (\mathbf{Z'Z} / \sigma_e^2 + (1/\sigma_a^2)\mathbf{A}^{-1})\mathbf{\hat{a}} = \mathbf{Z'y} / \sigma_e^2 \quad [2.4.5]$$

Multiplying both sides of Equations [2.4.4] and [2.4.5] leads to:

$$\mathbf{X'X\hat{b}} + \mathbf{X'Z\hat{a}} = \mathbf{X'y} \quad [2.4.6]$$

$$\mathbf{Z'X\hat{b}} + (\mathbf{Z'Z} + (\sigma_e^2 / \sigma_a^2)\mathbf{A}^{-1})\mathbf{\hat{a}} = \mathbf{Z'y} \quad [2.4.7]$$

or,

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \lambda\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{\hat{b}} \\ \mathbf{\hat{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix} \quad [2.4.8]$$

where

$$\lambda = \sigma_e^2/\sigma_a^2 = (1-h^2)/h^2 \quad [2.4.9]$$

$$h^2 = \sigma_a^2 / \sigma_y^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2) = \text{additive heritability} \quad [2.4.10]$$

Equation [2.4.8] is a standard MME formula for the additive model.

**Example 2.4.2: additive and dominance values**

For the mixed model with additive and dominance values, $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_2\mathbf{u} + \mathbf{e} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e}$ (Equation [2.1.11]), and $\mathbf{V} = \sigma_a^2\mathbf{ZAZ'} + \sigma_d^2\mathbf{ZDZ'} + \sigma_e^2\mathbf{I}$ (Equation [2.1.12]), the BLUP of **a** and **d** by the CE method are:

$$\mathbf{\hat{a}} = \sigma_a^2\mathbf{AZ'V^{-1}}(\mathbf{y} - \mathbf{X\hat{b}}) \quad [2.4.11]$$

$$\mathbf{\hat{d}} = \sigma_d^2\mathbf{DZ'V^{-1}}(\mathbf{y} - \mathbf{X\hat{b}}) \quad [2.4.12]$$

The MME based on the general formula of Equation [2.3.1] are:

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2+\mathbf{G}^{-1} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix}=\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \quad [2.4.13]$$

Using Equations [2.1.8] and [2.1.10], Equation [2.4.13] becomes:

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix} \\ \begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix}'\mathbf{R}^{-1}\mathbf{X} & \begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix}'\mathbf{R}^{-1}\mathbf{Z}_2=\begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix}'\mathbf{R}^{-1}\begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix}+\begin{bmatrix}(\sigma_a^2\mathbf{A})^{-1} & \mathbf{0} \\ \mathbf{0} & (\sigma_d^2\mathbf{D})^{-1}\end{bmatrix} \end{pmatrix}\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} \quad [2.4.14]$$
$$=\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \begin{bmatrix}\mathbf{Z} & \mathbf{Z}\end{bmatrix}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

Multiflying both sizes of Equation [2.4.14] by $\sigma_e^2$, Equation [2.4.14] is reduced to:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}\mathbf{X}' & \mathbf{Z}\mathbf{Z}'+\lambda_a\mathbf{A}^{-1} & \mathbf{Z}\mathbf{Z}' \\ \mathbf{Z}\mathbf{X}' & \mathbf{Z}\mathbf{Z}' & \mathbf{Z}\mathbf{Z}'+\lambda_d\mathbf{D}^{-1} \end{bmatrix}\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{d}} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \quad [2.4.15]$$

where

$$\lambda_a=\sigma_e^2/\sigma_a^2 \quad [2.4.16]$$

$$\lambda_d=\sigma_e^2/\sigma_d^2 \quad [2.4.17]$$

The BLUP of **a** and **d** ($\hat{\mathbf{a}}$ and $\hat{\mathbf{d}}$) are obtained by solving the MME of Equation [2.4.15].

**<u>Numerical example of the CE and MME methods</u>**

This example shows the CE and MME methods have identical results assuming the additive model of Equations [2.1.6] and [2.1.7]. The data of this numerial example include:

$$\mathbf{y}=\begin{bmatrix}10\\11\\12\\13\end{bmatrix},\ \mathbf{X}=\begin{bmatrix}1\\1\\1\\1\end{bmatrix},\ \mathbf{Z}=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},\ \mathbf{A}=\begin{bmatrix}1&0&0.5&0.5\\0&1&0.5&0.5\\0.5&0.5&1&0.5\\0.5&0.5&0.5&1\end{bmatrix}$$

Number of individuals: 4. Number of observations: 4, one observation per individual.
$\sigma_a^2=2$, $\sigma_e^2=7$.

From the CE method,

$$\hat{b}=[11.433333]$$

$$\hat{\mathbf{a}}=\begin{bmatrix}-0.111111\\0.1111111\\0.0708333\\0.1958333\end{bmatrix},\ \hat{\mathbf{e}}=(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}-\mathbf{Z}\hat{\mathbf{a}})=\begin{bmatrix}-1.322222\\-0.544444\\0.4958333\\1.3708333\end{bmatrix},\ \hat{\mathbf{e}}=\mathbf{R}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})=\begin{bmatrix}-1.322222\\-0.544444\\0.4958333\\1.3708333\end{bmatrix}$$

From the MME method,

$$\mathbf{s}=\begin{bmatrix}\hat{b}\\\hat{\mathbf{a}}\end{bmatrix}=\begin{bmatrix}11.433333\\-0.111111\\0.1111111\\0.0708333\\0.1958333\end{bmatrix}$$

These numerical results show that CE and MME have identical results.

**SAS program for the numerical example**

```
/* ANSC8141: Calculation of BLUP using CE and MME */
/* program name: ce_mme.sas */
PROC IML;
RESET PRINT ; * Prints everything below this line;
Q = 4; * 4 ANIMALS;
P = 1; * 1 FIXED EFFECTS;
N = 4; * 4 OBERVATIONS, ONE OBS/ANIMAL;
/* additive relationship matrix */
A   = {1    0   .5  .5 ,
       0    1   .5  .5 ,
      .5   .5    1  .5 ,
      .5   .5   .5  1 };
IA = INV(A);
X  = {1,1,1,1};         * X matrix for the common mean;
Y  = {10, 11, 12, 13};  * phenotypic observations;
va = 2;
ve = 7;
IDQ = I(Q);
IDN = I(N);
Z = IDQ;    * model matrix for individual additive effects;
* --- BLUP FROM MME ----;
XX = X`*X; XZ = X`*Z;
ZZ = Z`*Z;
```

```
XY = X`*Y;
ZY = Z`*Y;
RHS = XY//ZY;
RATIO = VE/VA;
C_AA = ZZ + IA*RATIO;
C1 = XX||XZ;
C2 = XZ`||C_AA;
C = C1//C2;
IC = GINV(C);
SOL = IC*RHS;
* --- BLUP FROM CE ----;
G_A = A*VA;
V = Z*G_A*Z` + IDN*VE;
IV = INV(V);
XIVX = X`*IV*X;
XIVY = X`*IV*Y;
B_HAT = GINV(XIVX)*XIVY;
BLUP_A = G_A*IV*Z`*(Y-X*B_HAT);

* --- BLUP OF E ----;
E_HAT1 = Y - X*B_HAT - Z*BLUP_A;
R = IDN*VE;
E_HAT2 = R*IV*(Y-X*B_HAT);
RUN;
```

**Exercises**

1) Translate all SAS programs in this chapter into R programs.
2) For the mixed model with additive and dominance values,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_2\mathbf{u} + \mathbf{e} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e} \qquad \text{Equation [2.1.11]}$$

$$\mathbf{V} = \sigma_a^2\mathbf{ZAZ}' + \sigma_d^2\mathbf{ZDZ}' + \sigma_e^2\mathbf{I} \qquad \text{Equation [2.1.12]}$$

Assume the number of individuals is 100, and the number of fixed effects is 10. Define the following:

a. The sizes of the **V**, **A** and **D** matrices.
b. The size of the MME.

## 2.5 Equivalence between the CE and MME Methods of BLUP

The CE method of Equation [2.2.3] and the MME method of Equation [2.3.1] have identical BLUP results. The proof of this equivalence requires the proofs that the $\hat{\mathbf{u}}$ and $\hat{\mathbf{b}}$ from the CE method are identical to those from the MME method.

The first step is to prove the CE and MME methods have identical $\hat{\mathbf{u}}$. The $\hat{\mathbf{u}}$ from the CE method (Equation [2.2.3]) is:

$$\hat{\mathbf{u}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{GZ}'\mathbf{Py} \qquad [2.2.3]$$

The MME of Equation [2.3.1] can be re-written as two sets of equations:

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\mathbf{b}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{u}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \quad [2.5.1]$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\mathbf{b}} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \quad [2.5.2]$$

Therefore, the $\hat{\mathbf{u}}$ vector from the MME method can be expressed as:

$$\begin{aligned}\hat{\mathbf{u}} &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\mathbf{b}}) = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \\ &= \boldsymbol{\Lambda}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})\end{aligned} \quad [2.5.3]$$

where

$$\boldsymbol{\Lambda} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1} \quad [2.5.4]$$

$$= \sigma_e^2(\mathbf{Z}'\mathbf{Z} + \sigma_e^2/\sigma_a^2\mathbf{A}^{-1})^{-1} \text{ for the additive model} \quad [2.5.5]$$

To prove the equivalence between the CE method of Equation [2.2.3] and the MME method of Equation [2.3.1] is to prove that the $\hat{\mathbf{u}}$ of Equation [2.2.3] and the $\hat{\mathbf{u}}$ of Equation [2.5.3] are identical:

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \boldsymbol{\Lambda}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \quad [2.5.6]$$

For Equation [2.5.6] to hold, the following equation must hold:

$$\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} = \boldsymbol{\Lambda}\mathbf{Z}'\mathbf{R}^{-1} \quad [2.5.7]$$

The following proves Equation [2.5.56 by proving Equation [2.5.7]. The key for this proof is to express $\mathbf{V}^{-1}$ of the CE method in terms of the quantities of the MME method. The Schur complement (Searle et al., 1992) is:

$$(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{C}(\mathbf{B}\mathbf{D}^{-1}\mathbf{C} + \mathbf{A})^{-1}\mathbf{B}\mathbf{D}^{-1} \quad [2.5.8]$$

In the Schur compliment, let $\mathbf{D} = \mathbf{R}$, $\mathbf{C} = \mathbf{Z}$, $\mathbf{B} = \mathbf{Z}'$, and $\mathbf{G} = \mathbf{A}^{-1}$ . Then, the $\mathbf{V}^{-1}$ in Equation [2.5.7] can be expressed as:

$$\begin{aligned}\mathbf{V}^{-1} &= (\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1} \\ &= \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}'\mathbf{R}^{-1}\end{aligned} \quad [2.5.9]$$

Applying Equation [2.5.9] to Equation [2.5.7],

$$\begin{aligned}\mathbf{GZ'V}^{-1} &= \mathbf{GZ'}[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1}] \\ &= \mathbf{G}[\mathbf{Z'R}^{-1} - \mathbf{Z'R}^{-1}\mathbf{Z}(\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1}] \\ &= \mathbf{GZ'R}^{-1} - \mathbf{G}(\mathbf{Z'R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} - \mathbf{G}^{-1})(\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1} \\ &= \mathbf{GZ'R}^{-1} - \mathbf{G}(\mathbf{Z'R}^{-1} - \mathbf{G}^{-1}(\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1} \\ &= \mathbf{GZ'R}^{-1} - \mathbf{GZ'R}^{-1} + \mathbf{GG}^{-1}(\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1} \\ &= (\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1} \\ &= \mathbf{\Lambda Z'R}^{-1}\end{aligned} \qquad [2.5.10]$$

Equation [2.5.10] proves Equation [2.5.7], and hence proves Equation [2.5.6] that states the CE method of Equation [2.2.3] and the MME method of Equation [2.5.3] have identical BLUP results.

The next step is to prove the CE and MME methods have identical $\hat{\mathbf{b}}$. Substituting Equation [2.5.6] into Equation [2.5.1] for $\hat{\mathbf{u}}$ yields:

$$\begin{aligned}&\mathbf{X'R}^{-1}\mathbf{X}\hat{\mathbf{b}} + \mathbf{X'R}^{-1}\mathbf{Z}[\mathbf{\Lambda Z'R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})] = \mathbf{X'R}^{-1}\mathbf{y} \\ &\mathbf{X'R}^{-1}\mathbf{X}\hat{\mathbf{b}} + \mathbf{X'R}^{-1}\mathbf{Z\Lambda Z'R}^{-1}\mathbf{y} - \mathbf{X'R}^{-1}\mathbf{Z\Lambda Z'R}^{-1}\mathbf{X}\hat{\mathbf{b}}) = \mathbf{X'R}^{-1}\mathbf{y} \\ &\mathbf{X'}(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z\Lambda Z'R}^{-1})\mathbf{X}\hat{\mathbf{b}} = \mathbf{X'}(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z\Lambda Z'R}^{-1})\mathbf{y}\end{aligned} \qquad [2.5.11]$$

Using $\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z\Lambda Z'R}^{-1}$ (Equation [2.5.8]), Equation [2.5.10] becomes the GLS equations:

$$\mathbf{X'V}^{-1}\mathbf{X}\hat{\mathbf{b}} = \mathbf{X'V}^{-1}\mathbf{y} \qquad [2.5.12]$$

Solving Equations [2.5.11] and [2.5.12] for $\hat{\mathbf{b}}$ yields:

$$\hat{\mathbf{b}} = [\mathbf{X'}(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z\Lambda Z'R}^{-1})\mathbf{X}]^{-}\mathbf{X'}(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z\Lambda Z'R}^{-1})\mathbf{y} \qquad [2.5.13]$$

$$\hat{\mathbf{b}} = (\mathbf{X'V}^{-1}\mathbf{X})^{-}\mathbf{X'V}^{-1}\mathbf{y} \qquad [2.5.14]$$

Since Equations [2.5.113] and [2.5.14] are identical, the BLUE from the MME method of Equation [2.2.12] and the GLS estimator from the CE method of Equation [2.5.13] are identical.

**Exercises**

1) Assume the additive model, modify the SAS program 'ce_mme.sas' or your R program to verify:
   a. $\mathbf{\Lambda} = (\mathbf{Z'R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1} = \sigma_e^2(\mathbf{Z'Z} + \sigma_e^2/\sigma_a^2\mathbf{A}^{-1})^{-1}$
   b. $\mathbf{V}^{-1} = (\mathbf{ZGZ'}+\mathbf{R})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z'R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z'R}^{-1}$
   c. $\mathbf{V}^{-1} = (\mathbf{ZGZ'}+\mathbf{R})^{-1} = \{\mathbf{I} - \mathbf{Z}[\mathbf{Z'Z}+(\sigma_e^2/\sigma_a^2)\mathbf{A}^{-1}]^{-1}\mathbf{Z'}\}/\sigma_e^2$

## 2.6 BLUP for Animals Without Phenotypic Observations

An advantage of BLUP is to predict the genetic merit for individuals without phenotypic observations such as the prediction of the adult body weight of an animl soon after birth for early genetic selection. The genetic merit of animals without phenotypic observations can be predicted through their relationships with phenotyped relatives (Henderson, 1977). This approach has an important application in genomic prediction of individuals without phenotypic observations.

Assuming the additive model of Equations [2.1.6] and [2.1.7] for an animal population, the prediction of animals without phenotypic observations starts with partitioning the **Z** matrix and the **a** vector into submatrices corresponding to animals with and without phenotypic observations:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \end{bmatrix} \qquad [2.6.1]$$

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_0 \end{bmatrix} \qquad [2.6.2]$$

where

$\mathbf{Z}_1 = \mathrm{N} \times \mathrm{n}_1$ incidence matrix for animals with observations = incidence matrix for $\mathbf{a}_1$ ,

$\mathbf{a}_1 = \mathrm{n}_1 \times 1$ vector of additive genetic values for animals with observations,

$\mathrm{n}_1$ = number of animals with observations,

$\mathbf{a}_0 = \mathrm{n}_0 \times 1$ vector of additive genetic values for animals without phenotypic observations,

$\mathbf{0} = \mathrm{N} \times \mathrm{n}_0$ matrix of zeros,

$\mathrm{n}_0$ = number of animals without observations, and $\mathrm{n}_1 + \mathrm{n}_0 = \mathrm{n}$ .

The additive model of Equation [2.1.6] to include animals without phenotypic observations through Equations [2.7.1] and [2.7.2] is:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_0 \end{bmatrix} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} \qquad [2.6.3]$$

where **Z** and **a** are defined by Equations [2.6.1] and [2.6.2].

Let the **A**-matrix and **A**-inverse be partitioned into submatrices corresponding to animals with and without phenotypic observations as:

$$\mathbf{G} = \mathrm{var}(\mathbf{a}) = \mathrm{var}\begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_0 \end{bmatrix} = \begin{bmatrix} \mathrm{var}(\mathbf{a}_1) & \mathrm{cov}(\mathbf{a}_1, \mathbf{a}_0\,') \\ \mathrm{cov}(\mathbf{a}_0, \mathbf{a}_1\,') & \mathrm{var}(\mathbf{a}_0) \end{bmatrix} = \sigma_a^2 \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{10} \\ \mathbf{A}_{01} & \mathbf{A}_{00} \end{pmatrix} = \sigma_a^2 \mathbf{A} \qquad [2.6.4]$$

$$\mathbf{G}^{-1} = \mathbf{A}^{-1} / \sigma_a^2 = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{10} \\ \mathbf{A}^{01} & \mathbf{A}^{00} \end{pmatrix} / \sigma_a^2 \qquad [2.6.5]$$

where $\mathbf{A}_{11} = n_1 \times n_1$ matrix of additive relationships among individuals with phenotypic records, $\mathbf{A}_{10} = n_1 \times n_0$ matrix of additive relationships between individuals with phenotypic observations and individuals without phenotypic observations, $\mathbf{A}_{00} = n_0 \times n_0$ matrix of additive relationships among individuals without phenotypic observations, and

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{10} \\ \mathbf{A}_{01} & \mathbf{A}_{00} \end{pmatrix} \quad [2.6.6]$$

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{10} \\ \mathbf{A}^{01} & \mathbf{A}^{00} \end{pmatrix} \quad [2.6.7]$$

With the above preparations, the CE and MME methods for BLUP of animals without phenotypic observation can be formulated.

**BLUP for animals without phenotypic records using CE**

The BLUP of $\mathbf{a}_0$ using CE methods is obtained as:

$$\hat{\mathbf{a}}_0 = E(\mathbf{a}_0 / \mathbf{y}, \hat{\mathbf{b}}) = \text{cov}(\mathbf{a}_0, \mathbf{y}')[\text{var}(\mathbf{y})]^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

$$= \sigma_a^2 \mathbf{A}_{01}\mathbf{Z}_1\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_a^2 \mathbf{A}_{01}\mathbf{Z}_1\mathbf{P}\mathbf{y} \quad [2.6.7]$$

$$= \mathbf{A}_{01}\mathbf{A}_{11}^{-1}\hat{\mathbf{a}}_1 \quad [2.6.8]$$

$$= \text{cov}(\mathbf{a}_0, \mathbf{a}_1')\left[\text{var}(\mathbf{a}_1)\right]^{-1}\hat{\mathbf{a}}_1 \quad [2.6.9]$$

= 'regression' of additive effects of animals without records on additive effects of animals with records

= **0** if animals without observations do not have relatives with observations

where **P** is defined by Equation [2.2.8], and $\hat{\mathbf{a}}_1$ is calculated as:

$$\hat{\mathbf{a}}_1 = \sigma_a^2 \mathbf{A}_{11}\mathbf{Z}_1\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_a^2 \mathbf{A}_{11}\mathbf{Z}_1\mathbf{P}\mathbf{y} \quad [2.6.10]$$

Equations [2.6.7]-[2.6.9] shows the mechanism of calculating BLUP for animals without phenotypic observations: the BLUP of animals without phenotypic observations is a regression on the BLUP of animals with phenotypic observations.

Equations [2.6.7]-[2.6.8] and [2.6.10] offer a computing strategy to calculate $\hat{\mathbf{a}}_0$ and $\hat{\mathbf{a}}_1$ separately. This computing strategy can be appealing if the number of animals without observations is large because the $\mathbf{A}_{11}$ required for the two separate calculations can be much smaller than the **A** matrix of Equation [2.6.6] for all animals. If the number of animals without phenotypic observations is mangeable, calculating $\hat{\mathbf{a}}_0$ and $\hat{\mathbf{a}}_1$ simultaneously in one system of euqations (Equation [2.4.1] for additive model) is convenient, as shown by the SAS program for the numerical example at the end of this section.

For calculating $\hat{\mathbf{a}}_0$ using the CE method, Equation [2.6.7] rather than [2.6.8] should be used because Equation [2.6.7] requires $\mathbf{A}_{11}$ only that is in the **P** and **V** matrices and does not use $\mathbf{A}_{11}^{-1}$ whereas Equation [2.6.8] requires both $\mathbf{A}_{11}$ and $\mathbf{A}_{11}^{-1}$. To use $\mathbf{A}_{11}^{-1}$ only and avoid the $\mathbf{A}_{11}$, the MME method should be used.

**BLUP for animals without phenotypic records using MME**

The BLUP of $\mathbf{a}_0$ using the MME method is obtained by the following equations:

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z}_1 & \mathbf{0} \\ \mathbf{Z}_1\mathbf{'X} & \mathbf{Z}_1\mathbf{'Z}_1+\lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{10} \\ \mathbf{0} & \lambda\mathbf{A}^{01} & \lambda\mathbf{A}^{00} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z}_1\mathbf{'y} \\ \mathbf{0} \end{pmatrix} \qquad [2.6.10]$$

The $\hat{\mathbf{a}}_0$ from the CE method of Equations [2.6.7]-[2.6.9] and the $\hat{\mathbf{a}}_0$ from the MME method Equation [2.6.10] are identical. The CE method has a direct and intuitive interpretation than the MME method, because the CE method of Equation [2.6.9] offers the interpretation that BLUP for animals without observations is a regression on the BLUP of animals with observations, whereas the MME method of Equation [2.6.10] does not offer such a direct and intuitive interpretation.

**Numerical example of BLUP for animals without phenotypic records**

The dataset is the same as the numerical example for the CE and MME method except that the phenotypic observation is assumed missing for animal #1. Under this assumption, the y, X and A matrices are:

$$\mathbf{y} = \begin{bmatrix} y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 11 \\ 12 \\ 13 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{bmatrix}$$

The **a** and **Z** matrices are:

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}, \quad \mathbf{Z}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} \mathbf{0} & \mathbf{Z}_1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The SAS program below shows the CE and MME methods have identical results for $\hat{\mathbf{a}}_0$ and $\hat{\mathbf{a}}_1$.

## SAS program for BLUP of animals without phenotypic records

```
/* ANSC8141: BLUP for individuals without observations */
/* ANSC8141: Calculation of BLUP using CE and MME */
/* for animals without phenotypic observations */
/* program name: blup_mis_obs.sas */
PROC IML;
RESET PRINT ;
Q = 4; * 4 ANIMALS;
P = 1; * 1 FIXED EFFECT;
N = 3; * 3 OBERVATIONS, ONE OBS/ANIMAL;
/* additive relationship matrix */
A   = {1    0   .5  .5 ,
       0    1   .5  .5 ,
      .5   .5    1  .5 ,
      .5   .5   .5  1 };
IA = INV(A);
X  = {1,1,1};    * X matrix for the common mean;
*Y  = {10, 11, 12, 13};  * phenotypic observations;
Y  = {11, 12, 13};  * individual 1 has no observation;
va = 2;
ve = 7;
Z1 = I(3);  * identity matrix for individuals with records;
imiss = j(3,1,0);  * 0 for individual without observation in Z;
IDN = I(N);
Z = imiss||Z1;     * Z matrix, note the position of imiss;
* --- BLUP FROM MME ----;
XX = X`*X; XZ = X`*Z;
ZZ = Z`*Z;
XY = X`*Y;
ZY = Z`*Y;
RHS = XY//ZY;
RATIO = VE/VA;
C_AA = ZZ + IA*RATIO;
C1 = XX||XZ;
C2 = XZ`||C_AA;
C = C1//C2;
IC = GINV(C);
SOL = IC*RHS;
* --- BLUP FROM CE ----;
BLUP1 = sol(|3:5,*|);
A01 = A(|1,2:4|);
A11 = A(|2:4,2:4|);
IA11 = inv(A11);
G_A11 = A11*VA;
V = Z1*G_A11*Z1` + IDN*VE;
IV = INV(V);
XIVX = X`*IV*X;
XIVY = X`*IV*Y;
B_HAT = GINV(XIVX)*XIVY;
BLUP_A1 = G_A11*IV*Z1`*(Y-X*B_HAT);
```

```
a0_a = a01*IA11*BLUP1;
a0_b = a01*IA11*BLUP_A1;
* --- BLUP OF E ----;
E_HAT1 = Y - X*B_HAT - Z1*BLUP_A1;
R = IDN*VE;
E_HAT2 = R*IV*(Y-X*B_HAT);
RUN;
```

**Exercises**

1) Translate all SAS programs in this chapter into R programs.

2) For the mixed model with additive values and individuals with missing phenotypic observations,

$$\mathbf{y} = \mathbf{Xb} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_0 \end{bmatrix} + \mathbf{e} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad \text{Equation [2.6.3]}$$

Assume the number of individuals with phenotypic observations is 100, and the number of individuals without phenotypic observations is 200. Define the following:
   a. The sizes of the **V**, $\mathbf{A}_{11}$ and $\mathbf{A}_{01}$ matrices.
   b. The size of the MME.

3) For the SAS program in Section 2.6, modify the SAS program 'blup_mis_obs.sas' or your R program to define individual #3 with missing phenotypic observation, and calculate:
   a. BLUP using the CE method.
   b. BLUP using the MME method
   c. Verify that $\hat{\mathbf{a}}_0$ from Equation [2.6.7] is identical to that from the CE method in the SAS program

## 2.7 BLUP with Repeated Records

'Repeated records' refers to two or more observations per individual. The mixed model with repeated records is often referred to as the 'repeatability model' that includes a random effect of 'permanent environment' (PE). A PE effect is assumed to be a lifetime effect and is not inheritable. The repeatability model allowing individuals without observations can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{p}_1 + \mathbf{Za} + \mathbf{e} \qquad [2.7.1]$$

where

$\mathbf{Z} = (\mathbf{Z}_1, \mathbf{0})$ as defined by Equation 2.6.1.
$\mathbf{p}_1$ = $n_1 \times 1$ vector of permanent environment effects (PE effects)
$n_1$ = number of animals with observations.

The first moment and second moments of Equation [2.7.1] are:

$$\begin{aligned}\mathbf{E(y)} &= \mathbf{Xb}\\ \mathbf{var(y)} &= \mathbf{Z}_1\text{var}(\mathbf{p})\mathbf{Z}_1' + \mathbf{Z}\text{var}(\mathbf{a})\mathbf{Z}' + \text{var}(\mathbf{e})\\ &= \sigma_p^2\mathbf{Z}_1\mathbf{Z}_1' + \sigma_a^2\mathbf{ZAZ}' + \sigma_e^2\mathbf{I}_N \qquad [2.7.2]\\ &= \sigma_p^2\mathbf{Z}_1\mathbf{Z}_1' + \sigma_a^2\mathbf{Z}_1\mathbf{A}_{11}\mathbf{Z}_1' + \sigma_e^2\mathbf{I}_N \qquad [2.7.3]\end{aligned}$$

where $\sigma_p^2$ = variance of PE effects, and $\mathbf{A}_{11}$ = additive relationship matrix for animals with observations, which is a submatrix of **A** defined by Equation [2.6.6]. The MME for Equation [2.7.1] are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_1 & 0 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1+\gamma\mathbf{I} & \mathbf{Z}_1'\mathbf{Z}_1 & 0 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1 & \mathbf{Z}_1'\mathbf{Z}_1+\lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{10} \\ 0 & 0 & \lambda\mathbf{A}^{01} & \lambda\mathbf{A}^{00} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{p}}_1 \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_0 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ 0 \end{bmatrix} \qquad [2.7.4]$$

where $\gamma=\sigma_e^2/\sigma_p^2$, $= \lambda=\sigma_e^2/\sigma_a^2$, and the **A** matrices with superscripts are defined by Equation [2.6.7]. Repeatability is defined as $r^2 = (\sigma_p^2+\sigma_a^2)/(\sigma_p^2+\sigma_a^2+\sigma_e^2)$. Predicted producing ability (PPA) predicts the individual's total merit of additive and PE effects and is calculated as:

$$\text{PPA} = \hat{\mathbf{p}}_1 + \hat{\mathbf{a}}_1 \qquad [2.7.5]$$

PPA is a measure of an individual's lifetime performance, and does not measure the genetic value to be transmitted to the next generation. The measure of the genetic value to be transmitted to the next generation is predicted transmitting ability (PTA), which is half of the individual's additive or breeding value, i.e.,

$$\text{PTA} = \frac{1}{2}\hat{\mathbf{a}} \qquad [2.7.6]$$

Equation [2.7.5] shows PPA is unavailable for animals without phenotypic observations, and Equation [2.7.6] shows PTA is available for all animals with and without phenotypic observations with the condition that correlated individuals exist.

The PPA of Equation [2.7.5] is from the mixed model with additive effects only. Since additive and nonadditive effects may all affect an individual's lifetime performance, the PPA based on genetic values can be defined:

$$\text{PPA} = \hat{\mathbf{p}}_1 + \hat{\mathbf{g}}_1 \qquad [2.7.7]$$

where $\hat{\mathbf{g}}_1$ = genotypic value as a summation of additive and nonadditive effects for individuals with phenotypic observations. Note that PTA of Equation [2.7.6] cannot be extended to include nonadditive effects because nonadditive effects such as dominance and epistasis effects do not transmit to the next generation.

**SAS program for BLUP without phenotypic records**

```
/* ANSC8141: REPEATED OBSERVATIONS */
/* program name: blup_pe.sas */
PROC IML;
RESET PRINT ;
Q = 4; * 4 ANIMALS;
P = 2; * 2 FIXED EFFECTS;
N = 9; * 9 OBERVATIONS, 2,3,4 OBS/ANIMAL;
/* additive relationship matrix */
A   = {1    0   .5   .5 ,
       0    1   .5  .5  ,
      .5   .5    1  .5  ,
       .5   .5   .5  1  };
IA = INV(A);
Y  = {11,13, 12,10,14, 9,13,11,15};  * individual 1 has no
observation;
X  = {1,2, 1,1,2, 1,1,2,2};   * X matrix for fixed effects;
dz ={1,1,2,2,2,3,3,3,3};      * assign observations to each
individual;
z1 = design(dz);
va = 2;
vp = 3;
ve = 7;
n1=3;
ID_pe = I(n1);
GP = ID_pe*vp;
GA = A*va;
G = block(GP,GA);
IG = inv(G);
IDN = I(N);
R = IDN*ve;
IR = inv(R);
imiss = j(N,1,0);     * 0 for individual 1 without observation in Z;
Z = Z1||imiss||Z1;    * note position of imiss;
* --- BLUP FROM MME ----;
XX = X`*IR*X; XZ = X`*IR*Z;
ZZ = Z`*IR*Z;
XY = X`*IR*Y;
ZY = Z`*IR*Y;
RHS = XY//ZY;
C_uu = ZZ + IG;
C1 = XX||XZ;
C2 = XZ`||C_uu;
C = C1//C2;
IC = GINV(C);
SOL = IC*RHS;
PE = sol(|2:4,*|);
BLUP1 = sol(|6:8,*|);
BLUP0 = sol(|5,*|);
PPA = PE + BLUP1;
```

```
* --- BLUP FROM CE ----;
A01 = A(|1,2:4|);
A11 = A(|2:4,2:4|);
IA11 = inv(A11);
V = Z*G*Z` + IDN*VE;
IV = INV(V);
XIVX = X`*IV*X;
XIVY = X`*IV*Y;
B_HAT = GINV(XIVX)*XIVY;
BLUP = G*Z`*IV*(Y-X*B_HAT);
BLUP_A1 = blup(|5:7,*|);
a0_a = a01*IA11*BLUP1;
a0_b = a01*IA11*BLUP_A1;
* --- BLUP OF E ----;
E_HAT1 = Y - X*B_HAT - Z*BLUP;
R = IDN*VE;
E_HAT2 = R*IV*(Y-X*B_HAT);
RUN;
```

**Exercises**

1) Translate all SAS programs in this chapter into R programs.

2) For the mixed model with additive values only, repeated observations, and individuals with missing phenotypic observations,

   $\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{p}_1 + \mathbf{Za} + \mathbf{e}$ Equation [2.7.1]

   Assume the number of individuals with phenotypic observations is 100, each individual has 2 observation and no individual has missing phenotypic observations. Define the following:
   a. The sizes of the **V** and **A** matrices.
   b. The size of the MME.
   c. The sizes of the $\mathbf{Z}_1$ and **Z** matrices.

## 2.8 Mixed Model with Multiple Genetic Factors: Multifactorial Mixed Model

Previous sections have two examples of multiple genetic factors: Equation [2.1.1] with additive and dominance effects, Equation [1.5.1] with additive, dominance and epistasis effects, and the repeatability model of Equation [2.7.1]. Such models with multiple random factors can be referred to as multifactorial mixed models. In general, a multifactorial mixed model can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{u}_1 + \ldots + \mathbf{Z}_k\mathbf{u}_k + \mathbf{e} = \mathbf{Xb} + \sum_{i=1}^{k}\mathbf{Z}_i\mathbf{u}_i + \mathbf{e} \quad [2.8.1]$$

$$= \mathbf{Xb} + (\mathbf{Z}_1, \ldots, \mathbf{Z}_k)(\mathbf{u}_1', \ldots, \mathbf{u}_k')' + \mathbf{e} \quad [2.8.2]$$

$$= = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \quad [2.8.3]$$

where

$$\mathbf{Z}=(\mathbf{Z}_1,...,\mathbf{Z}_k) \quad [2.8.4]$$

$$\mathbf{u}=(\mathbf{u}_1',...,\mathbf{u}_k')' \quad [2.8.5]$$

Under the assumption of unrelated genetic factors, the variance-covariance matrix of all genetic factors is:

$$\mathbf{G}=\text{var}(\mathbf{u})=\text{var}\begin{bmatrix}\mathbf{u}_1\\ \vdots\\ \mathbf{u}_k\end{bmatrix}=\begin{bmatrix}\text{var}(\mathbf{u}_1) & \vdots & \mathbf{0}\\ \vdots & \vdots & \vdots\\ \mathbf{0} & \vdots & \text{var}(\mathbf{u}_k)\end{bmatrix}=\begin{bmatrix}\mathbf{G}_1 & \vdots & \mathbf{0}\\ \vdots & \vdots & \vdots\\ \mathbf{0} & \vdots & \mathbf{G}_k\end{bmatrix}=\oplus_{i=1}^{k}\mathbf{G}_i \quad [2.8.6]$$

where '$\oplus$' denotes direct sum that defines a block diagonal matrix. With the definitions of Equations [2.9.1]-[2.9.6], the mathematical expectation and variance-covariance matrix are:

$$\text{E}(\mathbf{y})=\mathbf{Xb}, \qquad \text{E}\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix}=\mathbf{0} \quad [2.8.7]$$

$$\text{var}\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix}=\begin{pmatrix}\text{var}(\mathbf{y}) & \text{cov}(\mathbf{y},\,\mathbf{u}') & \text{cov}(\mathbf{y},\,\mathbf{e}')\\ \text{cov}(\mathbf{u},\,\mathbf{y}') & \text{var}(\mathbf{u}) & \text{cov}(\mathbf{u},\,\mathbf{e}')\\ \text{cov}(\mathbf{u},\,\mathbf{y}') & \text{cov}(\text{e},\,\mathbf{u}') & \text{var}(\mathbf{e})\end{pmatrix}=\begin{pmatrix}\mathbf{V} & \mathbf{ZG} & \mathbf{R}\\ \mathbf{GZ}' & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix} \quad [2.8.8]$$

Using Equations [2.8.1]-[2.8.8], BLUP of Equation [2.5.1] can be obtained using the CE method or the MME method. The general expression of the CE and MME methods for the multifactorial model of Equations [2.8.1]-[2.8.8] are:

$$\hat{\mathbf{u}}_i=\mathbf{G}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})=\mathbf{G}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}, \quad i=1,\ldots,k \quad [2.8.9]$$

$$\hat{\mathbf{b}}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad [2.2.7]$$

$$\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 & ... & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_k\\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1+\mathbf{G}_1^{-1} & ... & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_k\\ ... & ... & ... & ...\\ \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{Z}_1 & ... & \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{Z}_k+\mathbf{G}_k^{-1}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}_1\\ ...\\ \hat{\mathbf{u}}_k\end{bmatrix}=\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{y}\\ ...\\ \mathbf{Z}_k'\mathbf{R}^{-1}\mathbf{y}\end{bmatrix} \quad [2.8.10]$$

Although the CE method (Equations [2.8.9] and [2.2.7]) and the MME method (Equation [2.8.10]) have identical results, the computing difficulties of these two methods can differ drastically, as discussed next.

## 2.9 Discussions

**Computing difficulties of the CE and MME methods of BLUP**

The CE and MME methods each has advantages and disadvantaces, as summarized in the table blow.

| | CE method | MME method |
|---|---|---|
| Matrix size | • **V** defined by Equation [2.1.2] is a N×N matrix and is the largest matrix<br>• The size of **V** does not change for different models<br>• Repeated observations increase the size of the **V** matrix | • The coefficient matrix of MME defined by Equation [2.3.1] is the larges matrix<br>• Size of MME is a $(c+\sum_{i=1}^{k} n_i)$, c = number of fixed effects, $n_i$ = number of levels per genetic factor<br>• Size of MME can change for different genetic effects, e.g., sire effects can be much fewer than animal effects<br>• Size of MME increases as the number of genetic factors increases<br>• Repeated observations do not increase the size of MME |
| Relationship matrix | • Calculate but does not invert each relationship matrix | • invert each relationship matrix<br>• $\mathbf{A}^{-1}$ can be calculated based on pedigree information without the **A** matrix |
| Computing efficiency | • Best for mixed models with multiple genetic factors | • MME size can be made smaller than the **V** matrix, e.g., sire model, to become more efficient than CE |

**Invertibility of pedigree relationship matrices**

The **A** matrix cannot be inverted if identical twins are in the dataset. In this case, the MME method does not work unless the MME formulations are modified to avoid $\mathbf{A}^{-1}$. The CE method does not use $\mathbf{A}^{-1}$ and is appliable to data with identical twins.

**Properties of BLUP (example of additive values)**

BLUP has the following useful properties:

1) Invariance to choice of generalized inverse,
2) BLUP of $\mathbf{m'a} = \mathbf{m'\hat{a}}$, where **m** typically is a vector,
3) Sum of BLUP = zero if var(**a**) = $\mathbf{I}\sigma_a^2$, and sum of $\mathbf{A}^{-1}\hat{\mathbf{a}}$ = zero if var(**a**) = $\sigma_a^2\mathbf{A}$.

# CHAPTER 3. MULTIVAIATE MIXED MODEL

A multivariate mixed model is necessary for estimating genetic covariances and correlations among traits. For BLUP, a multivariate mixed model uses covariances among traits and leads to simultaneous prediction of genetic values of all traits. The use of covariances among traits potentially could increase the prediction accuracy for each trait over a single-trait prediction. This chapter only covers the MME method of BLUP for multivariate mixed models with additive values only.

## 3.1 Model Writing and Mixed Model Equations

**Mixed models in two orders of phenotypic observations**

A multivariate animal model with t traits is assumed to have c fixed effects per trait, n animals, and one observation per animal. Phenotypic observations can be arranged in two orders:

Order 1: Animals ordered by traits
Order 2: Traits ordered by animals.

For Order 1, the mixed model for a phenotypic observation can be written as:

$$y_{ijk} = b_{ij} + a_{ik} + e_{ijk},\ i = 1, ..., t;\ j = 1, ..., c;\ k = 1, ..., n. \quad [3.1.1]$$

where

$y_{ijk}$ = observation on trait i of animal k within fixed effect j
$b_{ij}$ = fixed effect j on trait i
$a_{ik}$ = additive genetic effect of animal k on trait i.
$e_{ijk}$ = residual of observation on trait i of animal k within fixed effect j.

For Order 2, the mixed model for a phenotypic observation can be written as:

$$y_{kji} = b_{ji} + a_{ki} + e_{kji},\ i = 1, ..., c;\ j = 1, ..., t;\ k = 1, ..., n. \quad [3.1.2]$$

where

$y_{kji}$ = observation on trait i of animal k within fixed effect j
$b_{ji}$ = fixed effect j of trait i
$a_{ik}$ = additive genetic effect of animal k on trait i.
$e_{kji}$ = residual of observation on trait i of animal k within fixed effect j

Note that subscripts i and k changed positions if the order of the phenotypic observation is changed to another order. This change of subscript positions is a typical difference between ‘animals ordered by traits’ and ‘traits ordered by animals’.

**Example 3.1.1**: Assume two traits observed on each animal, one observation per trait, and a total of four animals. The first two animals are in herd 1 and the last two animals in herd 2, where 'herd' is the fixed factor in the mixed model.

In the following table, the mixed models for the eight observations are given for '**animals ordered by traits**' using Equation [3.1.1] (left column) and for '**traits ordered by animals**' using Equation [3.1.2] (right column), where observations of trait 2 are shaded in yellow.

| **Order 1: Animals ordered by traits** | | **Order 2: Traits ordered by animals** | |
|---|---|---|---|
| t1-b1-a1: | $y_{111} = b_{11} + a_{11} + e_{111}$ | a1-b1-t1: | $y_{111} = b_{11} + a_{11} + e_{111}$ |
| t1-b1-a2: | $y_{112} = b_{11} + a_{12} + e_{112}$ | a1-b1-t2: | $y_{112} = b_{12} + a_{12} + e_{112}$ |
| t1-b2-a3: | $y_{123} = b_{12} + a_{13} + e_{123}$ | a2-b2-t1: | $y_{211} = b_{11} + a_{21} + e_{211}$ |
| t1-b2-a4: | $y_{124} = b_{12} + a_{14} + e_{124}$ | a2-b2-t2: | $y_{212} = b_{12} + a_{22} + e_{212}$ |
| t2-b1-a1: | $y_{211} = b_{21} + a_{21} + e_{211}$ | a3-b1-t1: | $y_{321} = b_{21} + a_{31} + e_{321}$ |
| t2-b1-a2: | $y_{212} = b_{21} + a_{22} + e_{212}$ | a3-b1-t2: | $y_{322} = b_{22} + a_{32} + e_{322}$ |
| t2-b2-a3: | $y_{223} = b_{22} + a_{23} + e_{223}$ | a4-b2-t1: | $y_{421} = b_{21} + a_{31} + e_{421}$ |
| t2-b2-a4: | $y_{224} = b_{22} + a_{24} + e_{224}$ | a4-b2-t2: | $y_{422} = b_{22} + a_{32} + e_{422}$ |

Using matrix notations, the mixed model for the eight observations in the order of 'animals ordered by traits' is:

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{123} \\ y_{124} \\ y_{211} \\ y_{212} \\ y_{223} \\ y_{224} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \\ b_{21} \\ b_{22} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{14} \\ a_{21} \\ a_{22} \\ a_{23} \\ a_{24} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{123} \\ e_{124} \\ e_{211} \\ e_{212} \\ e_{223} \\ e_{224} \end{bmatrix} \qquad [3.1.3]$$

and the mixed model for the eight observations in the order of 'traits ordered by animals' is:

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{211} \\ y_{212} \\ y_{321} \\ y_{322} \\ y_{421} \\ y_{422} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \\ b_{21} \\ b_{22} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \\ a_{31} \\ a_{32} \\ a_{41} \\ a_{42} \end{bmatrix} + \begin{bmatrix} e_{111} \\ e_{112} \\ e_{211} \\ e_{212} \\ e_{321} \\ e_{322} \\ e_{421} \\ e_{422} \end{bmatrix} \qquad [3.1.4]$$

Either Equation [3.1.3] or [3.1.4] can be described using the general mixed model of Equation [2.1.1], i.e., $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$. The $\mathbf{X}$ and $\mathbf{Z}$ matrices can be written in terms of the model matrices for each trait. In $\mathbf{X}$ and $\mathbf{Z}$, the model matrices for each trait are:

$$\mathbf{X}_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{Z}_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Then, Equation [3.1.3] of 'animals ordered by traits' can be written as:

$$\mathbf{y} = (\mathbf{I_t} \otimes \mathbf{X}_0)\mathbf{b} + (\mathbf{I}_t \otimes \mathbf{Z}_0)\mathbf{a} + \mathbf{e} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [3.1.5]$$

and Equation [3.1.4] of 'traits ordered by animals' can be written as:

$$\mathbf{y} = \mathbf{X}_0 \otimes \mathbf{I_t})\mathbf{b} + (\mathbf{Z}_0 \otimes \mathbf{I}_t)\mathbf{a} + \mathbf{e} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [3.1.6]$$

where $\mathbf{I}_t$ = t × t identity matrix, $\otimes$ denotes the Kronecker product, and

$\mathbf{X} = \mathbf{I}_t \otimes \mathbf{X}_0$ and $\mathbf{Z} = \mathbf{I}_t \otimes \mathbf{Z}_0$ for 'animals ordered by traits' [3.1.7]

$\mathbf{X} = \mathbf{X}_0 \otimes \mathbf{I}_t$ and $\mathbf{Z} = \mathbf{Z}_0 \otimes \mathbf{I}_t$ for 'traits ordered by animals' [3.1.8]

The first and second moments can be expressed using the general notations for Equation 3.1.1, i.e.,

$E(\mathbf{y}) = \mathbf{Xb}$ = t(n× 1) matrix
$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ = t(n × n) matrix

where

$\mathbf{G} = \text{var}(\mathbf{a})$

$= \mathbf{G}_0 \otimes \mathbf{A}$ for 'animals ordered by traits' = t(n × n) matrix [3.1.9]

$= \mathbf{A} \otimes \mathbf{G}_0$ for 'traits ordered by animals' = t(n × n) matrix [3.1.10]

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

$= \mathbf{R}_0 \otimes \mathbf{I}_n$ for 'animals ordered by traits' without missing trait [3.1.11]

$= \mathbf{I_n} \otimes \mathbf{R}_0$ for 'traits ordered by animals' without missing trait [3.1.12]

$$\mathbf{G}_0 = \begin{bmatrix} g_{11} & \cdots & g_{1t} \\ \cdots & \cdots & \cdots \\ g_{t1} & \cdots & g_{tt} \end{bmatrix} \qquad [3.1.13]$$

= t × t covariance matrix of additive effects with $g_{ij}$ as element ij

$$\mathbf{R}_0 = \begin{bmatrix} r_{11} & \cdots & r_{1t} \\ \cdots & \cdots & \cdots \\ r_{t1} & \cdots & r_{tt} \end{bmatrix} \qquad [3.1.14]$$

$= t \times t$ covariance matrix of residuals with $r_{ij}$ as element ij

$\mathbf{I}_n$ $= n \times n$ identity matrix.

The structure of **G** defined by Equations [3.1.9] and [3.1.10] applies whether or not missing data are present. However, the structure of **R** defined by Equations [3.1.11] and [3.1.12] does not apply when missing data are present.

**Example 3.1.2**: Structure of the **X**, **Z**, **G** and **R** matrices for Equations [3.1.5] and [3.1.6] with four individuals and two traits.

$$\mathbf{X} = \mathbf{I}_t \otimes \mathbf{X}_0 = \begin{bmatrix} \mathbf{X}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_0 \end{bmatrix}, \quad \mathbf{Z} = \mathbf{I}_t \otimes \mathbf{Z}_0 = \begin{bmatrix} \mathbf{Z}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_0 \end{bmatrix}$$

$$\mathbf{X} = \mathbf{X}_0 \otimes \mathbf{I}_t = \begin{bmatrix} \mathbf{I}_t & \mathbf{0} \\ \mathbf{I}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_t \\ \mathbf{0} & \mathbf{I}_t \end{bmatrix}, \quad \mathbf{Z} = \mathbf{Z}_0 \otimes \mathbf{I}_t = \begin{bmatrix} \mathbf{I}_t & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_t & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_t & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_t \end{bmatrix}$$

$$\mathbf{G} = G_0 \otimes \mathbf{A} = \begin{bmatrix} g_{11}\mathbf{A} & g_{12}\mathbf{A} \\ g_{21}\mathbf{A} & g_{22}\mathbf{A} \end{bmatrix}, \quad \mathbf{R} = R_0 \otimes \mathbf{I}_n = \begin{bmatrix} r_{11}\mathbf{I}_n & r_{12}\mathbf{I}_n \\ r_{21}\mathbf{I}_n & r_{22}\mathbf{I}_n \end{bmatrix}$$

$$\mathbf{G} = \mathbf{A} \otimes \mathbf{G}_0 = \begin{bmatrix} a_{11}\mathbf{G}_0 & a_{12}\mathbf{G}_0 & a_{13}\mathbf{G}_0 & a_{14}\mathbf{G}_0 \\ a_{21}\mathbf{G}_0 & a_{22}\mathbf{G}_0 & a_{23}\mathbf{G}_0 & a_{24}\mathbf{G}_0 \\ a_{31}\mathbf{G}_0 & a_{32}\mathbf{G}_0 & a_{33}\mathbf{G}_0 & a_{34}\mathbf{G}_0 \\ a_{41}\mathbf{G}_0 & a_{42}\mathbf{G}_0 & a_{43}\mathbf{G}_0 & a_{44}\mathbf{G}_0 \end{bmatrix}, \quad \mathbf{R} = \mathbf{I}_n \otimes \mathbf{R}_0 = \begin{bmatrix} \mathbf{R}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_0 \end{bmatrix}$$

**Multivariate mixed model equations**

The mixed model equations have the general expression of Equation [2.3.1]. The Kronecker product has the following properties useful for writing MME: $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD})$ for conforming matrices, $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$, and $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$. Applying these results to Equation [2.3.1], the MME for 'animals ordered by traits' are

$$\begin{pmatrix} \mathbf{R}_0^{-1} \otimes \mathbf{X}_0'\mathbf{X}_0 & \mathbf{R}_0^{-1} \otimes \mathbf{X}_0'\mathbf{Z}_0 \\ \mathbf{R}_0^{-1} \otimes \mathbf{Z}_0'\mathbf{X}_0 & \mathbf{R}_0^{-1} \otimes \mathbf{Z}_0'\mathbf{Z}_0 + \mathbf{G}_0^{-1} \otimes \mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} (\mathbf{R}_0^{-1} \otimes \mathbf{X}_0')\mathbf{y} \\ (\mathbf{R}_0^{-1} \otimes \mathbf{Z}_0')\mathbf{y} \end{pmatrix} \qquad [3.1.15]$$

and the MME for 'traits ordered by animals' are

$$\begin{pmatrix} \mathbf{X}_0'\mathbf{X}_0 \otimes \mathbf{R}_0^{-1} & \mathbf{X}_0'\mathbf{Z}_0 \otimes \mathbf{R}_0^{-1} \\ \mathbf{Z}_0'\mathbf{X}_0 \otimes \mathbf{R}_0^{-1} & \mathbf{Z}_0'\mathbf{Z}_0 \otimes \mathbf{R}_0^{-1} + \mathbf{A}_0^{-1} \otimes \mathbf{G}_0^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}_0' \otimes \mathbf{R}_0^{-1})\mathbf{y} \\ (\mathbf{Z}_0' \otimes \mathbf{R}_0^{-1})\mathbf{y} \end{pmatrix} \quad [3.1.16]$$

The size of the MME of either Equation [3.1.15] or [3.1.16] is t times as large as the single-trait MME. Two techniques provide some simplifications for computational efficiency as discussed after the numerical example and SAS program.

**Numerical example of multivariate BLUP from Equation [3.1.15]**

| |
|---|
| 34.814563 |
| 38.006815 |
| 11.667267 |
| 10.330131 |
| 2.644141 |
| 5.0296877 |
| -1.484297 |
| 0.8551706 |
| -0.927813 |
| 1.3236434 |
| -0.416275 |
| -0.183908 |
| -0.150626 |
| -0.27792 |
| -0.245152 |
| 0.5326797 |
| -0.469531 |
| 0.1812492 |
| -0.408074 |
| 0.4499998 |
| -0.130989 |

The $\hat{\mathbf{b}}$ vector (the top 6 elements) and $\hat{\mathbf{a}}$ vector (bottom 15 slelements in yellow) of the SAS output were from the MME of Equation [3.1.15] using the data in the SAS program below. These results serves as a comparison to those from canonical transformation to be discussed with the expectation that these two methods have identical results.

## SAS program for the MME method of multivariate BLUP

```
/* ANSC8141: MME method of multivariate BLUP */
/* animals ordered by traits */
/* program name: blup_mt.sas */
PROC IML;
RESET PRINT ;
Q = 5;  T = 3; P = 2;         N = Q;
G0 = { 2.02  .06   -.4,
       .06  1.83   -.03,
      -.4   -.03  .98 };
R0 = { 3.2   .77   -.8,
       .77   4.32  -.1,
       -.8  -.1   1.4};
Y1  = {32, 37, 35, 41, 38};
Y2  = {11, 12,  8, 10, 13};
Y3  = {2,   3,  4,  6,  5};
A   = {1    0   .5   0   .25 ,
       0    1   .5  .5   .25 ,
      .5   .5    1  .25  .5  ,
       0   .5   .25  1   .125,
      .25  .25  .5  .125  1};
DX0  = {1,1,2,2,2};
Y    = Y1//Y2//Y3;
IDQ   = I(Q);
IDT   = I(T);
X0   = DESIGN(DX0);
Z0   = IDQ;
X    = IDT@X0;
Z    = IDT@Z0;
*START;
*-------------------------------------------------------------;
IG0  = INV(G0);           IR0  = INV(R0);
IA   = INV(A);
*-------------------------------------------------------------;
IR0  = INV(R0);           IG0  = INV(G0);
IR   = IR0@IDQ;           IG   = IG0@IA;
XRX  = X`*IR*X;           XRZ  = X`*IR*Z;
ZRZ  = Z`*IR*Z;           ZRZG = ZRZ + IG;
XRY  = X`*IR*Y;           ZRY  = Z`*IR*Y;
*-------------------------------------------------------------;
C1   = XRX||XRZ;          C2   = XRZ`||ZRZG;
C    = C1//C2;            IC   = INV(C);
CY   = XRY//ZRY;          SOL  = IC*CY;
*-------------------------------------------------------------;
*FINISH;
RUN;
*PRINT SOL
```

## 3.2 Triangular and Canonical Transformations

A triangular transformation may transform the residual variance-covariance matrix of multiple traits into an identity matrix to simplify the mixed model equations if missing data have a specific pattern (Pollak, 1984). A canonical transformation can be used to transform correlated traits into uncorrelated canonical traits, so that mixed model equations for a multivariate model can be solved based on a set of mixed model equations for the uncorrlated canonical traits (Thompson, 1976; Meyer, 1985).

**A General Formula for Canonical and Triangular Transformations**

The mixed model is assumed to be that of Equation [3.1.6] for 'traits by animals' for triangular transformation and is assumed to be Equation [3.1.5] for 'animals by traits' for canonical transformation. .

Let $\mathbf{Q}_y$ be the transformation matrix for canonical or triangular transformation. Then, a canonical or triangular transformation for Equation [3.1.5] or [3.1.6] can be performed if:

$$\mathbf{Q}_y'\mathbf{y} = \mathbf{Q}_y'(\mathbf{Xb}) + \mathbf{Q}_y'(\mathbf{Za}) + \mathbf{Q}_y'\mathbf{e} \quad [3.2.1]$$
$$= \mathbf{X}(\mathbf{Q}_b'\mathbf{b}) + \mathbf{Z}(\mathbf{Q}_a'\mathbf{a}) + \mathbf{Q}_y'\mathbf{e} \quad [3.2.2]$$
$$= \mathbf{Xb}^* + \mathbf{Za}^* + \mathbf{e}^* \quad [3.2.3]$$

where $\mathbf{b}^* = \mathbf{Q}_b'\mathbf{b}$ , $\mathbf{a}^* = \mathbf{Q}_a'\mathbf{a}$ , $\mathbf{e}^* = \mathbf{Q}_y'\mathbf{e}$ , and the $\mathbf{Q}'$s are transformation matrices. Under the original model, estimation of fixed effects ($\hat{\mathbf{b}}$) and prediction of random effects ($\hat{\mathbf{a}}$ and $\hat{\mathbf{e}}$), can be obtained by:

$$\hat{\mathbf{b}} = \mathbf{Q}_b^{-1\prime}\hat{\mathbf{b}}^* \quad [3.2.4]$$
$$\hat{\mathbf{a}} = \mathbf{Q}_a^{-1\prime}\hat{\mathbf{a}}^* \quad [3.2.5]$$
$$\hat{\mathbf{e}} = \mathbf{Q}_y^{-1\prime}\hat{\mathbf{e}}^* \quad [3.2.6]$$

Equations [3.2.1]-[3.2.2] show that model matrices and the transformation matrix must satisfy the following condition:

$$\mathbf{Q}_y'\mathbf{X} = \mathbf{X}\mathbf{Q}_b' \quad [3.2.7]$$
$$\mathbf{Q}_y'\mathbf{Z} = \mathbf{Z}\mathbf{Q}_a'. \quad [3.2.8]$$

**Triangular Transformation**

Under the multivariate mixed model with missing data, a triangular transformation can transform the data so that the residual covariance matrix ($\mathbf{R}$) becomes an identity matrix. Such a transformation can be used if the missing data have an orderly pattern, i.e., if trait i is the last trait not missing, then none of the $(i-1)$ traits preceding it can be missing. When missing data have this orderly pattern, covariance matrices of residuals have only t possible structures. For example, the three structures for the residual covariance matrix for a record of three traits with none, one, or two of the three traits missing are:

$$\mathbf{R}_i = \mathrm{var}(\mathbf{e}_i) = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}$$ without missing data for any trait

$$= \begin{pmatrix} r_{11} & r_{12} & 0 \\ r_{21} & r_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{trait 3 is missing}$$

$$= \begin{pmatrix} r_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{traits 2 and 3 are missing}$$

In general, the structure of the residual covariance matrix for a record on the t traits can be written as:

$\mathbf{R}_j$ = the $\mathbf{R}_0$ with the last (t − j) rows and columns replaced by zeros, j = 1,..,t. [3.2.9]

The triangular transformation uses a triangular matrix that diagonalizes $\mathbf{R}_j$ and satisfies:

$\mathbf{R}_0 = \mathbf{L}'\mathbf{L}$ .

Let $\mathbf{L}_j$ be the j × j submatrix of $\mathbf{L}$ for the j trait with observations, $\mathbf{X}_j$ and $\mathbf{Z}_j$ be incidence matrices for a t traits of one individual, with the last (t − j) rows corresponding to missing traits set to zeros, j = 1,..,t, and let $\mathbf{L}_{j0}$ be the $\mathbf{L}$ matrix with the same rows set to zeros. Then, the $\mathbf{L}$ matrix has the following properties that make the triangular transformation possible:

1) If $\mathbf{L}$ is an upper triangular matrix, $\mathbf{L}^{-1}$ is also an upper triangular matrix
2) $\mathbf{L}_j^{-1}$ is the first j × j submatrix of $\mathbf{L}^{-1}$ corresponding to traits with observations
3) $(\mathbf{L}_{j0}')^{-1}\mathbf{X}_j = (\mathbf{L}')^{-1}\mathbf{X}_j = \mathbf{X}_j(\mathbf{L}')^{-1}$ and $(\mathbf{L}_{j0}')^{-1}\mathbf{Z}_j = (\mathbf{L}')^{-1}\mathbf{Z}_j = \mathbf{Z}_j(\mathbf{L}')^{-1}$
4) $\mathbf{L}(\mathbf{R}_j)^{-1}\mathbf{L}' = \mathbf{I}_j$

where $\mathbf{R}_j$ is defined by Equation 3.2.9, $\mathbf{I}_j$ = t × t identity matrix with the last (t − j) diagonal elements corresponding to missing traits set to zero. Assuming 'traits ordered by animals', the transformation matrix for a triangular transformation can be defined as:

$$\mathbf{Q}_y' = \mathbf{I}_q \otimes (\mathbf{L}')^{-1} = \mathbf{Q}_a'$$

Then, the triangular transformation on the data is:

$$\begin{aligned}
&\mathbf{Q}_y'\mathbf{y} = \mathbf{y}^* = \mathbf{Q}_y'(\mathbf{X}\mathbf{b}) + \mathbf{Q}_y'(\mathbf{Z}\mathbf{a}) + \mathbf{Q}_y'\mathbf{e} \\
&= [\otimes_{i=1}^{n} (\mathbf{L}')^{-1}\mathbf{X}_i]\mathbf{b} + [\otimes_{i=1}^{n} (\mathbf{L}')^{-1}\mathbf{Z}_i]\mathbf{a} + \mathbf{Q}_y'\mathbf{e} \\
&= [\otimes_{i=1}^{n} \mathbf{X}_i (\mathbf{L}')^{-1}]\mathbf{b} + [\otimes_{i=1}^{n} \mathbf{Z}_i (\mathbf{L}')^{-1}]\mathbf{a} + \mathbf{Q}_y'\mathbf{e} \\
&= (\otimes_{i=1}^{n} \mathbf{X}_i)[(\mathbf{L}')^{-1} \otimes \mathbf{I}_n]\mathbf{b} + (\otimes_{i=1}^{n} \mathbf{Z}_i)[(\mathbf{L}')^{-1} \otimes \mathbf{I}_n]\mathbf{a} + \mathbf{Q}_y'\mathbf{e} \\
&= \mathbf{X}(\mathbf{Q}_y'\mathbf{b}) + \mathbf{Z}(\mathbf{Q}_y'\mathbf{a}) + \mathbf{Q}_y'\mathbf{e} = \mathbf{X}\mathbf{b}^* + \mathbf{Z}\mathbf{a}^* + \mathbf{e}^* \qquad [3.2.10]
\end{aligned}$$

With the above transformation, the residual covariance matrix for the transformed data becomes an identity matrix. The mixed model equations for the triangular-transformed mixed model are:

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{B}^{-1} \otimes \mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}}^* \\ \hat{\mathbf{a}}^* \end{pmatrix} = \begin{pmatrix} \mathbf{X'y}^* \\ \mathbf{Z'y}^* \end{pmatrix}$$

where $\mathbf{b}^* = \mathbf{Q}_y'\mathbf{b}$, $\mathbf{a}^* = \mathbf{Q}_y'\mathbf{a}$, $\mathbf{y}^* = \mathbf{Q}_y'\mathbf{y}$, and

$$\mathbf{B} = \left(\mathbf{L}'\right)^{-1} \mathbf{G}_0 \mathbf{L}^{-1} \qquad [3.2.11]$$

.

**Canonical transformation**

The method of canonical transformation transforms the covariance matrices of residuals ($\mathbf{R}_0$) and additive genetic values ($\mathbf{G}_0$) into diagonal matrices. Then, the t traits are analyzed as t single traits and the results of the t single traits are transformed backed to the results of the multivariate analysis. The canonical transformation described below first factorizes $\mathbf{R}_0$ as $\mathbf{L}'\mathbf{L}$, the same as for the triangular transformation, and then compute eigenvalues and eigenvectors to factorize the **B** matrix of Equation [3.2.11] as:

$$\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}' \qquad [3.2.12]$$

where

$$\mathbf{U} = \text{orthogonal matrix formed by eigenvectors of } \mathbf{B} \qquad [3.2.13]$$
$$\mathbf{\Lambda} = \text{diagonal matrix of eigenvalues of } \mathbf{B} \qquad [3.2.14]$$

Define

$$\mathbf{M}' = \mathbf{U}'(\mathbf{L}')^{-1} . \qquad [3.2.15]$$

Then,

$$\mathbf{M}'\mathbf{R}_0\mathbf{M} = \mathbf{I}_t , \qquad \mathbf{M}'\mathbf{G}_0\mathbf{M} = \mathbf{\Lambda} \qquad [3.2.16]$$

Assuming animals are ordered by traits, the transformation matrices in Equation 3.2.10 for a canonical transformation are:

$$\mathbf{Q}_y' = \mathbf{M}' \otimes \mathbf{I}_n = \mathbf{Q}_a', \qquad \mathbf{Q}_b' = \mathbf{M}' \otimes \mathbf{I}_c \, \mathbf{Q}_b'$$

and the canonical transformation is:

$$\mathbf{y}^* = \mathbf{Q}_y'\mathbf{y} = \mathbf{X}(\mathbf{Q}_b'\mathbf{b}) + \mathbf{Z}(\mathbf{Q}_a'\mathbf{a}) + \mathbf{Q}_y'\mathbf{e}$$

with $\text{var}\left(\mathbf{Q}_a'\mathbf{a}\right) = \mathbf{\Lambda} \otimes \mathbf{A}$, and $\text{var}\left(\mathbf{Q}_y'\mathbf{e}\right) = \mathbf{I}_t \otimes \mathbf{I}_n = \mathbf{I}_{tn}$. The mixed model equations for trait i under the canonical transformation are:

$$\begin{pmatrix} \mathbf{X}_0'\mathbf{X}_0 & \mathbf{X}_0'\mathbf{Z}_0 \\ \mathbf{Z}_0'\mathbf{X}_0 & \mathbf{Z}_0'\mathbf{Z}_0 + \mathbf{A}^{-1}\lambda_i \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}}^* \\ \hat{\mathbf{a}}^* \end{pmatrix} = \begin{pmatrix} \mathbf{X}_0'\mathbf{y}^* \\ \mathbf{Z}_0'\mathbf{y}^* \end{pmatrix} \quad [3.2.17]$$

where $\lambda_i$ = eigenvalue i in $\mathbf{\Lambda}$ of Equation [3.2.14].

Prediction of random effects and estimation of fixed effects from the untransformed mixed model equations can be obtained from solutions from Equation [3.2.17]. Let

$$\hat{\mathbf{b}}^* = (\hat{b}_1^{*}{}',\ldots,\hat{b}_t^{*}{}')'$$
$$\hat{\mathbf{a}}^* = (\hat{a}_1^{*}{}',\ldots,\hat{a}_t^{*}{}')'$$
$$\hat{\mathbf{e}}^* = (\hat{e}_1^{*}{}',\ldots,\hat{e}_t^{*}{}')'$$

Then, solutions for fixed and random effects from the untransformed multivariate mixed model equations are obtained as:

$$\hat{\mathbf{b}} = (\mathbf{M}' \otimes \mathbf{I}_p)^{-1} \hat{\mathbf{b}}^* \quad [3.2.18]$$
$$\hat{\mathbf{a}} = (\mathbf{M}' \otimes \mathbf{I}_q)^{-1} \hat{\mathbf{a}}^* \quad [3.2.19]$$
$$\hat{\mathbf{e}} = (\mathbf{M}' \otimes \mathbf{I}_q)^{-1} \hat{\mathbf{e}}^* . \quad [3.2.20]$$

The numerical example below shows that the $\hat{\mathbf{b}}$ of Equation [3.2.18] and the $\hat{\mathbf{a}}$ of Equation [3.2.19] from canonical transformation are identical to those from the original MME of Equation [3.1.15].

**Numerical example of canonical transformation**

$$\hat{\mathbf{b}} = \begin{bmatrix} 34.814563 \\ 38.006815 \\ 11.667267 \\ 10.330131 \\ 2.644141 \\ 5.0296877 \end{bmatrix}, \quad \hat{\mathbf{a}} = \begin{bmatrix} -1.484297 \\ 0.8551706 \\ -0.927813 \\ 1.3236434 \\ -0.416275 \\ -0.183908 \\ -0.150626 \\ -0.27792 \\ -0.245152 \\ 0.5326797 \\ -0.469531 \\ 0.1812492 \\ -0.408074 \\ 0.4499998 \\ -0.130989 \end{bmatrix}$$

These results are identical to those from the original MME of Equation [3.1.15].

## SAS program for the multivariate MME method using canonical transformation

```
/* ANSC8141: MME method of multivariate BLUP */
/* canonical transformation */
/* program name: blup_ct.sas */
PROC IML;
RESET PRINT ;
Q = 5;  T = 3; P = 2;         N = Q;
G0 = { 2.02  .06   -.4,
       .06  1.83   -.03,
      -.4   -.03  .98 };
R0 = { 3.2   .77   -.8,
       .77   4.32  -.1,
       -.8  -.1   1.4};
Y1  = {32, 37, 35, 41, 38};
Y2  = {11, 12,  8, 10, 13};
Y3  = {2,   3,  4,  6,  5};
A   = {1    0   .5   0   .25 ,
       0    1   .5  .5   .25 ,
      .5   .5    1  .25  .5  ,
       0   .5   .25  1   .125,
      .25  .25  .5  .125  1};
DX0  = {1,1,2,2,2};
Y    = Y1//Y2//Y3;
IDQ   = I(Q);
IDT   = I(T);
X0   = DESIGN(DX0);
Z0   = IDQ;
X    = IDT@X0;
Z    = IDT@Z0;
DG0  = DET(G0);   DR0 = DET(R0);
*-------------------------------------------------------------;
LU   = ROOT(R0);          LL   = LU`;        ILL = INV(LL);
B    = ILL*G0*ILL`;
EVAL = EIGVAL(B);         EVEC = EIGVEC(B);
M    = ILL`*EVEC;
MG0M = M`*G0*M;           MR0M = M`*R0*M;
IA   = INV(A);
*-------------------------------------------------------------;
IR0  = INV(MR0M);         IG0  = INV(MG0M);
IR   = IR0@IDQ;           IG   = IG0@IA;
XRX  = X`*IR*X;           XRZ  = X`*IR*Z;
ZRZ  = Z`*IR*Z;           ZRZG = ZRZ + IG;
QY   = M`@IDQ;            YC   = QY*Y;
XRY  = X`*IR*YC;          ZRY  = Z`*IR*YC;
*-------------------------------------------------------------;
C1   = XRX||XRZ;          C2   = XRZ`||ZRZG;
C    = C1//C2;            IC   = INV(C);
CY   = XRY//ZRY;          SOL  = IC*CY;
NB   = T*P;               NA   = T*Q;
SOLB = SOL(|1:NB, |);     SOLA = SOL(|NB+1:NB+NA, |);
IDP  = I(P);              IQB  = INV(M)`@IDP;
IQA  = INV(M)`@IDQ;
BHAT = IQB*SOLB;          AHAT = IQA*SOLA;
*-------------------------------------------------------------;
RUN;
```

# CHAPTER 4: SELECTION INDEX

## 4.1 Selection Index Theory

### Three methods for selection of multiple traits

Three methods can be used for selection of multiple traits.

*Tandem Selection*: This method selects one trait per generation until the trait reached a satisfactory level, and then selects another trait. The advantage of this method is the maximum selection differential or intensity for the trait being selected at each generation. The disadvantage is potential undesirable correlated responses of correlated traits.

*Independent Culling*: This method sets a minimum level for each trait, and selects the individual with satisfactory levels for all traits. This method considers all traits in each generation so that the problem of undesirable correlated response could be addressed to some degree, but the economic gain of this selection method may not be optimal.

*Selection Index*: Selection index (Smith, 1936; Hazel, 1943). This method selects individuals according an index that maximizes the correlation between the index and the overall economic value of all traits. Theoretically, selection index is the best among the above three methods.

### Selection index

$$I = b_1y_1 + b_2y_2 + \cdots + b_ty_t = \text{selection index} \quad [4.1.1]$$

$$H = w_1g_1 + w_2g_2 + \cdots + w_tg_t = \text{aggregate genetic value (true merit of the t traits)} \quad [4.1.2]$$

where $b_i$ = coefficient of phenotypic value for trait i, $y_i$ = phenotypic value for trait i, and t = number of traits; $w_i$ = economic value of the genetic value for trait i, and $g_i$ = the genetic value of trait i. Selection index (I) is used to predict H. The weights of selection index, the $b_i$'s, are obtained by maximizing the correlation between I and H, or by minimizing the error variance of the index, an approach to be used below. In matrix notations, I and H can be expressed as:

$$I = \mathbf{b'y} \quad [4.1.3]$$

$$H = \mathbf{w'g} \quad [4.1.4]$$

where $\mathbf{b} = (b_1, \cdots, b_t)'$, $\mathbf{w} = (w_1, \cdots, w_t)'$ = $1 \times t$ column vector of economic values for the t traits, $\mathbf{y} = (y_1, \cdots, y_t)'$, $\mathbf{g} = (g_1, \cdots, g_t)'$. Then, $\mathbf{b}$ is obtained by maximizing the correlation between I and H or by minimizing the error variance of the index. The error variance of the index is:

$$\begin{aligned} \text{var}(I - H) &= \text{var}(\mathbf{b'y} - \mathbf{w'g}) = \mathbf{b'}[\text{var}(\mathbf{y})]\mathbf{b} - 2\mathbf{cov(b'y, g'w)} + \mathbf{w'}[\text{var}(\mathbf{g})]\mathbf{w} \\ &= \mathbf{b'Vb} - 2\mathbf{b'Gw} + \mathbf{w'Gw} \end{aligned}$$

where

$$\mathbf{V} = \text{var}(\mathbf{y}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1t} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{t1} & \sigma_{t2} & \cdots & \sigma_{tt} \end{bmatrix} \quad [4.1.5]$$

$$\mathbf{G} = \text{var}(\mathbf{g}) = \begin{bmatrix} \sigma_{(11)} & \sigma_{(12)} & \cdots & \sigma_{(1t)} \\ \sigma_{(21)} & \sigma_{(22)} & & \sigma_{(2t)} \\ & & & \\ \sigma_{(t1)} & \sigma_{(t2)} & & \sigma_{(tt)} \end{bmatrix} \quad [4.1.6]$$

and

$\sigma_{ij}$ = phenotypic covariance between traits i and j,

$\sigma_{(ij)}$ = genetic covariance between traits i and j.

The **V** matrix defined by Equation [4.1.5] and the **G** matrix defined by Equation [4.1.6] have the following important applications:

$r_{ij}$ = phenotypic correlation between traits i and j = $\sigma_{ij} / [\sigma_{ii}\ \sigma_{jj}]^{½}$

$r_{(ij)}$ = genetic correlation between traits i and j = $\sigma_{(ij)} / [\sigma_{(ii)}\ \sigma_{(jj)}]^{½}$

$h_i^2$ = heritability in the broad sense of trait i = $\sigma_{(ii)} / \sigma_{ii}$

Taking the partial derivative of var(I − H) and setting the partial derivative to zero lead to:

$$\frac{\partial \text{ var(I-H)}}{\partial \mathbf{b}} = \frac{\partial(\mathbf{b}'\mathbf{V}\mathbf{b} - 2\mathbf{b}'\mathbf{G}\mathbf{w} + \mathbf{w}'\mathbf{G}\mathbf{w})}{\partial \mathbf{b}} = \frac{\partial(\mathbf{b}'\mathbf{V}\mathbf{b} - 2\mathbf{b}'\mathbf{G}\mathbf{w})}{\partial \mathbf{b}} = 2\mathbf{V}\mathbf{b} - 2\mathbf{G}\mathbf{w} = \mathbf{0}$$

Solving the above equation for **b** yields:

$$\mathbf{b} = \mathbf{V}^{-1}\mathbf{G}\mathbf{w} \quad [4.1.7]$$

For a single trait, Equation [4.1.7] reduces to: $b = h^2 w$, and the selection index defined by Equation [4.1.1] reduces to:

$$I = (b)(y) = (h^2 w)(y) \propto y \quad [4.1.8]$$

where '$\propto$' stands for 'proportional to'.

**Accuracy of selection index**

The accuracy of the index is defined as the correlation between the index and the aggregate genetic value, denoted by $R_{IH}$. The variance of I ($\sigma_I^2$), the variance of H ($\sigma_H^2$) and the covariance between I and H ($\sigma_{IH}$) are:

$$\sigma_I^2 = \text{var}(I) = \text{var}(\mathbf{b}'\mathbf{y}) = \mathbf{b}'\mathbf{V}\mathbf{b} = \mathbf{b}'\mathbf{V}\mathbf{V}^{-1}\mathbf{G}\mathbf{w} = \mathbf{b}'\mathbf{G}\mathbf{w} \quad [4.1.9]$$

$$\sigma_H^2 = \text{var}(H) = \text{var}(\mathbf{w}'\mathbf{g}) = \mathbf{w}'\mathbf{G}\mathbf{w} \quad [4.1.10]$$

$$\sigma_{IH} = \text{cov}(I,H) = \text{cov}[\mathbf{b}'\mathbf{y}, (\mathbf{w}'\mathbf{g})'] = \mathbf{b}'\text{cov}(\mathbf{y}, \mathbf{g}')\mathbf{w} = \mathbf{b}'\mathbf{G}\mathbf{w} = \sigma_I^2 \quad [4.1.11]$$

$$R_{IH} = \frac{\text{cov}(I,H)}{\sqrt{\text{var}(I)\,\text{var}(H)}} = \frac{\sqrt{\text{var}(I)}}{\sqrt{\text{var}(H)}} = \frac{\sqrt{\sigma_I^2}}{\sqrt{\sigma_H^2}} = \frac{\sigma_I}{\sigma_H} \quad . \quad [4.1.12]$$

**Selection response of the aggregate genetic value**

Selection response of aggregate genetic value to index selection ($\Delta H$) is defined as the regression of the aggregate genetic value on the selection index:

$$\Delta H = b_{HI} S_I = b_{HI}(i_I)\sigma_I = (i_I)\sigma_I$$

where $b_{HI}$ = regression coefficient of H on I, $S_I$ = selection differential of I, $i_I$ = the selection intensity of $I = S_I/\sigma_I$.

**Note:** Selection index can be used for other purposes of selection. Two of such examples are:
1) Selection for one trait using information on several other traits,
2) Selection for the additive genetic value of an individual using information from relatives.

**Problem of selection index theory**

The selection index theory has a problem: the selection index theory does not have an approach to account for the effects of various factors on the phenotypic values or to estimate variance-covariance components for calculating selection index.

**4.2 Selection Index Based on BLUP**

The BLUP theory provides the most ideal approach to calculate selection index, because BLUP directly predicts the aggregate genetic value (H), and has a methodology to estimate non-genetic fixed effect using BLUE and to estimate variance-covariance components using REML or a similar method. The selection index of an individual calculated from BLUP can be expressed as:

$$\hat{H}_j = \mathbf{w}'\hat{\mathbf{g}}_j = \text{BLUP of } H_j = \mathbf{w}'(\text{BLUP of } \mathbf{g}_j) \quad [4.2.1]$$

where $\mathbf{w} = (w_1, \cdots, w_t)' = 1 \times t$ column vector of economic values for the t traits, $\hat{\mathbf{g}}_j$ = BLUP of genotypic values (e.g., breeding values) of t traits of individual j.

**Example 4.2.1: Net merit (NM$) – USDA's dairy cattle selection index**

USDA's genetic evaluation program calculates a selection index named Net Merit (NM$) as the total dollar value of all phenotypes of an individuals. The formula of NM$ is

$$NM\$ = \mathbf{w}'\hat{\mathbf{a}} \qquad [4.2.2]$$

where $\hat{\mathbf{a}}$ = predicted transmitting ability (PTA) = 0.5(BLUP of breeding value). The 2010 NM$ considered 10 traits, protein yield, fat yield, milk yield, productive life (PL), somatic cell score (SCS), udder, feet/legs, body size, daughter pregnancy rate (DPR), calving ability (CA$) (Table 4.2.1). As a linear function of the PTA for 10 traits, NM$ is a BLUP method.

**Table 4.2.1 Example of calculating NM$ for an individual**

| Trait | Units | Standard deviation (SD) | NM$ calculation | | |
|---|---|---|---|---|---|
| | | | $/PTA unit | PTA of the individual | $ × PTA |
| Protein | Pounds | 19 | 3.41 | +70 | 238.7 |
| Fat | Pounds | 27 | 2.89 | +80 | 231.2 |
| Milk | Pounds | 723 | 0.001 | +2,000 | 2 |
| PL | Months | 2.5 | 35 | +2.5 | 87.5 |
| SCS | Log | 0.23 | −182 | 2.95 (− 3.00) | 9.1 |
| Udder | Composite | 0.90 | 32 | +1.5 | 48 |
| Feet/legs | Composite | 1.03 | 15 | +0.5 | 7.5 |
| Body size | Composite | 1.03 | −23 | −1.0 | 23 |
| DPR | Percent | 1.70 | 27 | +0.3 | 8.1 |
| CA$ | Dollars | 20 | 1 | +30 | 30 |
| Total | | | | | **NM$ = $684.8** |

Source: http://aipl.arsusda.gov/reference/nmcalc.htm

# CHAPTER 5: PREDICTION ERROR VARIANCE AND RELIABILITY

Prediction error variance (PEV) is the variance of prediction errors, where each prediction error for an individual is the deviation of the BLUP from its true value and is a measure of prediction accuracy (the smaller the PEV, the better the prediction). PEV is used in some formulations for reliability of BLUP and REML estimation of variance and covariance components based on mixed model equations.

## 5.1 Variance-Covariance Matrix of BLUE and Prediction Errors Based on MME

This chapter assume the additive model of Equations [2.1.6] and [2.1.7]. Under this additive model, prediction errors are $\hat{\mathbf{a}} - \mathbf{a}$, and PEV is: $\text{PEV} = \text{var}(\hat{\mathbf{a}} - \mathbf{a})$. PEV for BLUP using MME is obtained as a function of the inverse of the coefficient matrix of the mixed model equations. The exact formula depends on how the mixed model equations are written, the original MME of Equation [2.3.1], or the simplified MME of Equation [2.4.8]. The coefficient matrix for the MME of Equation [2.3.1] or [2.4.8] is:

$$\mathbf{C} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix} \quad \text{for MME of Equation [2.3.1]} \qquad [2.3.3]$$

$$\mathbf{C} = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{pmatrix} \quad \text{for MME of Equation [2.4.8]} \qquad [5.1.1]$$

where $\lambda = \sigma_e^2/\sigma_a^2 = (1-h^2)/h^2$ (Equation [2.4.9]). Let the $\mathbf{C}$ matrix and its generalized inverse be partitioned as:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{bb} & \mathbf{C}_{ba} \\ \mathbf{C}_{ab} & \mathbf{C}_{aa} \end{pmatrix} \qquad [5.1.2]$$

$$\mathbf{C}^{-} = \begin{pmatrix} \mathbf{C}^{bb} & \mathbf{C}^{ba} \\ \mathbf{C}^{ab} & \mathbf{C}^{aa} \end{pmatrix} \qquad [5.1.3]$$

where $\mathbf{C}_{bb}$ and $\mathbf{C}^{bb}$ are c × c submatrices, $\mathbf{C}_{ba}$ and $\mathbf{C}^{ba}$ are c × n submatrices, $\mathbf{C}^{ab} = (\mathbf{C}^{ba})'$, and $\mathbf{C}_{aa}$ and $\mathbf{C}^{aa}$ are n × n submatrices. For the MME of Equation [2.3.1], the $\mathbf{C}$ is defined by Equation [2.3.3]. For the MME of Equation [2.4.8], the $\mathbf{C}$ is defined by Equation [5.1.1]:

The general expression of the variance of BLUE of $\mathbf{b}$, prediction errors and the covariance between $\mathbf{b}$ and $(\hat{\mathbf{a}} - \mathbf{a})$ is:

$$\text{var}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{a}}-\mathbf{a}\end{bmatrix}=\begin{pmatrix}\text{var}(\hat{\mathbf{b}}) & \text{cov}[\hat{\mathbf{b}},(\hat{\mathbf{a}}-\mathbf{a})']\\ \text{cov}[(\hat{\mathbf{a}}-\mathbf{a}),\hat{\mathbf{b}}'] & \text{var}(\hat{\mathbf{a}}-\mathbf{a})\end{pmatrix} \quad [5.1.4]$$

$$=\mathbf{C}^{-}=\begin{pmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba}\\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{pmatrix} \quad \text{if } \mathbf{C} \text{ is from Equation [2.3.3]} \quad [5.1.5]$$

$$=\mathbf{C}^{-}\sigma_e^2=\begin{pmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba}\\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{pmatrix}\sigma_e^2 \quad \text{if } \mathbf{C} \text{ is from Equation [5.1.1]} \quad [5.1.6]$$

From Equations [5.1.4]-[5.1.6], $\hat{\mathbf{a}}-\mathbf{a}$ = prediction errors, and the PEV matrix is:

$$\text{PEV} = \text{var}(\hat{\mathbf{a}}-\mathbf{a})$$

$$= \mathbf{C}^{aa} \quad \text{if } \mathbf{C} \text{ is the coefficient matrix of Equation [2.3.1]} \quad [5.1.7]$$

$$= \mathbf{C}^{aa}\sigma_e^2 \quad \text{if } \mathbf{C} \text{ is the coefficient matrix of Equation [2.4.8]} \quad [5.1.8]$$

The PEV for individual j is the $j^{th}$ diagonal element of Equation [5.1.7] or [5.1.8].

## 5.2 Variance-Covariance Matrix of BLUE and Prediction Errors Based on CE

Under the CE method, $\hat{\mathbf{a}}$ is calculated using Equation [2.4.11]:

$$\hat{\mathbf{a}}=\mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})=\mathbf{GZ}'\mathbf{P}\mathbf{y} \quad [2.4.11]$$

where $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$ (Equation 2.1.10), and $\mathbf{G}=\text{var}(\mathbf{a})=\sigma_a^2\mathbf{A}$. Note that

$$\mathbf{PVP}=\mathbf{P} \quad [5.1.9]$$

Using these result and Equation 2.4.11,

$$\text{var}(\hat{\mathbf{a}}) = \mathbf{GZ}'\mathbf{PZG} \quad [5.2.1]$$

$$\text{cov}(\hat{\mathbf{a}},\mathbf{a}') = \text{cov}\left(\mathbf{GZ}'\mathbf{P}\mathbf{y},\mathbf{a}'\right) = \mathbf{GZ}'\mathbf{PZG} = \text{cov}(\mathbf{a},\hat{\mathbf{a}}') = \text{var}(\hat{\mathbf{a}}) \quad [5.2.2]$$

$$\text{PEV} = \text{var}(\hat{\mathbf{a}}-\mathbf{a}) = \text{var}(\hat{\mathbf{a}}) - \text{cov}(\hat{\mathbf{a}},\mathbf{a}') - \text{cov}(\mathbf{a},\hat{\mathbf{a}}') + \text{var}(\mathbf{a})$$

$$= \text{var}(\mathbf{a}) - \text{var}(\hat{\mathbf{a}}) \quad [5.2.3]$$

$$= \mathbf{G} - \mathbf{GZ}'\mathbf{PZG} \quad [5.2.4]$$

$$\begin{aligned}\text{cov}(\hat{\mathbf{a}},\hat{\mathbf{b}}') &= \text{cov}[\mathbf{GZ}'\mathbf{P}\mathbf{y},\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}]\\ &= \mathbf{GZ}'\mathbf{P}\,\text{cov}(\mathbf{y},\mathbf{y}')\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\\ &= \mathbf{GZ}'\mathbf{PVV}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\\ &= \mathbf{GZ}'(\mathbf{PX})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}] = 0\end{aligned} \quad [5.2.5]$$

$$\begin{aligned}\text{cov}(\mathbf{a},\hat{\mathbf{b}}') &= \text{cov}[\mathbf{a},\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}] = \text{cov}(\mathbf{a},\mathbf{y}')\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} \\ &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}]\end{aligned} \quad [5.2.6]$$

$$\text{cov}(\hat{\mathbf{a}}-\mathbf{a},\hat{\mathbf{b}}') = -\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} \quad [5.2.7]$$

$$\text{var}(\hat{\mathbf{b}}) = \mathbf{C}^{bb} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} \quad [5.2.8]$$

## 5.3 Variance-Covariance Matrix of BLUE and Prediction Errors for CE and MME

Combining the CE and MME results, the variance-covariance matrix of BLUE and prediction errors are:

$$\text{var}\begin{bmatrix}\hat{\mathbf{b}} \\ \hat{\mathbf{a}}-\mathbf{a}\end{bmatrix} = \begin{pmatrix}\text{var}(\hat{\mathbf{b}}) & \text{cov}[\hat{\mathbf{b}},(\hat{\mathbf{a}}-\mathbf{a})'] \\ \text{cov}[(\hat{\mathbf{a}}-\mathbf{a}),\hat{\mathbf{b}}') & \text{var}(\hat{\mathbf{a}}-\mathbf{a})\end{pmatrix}$$

$$= \begin{pmatrix}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} & -(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \\ -\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} & \mathbf{G}-\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}\end{pmatrix} \quad [5.3.1]$$

$$= \mathbf{C}^{-} = \begin{pmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba} \\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{pmatrix} \quad \text{if } \mathbf{C} \text{ is from Equation [2.3.3]} \quad [5.1.5]$$

$$= \mathbf{C}^{-}\sigma_e^2 = \begin{pmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba} \\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{pmatrix}\sigma_e^2 \quad \text{if } \mathbf{C} \text{ is from Equation [5.1.1]} \quad [5.1.6]$$

Equations [5.1.4], [5.1.5] and [5.3.1] provide useful results of matrix algebra of mixed models for the variances and covariances of BLUE, BLUP and prediction errors of BLUP.

## 5.4 Accuracy and Reliability of BLUP

The ccuracy of BLUP for predicting genetic values is the correlation between the BLUP of the genetic values and the unknown true genetic values, and reliability of BLUP is the squared accuracy. The BLUP accuracy for predicting the additive value of the $i^{th}$ individual is:

$$\begin{aligned}R_i &= \text{cov}(\hat{a}_i, a_i') / \sqrt{\text{var}(\hat{a}_i)\,\text{var}(a_i)} \\ &= \text{cov}(\hat{\mathbf{a}},\mathbf{a}')_{ii} / \sqrt{\text{var}(\hat{\mathbf{a}})_{ii}\,\text{var}(\mathbf{a})_{ii}} \\ &= \sqrt{\text{var}(\hat{\mathbf{a}})_{ii} / \text{var}(\mathbf{a})_{ii}}\end{aligned} \quad [5.4.1]$$

The BLUP reliability for predicting the additive value of the $i^{th}$ individual is the squared accuracy:

$$R_i^2 = \text{var}(\hat{\mathbf{a}})_{ii} / \text{var}(\mathbf{a})_{ii} \qquad [5.4.2]$$

where $\text{var}(\hat{\mathbf{a}})_{ii}$ = the $i^{th}$ diagonal element in $\text{var}(\hat{\mathbf{a}}) = \mathbf{GZ'PZG}$ (Equation 5.2.1), $\text{var}(\mathbf{a})_{ii}$ = the $i^{th}$ diagonal element in $\mathbf{G} = \sigma_a^2\mathbf{A}$ for the additive model.

For the CE method, using $\mathbf{G} = \sigma_a^2\mathbf{A}$ and the result of $\text{var}(\hat{\mathbf{a}}) = \mathbf{GZ'PZG}$ (Equation 5.2.1), the reliability for the $i^{th}$ individual is:

$$R_i^2 = \sigma_a^2(\mathbf{AZ'PZA})_{ii}/A_{ii} \qquad [5.4.3]$$

where $A_{ii}$ = additive relationship of the $i^{th}$ individual = the $i^{th}$ diagonal element in the **A** matrix.

For the MME method, $\text{var}(\hat{\mathbf{a}} - \mathbf{a}) = \text{var}(\mathbf{a}) - \text{var}(\hat{\mathbf{a}})$ (Equation [5.2.3]. Rearranging Equation [5.2.3] leads to:

$$\text{var}(\hat{\mathbf{a}}) = \text{var}(\mathbf{a}) - \text{var}(\hat{\mathbf{a}} - \mathbf{a}) = \mathbf{G} - \text{PEV} \qquad [5.4.4]$$

$$= \mathbf{G} - \mathbf{C}^{aa} \qquad \text{if } \mathbf{C} \text{ is from Equation [2.3.3]} \qquad [5.4.5]$$

$$= \mathbf{G} - \mathbf{C}^{aa}\sigma_e^2 \qquad \text{if } \mathbf{C} \text{ is from Equation [5.1.1]} \qquad [5.4.6]$$

Using the results of Equations [5.4.4]-[5.4.6] and the reliability definition of Equation [5.4.2], the BLUP reliability using the MME method for the $i^{th}$ individual is:

$$R_i^2 = 1 - \text{PEV}_{ii}/(A_{ii}\sigma_a^2) \qquad [5.4.7]$$

$$= 1 - \mathbf{C}_{ii}^{aa}/(A_{ii}\sigma_a^2) \qquad \text{if } \mathbf{C} \text{ is from Equation [2.3.3]} \qquad [5.4.8]$$

$$= 1 - \lambda\mathbf{C}_{ii}^{aa}/A_{ii} \qquad \text{if } \mathbf{C} \text{ is from Equation [5.1.1]} \qquad [5.4.9]$$

where $\mathbf{C}_{ii}^{aa}$ is the $i^{th}$ diagonal element of Equation [5.1.6] if the MME are Equation [2.3.3], or is the $i^{th}$ diagonal element of Equation [5.1.7] if the MME are Equation [2.4.8].

**Exercises**

1) Prove **PVP = P**.
2) Prove $\text{var}(\hat{\mathbf{b}}) = (\mathbf{X'V}^{-1}\mathbf{X})^{-}$.

# CHAPTER 6: MAXIMUM LIKELIHOOD ESTIMATION OF VARIANCE-COVARIANCE COMPONENTS

Variance components are required for using mixed models and for estimating hertitability and repeatability. Maximum likelihood (ML) estimation (Hartley and Rao, 1967) is a likelihood based method for estimating variance and covariance components.

All mixed models including multifactor and multivariate models can be described by the general notations of Equations [2.1.1] and [2.1.2]:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad [2.1.1]$$

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R} \qquad [2.1.2]$$

with $E(\mathbf{y}) = \mathbf{Xb}$ . For estimating variance-covariance components using the ML method, the phenotypic observations are assumed to have a normal distribution, $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V})$.

## 6.1 Structure of Variance-Covariance Matrix as a Function of Partial Derivatives

Each variance-covariance matrix, such as **V**, **G** and **R**, is assumed to be a linear function of variance-covariance components and can be expressed as a linear function of partial derivatives with respect to each variance or covariance component. Such expressions are convenient for deriving formulations of computing algorithms for mixed models with multiple genetic factors and multiple traits. Assume k variance-covariance components for **V**, h variance-covariance components for **G**, and r variance components for **R** such that k = h + r. Then,

$$\mathbf{V} = \frac{\partial \mathbf{V}}{\partial \sigma_1^2}\sigma_1^2 + \cdots + \frac{\partial \mathbf{V}}{\partial \sigma_k^2}\sigma_k^2 = \sum_{j=1}^{k} \frac{\partial \mathbf{V}}{\partial \sigma_j^2}\sigma_j^2 = \sum_{j=1}^{k} \mathbf{V}_j^* \sigma_j^2 \qquad [6.1.1]$$

where $\sigma_j^2$ is a variance or covariance in **V**, and

$$\mathbf{V}_j^* = \frac{\partial \mathbf{V}}{\partial \sigma_j^2} \qquad [6.1.2]$$

To distinguish genetic variance-covariance components ($\theta_i$) form residual variance-covariance components ($\gamma_i$), the **G** and **R** matrices can be expressed as:

$$\mathbf{G} = \frac{\partial \mathbf{G}}{\partial \theta_1}\theta_1 + \cdots + \frac{\partial \mathbf{G}}{\partial \theta_h}\theta_h = \sum_{i=1}^{h} \frac{\partial \mathbf{G}}{\partial \theta_i}\theta_i = \sum_{i=1}^{h} \mathbf{G}_i^* \theta_i \qquad [6.1.3]$$

$$\mathbf{R} = \frac{\partial \mathbf{R}}{\partial \gamma_1}\gamma_1 + \cdots + \frac{\partial \mathbf{R}}{\partial \gamma_r}\gamma_r = \sum_{i=1}^{r} \frac{\partial \mathbf{R}}{\partial \gamma_i}\gamma_i = \sum_{i=1}^{r} \mathbf{R}_i^* \gamma_i \qquad [6.1.4]$$

where

$$\mathbf{G}_i^* = \frac{\partial \mathbf{G}}{\partial \theta_i} \qquad [6.1.5]$$

$$\mathbf{R}_i^* = \frac{\partial \mathbf{R}}{\partial \gamma_i}\gamma_i \qquad [6.1.6]$$

Equations [6.1.3]-[6.1.6] are useful for deriving general formulations for estimating variance-covariance components.

**Example 6.1.1**

A mixed model with additive and dominance values can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}(\mathbf{a} + \mathbf{d}) + \mathbf{e} \qquad [2.1.12]$$
$$\mathbf{V} = \sigma_a^2\mathbf{ZAZ}' + \sigma_d^2\mathbf{ZDZ}' + \sigma_e^2\mathbf{I} \qquad [2.1.13]$$

In terms of partial derivatives with respect to variance components, the **G** and **V** matrices can be expressed as:

$$\mathbf{G} = \text{var}\begin{bmatrix}\mathbf{a}\\ \mathbf{d}\end{bmatrix} = \begin{bmatrix}\text{var}(\mathbf{a}) & \mathbf{0}\\ \mathbf{0} & \text{var}(\mathbf{d})\end{bmatrix} = \begin{pmatrix}\sigma_a^2\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \sigma_d^2\mathbf{D}\end{pmatrix} \qquad [2.1.11]$$

$$= \begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\sigma_a^2 + \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{pmatrix}\sigma_d^2 \qquad [6.1.7]$$
$$= \mathbf{G}_a^*\sigma_a^2 + \mathbf{G}_d^*\sigma_d^2$$

$$\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}\,\mathbf{V} = \sigma_a^2\mathbf{ZAZ}' + \sigma_d^2\mathbf{ZDZ}' + \sigma_e^2\mathbf{I} \qquad [2.1.13]$$
$$= \mathbf{V}_a^*\sigma_a^2 + \mathbf{V}_d^*\sigma_d^2 + \mathbf{V}_e^*\sigma_e^2 \qquad [6.1.8]$$

where $\mathbf{G}_a^* = \begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}$, $\mathbf{G}_d^* = \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{pmatrix}$, $\mathbf{V}_a^* = \mathbf{ZAZ}'$, $\mathbf{V}_d^* = \mathbf{ZDZ}'$, and $\mathbf{V}_e^* = \mathbf{I}$.

**Example 6.1.2**

Assuming 'animals ordered by traits' for two traits, the **G** matrix (Equation 3.1.9) and **R** matrix (Equation 3.1.11) have the following structure:

$$\mathbf{G} = \text{var}\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{pmatrix} g_{11}\mathbf{A} & g_{12}\mathbf{A} \\ g_{21}\mathbf{A} & g_{22}\mathbf{A} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} g_{11} + \begin{pmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A} & \mathbf{0} \end{pmatrix} g_{12} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A} \end{pmatrix} g_{22}$$

$$= \mathbf{G}_{11}^{*} g_{11} + \mathbf{G}_{12}^{*} g_{12} + \mathbf{G}_{22}^{*} g_{22}$$

$$\mathbf{R} = \text{var}\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} == \begin{pmatrix} r_{11}\mathbf{I} & r_{12}\mathbf{I} \\ r_{21}\mathbf{I} & r_{22}\mathbf{I} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} r_{11} + \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix} r_{12} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} r_{22}$$

$$= \mathbf{R}_{11}^{*} r_{11} + \mathbf{R}_{12}^{*} r_{12} + \mathbf{R}_{22}^{*} r_{22}$$

## 6.2 ML Estimation of Variance-Covariance Components

### Likelihood function

The likelihood function for the mixed model of Equations [2.1.1] and [2.1.2] is:

$$f(\mathbf{V};\mathbf{y}) = \frac{1}{(2\pi)^{N/2} |\mathbf{V}|^{1/2}} e^{-(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})/2} \qquad [6.2.1]$$

The log likelihood function is:

$$L(\mathbf{V};\mathbf{y}) \propto -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}) \qquad [6.2.2]$$

### General iteration algorithm for ML estimation of variance-covariance components

Taking derivatives of Equation 6.2.2 with respect to the variance and covariance components and setting the resulting equations to zero yield the following general formula for estimating variance and covariance components:

$$\text{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_j^2}\right) = (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_j^2}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}) \qquad [6.2.3]$$

or,

$$\text{tr}\left(\mathbf{V}^{-1}\mathbf{V}_j^{*}\right) = (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}\mathbf{V}_j^{*}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}) \qquad [6.2.4]$$

The ML estimator of $\mathbf{b}$ is:

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad [2.2.7]$$

Substituting $\hat{\mathbf{b}}$ for $\mathbf{b}$ in Equation 6.2.4, $\sigma_j^2$ can be estimated by the following iterative solution:

$$\begin{aligned}\sigma_j^{2(i+1)} &= \sigma_j^{2(i)}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^{(i)})'\mathbf{V}^{-1(i)}\mathbf{V}_j^*\mathbf{V}^{-1(i)}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}^{(i)}) / \mathrm{tr}(\mathbf{V}^{-1(i)}\mathbf{V}_j^*) \\ &= \sigma_j^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{V}_j^*\mathbf{P}^{(i)}\mathbf{y} / \mathrm{tr}(\mathbf{V}^{-1(i)}\mathbf{V}_j^*)\end{aligned} \qquad [6.2.5]$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$ (Equation 2.2.8). Equation [6.2.5] is generally applicable to a wide range of mixed models including the multifactorial (Equation 2.8.1] and multivariate (Equations 3.1.5] and [3.1.6] models.

**Iterative algorithm for the additive model using A matrix**

For the additive model of Equations [2.1.6] and [2.1.7], $\mathbf{G} = \sigma_a^2\mathbf{A}$ (Equation 2.1.8), and $\mathbf{R} = \sigma_e^2\mathbf{I}_N$ (Equation 2.1.7), the two variance components can be estimated using the following iterative solutions:

$$\sigma_a^{2(i+1)} = \sigma_a^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{V}_j^*\mathbf{P}^{(i)}\mathbf{y} / \mathrm{tr}(\mathbf{V}^{-1(i)}\mathbf{Z}\mathbf{A}\mathbf{Z}') \qquad [6.2.6]$$

$$\sigma_e^{2(i+1)} = \sigma_e^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{P}^{(i)}\mathbf{y} / \mathrm{tr}(\mathbf{V}^{-1(i)}) \qquad [6.2.7]$$

Equations [6.2.5]-[6.2.7] do not calculate the BLUP of genetic effects but requires the calculation of the **A** matrix.

**Iterative algorithm for the additive model using $\mathbf{A^{-1}}$**

The **A** matrix using pedigree information is much more difficult to calculate than the $\mathbf{A}^{-1}$ using Henderson's simple method for calculating $\mathbf{A}^{-1}$ (Henderson, 1975). If the $\mathbf{A}^{-1}$ is to be used, Equations [6.2.6-6.2.7] can be expressed in terms of $\hat{\mathbf{a}}$. Using Equation [2.4.1],

$$\begin{aligned}&(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}) \\ &= [(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}]\mathbf{A}^{-1}[\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})] / (\sigma_a^2)^2 = \hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} / (\sigma_a^2)^2\end{aligned} \qquad [6.2.8]$$

Using the result of $\mathbf{G}\mathbf{Z}\mathbf{V}^{-1}$ of Equation [2.5.10],

$$\mathrm{tr}(\mathbf{V}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}') = \mathrm{tr}(\mathbf{Z}'\mathbf{G}\mathbf{Z}\mathbf{V}^{-1}) / \sigma_a^2 = \mathrm{tr}[\mathbf{Z}'(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}] / \sigma_a^2 \qquad [6.2.9]$$

Therefore, Equation [6.2.7] for the additive model can be expressed as:

$$\sigma_a^{2(i+1)} = \hat{\mathbf{a}}'^{(i)}\mathbf{A}^{-1}\hat{\mathbf{a}} / \mathrm{tr}\{\mathbf{Z}'[\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + (1/\sigma_a^2)\mathbf{A}^{-1}\mathbf{Z}'\mathbf{R}^{-1}\} \qquad [6.2.10]$$

Applying the result of Equation [2.2.5] to Equation [6.2.4] yields:

$$(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}) = (\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\mathbf{R}\mathbf{R}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}) / (\sigma_e^2)^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}} / (\sigma_e^2)^2 \qquad [6.2.11]$$

Using Equation [2.5.9], $\mathbf{V}^{-1}$ can be expressed in terms of $\mathbf{A}^{-1}$ so that $\text{tr}(\mathbf{V}^{-1})$ in Equation [6.2.7] can be expressed as:

$$\begin{aligned}
&\text{tr}(\mathbf{V}^{-1}) = \text{tr}(\mathbf{Z}\mathbf{G}\mathbf{Z}'+\mathbf{R})^{-1} = \text{tr}[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}] \\
&= \text{tr}[\mathbf{I}_N - \mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}'] / \sigma_e^2 \qquad [6.2.12] \\
&= \{N - \text{tr}[\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\} / \sigma_e^2
\end{aligned}$$

where $\lambda = \sigma_e^2 / \sigma_a^2$. Substituting the result of Equation [6.2.12] in Equation [6.2.7] yields:

$$\sigma_e^{2(i+1)} = \hat{\mathbf{e}}^{(i)}{}'\hat{\mathbf{e}}^{(i)} / \text{tr}\{N - \text{tr}[\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda^{(i)})^{-1}]\} \qquad [6.2.13]$$

From Equation 2.1.12a, $\hat{\mathbf{b}}$ in Equation 6.2.12 can be calculated as:

$$\hat{\mathbf{b}}^{(i+1)} = [\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}'\mathbf{X}]^{-}[\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda)^{-1}\mathbf{Z}'\mathbf{y}] \qquad [6.2.14]$$

Equations [6.2.10], [6.2.13] and [6.2.14] are the iterative solutions using $\mathbf{A}^{-1}$ from pedigree data. These equations yield identical results as those from Equations [6.2.6-6.2.7].

# CHAPTER 7: RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION OF VARIANCE-COVARIANCE COMPONENTS

Restricted (or residual) maximum likelihood estimation (REML) (Patterson and Thompson, 1971) has been a popular choice for estimating variance components. REML can use either the CE method of Equation [2.2.3] or the MME method of Equation [2.3.1]. REML formulations using CE and MME have the same results but have different computing properties, particularly for genomic prediction and estimation described in later chapters.

## 7.1 General REML Equations

**Residual likelihood function**

The phenotypic observations with the mixed model of Equations [2.1.1] and [2.1.2] are assumed to follow a normal distribution, i.e., $\mathbf{y} \sim N(\mathbf{Xb}, \mathbf{V})$. REML estimates of variance-covariance components are obtained by maximizing the residual likelihood of $\mathbf{K'y}$, where $\mathbf{K'}$ satisfies:

$$\mathbf{K'Xb} = \mathbf{0} \qquad [7.1.1]$$

With the requirement of Equation [7.1.1], $\mathbf{K'y}$ is unaffected by $\mathbf{b}$, and $\mathbf{K'y} \sim N(\mathbf{0}, \mathbf{K'VK})$. The residual likelihood function is:

$$f(\mathbf{V};\mathbf{y},\mathbf{K}) = \frac{1}{(2\pi)^{(N-r_x)/2} \,|\, \mathbf{K'VK} \,|^{1/2}} e^{-[\mathbf{y'K}(\mathbf{K'VK})^{-1}\mathbf{K'y}]/2} \qquad [7.1.2]$$

where $r_x$ = rank of $\mathbf{X}$.

The log residual likelihood function is:

$$L(\mathbf{V};\mathbf{y},\mathbf{K}) \propto -\frac{1}{2}\log|\,\mathbf{K'VK}\,| - \frac{1}{2}\mathbf{y'K}(\mathbf{K'VK})^{-1}\mathbf{K'y} \qquad [7.1.3]$$

An important result for deriving REML is:

$$\mathbf{P} = \mathbf{K}(\mathbf{K'VK})^{-1}\mathbf{K'} \qquad \text{(Searle, 1982)} \qquad [7.1.4]$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X'V}^{-1}\mathbf{X}\right)^{-}\mathbf{X'V}^{-1}$ (Equation [2.2.8]). Therefore, Equation [7.1.3] can be written as:

$$L(\mathbf{V};\mathbf{y},\mathbf{K}) \propto -\frac{1}{2}\log|\,\mathbf{K'VK}\,| - \frac{1}{2}\mathbf{y'Py} \qquad [7.1.5]$$

## General REML formulations

The derivation of REML formulations requires Equation [7.1.4] and two formulae of matrix derivatives in Appendix 1, including:

$$\frac{\partial(\mathbf{A}^{-1})}{\partial x} = -\mathbf{A}^{-1}\frac{\partial(\mathbf{A})}{\partial x}\mathbf{A}^{-1} \qquad [A1.15]$$

$$\frac{\partial(\log|\mathbf{A}|)}{\partial x} = \mathrm{tr}[\mathbf{A}^{-1}\frac{\partial(\mathbf{A})}{\partial x}] \qquad [A1.16]$$

Where the **A** matrix is a function of $x_j$, and is not the additive relationship matrix. Based on Equations [7.1.4] and [A1.15], the first derivative of **P** (Searle, 1982) with respect to $\sigma_j^2$ is:

$$\frac{\partial(\mathbf{P})}{\partial\sigma_j^2} = -\mathbf{P}\frac{\partial(\mathbf{V})}{\partial\sigma_j^2}\mathbf{P} = -\mathbf{P}\mathbf{V}_j^*\mathbf{P} \qquad [7.1.6]$$

where $\sigma_j^2$ is a variance or covariance in the **V** matrix defined by Equation [6.1.1], and $\mathbf{V}_j^*$ is defined by Equation [6.1.2]. The first derivative of $\mathbf{y}'\mathbf{P}\mathbf{y}$ is:

$$\frac{\partial(\mathbf{y}'\mathbf{P}\mathbf{y})}{\partial\sigma_j^2} = -\mathbf{y}'\frac{\partial(\mathbf{P})}{\partial\sigma_j^2}\mathbf{y} = -\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{y} \qquad [7.1.7]$$

The derivative of $\log|\mathbf{K}'\mathbf{V}\mathbf{K}|$ based on Equation [A1.16] is:

$$\frac{\partial(\log|\mathbf{K}'\mathbf{V}\mathbf{K}|)}{\partial\sigma_j^2} = \mathrm{tr}[(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\frac{\partial(\mathbf{K}'\mathbf{V}\mathbf{K})}{\partial\sigma_j^2}] = \mathrm{tr}[(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{V}_j^*\mathbf{K}] = \mathrm{tr}(\mathbf{P}\mathbf{V}_j^*) \qquad [7.1.8]$$

Combining Equations [7.1.6]-[7.1.8], the first derivative of the likelihood function of Equation [7.1.5] is:

$$\frac{\partial L(\mathbf{V};\mathbf{y},\mathbf{K})}{\partial\sigma_j^2} = -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{V}_j^*) + \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{y} \qquad [7.1.9]$$

Setting Equation [7.1.9] to zero, the REML estimate of $\sigma_j^2$ is obtained by solving the following equation:

$$\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{y} = \mathrm{tr}(\mathbf{P}\mathbf{V}_j^*) \qquad [7.1.10]$$

where $\mathbf{V}_j^*$ is defined by Equation [6.1.2]. From Equation [7.1.10], $\sigma_j^2$ can be estimated using the following EM type iterative algorithm:

$$\sigma_j^{2(i+1)} = \frac{\sigma_j^{2(i)} \mathbf{y}' \mathbf{P}^{(i)} \mathbf{V}_j^{*(i)} \mathbf{P}^{(i)} \mathbf{y}}{\text{tr}(\mathbf{P}^{(i)} \mathbf{V}_j^{*(i)})} \quad [7.1.12]$$

Equation [7.1.12] is a general REML formula applicable to a wide-range of mixed models including multiple genetic factors and multiple traits.

To calculate REML using MME, mathematical terms in Equation [7.1.12] needs to be converted into those from MME. The **P** matrix of Equation [2.2.8] can be expressed in terms of the MME results (Gilmour et al., 1995):

$$\mathbf{P} = \mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})\mathbf{C}^{-}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1} \quad [7.1.13]$$

where

$$\mathbf{C}^{-} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-} \quad [7.1.14]$$

General formulations for REML estimation of variance and covariance components using the MME of Equation [2.3.1] (Henderson, 1986) are:

$$\hat{\mathbf{u}}'\mathbf{Q}_i\hat{\mathbf{u}} = \text{tr}\left(\mathbf{Q}_i\mathbf{G}\right) - \text{tr}[\mathbf{Q}_i \text{var}(\hat{\mathbf{u}} - \mathbf{u})] \quad [7.1.15]$$

$$\hat{\mathbf{e}}'\mathbf{S}_i\hat{\mathbf{e}} = \text{tr}\left(\mathbf{S}_i\mathbf{R}\right) - \text{tr}\left(\mathbf{S}_i\mathbf{F}\mathbf{C}^{-}\mathbf{F}'\right) \quad [7.1.16]$$

where

$\hat{\mathbf{u}}$ = BLUP of all genetic effects in the mixed model of Equation [2.1.1]
$\mathbf{Q}_i = \mathbf{G}^{-1}\mathbf{G}_i^{*}\mathbf{G}^{-1}$
$\mathbf{G}$ = genetic variance-covariance matrix = var($\mathbf{u}$)
var($\hat{\mathbf{u}} - \mathbf{u}$)) = covariance matrix of prediction errors
$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}}$ = BLUP of random residuals (Equation [2.2.6]
$\mathbf{S}_i = \mathbf{R}^{-}\mathbf{R}_i^{*}\mathbf{R}^{-}$
$\mathbf{F}$ = ($\mathbf{X}$, $\mathbf{Z}$)
$\mathbf{R}^{-}$ = a generalized inverse of $\mathbf{R}$ allowing missing observations
$\mathbf{C}^{-}$ = a generalized inverse of $\mathbf{C}$ of the coefficient matrix for the mixed model equations
$\mathbf{G}_i^{*}$ = partial derivative of $\mathbf{G}$ with respect to $\theta_i$ defined by Equation [6.1.5]
$\mathbf{R}_i^{*}$ = partial derivative of $\mathbf{R}$ with respect to $\gamma_i$ defined by Equation [6.1.6].

An EM-REML iteration algorithm based on Equations [7.1.15] and [7.1.16] can be expressed as:

$$\sigma_j^{2(i+1)} = \frac{\sigma_j^{2(i)}\hat{\mathbf{u}}^{(i)\prime}\mathbf{Q}_j^{(i)}\hat{\mathbf{u}}^{(i)}}{\mathrm{tr}\{\mathbf{Q}_j^{(i)}[\mathbf{G} - \mathrm{var}(\hat{\mathbf{u}}^{(i)} - \mathbf{u})]\}} \quad [7.1.17]$$

$$\sigma_{rj}^{2(i+1)} = \frac{\sigma_{rj}^{2(i)}\hat{\mathbf{e}}^{(i)\prime}\mathbf{S}_j^{(i)}\hat{\mathbf{e}}^{(i)}}{\mathrm{tr}\{\mathbf{S}_j^{(i)}[\mathbf{R}^{(i)} - \mathbf{F}\mathbf{C}^{-(i)}\mathbf{F}']\}} \quad [7.1.18]$$

From Equations [7.1.17] and [7.1.18], formulations for estimating variance-covariance components for a given model using the MME method can be derived.

## 7.2 REML Using the CE and the MME Methods of BLUP for Additive Model

For the single traitadditive model of Equations [2.1.6] and [2.1.7] with additive and residual variances, The CE method for REML can be derived from Equations [7.1.12] and the MME method for REML can be derived from Equations 7.1.17-7.1.18.

The EM type of iteration algorithm using the CE method can be expressed as:

$$\sigma_a^{2(i+1)} = \frac{\sigma_a^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{P}^{(i)}\mathbf{y}}{\mathrm{tr}(\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}\mathbf{Z}')} \quad [7.2.1]$$

$$\sigma_e^{2(i+1)} = \frac{\sigma_e^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{P}^{(i)}\mathbf{y}}{\mathrm{tr}(\mathbf{P}^{(i)})} \quad [7.2.2]$$

Although Equations [7.2.1] and [7.2.2] do not need to calculate BLUP, these equations are referred as the CE method of REML estimation because of the use of the $\mathbf{V}^{-1}$ and $\mathbf{P}$ matrix rather than the MME and the use of the $\mathbf{A}$ matrix rather than the $\mathbf{A}^{-1}$.

The EM type iteration algorithm using MME method can be expressed as:

$$\sigma_a^{2(i+1)} = \frac{\hat{\mathbf{a}}^{(i)\prime}\mathbf{A}^{-1}\hat{\mathbf{a}}^{(i)}}{n - \mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}^{aa(i)})\lambda^{(i)}} \quad [7.2.3]$$

$$\sigma_e^{2(i+1)} = \frac{\hat{\mathbf{e}}^{(i)\prime}\hat{\mathbf{e}}^{(i)}}{N - [r - \mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}^{aa(i)})\lambda^{(i)}]} \quad [7.2.4]$$

where n = number of individuals, r = rank of $\mathbf{C}$, $\lambda = \sigma_e^2/\sigma_a^2 = (1-h^2)/h^2$ (Equation [2.4.9], and $\mathbf{C}^{aa}$ is form the $\mathbf{C}^-$ of the simplified MME (Equation 2.4.8):

$$\mathbf{C}^- = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{A}^{-1} \end{pmatrix}^- = \begin{pmatrix} \mathbf{C}^{bb} & \mathbf{C}^{ba} \\ \mathbf{C}^{ab} & \mathbf{C}^{aa} \end{pmatrix} \quad [7.2.5]$$

The CE set of Equations [7.2.1]-[7.2.2] and the MME set of Equations [7.2.3]-[7.2.4] have identical results at every iteration.

**Numerical example and SAS program**

```
/* ANSC8141: REML using MME method */
/* program name: reml_mme.sas */

PROC IML;
*RESET PRINT ;
VA=1    ;          VE=2    ;    * STARTING VALUES;
NB=2;    NU=5;   NREC=9;
NU0=NB+1;  NC=NB+NU;
TOL=.00000001;
Y  = {10, 6,8, 10,8, 5,7, 9,12};
Z  = {1 0 0 0 0,
      0 1 0 0 0,
      0 1 0 0 0,
      0 0 1 0 0,
      0 0 1 0 0,
      0 0 0 1 0,
      0 0 0 1 0,
      0 0 0 0 1,
      0 0 0 0 1};
DX = {1,1,1,1,1,2,2,2,2};
X = DESIGN(DX);
A  = {1    0   .5   0   .25 ,
      0    1   .5  .5   .25 ,
     .5   .5    1  .25  .5  ,
      0   .5   .25  1   .125,
     .25  .25  .5  .125  1};
IA=INV(A);
XX=X`*X; ZZ=Z`*Z;     XY=X`*Y;   ZY=Z`*Y;
CY=XY//ZY;
CBB=XX;      CBU=X`*Z;      CB=CBB||CBU;

K=0;         DIFA=100;          DIFE=100;
*----- ITERATION STARTS ---------------------------------------;
START;
DO WHILE(DIFA>TOL | DIFE>TOL & K <40);
RATIO=VE/VA;
CUU=ZZ+IA*RATIO;
CU=CBU`||CUU;
C=CB//CU;
IC=GINV(C);
rankc = trace(c*ic);
SOL=IC*CY;
BHAT=SOL(|1:NB,* |);
UHAT=SOL(|NU0:NC,*|);
EHAT = Y - X*BHAT - Z*UHAT;
```

```
ICUU = IC(|NU0:NC,NU0:NC|);
UIAU = UHAT`*IA*UHAT;
EE = EHAT`*EHAT;
TR = TRACE(IA*ICUU)*RATIO;
   REMLA = UIAU/(NU - TR);
   REMLE = EE/(NREC - rankc + TR);
DIFA=ABS(REMLA - VA);         DIFE=ABS(REMLE - VE);
VA=REMLA;         VE=REMLE;          K=K+1;
PRINT K VA VE DIFA DIFE;
*---- END OF ITERATIONS ------------------------------------;
END;
FINISH;
RUN;
```

# CHAPTER 8: NEWTON-RAPHSON, SCORING, AI-REML ALGORITHMS

The EM-type of iterative algorithms for estimation of variance-covariance components is known to be slow. The AI-REML algorithm (Gilmour et al., 1995; Johnson and Thompson, 1995; Lee and van der Werf, 2006) has been shown to be able to reduce the number of iterations significantly. The AI-REML algorithm involves the theory of the Newton-Raphson and scoring algorithms. In describing these algorithms, the phenotypic variance-covariance matrix (**V**) is assumed to be a linear function of variance-covariance components and has the general expression of Equation 6.1.1, i.e.,

$$\mathbf{V} = \frac{\partial \mathbf{V}}{\partial \sigma_1^2}\sigma_1^2 + \cdots + \frac{\partial \mathbf{V}}{\partial \sigma_k^2}\sigma_k^2 = \sum_{j=1}^{k} \frac{\partial \mathbf{V}}{\partial \sigma_j^2}\sigma_j^2 = \sum_{j=1}^{k} \mathbf{V}_j^* \sigma_j^2 \qquad [6.1.1]$$

The Newton-Raphson, scoring, and AI-REML algorithms will be described for the CE methods only. MME version of AI-REML can be established from the CE version of AI-REML, e.g., a MME version of AI-REML for genomic estimation of additive and dominance variances (Da et al., 2014).

## 8.1 The Newton-Raphson Algorithm

The Newton-Raphson algorithm for REML estimation can be formulated as:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \left(\mathbf{H}^{(i)}\right)^{-1} \boldsymbol{\Delta}^{(i)} \qquad [8.1.1]$$

where $\boldsymbol{\theta} = \{\sigma_1^2, \sigma_2^2, ..., \sigma_k^2\}'$ = k × 1 column vector of variance-covariance components, **H** = k × k Hessian matrix consisting of second derivatives of the REML log-likelihood of Equation 7.1.5, $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, ..., \Delta_k)'$ = k × 1 column vector of the partial derivatives of the log-likelihood function of Equation [7.1.3] with respect to each variance-covariance component, and i = iteration number. Note that '$-\mathbf{H}$' can be referred to as the 'observed information matrix' (Gilmour et al., 1995).

A typical term of the first derivative in $\boldsymbol{\Delta}$ is given by Equation [7.1.9]:

$$\Delta_j = \frac{\partial L(\mathbf{V}; \mathbf{y}, \mathbf{K})}{\partial \sigma_j^2} = -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{V}_j^*) + \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{y} \qquad [7.1.9]$$

where **P** is defined by Equation [2.2.8], and $\mathbf{V}_j^*$ is defined in Equation [6.1.2]. The second derivatives of the **H** matrix requires two formulae in Appendix 1:

$$\frac{\partial \mathbf{F}\mathbf{G}}{\partial x_j} = \mathbf{F}\frac{\partial(\mathbf{G})}{\partial x_j} + \frac{\partial(\mathbf{F})}{\partial x_j}\mathbf{G} \qquad \text{(Harville, 1997)} \qquad [A1.17]$$

$$\frac{\partial[\mathrm{tr}(\mathbf{F})]}{\partial x_j} = \mathrm{tr}\frac{\partial \mathbf{F}}{\partial x_j} \qquad [A1.18]$$

where **F** and **G** each is a function of $x_j$. Based on Equation [A1.17],

$$\frac{\partial(\mathbf{PV}_j^*\mathbf{P})}{\partial \sigma_k^2} = -2\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{P} \qquad [8.1.2]$$

Based on Equation [A1.18],

$$\frac{\partial[\mathrm{tr}(\mathbf{PV}_j^*)]}{\partial \sigma_k^2} = \mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) \qquad [8.1.3]$$

Combining the results of Eqauations [8.1.2] and [8.1.3], a typical term of the second derivatives in **H** of Equation [8.1.1] is:

$$\begin{aligned}\frac{\partial^2 L(\mathbf{V};\mathbf{y},\mathbf{K})}{\partial \sigma_j^2 \partial \sigma_k^2} &= \mathrm{tr}\{\frac{\partial[-\frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*)]}{\partial \sigma_k^2}\} + \frac{\partial(\frac{1}{2}\mathbf{y}'\mathbf{PV}_j^*\mathbf{Py})}{\partial \sigma_k^2} \\ &= \frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) - \mathbf{y}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{Py}\end{aligned} \qquad [8.1.4]$$

Substituiting Equations [7.1.9] and [8.1.4] in Equation [8.1.1] completes the Newton-Raphson algorithm for estimating variance-covariance components.

## 8.2 The Scoring Algorithm

The scoring algorithm uses the Fisher's information matrix in place of $-\mathbf{H}$ in the iterative algorithm of Equation 8.1.1. Fisher's information matrix, $\mathbf{I}(\boldsymbol{\theta})$, is the mathematical expectation of $-\mathbf{H}$, i.e.,

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}(\mathbf{H}) \qquad [8.2.1]$$

Fisher's information matrix is the approximate variance-covariance matrix of the estimates of variance-covariance components. The derivation of the mathematical expectation in Fisher's information matrix requires a formula of the mathematical expectation of a quadratic form in Appendix 1:

$$\mathrm{E}(\mathbf{x}'\mathbf{Ax}) = \mathrm{E}[\mathrm{tr}(\mathbf{Axx}')] = \mathrm{tr}[\mathbf{A}\mathrm{E}(\mathbf{xx}')] = \mathrm{tr}\{\mathbf{A}[\mathrm{var}(\mathbf{x}) + \mathrm{E}(\mathbf{x})\mathrm{E}(\mathbf{x})']\} \qquad [A1.6]$$

In addition to Equation [A1.6], the results of **PVP** = **P** (Equation [5.1.9], **PX** = **0** at the end of Section 2.2, E(**Py**) = **PXb = 0**, and Var(**Py**) = **PVP** = **P** are also needed. Applying these results and Equation [A1.6] to Equation [8.1.4], a typical term in the $\mathbf{I}(\boldsymbol{\theta})$ matrix is:

$$\begin{aligned}
&\mathrm{E}\left(-\frac{\partial^2 \mathrm{L}}{\partial\sigma_j^2\partial\sigma_k^2}\right) = -\mathrm{E}\left[-\frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) + \mathbf{y}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{Py}\right] \\
&= \frac{1}{2}\mathrm{E}[\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*)] - \mathrm{E}(\mathbf{y}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{Py}) \\
&= \frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) - \mathrm{tr}\{\mathbf{V}_j^*\mathbf{PV}_k^*[\mathrm{Var}(\mathbf{Py}) + \mathrm{E}(\mathbf{Py})\mathrm{E}(\mathbf{Py})']\} \\
&= \frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) - \mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) = -\frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*)
\end{aligned} \qquad [8.2.2]$$

An alternative proof for $\mathrm{E}(\mathbf{y}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{Py}) = \mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*)$ is:

$$\begin{aligned}
&\mathrm{E}(\mathbf{y}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{Py}) = \mathrm{tr}\{\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{P}[\mathrm{Var}(\mathbf{y}) + \mathrm{E}(\mathbf{y})\mathrm{E}(\mathbf{y})']\} \\
&= \mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{PV}) + \mathrm{tr}(\mathbf{b}'\mathbf{X}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{PXb}) = \mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*)
\end{aligned}$$

With Equations [8.2.1] and [8.2.2], the scoring algorithm for REML estimation is:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \left(\mathbf{I}(\boldsymbol{\theta})^{(i)}\right)^{-1}\boldsymbol{\Delta}^{(i)} \qquad [8.2.3]$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}$ are the same as in Equation [8.1.1]. The Fisher's information matrix, $\mathbf{I}(\boldsymbol{\theta})$, can be referred to as the 'expected information matrix'.

### 8.3 The AI-REML Algorithm

The average information REML (AI-REML) algorithm uses the average of the observed and expected information matrices in the iteration algorithm of Equation [8.1.1] and [8.2.3]:

$$\mathbf{AI} = \frac{1}{2}[\mathbf{I}(\boldsymbol{\theta}) - \mathbf{H}] = \frac{1}{2}[-\mathrm{E}(\mathbf{H}) - \mathbf{H}] \qquad [8.3.1]$$

A typical term in the $\mathbf{AI}$ matrix is:

$$\frac{1}{2}\left[-\mathrm{E}\left(\frac{\partial^2 \mathrm{L}(\mathbf{V};\mathbf{y},\mathbf{K})}{\partial\sigma_j^2\partial\sigma_k^2}\right) - \frac{\partial^2 \mathrm{L}(\mathbf{V};\mathbf{y},\mathbf{K})}{\partial\sigma_j^2\partial\sigma_k^2}\right] = \frac{1}{2}\mathbf{y}'\mathbf{PV}_j^*\mathbf{PV}_k^*\mathbf{Py} \qquad [8.3.2]$$

With Equations 8.3.1 and 8.3.2, the AI algorithm for REML estimation is:

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \left(\mathbf{AI}^{(i)}\right)^{-1}\boldsymbol{\Delta}^{(i)} \qquad [8.3.3]$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\Delta}$ are the same as in Equation 8.1.1. In Equation 8.3.3, the '½' in the **AI** and $\boldsymbol{\Delta}$ matrices cancels and can be dropped from the first and second derivatives of Equations 7.1.9 and 8.1.2. For estimating additive and residual variances using the CE method, the AI-REML algorithm is:

$$\begin{bmatrix}\sigma_a^2\\ \sigma_e^2\end{bmatrix}^{(i+1)} = \left(\begin{bmatrix}\sigma_a^2\\ \sigma_e^2\end{bmatrix} + \begin{bmatrix}\mathbf{y'PV_a^*PV_a^*Py} & \mathbf{y'PV_a^*PPy}\\ \mathbf{y'PPV_a^*Py} & \mathbf{y'PPPy}\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{tr}(\mathbf{PV_a^*}) - \mathbf{y'PV_a^*Py}\\ \mathrm{tr}(\mathbf{P}) - \mathbf{y'PPy}\end{bmatrix}\right)^{(i)} \qquad [8.3.4]$$

where $\mathbf{V}_a^* = \mathbf{ZAZ}'$. The AI-algorith of Equation [8.3.4] and the EM-REML of Equations [7.2.1]-[7.2.2] can be used jointly: use AI-REML first and switch to EM-REML when AI-REML fails.

## 8.4 Comparison between EM-REML and AI-REML

Numerical evaluations of EM-REML and AI-REML algorithms (Table 8.4.1) showed that AI-REML was much faster than EM-REML but had trouble with extreme heritability levels. EM-REML was slow but was reliable and converged in all cases including cases with null heritabilities, where AI-REML failed most of the time. When successful, AI-REML reduced the number of iteration of EM-REML by 93-98%. Therefore, a combination of EM and AI is ideal: use AI-REML when AI-REML produces positive estimates for variance components, and switch to EM-REML when AI-REML produced negative estimates of variance components (Wang et al., 2014).

Table 8.4.1 Comparison of iteration numbers of EM-REML and AI-REML (tolerance = $10^{-8}$) using simulated data with different heritability levels (Wang et al., 2014)

| Replication | $h_\alpha^2 = 0.0, h_\delta^2 = 0.0$ | | $h_\alpha^2 = 0.3, h_\delta^2 = 0.3$ | |
|---|---|---|---|---|
| | EM-REML | AI-REML | EM-REML | AI-REML |
| 1 | 173 | -[1] | 322 | 9 |
| 2 | 231 | - | 386 | 12 |
| 3 | 348 | - | 348 | 9 |
| 4 | 359 | - | 354 | 8 |
| 5 | 481 | 18 | 458 | 10 |
| 6 | 138 | - | 295 | 10 |
| 7 | 871 | - | 416 | 8 |
| 8 | 134 | - | 353 | 9 |
| 9 | 291 | 16 | 336 | 12 |
| 10 | 1000 | 1000[1] | 431 | 11 |

[1]AI-REML failed. $h_\alpha^2$ = genomic additive heritability, $h_\delta^2$ = genomic dominance heritability.

# CHAPTER 9: CONCEPTS OF GENOMIC SELECTION

Genomic prediction of an individual's genetic merit uses a large number of SNP markers covering the entire genome (Meuwissen et al., 2001). This approach has been shown to be more accurate than traditional genetic evaluation using pedigree information in numerous animal and plant studies, and has rapidly become widely accepted approach for genetic improvement in agricultural species. At the core of genomic selection is genomic evaluation using genome-wide SNP markers.

## 9.1 The Procedure of Genomic Selection

Genomic selection typically involves two types of populations: a training population and a validation population. The training population is also termed as reference population and consists of individuals with SNP data and phenotypic data, and provides genomic evaluation for both the training and validation populations. Validation population is also termed as selection candidates, consists of individuals with SNP data but without phenotypic data, and is the main interest of genomic selection (Figure 9.1.1).

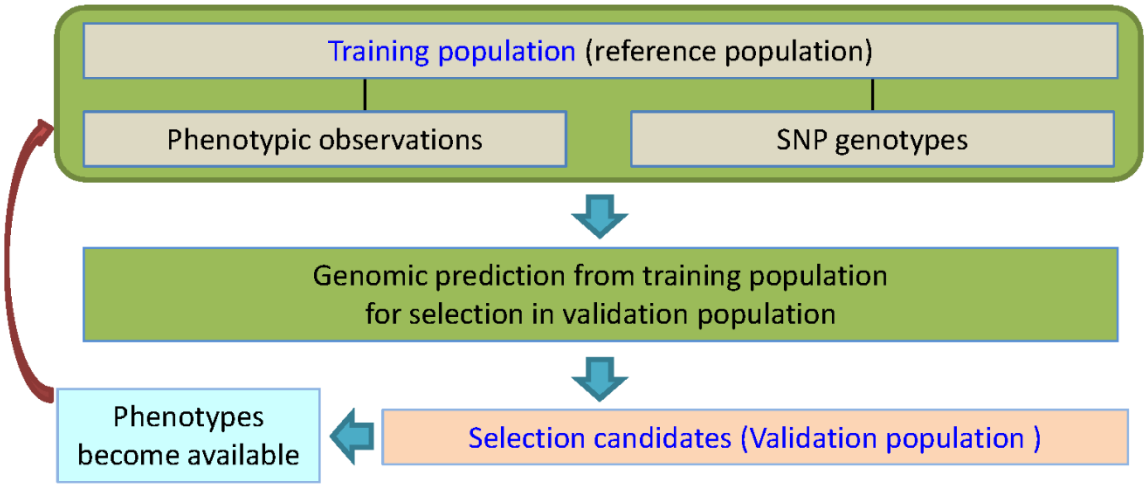



**Figure 9.1.1** Implementation and benefits of genomic selection.

## 9.2 Mechanism of Genomic Prediction

Genomic similarity between the training and validation populations is the fundamental mechanism of genomic prediction. The more similar between the training and validation populations, the more accurate the genomic prediction of the validation population based on SNP information from the training population.

The genomic similarity between the training and validation populations can be used by two methods: the method of relationships calculated from SNPs or the method of SNP similarity between the training and validation populations for predicting the validation population. Tables 9.2.1 and 9.2.2 are intuitive but not sicentic examples illustrating the two methods utilizing genmic similartity for genomic prediction of the validation population.

**Table 9.2.1** Intuitive illustration of the mechanism of genomic prediction of the validation population using relationships between the training and validation populations.

| | Cow 1 | Cow 2 | Cow 3 | Genomic prediction of validation population |
|---|---|---|---|---|
| Estimated genomic **additive value** from training population | 3 | 2 | 1 | |
| Relationship with Cow 4 | 1/2 | 1/4 | 1/8 | 17/8 |
| Relationship with Cow 5 | 1/8 | 1/4 | 1/2 | 11/8 |

**Table 9.2.2** Intuitive illustration of the mechanism of genomic prediction of the validation population using SNP similarity between the training and validation populations.

| | SNP 1 | SNP 2 | SNP 3 | Genomic prediction of validation population |
|---|---|---|---|---|
| Estimated **SNP effect** from training population | 3 | 2 | 1 | |
| Genotype coding of individual 1 | 2 | 1 | 0 | 8 |
| Genotype coding of individual 2 | 0 | 1 | 2 | 4 |

Table 9.2.1 is an intuitive example of predicting the validation population using the relationships between the training population (cows 1-3) and validation population (cows 4 and 5) for predicting the validation population (cows 4 and 5). Cow 4 had the highest relationship with cow 1 that had the highest estimated additive value, and had the lowest relationship with cow 3 that had the lowest additive value. The relationships of Cow 5 with the training population was the opposite of those of cow 4. Consequently, cow 4 had higher estimated genomic value than cow 5.

Table 9.2.2 is an intuitive example of predicting the validation population using the SNP similarity between the training and validation populations. The SNP genotype codings of individual 1 were in the same order as the estimated SNP additive values from the training population, whereas the SNP genotype codings of individual 2 were opposite to to the order of the SNP additive values from the training population. Consequently, individual 1 had higher estimated genomic value than that of individuals.

These two intuitive examples are for illustration of the mechanism of genomic prediction but are not rigorous theory of genomic prediction. However, the theory and methods of genomic prediction described in the next chapter utilize the the main idea of the two intuitive examples: relationship and SNP similarity between the training populations.The theory will show methods developed based relationship and SNP similarity in fact have identical results of genomic prediction.

# CHAPTER 10: QUANTITATIVE GENETICS MODEL FOR GENOMIC PREDICTION

Genomic prediction of an individual's genetic merit uses SNP markers covering the entire genome. Genomic BLUP (GBLUP) is the genomic version of BLUP and is widely used method for genomic prediction. This chapter describes the quantitative genetics (QG) model for genomic prediction that will be used for defining genomic relationships and for establishing the reparameterized QG model due to the use of genomic relationships for GBLUP.

## 10.1 Model and Assumptions

The QG model for genomic prediction is based on the single-locus genetic partion and the definitions of genetic values and variances for multiple loci in Chapter 1. Assuming n individuals with N observations and m SNPs, the QG model for genomic prediction of additive values can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZW}_{\alpha}\boldsymbol{\alpha}_{o} + \mathbf{e} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [10.1.1]$$

where

$\mathbf{y}$ = N × 1 vector of phenotypic observations

$\mathbf{Z}$ = N×n model matrix allocating phenotypic observations to SNP marker genotypes of individuals = identity matrix if N = n

$\mathbf{W}_{\alpha}$ = n×m model matrix for gene substitution effects of SNP markers

$\mathbf{X}$ = N×c model matrix for fixed non-genetic effects such as herd-year-season in dairy cattle,

$\mathbf{b}$ = c×1vector of fixed effects

$\boldsymbol{\alpha}_{o}$ = m×1 column vector of additive effects of m SNP markers [10.1.2]

$\mathbf{a} = \mathbf{W}_{\alpha}\boldsymbol{\alpha}_{o}$ = n×1 column vector of genomic additive values of n individuals [10.1.3]

$\mathbf{e}$ = N × 1 vector of random residuals

Subscript 'o' of $\boldsymbol{\alpha}_{o}$ indicates the additive effects are from the original genetic partion of the QG model, and will be dropped for the RQG model. Assumptions for the first and second moments of the QG model are:

$$E(\mathbf{y}) = \mathbf{Xb} \qquad [10.1.4]$$

$$\text{var}(\boldsymbol{\alpha}_{o}) = \mathbf{G}_{\alpha o} = \sigma^2_{\alpha o}\mathbf{I}_{m} \qquad [10.1.5]$$

$$\text{var}(\mathbf{a}) = \mathbf{G}_{a} = \text{Var}(\mathbf{W}_{\alpha}\boldsymbol{\alpha}_{o}) = \mathbf{W}_{\alpha}\text{Var}(\boldsymbol{\alpha}_{o})\mathbf{W}_{\alpha}' = \mathbf{W}_{\alpha}\mathbf{G}_{\alpha o}\mathbf{W}_{\alpha}' = \sigma^2_{\alpha o}\mathbf{W}_{\alpha}\mathbf{W}_{\alpha}' \qquad [10.1.6]$$

$$\text{var}(\mathbf{e}) = \mathbf{R} = \sigma^2_{e}\mathbf{I}_{N} \qquad [10.1.7]$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZG}_{a}\mathbf{Z}' + \mathbf{R} = \sigma^2_{\alpha o}\mathbf{ZW}_{\alpha}\mathbf{W}_{\alpha}'\mathbf{Z}' + \sigma^2_{e}\mathbf{I}_{N} \qquad [10.1.8]$$

where

$\sigma^2_{\alpha o}$ = common variance of additive effects of all SNPs [10.1.9]

$\mathbf{I}_{m}$ = m×m identity matrix

$\mathbf{G}_{\alpha o}$ = m×m variance-covariance matrix of additive SNP effects

$\mathbf{G}_{a}$ = n×n variance-covariance matrix of additive values

$\sigma_e^2$ = residual variance
$\mathbf{V}$ = phenotypic variance-covariance matrices
$\mathbf{I}_N$ = N×N identity matrix

Details of QG model are discussed below.

## 10.2 SNP Additive Effects, Model matrix, and Additive values

Detailed notations of the SNP additive effects of Eqaution [10.1.2], its model matrix, and additive values of n individuals are:

$$\boldsymbol{\alpha}_o = (\alpha_{o1},...,\alpha_{om})' \qquad [10.2.1]$$

$$\mathbf{W}_\alpha = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{21} & \cdots & w_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \qquad [10.2.2]$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} a_{1j} \\ \sum_{j=1}^{m} a_{2j} \\ \cdots \\ \sum_{j=1}^{m} a_{nj} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} w_{1j}\alpha_{oj} \\ \sum_{j=1}^{m} w_{2j}\alpha_{oj} \\ \cdots \\ \sum_{j=1}^{m} w_{nj}\alpha_{oj} \end{bmatrix} \qquad [10.2.3]$$

where $\alpha_{oj}$ = additive effect of the $j^{th}$ SNP, $w_{ij}$ = SNP addive coding for the $i^{th}$ individual and the $j^{th}$ SNP, and

$$a_{ij} = w_{ij}\alpha_{oj} = \text{addive value for the } i^{th} \text{ individual and the } j^{th} \text{ SNP} \qquad [10.2.4]$$

Equation [10.2.4] is the same as the additive values from the genetic partition of Equation [1.2.25]. The genomic additive value for the $i^{th}$ individual ($a_i$) in Equation [10.2.3] is the sum of the additive values of all SNPs. In the $\mathbf{W}_\alpha$ matrix of Equation [10.1.10], each element is a SNP coding based on Equation [1.2.7] from the signle-locus partition can be summarized as:

$w_{ij} = 2p$ for genotype 0 ($A_1A_1$) [10.2.5]
$= q - p$ for genotype 1 ($A_1A_2$) [10.2.6]
$= -2p$ for genotype 2 ($A_2A_2$) [10.2.7]

where $w_{ij}$ = the SNP addive coding for the $i^{th}$ individual and the $j^{th}$ SNP.

In a dataset of SNP genotype, the three genotypes of each SNP typically is represented by 0,1 and 2, where '0' and '2' are the two homozygous genotypes, and '1' is the heterozygous genotypes (Table 10.2.1).

**Table 10.2.1** Example of SNP data. '0' indicates one homozygous genotype. '2' indicates the other homozygous genotype. '1' indicates the heterozygous genotype. '5' indicates a missing SNP genotype.

| Animal | SNP1 | SNP2 | SNP3 | SNP4 | SNP5 | SNP6 | SNP7 | SNP8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 2 | 1 | 0 | 1 |
| 2 | 2 | 5 | 1 | 2 | 0 | 2 | 1 | 1 |
| 3 | 2 | 0 | 0 | 2 | 1 | 1 | 0 | 2 |
| 4 | 5 | 0 | 0 | 2 | 1 | 1 | 0 | 2 |
| 5 | 2 | 0 | 1 | 2 | 0 | 2 | 1 | 1 |
| 6 | 2 | 0 | 1 | 1 | 1 | 1 | 0 | 2 |
| 7 | 1 | 1 | 2 | 2 | 0 | 2 | 1 | 0 |
| 8 | 1 | 2 | 2 | 2 | 0 | 2 | 1 | 0 |
| 9 | 1 | 0 | 1 | 2 | 1 | 1 | 0 | 1 |
| 10 | 2 | 0 | 5 | 1 | 0 | 2 | 1 | 1 |

## 10.3 Centered SNP Coding and QG SNP Coding of the Model Matrix for Additive Effects

**Alternative SNP codings for $W_\alpha$ matrix**

In addition to the QG coding of Equations [10.2.5]-[10.2.7] for the $\mathbf{W}_\alpha$ matrix, other SNP coding were used in the literature including the the 1-0-(-1) or 2-1-0 coding and their 'centered coding' (Table 10.3.1). For the 1-0-(-1) coding, the mean of the three codes is p – q, and the mean of the 2-1-0 codes is 2p. The 'centered coding' subtracts the mean of the 1-0-(-1) or 2-1-0 coding from each code and results in the same coding as the QG coding.

**Table 10.3.1** Two methods of centered SNP coding and the QG coding for $\mathbf{W}_\alpha$ matrix.

| SNP genotype | $A_1A_1$ (0) | $A_1A_2$ (1) | $A_2A_2$ (2) | Mean |
|---|---|---|---|---|
| Genotypic frequency | $P_{11}=p^2$ | $P_{12}=2pq$ | $P_{22}=q^2$ | |
| 1-0-(–1) coding | 1 | 0 | –1 | p–q |
| Centered 1-0-(–1) coding | 1– (p–q) | 0– (p–q) | –1– (p–q) | 0 |
| | $= 2q = w_{\alpha 11}$ | $= q–p = w_{\alpha 12}$ | $= –2p = w_{\alpha 22}$ | |
| 2-1-0 coding | 2 | 1 | 0 | 2p |
| Centered 2-1-0 coding | 2– 2p | 1–2p | 0– 2p | 0 |
| | $= 2q = w_{\alpha 11}$ | $= q–p = w_{\alpha 12}$ | $= – 2p = w_{\alpha 22}$ | |
| QG coding | $w_{\alpha 11}= 2q$ | $w_{\alpha 11}= q–p$ | $w_{\alpha 22} = – 2p$ | 0 |

p = frequency of the $A_1$ allele, and q = frequency of the $A_2$ allele = 1 – p.

## Examples of $W_\alpha$ matrix

**Table 10.3.2** Example of SNP data for defining the $\mathbf{W}_\alpha$ matrix

| Individuals | Cow #1 | Cow #2 | Cow #3 | Cow #4 | Bull #1 | p | 2pq |
|---|---|---|---|---|---|---|---|
| Phenotypic observations | 1.7, 1.2, 1.3 | 2.1, 2.3 | 3.1 | 4.2, 4.3 | None | | |
| SNP #1 | AA, 0 | AA, 0 | AC, 1 | AA, 0 | AC, 1 | 4/5 | 8/25 |
| SNP #2 | GG, 2 | AA, 0 | AG, 1 | AG, 1 | AG, 1 | 1/2 | 1/2 |
| SNP #3 | AA, 0 | AA, 0 | AT, 1 | AA, 0 | AT, 1 | 4/5 | 8/25 |
| SNP #4 | CG, 1 | CG, 1 | GG, 2 | CC, 0 | CG, 1 | 1/2 | 1/2 |
| SNP #5 | CC, 0 | TT, 2 | CC, 0 | CT, 1 | CT, 1 | 3/5 | 12/25 |
| SNP #6 | TT, 2 | GG, 0 | GT, 1 | GT, 1 | GG, 0 | 3/5 | 12/25 |
| SNP #7 | AA, 0 | CC, 2 | CC, 2 | AA, 0 | AA, 0 | 2/5 | 12/25 |
| SNP #8 | AG, 1 | AA, 0 | GG, 2 | GG, 2 | AA, 0 | 1/2 | 1/2 |
| SNP #9 | AT, 1 | TT, 2 | TT, 2 | TT, 2 | AA, 0 | 3/10 | 42/100 |
| SNP #10 | GG, 2 | GG, 2 | CG, 1 | CG, 1 | CC, 0 | 2/5 | 12/25 |
| Sum | | | | | | | 448/100 |
| Average | | | | | | | 0.448 |

The allele with lower alphabetical order is considered as $A_1$ allele. For example, for CG SNP alleles, 'C' is $A_1$ and 'G' is $A_2$. $A_1A_1 = 0$, $A_1A_2 = 1$, $A_2A_2 = 2$. p = frequency of $A_1$ allele.

Table 10.3.2 is an example of the original SNP genotypes in A, C, G, T nucleotides, the corresponding 0, 1, or 2 code for each SNP genotype of 10 SNPs, and the phenotypic observations for 4 cows. One bull has genotype data but does not have phenotypic data. The frequency of '$A_1$' allele and the heterozygosity of each SNP are calculated in the last two columns.

Using the 1-0-(–1) coding in Table 10.3.1, the $\mathbf{W}_\alpha$ matrix is:

$$\mathbf{W}_\alpha = \begin{bmatrix} 1 & -1 & 1 & 0 & 1 & -1 & 1 & 0 & 0 & -1 \\ 1 & 1 & 1 & 0 & -1 & 1 & -1 & 1 & -1 & -1 \\ 0 & -1 & 0 & -1 & 1 & 0 & -1 & -1 & -1 & 0 \\ 1 & -1 & 1 & 1 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{matrix} \text{Cow \#1} \\ \text{Cow \#2} \\ \text{Cow \#3} \\ \text{Cow \#4} \\ \text{Bull \#1} \end{matrix}$$

Using the 2-1-0 coding in Table 10.3.1, the $\mathbf{W}_\alpha$ matrix is:

$$\mathbf{W}_\alpha = \begin{bmatrix} 2 & 0 & 2 & 1 & 2 & 0 & 2 & 1 & 1 & 0 \\ 2 & 2 & 2 & 1 & 0 & 2 & 0 & 2 & 0 & 0 \\ 1 & 1 & 1 & 0 & 2 & 1 & 0 & 0 & 0 & 1 \\ 2 & 1 & 2 & 2 & 1 & 1 & 2 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 2 & 2 & 2 & 2 \end{bmatrix} \begin{matrix} \text{Cow\#1} \\ \text{Cow\#2} \\ \text{Cow\#3} \\ \text{Cow\#4} \\ \text{Bull\#1} \end{matrix}$$

Using centered 1-0-(–1) or 2-1-0 codings or the QG coding in Table 10.3.1, the $\mathbf{W}_\alpha$ matrix is:

$$\mathbf{W}_\alpha = \begin{bmatrix} 2q & -2p & 2q & q-p & 2q & -2p & 2q & q-p & q-p & -2p \\ 2q & 2q & 2q & q-p & -2p & 2q & -2p & 2q & -2p & -2p \\ q-p & q-p & q-p & -2p & 2q & q-p & -2p & -2p & -2p & q-p \\ 2q & q-p & 2q & 2q & q-p & q-p & 2q & -2p & -2p & q-p \\ q-p & q-p & q-p & q-p & q-p & 2q & 2q & 2q & 2q & 2q \end{bmatrix} \begin{matrix} \text{Cow\#1} \\ \text{Cow\#2} \\ \text{Cow\#3} \\ \text{Cow\#4} \\ \text{Bull\#1} \end{matrix}$$

The QG coding is preferred because this coding has readily available interpretations of additive effect, value and variance, whereas such QG interpretations are unavailable for the 1-0-(–1) or 2-1-0 codings unless some assumptions are made. The centered 1-0-(–1) or 2-1-0 codings are equally good as the QG coding because they are identical to the QG coding although the centered codings do not have a theory leading to the QG interpretations of additive effect, value and variance.

## 10.4 The Variance-covariance Matrix of Additive Values

The variance-covariance matrix of additive effects of Equation [10.1.5] assumes the additive effects of different SNPs are independent of each other and have the same variance, and these assumptions results in a diagonal structure of the variance-covariance matrix:

$$\text{var}(\boldsymbol{\alpha}_o) = \mathbf{G}_{\alpha o} = \sigma^2_{\alpha o}\mathbf{I}_m = \begin{bmatrix} \sigma^2_{\alpha o} & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \sigma^2_{\alpha o} \end{bmatrix} \qquad [10.4.1]$$

Elements in the variance-covariance matrix of additive vlues of Equation [10.1.6] are:

$$\mathbf{G}_a = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}_\alpha' = \begin{bmatrix} \text{var}(a_1) & \text{cov}(a_1,a_2) & \cdots & \text{cov}(a_1,a_n) \\ \text{cov}(a_2,a_1) & \text{var}(a_2) & \cdots & \text{cov}(a_2,a_n) \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(a_n,a_2) & \text{cov}(a_n,a_2) & \cdots & \text{var}(a_n) \end{bmatrix} = \begin{bmatrix} G_a^{11} & G_a^{12} & \cdots & G_a^{1n} \\ G_a^{21} & G_a^{22} & \cdots & G_a^{2n} \\ \cdots & \cdots & \cdots & \cdots \\ G_a^{n1} & G_a^{n2} & \cdots & G_a^{nn} \end{bmatrix}$$

$$= \sigma^2_{\alpha o}\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix}_{n\times m}\begin{bmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ w_{1m} & w_{2m} & \cdots & w_{nm} \end{bmatrix}_{m\times n}$$

$$= \sigma^2_{\alpha o}\begin{bmatrix} \sum_{j=1}^m w_{1j}^2 & \sum_{j=1}^m w_{1j}w_{2j} & \cdots & \sum_{j=1}^m w_{1j}w_{nj} \\ \sum_{j=1}^m w_{2j}w_{1j} & \sum_{j=1}^m w_{2j}^2 & \cdots & \sum_{j=1}^m w_{2j}w_{nj} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{j=1}^m w_{nj}w_{1j} & \sum_{j=1}^m w_{nj}w_{2j} & \cdots & \sum_{j=1}^m w_{nj}^2 \end{bmatrix}_{n\times n} \qquad [10.4.2]$$

From Equation [10.4.2], a typical element in $\mathbf{G}_a$ is:

$$\text{cov}(a_i,a_k) = G_a^{ik} = \sigma^2_{\alpha o}\sum_{j=1}^m w_{ij}w_{kj} \qquad [10.4.3]$$

With the understanding the QG model leading to Equation [10.4.3], methods of genomic prediction can be formulated based on the QG model of Equation [10.1.1].

## 10.5 One Model, Two Versions, Four Methods of Prediction

The QG model of Equation [10.1.1] has two versions, QG1 and QG2:

QG1: $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$ [10.5.1]

$\mathbf{V} = \mathbf{ZG}_a\mathbf{Z}' + \mathbf{R} = \sigma^2_{\alpha o}\mathbf{ZW}_\alpha\mathbf{W}_\alpha'\mathbf{Z}' + \sigma^2_e\mathbf{I}_N$ [10.5.2]

QG2: $\mathbf{y} = \mathbf{Xb} + \mathbf{ZW}_\alpha\boldsymbol{\alpha}_o + \mathbf{e} = \mathbf{Xb} + (\mathbf{ZW}_\alpha)\boldsymbol{\alpha}_o + \mathbf{e} = \mathbf{Xb} + \mathbf{Z}_\alpha\boldsymbol{\alpha}_o + \mathbf{e}$ [10.5.3]

$\mathbf{V} = \mathbf{ZG}_a\mathbf{Z}' + \mathbf{R} = \sigma^2_{\alpha o}\mathbf{Z}_\alpha\mathbf{Z}_\alpha' + \sigma^2_e\mathbf{I}_N$ [10.5.4]

In Equations [10.5.3] and [10.5.4],

$$\mathbf{Z}_\alpha = \mathbf{ZW}_\alpha \qquad [10.5.5]$$

QG1 and QG2 are identical models but have differences in prediction methods. QG1 predicts additive values (Equation [10.1.3]) with $\mathbf{G}_a = \sigma^2_{\alpha o}\mathbf{Z}\mathbf{W}_\alpha\mathbf{W}_\alpha'$ (Equation 10.4.2]), whereas QG2 predicts additive effects (Equation [10.1.2]) with $\mathbf{G}_{\alpha o} = \sigma^2_{\alpha o}\mathbf{I}_m$ (Equation 10.1.5]).

Each version (QG1 or QG2) can use two methods for genomic prediction: the CE and MME methods, yielding a total of four methods for genomic prediction with identical results: QG1-CE, QG1-MME, QG2-CE and QG2-MME (Table 10.5.1).

**Table 10.5.1** Four methods of genomic prediction from the QG1 and QG2 models. (Shaded methods are not recommended for implementation)

| CE method | MME method |
|---|---|
| QG1: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e}$ [10.5.1], $\mathbf{V} = \sigma^2_{\alpha o}\mathbf{Z}\mathbf{W}_\alpha\mathbf{W}_\alpha'\mathbf{Z}' + \sigma^2_e\mathbf{I}_N$ [10.5.2] | |
| QG1-CE: most efficient when N<m, N=n<br><br>$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$<br>$\hat{\mathbf{a}} = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}_\alpha'\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$ | QG1-MME: can be efficient when n<m, N>n<br><br>$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \lambda_{\alpha o}(\mathbf{W}_\alpha\mathbf{W}_\alpha')^{-1}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}} \\ \hat{\mathbf{a}}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y}\end{pmatrix}$<br>$\lambda_{\alpha o} = \sigma^2_e / \sigma^2_{\alpha o}$ |
| QG2: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_\alpha\boldsymbol{\alpha} + \mathbf{e}$ [10.5.3], $\mathbf{V} = \sigma^2_{\alpha o}\mathbf{Z}_\alpha\mathbf{Z}_\alpha' + \sigma^2_e\mathbf{I}_N$ [10.5.4] | |
| QG2-CE: no computing advantage<br><br>$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$<br>$\hat{\boldsymbol{\alpha}}_o = \sigma^2_{\alpha o}\mathbf{Z}_\alpha'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$<br>$\hat{\mathbf{a}} = \mathbf{W}_\alpha\hat{\boldsymbol{\alpha}}_o$ | QG2-MME: generally is most efficient for n>m<br><br>$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_\alpha \\ \mathbf{Z}_\alpha'\mathbf{X} & \mathbf{Z}_\alpha'\mathbf{Z}_\alpha + \lambda_\alpha\mathbf{I}_m\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}} \\ \hat{\boldsymbol{\alpha}}\end{pmatrix} = \begin{pmatrix}\mathbf{X}'\mathbf{y} \\ \mathbf{Z}_\alpha'\mathbf{y}\end{pmatrix}$<br>$\mathbf{Z}_\alpha = \mathbf{Z}\mathbf{W}_\alpha,\ \hat{\mathbf{a}} = \mathbf{W}_\alpha\hat{\boldsymbol{\alpha}}_o$ |

QG1-CE is the best when each individual has one observation (N = n) and the number of observations is less than the number of SNPs (N<m).

QG1-MME can be more efficient than QG1-CE when the number of individuals is less than the number of SNPs (n<m) and repeated observations per individual exist (N>n). However, QG1-MME is not discussed further because the current QG model does not include the permanent environment effects for repeated observations. For repeated observations, QG1-MME could be computationally competitive if the number of individuals is manageable. QG1-MME is the only method requiring the inversion of the $\mathbf{W}_\alpha\mathbf{W}_\alpha'$ matrix, which is nonivertible if the number of individuals is greater than the number of SNPs. This is a unique negative aspect of QG1-MME.

QG2-CE has no computing advantage and is not recommended for implementation.

QG2-MME is the best when the number of observations is greater than the number of SNPs (N>m).

The above analysis leads to the conclusion that QG1-CE and QG2-MME are two complementary prediction methods, with QG1-CE for small number of individuals and large number of SNPs, and QG2-MME for small number of SNPs and large number of individuals.

# CHAPTER 11: GENOMIC RELATIONSHIP MATRIX

GBLUP uses genomic relationship matrix in place of the pedigree relationship matrix for BLUP. This chapter focuses on genomic additive relationship matrix and provide brief descriptions of other genomic relationship matrices.

## 11.1 Formulations of Genomic Additive Relationship Matrix

Genomic additive relationship matrix (VanRaden, 2008) is defined based on model matrix of the additive effects in the QG model of Equation [10.1.1]:

$$\mathbf{A}_g = \mathbf{W}_\alpha \mathbf{W}_\alpha' / k_1 \qquad [11.1.1]$$

$$k_1 = 2\sum_{i=1}^{m} p_i (1-p_i) = \text{total heterozygosity of m SNPs} \qquad [11.1.2]$$

The derivation used standardized (-1)-0-1 coding by dividing the centered codes by the standard deviation of the SNP coding (Table 10.1.2). This can be termed as 'across SNP standardization' because the same standard deviation of the SNP coding is used for all SNPs. VanRaden (2008) also proposed the standardization using the standard deviation of each SNP rather than a common standard deviation, a method that can be termed as 'within SNP standardization'.

A second definition of genomic additive relationship matrix uses a different denominator (Hayes and Goddard, 2010):

$$\mathbf{A}_g = \mathbf{W}_\alpha \mathbf{W}_\alpha' / k_2 \qquad [11.1.3]$$

$$k_2 = \text{tr}(\mathbf{W}_\alpha \mathbf{W}_\alpha')/n = \text{average of the diagonal elements of } \mathbf{W}_\alpha \mathbf{W}_\alpha' \qquad [11.1.4]$$

Numerical showed the two definitions of genomic additive relationships had similar results in a swine sample and a dairy cattle sample (Wang and Da, 2014). However, these two definitions have differences.

Definition I of Euations [11.1.1] and [11.1.2] is the genomic version of the pedigree additive relationships, and the additive variance estimated using Definition I is the total genomic additive variance of all SNPs. Definition II of Equations [11.1.3] and [1.1.4] is not the the genomic version of the pedigree additive relationships, and the additive variance estimated using Definition II is the the average of the additive variances of all individuals.

Both definitions have the same accuracy of genomic prediction of the QG methods, and the differences between Definitions I and II tend to diminish in populations without inbreeding.

The $\mathbf{A}_g$ matrix is noninvertible if the number of individuals is greater than the numbers of SNP (n>m), same condition as for the $\mathbf{W}_\alpha \mathbf{W}_\alpha'$. Presence of duplicated individuals or identical twince results in singular genomic and pedigree additive relationship matrices.

Although why standardization of SNP coding leads to genomic additive relationships was unclear, an alternative derivation and numerical results confirm that the genomic additive relationship matrix of Equations [11.1.1] and [11.1.2] is the genomic version of the pedigree additive relationship matrix.

## 11.2 Alternative Derivation of Genomic Additive Relationship Matrix

An alternative derivation to standardization of SNP coding is set equal the genomic variance-covariance of additive values defined by Equation [10.1.6] and the pedigree variance-covariance of additive values defined by Equation [2.1.8]:

$$\mathbf{G}_a = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}_\alpha' = \sigma^2_a\mathbf{A} \qquad [11.2.1]$$

From Equation [11.2.1],

$$\mathbf{A} = (\sigma^2_{\alpha o}/\sigma^2_a)\mathbf{W}_\alpha\mathbf{W}_\alpha' \qquad [11.2.2]$$

In Equation [11.2.2], $\sigma^2_a$ is the total additive variance of m SNPs:

$$\sigma^2_a = \sum_{i=1}^{m}\sigma^2_{ai} = 2\sum_{i=1}^{m} p_i(1-p_i)\alpha^2_{oi} \qquad [11.2.3]$$

where

$$\sigma^2_{ai} = 2p_i(1-p_i)\alpha^2_{oi} = \text{additive variance of the } i^{th} \text{ SNP} \qquad [11.2.4]$$

Since each additive value is a product between a function of the allele frequency and the additive effect according to Equations [1.2.13]-1.2.15], an additive value can be expressed as:

$$a = [f(p)]\alpha_o \qquad [11.2.5]$$

where f(p) is a function of allele frequency p. Therefore,

$$\text{var}(a) = \sigma^2_a = [f(p)]^2\,\text{var}(\alpha_o) = [f(p)]^2\sigma^2_{\alpha o} = 2p(1-p)\alpha^2_o \qquad [11.2.6]$$

Since the only term in Equation [11.2.6] as a function of p is $2p(1-p)$, $[f(p)]^2 = 2p(1-p)$, and:

$$\sigma^2_{\alpha o} = \alpha^2_o \qquad [11.2.7]$$

$$\sigma^2_{ai} = 2p_i(1-p_i)\sigma^2_{\alpha oi} = \text{additive variance of the } i^{th} \text{ SNP} \qquad [11.2.8]$$

$$= 2p_i(1-p_i)\sigma^2_{\alpha o} \text{ assuming equal } \sigma^2_{\alpha o} \text{ for all SNPs} \qquad [11.2.9]$$

Equation [11.2.8] leads to 'within SNP standardization', whereas Equation [11.2.9] leads to 'across SNP standardization'. Under the assumption of Equation [11.2.9], the total additive variance of m loci based on the definitions of Equation [11.2.3] becomes:

$$\sigma^2_a = \sum_{i=1}^{m}\sigma^2_{ai} = [2\sum_{i=1}^{m} p_i(1-p_i)]\sigma^2_{\alpha o} \qquad [11.2.10]$$

Substituting Equations [11.2.7] and [11.2.10] into Equation [11.2.2] yields genomic additive relationship Matrix:

$$\begin{aligned}\mathbf{A}_g &= (\sigma^2_{\alpha o} / \sigma^2_a)\mathbf{W}_\alpha \mathbf{W}_\alpha' \\ &= \mathbf{W}_\alpha \mathbf{W}_\alpha' / [2\sum_{i=1}^{m} p_i (1-p_i)] = \mathbf{W}_\alpha \mathbf{W}_\alpha' / k_1\end{aligned} \qquad [11.2.11]$$

where $\mathbf{A}_g$ denotes genomic additive relationship matrix and replaces the pedigree **A** matrix in Equation [11.2.2]. The same procedure leading to Equation [11.2.11] can be used to derive genomic dominance relationship matrix. The derivation leading to Equation [11.2.11] shows that genomic additive relationship matrix of Equation [11.1.1] through standardization of SNP coding is the genomic version of the pedigree additive relationship matrix.

## 11.3 Numerical Evaluation

The numerical results of Table 11.3.1 showed that the average genomic additive relationships of Equation [11.1.1] were remarkably consistent with the average of pedigree additive relationships defined by Equation [1.6.3]. The difference between genomic and pedigree additive relationships were 0.006, 0.013 and 0.005 for parent-offspring, fullsibs and halfsibs respectively. Such small differences confired that the genomic additive relationship matrix defined by Equation [11.1.1] is the genomic version of the pedigree additive relationships defined by Equation [1.6.3].

Table 11.3.1 also showed two major differences between genomic and pedigree relationships: genomic relationship had larger variations than pedigree relationships, and genomic relationships had values outside the [0,2] parameter space, e.g., the halfsibs had negative genomic additive relationships, and the parent-offspring and halfsib had negative genomic dominance relationships.

Table 11.3.1 Comparison between genomic and pedigree relationships using 3534 pigs with 45,376 SNPs with minor allele frequency > 0.05. (Wang and Da, 2014)

| Relatives | Additive relationship | | Dominance relationship | |
|---|---|---|---|---|
| | Mean±SD | Range | Mean±SD | Range |
| Parent-offspring (3518 pairs) | 0.534±0.090 | 0.334, 0.856 | 0.033±0.066 | -0.137, 0.368 |
| 2×(coancestry coefficient) | 0.528±0.031 | 0.500, 0.723 | 0.000 [a] | - |
| Difference in mean | 0.006 | | 0.033 | |
| Fullsibs (1441 pairs) | 0.543±0.100 | 0.300, 0.891 | 0.272±0.087 | 0.044, 0.598 |
| 2×(coancestry coefficient) | 0.530±0.037 | 0.500, 0.692 | 0.250 [a] | - |
| Difference in mean | 0.013 | | 0.022 | |
| Halfsibs (23,628 pairs) | 0.299±0.091 | -0.037, 0.830 | 0.028±0.049 | -0.136, 0.551 |
| 2×(coancestry coefficient) | 0.294±0.039 | 0.250, 0.525 | 0.000 [a] | - |
| Difference in mean | 0.005 | | 0.028 | |

[a] These values are expected pedigree dominance relationships, not observed dominance relationships calculated from the pedigree data.

.

# CHAPTER 12: GENOMIC BEST LINEAR UNBIASED PREDICTION (GBLUP)

Genomic best linear unbiased prediction (GBLUP) is the genomic version of BLUP with the pedigree relationship in BLUP replaced by the corresponding genomic relationships. Because of the use of genomic relationships, the QG model for genomic prediction needs to be reparameterized for GBLUP. The QG model and the reparameterized QG model for GBLUP have the same prediction accuracy but GBLUP has advantages over QG-BLUP.

## 12.1 Reparameterized QG Model For GBLUP Using Genomic Relationships

The mixed model for GBLUP using genomic relationships is a reparameterized QG (RQG) model. In the RQG model, the additive effects and their model matrix are reparameterized such that the variance-covariance matrix becomes a function of genomic additive relationship matrix. The QG and RQG models have the same additive values and the variance-covariance matrix of additive values and hence have the same accuracy of genomic prediction but the RQG model has advantages over the QG model to be discussed at the end of this chapter. The RQG model is:

$$\begin{aligned}\mathbf{y} &= \mathbf{Xb} + \mathbf{ZW}_\alpha\boldsymbol{\alpha}_o + \mathbf{e} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \\ &= \mathbf{Xb} + \mathbf{Z}(\mathbf{W}_\alpha / \sqrt{k_i})(\sqrt{k_i}\boldsymbol{\alpha}_o) + \mathbf{e} = \mathbf{Xb} + \mathbf{ZT}_\alpha\boldsymbol{\alpha} + \mathbf{e}\end{aligned} \quad [12.1.1]$$

where $k_i$ is defined by Equation [11.1.2] or [11.1.4],

$$\boldsymbol{\alpha} = \sqrt{k_i}\boldsymbol{\alpha}_o = \text{reprameterized additive effects of m SNP markers} \quad [12.1.2]$$

$$\alpha_j = \sqrt{k_i}\alpha_{oj} = \text{reprametrized additive effect of } j^{th} \text{ SNP} \quad [12.1.3]$$

$$\mathbf{a} = \mathbf{W}_\alpha\boldsymbol{\alpha}_o = \mathbf{T}_\alpha\boldsymbol{\alpha} = \text{genomic additive values} \quad [12.1.4]$$

$$\mathbf{T}_\alpha = \mathbf{W}_\alpha / \sqrt{k_i} = \text{model matrix of reparameterized additive effects} \quad [12.1.5]$$

With the reparameterized model of Equations [12.1.1]-[12.1.5], the $\mathbf{G}_a$ matrix of Equation [10.1.6] and the $\mathbf{A}_g$ matrix of Equation [11.1.1] or [11.1.3] can be expressed as:

$$\begin{aligned}\mathbf{G}_a &= \text{Var}(\mathbf{a}) = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}_\alpha' \\ &= \mathbf{T}_\alpha\text{Var}(\boldsymbol{\alpha})\mathbf{T}_\alpha' = \mathbf{T}_\alpha\mathbf{G}_{\boldsymbol{\alpha}}\mathbf{T}_\alpha' = \sigma^2_\alpha\mathbf{T}_\alpha\mathbf{T}_\alpha' = \sigma^2_\alpha\mathbf{A}_g\end{aligned} \quad [12.1.6]$$

where

$$\sigma^2_\alpha = \text{Var}(\alpha_j) = \text{Var}(\sqrt{k_i}\alpha_{oj}) = k_i\sigma^2_{\alpha o} = \text{genomic additive variance} \quad [12.1.7]$$

$$\mathbf{G}_\alpha = \text{Var}(\boldsymbol{\alpha}) = k_i\sigma^2_{\alpha o}\mathbf{I}_m = \sigma^2_\alpha\mathbf{I}_m \quad [12.1.8]$$

$$\mathbf{A}_g = \mathbf{T}_\alpha\mathbf{T}_\alpha' = \mathbf{W}_\alpha\mathbf{W}_\alpha / k_i \quad [12.1.9]$$

The genomic additive variance of Equation [12.1.7] is the total additive variance of all SNPs if $k_i = k_1$, or is the average of the additive variances of all individuals if $k_i = k_2$.

The phnotypic variance-covariance matrix is:

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}_a\mathbf{Z}' + \mathbf{R} = \sigma^2_\alpha \mathbf{Z}\mathbf{A}_g\mathbf{Z}' + \sigma^2_e \mathbf{I}_N \quad [12.1.10]$$

With the RQG model of Equations [12.1.1]-[12.1.10], two RQG models can be defined in parallel to the two QG models of Equations [10.5.1]-[10.5.4]. The two RQG models based on Equations [12.1.1] and [12.1.10] are:

RQG1:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} \quad [12.1.11]$$

$$\mathbf{V} = \mathbf{Z}\mathbf{G}_a\mathbf{Z}' + \mathbf{R} = \sigma^2_\alpha \mathbf{Z}\mathbf{A}_g\mathbf{Z}' + \sigma^2_e \mathbf{I}_N \quad [12.1.12]$$

RQG2:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{T}_\alpha\boldsymbol{\alpha} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{Z}_\alpha\boldsymbol{\alpha} + \mathbf{e} \quad [12.1.13]$$

$$\mathbf{V} = \mathbf{Z}\mathbf{G}_a\mathbf{Z}' + \mathbf{R} = \sigma^2_\alpha(\mathbf{Z}\mathbf{T}_\alpha)(\mathbf{T}_\alpha\mathbf{Z})' + \sigma^2_e\mathbf{I}_N = \sigma^2_\alpha\mathbf{Z}_\alpha\mathbf{Z}_\alpha' + \sigma^2_e\mathbf{I}_N \quad [12.1.14]$$

where

$$\mathbf{Z}_\alpha = \mathbf{Z}\mathbf{T}_\alpha \quad [12.1.15]$$

From the two RQG models, four GBLUP methods are possible in parallel to the four QG methods of Table 10.5.1.

## 12.2 GBLUP of Two Complementary Models

**Comparison of four GBLUP methods**

The RQG models yield four possible BLUP methods that parallel the QG methods for genomic prediction. These four methods have identical results but different computing requirements and hence provide an opportunity for selecting the best method for a given data structure. Detailed comparisons of these four methods are listed in Table 12.2.1, which essentially is the same as Table 10.5.1 with the RQG notations replacing the QG notations. Table 12.2.1 also has more details than in Table 10.5.1.

RQG1-CE and RQG2-MME are two GBLUP methods with complementary computing advantages: RQG1-CE is most efficient when the number of individuals is smaller than the number of SNPs, whereas RQG2-MME is most efficient when the number of SNPs is smaller than the number of individuals.

RQG1-MME has limited application: when the number of observations is greater than the number of individuals and the number of SNPs is greater than number of individuals. However, when repeated observations are present, the permanent environment (PE) effects should be

considered in the prediction model. However, adding PE effects to the RQG1 model may double the size of the MME of RQG1-MME. Therefore, RQG1-MME is not considered further.

RQG2-CE has no computing advantage because the numbers of individuals and observations affect the size of the **V** matrix, and those numbers along with the number of SNPs all affect the size of the $\mathbf{Z}_\alpha$ matrix.

**Table 12.2.1** Four methods of genomic prediction from the QG1 and QG2 models. (Shaded methods are not recommended for implementation)

| CE method | MME method |
|---|---|
| RQG1: $\mathbf{y}=\mathbf{Xb}+\mathbf{Za}+\mathbf{e}$ [12.1.11], $\mathbf{V}=\mathbf{ZG}_a\mathbf{Z}'+\mathbf{R}=\sigma^2_\alpha\mathbf{ZA}_g\mathbf{Z}'+\sigma^2_e\mathbf{I}_N$ [12.1.12] | |
| RQG1-CE: most efficient when N<m, N=n<br><br>$\hat{\mathbf{b}}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$<br>$\hat{\mathbf{a}}=\sigma^2_\alpha\mathbf{A}_g\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})$<br>Strength<br>• Number of SNPs does not change the size of **V**<br>• Number of genetic factors do not change the size of **V**<br>• Does not invert $\mathbf{A}_g$<br>Weakness<br>• Number of individuals affects the size of **V**<br>• Number of observations affects the size of **V** | RQG1-MME: can be efficient when n<m, N>n<br><br>$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\lambda_\alpha\mathbf{A}_g^{-1}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{a}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\end{pmatrix}$, $\lambda_\alpha=\sigma^2_e/\sigma^2_\alpha$<br>Strength<br>• Number of observations does not affect the size of MME<br>• Number of SNPs does not change the size of **V**<br>Weakness<br>• Number of individuals affects the size of MME<br>Number of genetic factors affects the size of MME<br>• Requires $\mathbf{A}_g^{-1}$ |
| RQG2: $\mathbf{y}=\mathbf{Xb}+\mathbf{Z}_\alpha\boldsymbol{\alpha}+\mathbf{e}$ [12.1.13], $\mathbf{V}=\sigma^2_\alpha\mathbf{Z}_\alpha\mathbf{Z}_\alpha'+\sigma^2_e\mathbf{I}_N$ [12.1.14] | |
| RQG2-CE: no computing advantage<br><br>$\hat{\mathbf{b}}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$<br>$\hat{\boldsymbol{\alpha}}=\sigma^2_\alpha\mathbf{Z}_\alpha'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})$, $\hat{\mathbf{a}}=\mathbf{T}_\alpha\hat{\boldsymbol{\alpha}}$ | RQG2-MME: generally is most efficient for n>m<br><br>$\begin{pmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_\alpha\\ \mathbf{Z}_\alpha'\mathbf{X} & \mathbf{Z}_\alpha'\mathbf{Z}_\alpha+\lambda_\alpha\mathbf{I}_m\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}}\\ \hat{\boldsymbol{\alpha}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}'\mathbf{y}\\ \mathbf{Z}_\alpha'\mathbf{y}\end{pmatrix}$, $\mathbf{Z}_\alpha=\mathbf{ZT}_\alpha$ $\hat{\mathbf{a}}=\mathbf{T}_\alpha\hat{\boldsymbol{\alpha}}$<br><br>Strength<br>• Number of individuals does not change the size of the MME<br>• Number of observations does not affect the size of MME<br>Weakness<br>• Number of SNP markers affects the size of MME<br>• Number of genetic effects affects the size of MME |

Based on the above comparison, RQG1-MME and RQG2-CE are not considered further. The two complementary methods of RQG1-CE and RQG2-MME are the two GBLUP methods to be described in details, and RQG1-CE will be referred to as 'GBLUP-CE' and RQG2-MME as 'GBLUP-MME'.

**Formulations of GBLUP-CE and GBLUP-MME**

Among the four possible GBLUP methods based on RQG1 and RGQ2 models, two GBLUP methods have complementary computing advantages: RQG1-CE and RQG2-MME. Formulations of these two methods are described below and these formulations including all individuals with and without phenotypic observation. Formulations for predicting validation individuals without phenotypic observations will be described separately.

For RQG1-CE, GBLUP, GBLUP reliability and BLUE are:

$$\hat{\mathbf{a}} = \sigma^2_\alpha \mathbf{A}_g \mathbf{Z}' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma^2_\alpha \mathbf{A}_g \mathbf{Z}' \mathbf{P}\mathbf{y} \quad [12.2.1]$$

$$R^2_{ai} = \sigma^2_\alpha \left( \mathbf{A}_g \mathbf{Z}' \mathbf{P} \mathbf{Z} \mathbf{A}_g \right)_{ii} \Big/ a_{ii} \quad [12.2.2]$$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^- \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad [2.2.7]$$

where $a_{ii}$ = the $i^{th}$ diagonal element of $\mathbf{A}_g$.

For RQG2-MME, BLUE, GBLUP, and GBLUP reliability are:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_\alpha \\ \mathbf{Z}_\alpha'\mathbf{X} & \mathbf{Z}_\alpha'\mathbf{Z}_\alpha + \lambda_\alpha \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\alpha}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_\alpha'\mathbf{y} \end{pmatrix} \text{ with } \lambda_\alpha = \sigma^2_e / \sigma^2_\alpha \quad [12.2.3]$$

$$\hat{\mathbf{a}} = \mathbf{T}_\alpha \boldsymbol{\alpha} \quad [12.2.4]$$

$$R^2_{ai} = 1 - \lambda_\alpha \left( \mathbf{T}_\alpha \mathbf{C}^{\alpha\alpha} \mathbf{T}_\alpha ' \right)_{ii} \Big/ a_{ii} \quad [12.2.5]$$

where $\mathbf{C}^{\alpha\alpha}$ is the submatrix in a generalized inverse of the coefficient matrix of Equation [12.2.3] corresponding to $\hat{\boldsymbol{\alpha}}$.

## 12.3 GBLUP for Individuals without Phenotypic Observations

The mechanism of genomic prediction is the similarity of genome-wide SNPs between the training and validation populations, and the similarity between the training and validation populations can be used by two methods: the method of relationships or the method of SNP similarity between the training and validation populations for predicting the validation population, as described in Chapter 9. Tables 9.2.1 and 9.2.2 are intuitive examples illustrating the two methods of genmic similartity for genomic prediction of the validation population, but had different numerical results for not using real prediction methods. The CE and MME methods described in this chapter are real prediction methods that have identical results, where CE uses

relationships whereas MME uses SNP similarity between the training and validation populations, and these two methods have identical results.

The GBLUP methods for individuals without phenotypic observations calculated separately from the GBLUP for individuals with phenotypic observations are similar to the BLUP methods in Section 2.6. The RQG1-CE method requires only minor notations changes from those in Section 2.6, whereas the the RQG2-MME method that calculates GBLUP of additive values based on the SNP additive effects is a new idea not covered in Section 2.6 although the formulations in this chapter and those in Section 2.6 bear some similarity.

For both RQG1-CE and RQG2-MME, Equations [2.6.1] and [2.6.2] are needed:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \end{bmatrix} \qquad [2.6.1]$$

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_0 \end{bmatrix} \qquad [2.6.2]$$

where

$\mathbf{Z}_1 = \mathrm{N} \times \mathrm{n}_1$ incidence matrix for animals with observations = incidence matrix for $\mathbf{a}_1$ ,

$\mathbf{a}_1 = \mathrm{n}_1 \times 1$ vector of genomic additive values for animals with observations,

$\mathrm{n}_1$ = number of animals with observations,

$\mathbf{a}_0 = \mathrm{n}_0 \times 1$ vector of genomic additive values for animals without phenotypic observations,

$\mathbf{0} = \mathrm{N} \times \mathrm{n}_0$ matrix of zeros,

$\mathrm{n}_0$ = number of animals without observations, and $\mathrm{n}_1 + \mathrm{n}_0 = \mathrm{n}$ .

In addition, the genomic additive relationship matrix needs to be partitioned as the partition of Equation [2.6.6]:

$$\mathbf{A}_g = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{10} \\ \mathbf{A}_{01} & \mathbf{A}_{00} \end{pmatrix} \qquad [12.3.1]$$

where $\mathbf{A}_{11} = \mathrm{n}_1 \times \mathrm{n}_1$ matrix of genomic additive relationships among individuals with phenotypic records, $\mathbf{A}_{10} = \mathrm{n}_1 \times \mathrm{n}_0$ matrix of genomic additive relationships between individuals with phenotypic observations and individuals without phenotypic observations, $\mathbf{A}_{00} = \mathrm{n}_0 \times \mathrm{n}_0$ matrix of genomic additive relationships among individuals without phenotypic observations, and $\mathrm{n}_1 + \mathrm{n}_0 = \mathrm{n}$ .

**The CE method (RQG1-CE) for individuals without phenotypic data**

With the partitions of Equations [2.6.1] and [2.6.2], the RQG1 model for individuals with phenotypic observations is:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{Z}_1\mathbf{a}_1 + \mathbf{e} \qquad [12.3.2]$$

$$\mathbf{V} = \sigma^2_\alpha \mathbf{Z}\mathbf{A}_g\mathbf{Z}' + \sigma^2_e \mathbf{I}_N = \sigma^2_\alpha \mathbf{Z}_1\mathbf{A}_g\mathbf{Z}_1' + \sigma^2_e \mathbf{I}_N \qquad [12.3.3]$$

The GBLUP for individuals with phenotypic observations ( $\hat{\mathbf{a}}_1$ ) is:

$$\hat{\mathbf{a}}_1 = \sigma_a^2 \mathbf{A}_{11} \mathbf{Z}_1 \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_a^2 \mathbf{A}_{11} \mathbf{Z}_1 \mathbf{P}\mathbf{y} \qquad [12.3.4]$$

where $\sigma_\alpha^2$ is the genomic additive variance, and $\mathbf{A}_{11}$ is genomic additive relationship matrix from Equation [12.3.1] rather than from pedigree additive relationships of Equation [2.6.6], $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ and $\mathbf{V}$ is defined by Equation [12.3.3].

The GBLUP for individuals without phenotypic observations ( $\hat{\mathbf{a}}_0$ ) is calculated through genomic relationships with individuals without phenotypic observations using any of the following two equations:

$$\hat{\mathbf{a}}_0 = \sigma_\alpha^2 \mathbf{A}_{01} \mathbf{Z}_1 \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_\alpha^2 \mathbf{A}_{01} \mathbf{Z}_1 \mathbf{P}\mathbf{y} \qquad [12.3.5]$$

$$\hat{\mathbf{a}}_0 = \mathbf{A}_{01} \mathbf{A}_{11}^{-1} \hat{\mathbf{a}}_1 \qquad [12.3.6]$$

where $\sigma_\alpha^2$ is the genomic additive variance, and $\mathbf{A}_{01}$ is genomic additive relationship matrix from Equation [12.3.1]. The GBLUP reliability for $i^{th}$ individual without phenotypic observations calculated by the CE method is:

$$R_{ai}^2 = \sigma_\alpha^2 \left( \mathbf{A}_{01} \mathbf{Z}' \mathbf{P} \mathbf{Z} \mathbf{A}_{10} \right)_{ii} / a_{0ii} \qquad [12.3.7]$$

where $a_{0ii}$ is the $i^{th}$ diagonal element in the $\mathbf{A}_{00}$ matrix defined by Equation [12.3.1].

**The MME method (RQG2-MME) for individuals without phenotypic data**

For the RQG2 model for individuals with phenotypic observations, the $\mathbf{Z}$ matrix has the same partition as that of Equation [2.6.1], $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \end{bmatrix}$. In addition, the $\mathbf{T}_\alpha$ matrix in Equation [12.1.12] need to be partitioned as:

$$\mathbf{T}_\alpha = \begin{bmatrix} \mathbf{T}_{\alpha 1} \\ \mathbf{T}_{\alpha 0} \end{bmatrix} \qquad [12.3.8]$$

where $\mathbf{T}_{\alpha 1}$ = $n_1 \times m$ model matrix of the SNP additive effects of individuals with phenotypic observation, $\mathbf{T}_{\alpha 1}$ = $n_0 \times m$ model matrix of the SNP additive effects of individuals without phenotypic observation. Then, the RQG2 model for individuals with phenotypic observations is:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_1 \mathbf{T}_{\alpha 1} \boldsymbol{\alpha}_1 + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{Z}_{\alpha 1} \boldsymbol{\alpha}_1 + \mathbf{e} \qquad [12.3.9]$$

$$\mathbf{V} = (\mathbf{Z}_1 \mathbf{T}_{\alpha 1}) \mathbf{G}_\alpha (\mathbf{Z}_1 \mathbf{T}_{\alpha 1})' + \mathbf{R} = \sigma_\alpha^2 \mathbf{Z}_{\alpha 1} \mathbf{Z}_{\alpha 1}' + \sigma_e^2 \mathbf{I}_N \qquad [12.3.10]$$

where $\boldsymbol{\alpha}_1$ = m×1 = column vector of SNP additive effects calculated from individuals with phenotypic observations, and $\mathbf{G}_\alpha = \sigma^2_\alpha \mathbf{I}_m$ (Equation [12.1.8]).

$$\mathbf{Z}_{\alpha 1} = \mathbf{Z}_1 \mathbf{T}_{\alpha 1} \qquad [12.3.11]$$

Using the MME method, the GBLUP of SNP additive effects for individuals with phenotypic observations ($\hat{\boldsymbol{\alpha}}_1$) is calculated from the following MME:

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_\alpha \\ \mathbf{Z}_{\alpha 1}'\mathbf{X} & \mathbf{Z}_{\alpha 1}'\mathbf{Z}_{\alpha 1} + \lambda_\alpha \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\boldsymbol{\alpha}}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_{\alpha 1}'\mathbf{y} \end{pmatrix} \qquad [12.3.12]$$

The GBLUP of additive values for individuals without phenotypic observations ($\hat{\mathbf{a}}_0$) is calculated based on the GBLUP of SNP additive effects of individuals with phenotypic observations:

$$\hat{\mathbf{a}}_0 = \mathbf{T}_{\alpha 0} \hat{\boldsymbol{\alpha}}_1 = \sigma^2_\alpha \mathbf{A}_{01} \mathbf{Z}_1' \mathbf{P}\mathbf{y} \qquad [12.3.13]$$

where $\hat{\boldsymbol{\alpha}}_1$ is the GBLUP of SNP additive effects of individuals with phenotypic observations calculated from the MME of Equation [12.3.12], $\mathbf{T}_{\alpha 0}$ is defined by Equation [12.3.8].

Let the coefficient matrix of the MME of Equation [12.3.12] be partitioned as:

$$\mathbf{C}^- = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_\alpha \\ \mathbf{Z}_{\alpha 1}'\mathbf{X} & \mathbf{Z}_{\alpha 1}'\mathbf{Z}_{\alpha 1} + \lambda_\alpha \mathbf{I}_m \end{pmatrix}^- = \begin{pmatrix} \mathbf{C}^{bb} & \mathbf{C}^{ba} \\ \mathbf{C}^{ab} & \mathbf{C}^{aa} \end{pmatrix} \qquad [12.3.14]$$

Then, the GBLUP reliability of $\hat{\mathbf{a}}_0$ is:

$$R^2_{ai} = 1 - \lambda_\alpha \left( \mathbf{T}_{\alpha 0} \mathbf{C}^{\alpha\alpha} \mathbf{T}_{\alpha 0}' \right)_{ii} / a_{ii} \qquad [12.3.15]$$

.

The CE and MME methods of GBLUP for predicting the additive values of individuals without phenotypic observations were different methods for using the genome similarity between the training and validation populations with identical results. The CE method (Equations [12.3.5] and [12.3.6]) uses relationships whereas the MME method (Equations [12.3.12] and [12.3.13]) uses SNP similarity between the training and validation populations. These methods calculate GBLUP separately for the training and validation populations mainly for the purpose of showing the mechanisms of using relationships and SNP similarity.

GBLUP for both training and validation populations can be calculated in one system of equations but the mechanisms of using relationships and SNP similarity are not as clear as shown by the separate calculations.

**Numerical example**

This example uses the data in Table 10.1.1 to show that RQG1-CE using relationships (Table 12.3.1) and RQG2-MME using SNP similarity (Table 12.3.1) indeed have identical GBLUP

results. Table [12.3.1] using genomic relationships is the scientific version of the intuitive example of Table [9.2.1], whereas Table [12.3.1] using SNP similarity is the scientific version of the intuitive example of Table [9.2.2].

Table 12.3.1 Example of exact solution for genomic prediction using estimated breeding values from the training population

| | Cow 1 | Cow 2 | Cow 3 | Cow 4 | $\hat{a}_0$ for Bull |
|---|---|---|---|---|---|
| $\sigma^2_\alpha \mathbf{Z}_1\mathbf{Py}$ | -0.072235 | 0.0722354 | -0.012737 | 0.0127375 | 0.001426 |
| $\mathbf{A}_{01}$ | -0.253731 | -0.253731 | -0.402985 | -0.291045 | |

Table 12.3.2 Example of exact solution for the mechanism of genomic prediction using estimated SNP effects from the training population

| | SNP effects from training population | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | SNP1 | SNP 2 | SNP 3 | SNP 4 | SNP 5 | SNP 6 | SNP 7 | SNP 8 | SNP 9 | SNP 10 | $\hat{a}_0$ for Bull |
| $\hat{\boldsymbol{\alpha}}_1$ | 0.0055018 | 0.0624019 | 0.0055018 | 0.0110035 | –0.067904 | 0.0624019 | –0.051398 | 0.0312009 | –0.031201 | 0 | 0.001426 |
| $\mathbf{T}_{\alpha 0}$ | –0.259161 | 0 | –0.259161 | 0 | –0.086387 | 0.3455474 | 0.3455474 | 0.4319342 | 0.6047079 | 0.5183211 | |

## 12.4 Validation Studies and Accuracy Measures of Genomic Prediction

The accuracy of genomic prediction can be evaluated using validation studies that offer an objective judgement of the performance of a prediction method. Cross validation is a commonly used validation method in research and progressive validation can also be used when available.

### 12.4.1 Cross Validation of Prediction Accuracy

A k-fold cross validation divides the entire population with phenotypic and SNP data into k subpopulations. Each subpopulation is treated as a validation population with phenotypic observation omitted when calculating GBLUP, is predicted by the remaining $k-1$ subpopulations, and serves as part of the training population $k-1$ times predicting each of the remaining $k-1$ subpopulation.

For the 3-fold validation in Figures 12.4.1, the entire population is divided into 3 subpopulations each serving as a validation population once being predicted by the remaing 2 subpopulations. Population 1 serves as the validation once and is predicted by populations 2 and 3, and serves as one of the two subpopulations as the training population predicting population 2 or 3; Population 2 as the validation once and is predicted by populations 1 and 3, and serves as one of the two subpopulations as the training population predicting population 1 or 3; Population 3 serves as the validation once and is predicted by populations 1 and 3, and serves as one of the two subpopulations as the training population predicting population 1 or 2.

Progressive validation resembles genomic selection practice that predicts new individuals using existing individuals in the population. In a long-term program of genomic selection, progressive validation offers an evaluation of practical impact of genomic evaluation.

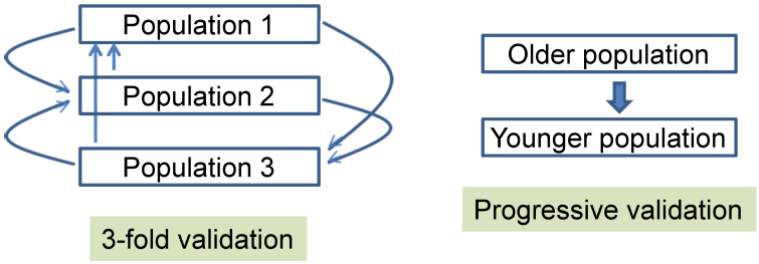



**Figure 12.4.1** Examples of k-fold validation and progressive validation. Left: 3-fold validation as an example of k-fold validation. Right: progressive validation that predicts a younger population using an older population as the training population.

### 12.4.2 Accuracy Measures in Cross Validation of Prediction Accuracy

The main accuracy measure in a k-fold validation study is the correlation between the phenotypic observations and GBLUP calculated using the training populations (predictive ability, Legarra et al., 2008). In addition, three other accuracy measures can be calculated for each validation population: including the calculated accuracy of predicting genetic values (square root of reliability) of the validation and/or training population, predicted correlation between phenotypic observations and GBLUP, and predicted accuracy of predicting genetic values. The

prediction accuracy of a progressive validation is the same as that of any validation population in a k-fold validation and therefore only the prediction accuracies of the k-fold cross validation is described. Each accuracy measure is calculated for each validation and the k estimates from the k validation populations is averaged. Let

$\hat{\mathbf{g}}_{0i}$ = genomic predicted genetic values of individuals in the $i^{th}$ validation population (GBLUP),

$\mathbf{y}_{0i}$ = phenotypic values of individuals in the $i^{th}$ validation population.

Note that 'genetic values' can be any type of genetic values such as additive and dominance values or the summation of additive and dominance values. Also not that the phenotypic values of individuals in validation populations are omitted when calculating the GBLUP for each validation population. Then, the accuracy of predicting phenotypic values in the $i^{th}$ validation population is the correlation between the GBLUP and the phenotypic observations:

$$\hat{R}_{0pi} = \text{corr}(\hat{\mathbf{g}}_{0i}', \mathbf{y}_{0i}) \qquad [12.4.1]$$

where 'corr' stands for 'correlation'. The average of the the accuracy of predicting phenotypic values across all validation populations is

$$\hat{R}_{0p} = (\sum_{i=1}^{k} \hat{R}_{0pi}) / k \qquad [12.4.2]$$

Equation [12.4.2] is the primary accuracy measure of genomic prediction in a k-fold cross validatiuon study.

In addition to the accuracy measure of Equation [12.4.2], three other accuracy measures can be calculated for each validation population: the accuracy of predicting genetic values, estimated accuracy of predicting phenotypic values, and estimated accuracy of predicting genetic values.

The accuracy of predicting genetic values is the square root of reliability defined by Equation [12.3.7] for RQG1-CE or Equation [12.3.15] for RQG2-MME. Since every individual has a calculated reliability, the accuracy as the square root of the reliability is available for every individuals. Therefore, the accuracy of predicting genetic values in a validation population is the average of all individual accuracies in the validation population:

$$R_{0i} = \sum_{j=1}^{n_i} \text{corr}(\hat{g}_{0ij}, g_{0ij}) / n_i = (\sum_{j=1}^{n_i} R_{0ij}) / n_i \qquad [12.4.3]$$

where $R_{0ij}$ is the square root of the reliability (Equation [12.3.7] or [12.3.15]) of the $j^{th}$ individual and in the $i^{th}$ validation population. Note that the calculation of $R_{0i}$ does not use the phenotypic data although its calculation uses the SNP data.

The accuracy predicting genetic values of the k validation populations is the average across all k validation populations:

$$R_0 = (\sum_{i=1}^{k} R_{0i}) / k \qquad [12.4.4]$$

The relationship between $\hat{R}_{0pi}$ and $R_{0i}$ in the $i^{th}$ validation population is:

$$R_{0pi} = R_{0i}\sqrt{h_i^2} \quad [12.4.5]$$

$$\tilde{R}_{0i} = \hat{R}_{0pi} / \sqrt{h_i^2} \quad [12.4.6]$$

where $R_{0pi}$ = estimated accuracy of predicting phenotypic values calculated from $R_{0i}$ and heritability ($h_i^2$), $\tilde{R}_{0i}$ = estimated accuracy of predicting genetic values calculated from $\hat{R}_{0pi}$ and heritability ($h_i^2$). From Equations [12.4.5] and [12.4.6], the average estimated accuracy of predicting phenotypic values ($R_{0p}$) and the everage of estimated accuracy of predicting genetic values ($\tilde{R}_0$) are:

$$R_{0p} = (\sum_{i=1}^{k} R_{0pi}) / k \quad [12.4.7]$$

$$\tilde{R}_0 = (\sum_{i=1}^{k} \tilde{R}_{0i}) / k \quad [12.4.8]$$

As seen from the above results, each validation can calculate four accuracy measures and one heritability estimate. Therefore, standard deviation for each of the five measures can be calculated to understand the variations among the k validation populations, and such standard deviations are often shown as the error bars on a figure showing the accuracy measures or the heritability estimates.

The phenotypic observations of each validation population ($\mathbf{y}_{0i}$) for calculating the accuracy of predicting phenpotypic values using Equations [12.4.1] and [12.4.2] can be the original phenotypic observations or the corrected phenotypic observations after removing fixed nongenetic effects. Using the original phenotypic observations of the validation populations provides an estimate of prediction accuracy for individuals without known fixed nongenetic effects such as herd-year-season effects for a newborn, whereas using the corrected phenotypic observations of the validation populations provides an accuracy estimate unaffected by fixed nongenetic effects.

Figure 12.4.2 is an example of reporting prediction accuracies with their standard deviations, using the original and corrected phenotypic observations of the validation populations in Duroc pigs. Each error bar stands for one standard deviation above and below the mean. The percentage for each trait is the (accuracy increase)% of the haplotype model over the best SNP model. For the same haplotype model, corrected phenotypic values had a higher prediction accuracy for four traits (LMA, LMD, BF and TN) and a lower prediction accuracy for four traits (AGW, ADG, BJS and FCR). On average across the eight traits, predictions were more accurate for the original phenotypic values than for the corrected phenotypic values for both the SNP and haplotype models. The average observed prediction accuracy under the SNP models was 0.316 for the original and 0.309 for the corrected phenotypic values, and the average observed prediction accuracy under the haplotype models was 0.327 for the original and 0.319 for the corrected phenotypic values.

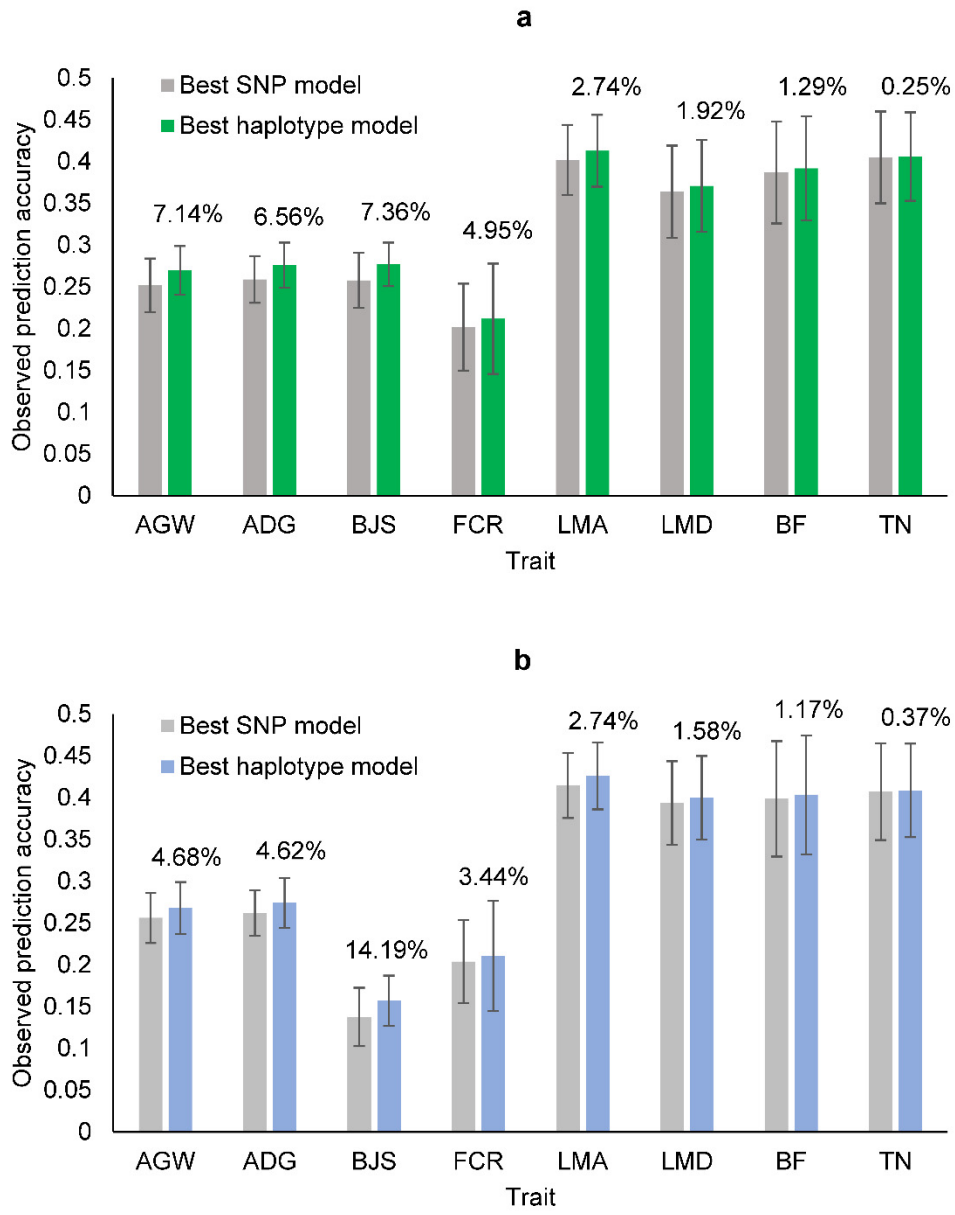



**Figure 12.4.2** Average prediction accuracy ( $\hat{R}_{0p}$ ) of the best haplotype model relative to the best SNP model for each phenotype from ten-fold validations (Bian et al., 2021). **a** Accuracy of predicting phenotypic values using the original phenotypic values of the validation populations. **b** Accuracy of predicting phenotypic values using the corrected phenotypic values of the validation populations. The error bar is one standard deviation above and below the average prediction accuracy, where standard deviation was calculated from tenfold validations. AGW is age at 100 kg live weight, ADG is average daily gain during 30–100 kg live weight, BJS is body judging score, FCR is feed conversion ratio, LMA is loin muscle area at 100 kg, LMD is loin muscle depth at 100 kg, BF back fat thickness at 100 kg, TN is teat number.

## 12.5 Ridge Regression for Genomic Prediction

Ridge regression is intended to overcome ill-conditioned situations where correlations between the various predictors in the model cause the $\mathbf{X'X}$ matrix to be close to singular, giving rise to unstable parameter estimates (Draper and Smith, 1998). The ridge regression solution to this potential problem is to redefine the fixed factors as random factors with a diagonal variance-covariance matrix (Hoerl, 1960; Hoerl and Kennard, 1970). Let the random model and its second moment be written as:

$$\mathbf{y} = \mathbf{Su} + \mathbf{e} \qquad [12.5.1]$$

$$\text{Var}(\mathbf{y}) = \sigma^2_u \mathbf{SS'} + \sigma^2_e \mathbf{I}_N = \mathbf{V} \qquad [12.5.2]$$

Then, the ridge regression solution is:

$$\hat{\mathbf{u}} = (\mathbf{S'S} + \lambda \mathbf{I})^{-1} \mathbf{S'y} \qquad [12.5.3]$$

where $\lambda = \sigma^2_e / \sigma^2_u$ .

For genomic prediction using a mixed model with fixed nongenetic factors, ridge regression becomes ridge regression BLUP (rrBLUP) because the SNP effects are solved within the BLUP framework. Assuming one phenotypic observation per individual, the mixed model and its first and second moments for ridge regression using SNP information can be written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Su} + \mathbf{e} \qquad [12.5.4]$$

$$\text{Var}(\mathbf{y}) = \sigma^2_u \mathbf{SS'} + \sigma^2_e \mathbf{I}_N = \mathbf{V} \qquad [12.5.5]$$

$$\text{E}(\mathbf{y}) = \mathbf{Xb} \qquad [12.5.6]$$

where **u** = random SNP effects, **S =** model matrix of **u**, and the other symbols have the same definitions as for GBLUP. The SNP coding in the **S** matrix can be 1-0-(-1). The MME for Equations [12.5.4]-[12.5.7] are:

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'S} \\ \mathbf{S'X} & \mathbf{S'S} + \lambda \mathbf{I}_m \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{S'y} \end{pmatrix} \qquad [12.5.7]$$

The rrBLUP of Equation [12.5.7] is similar to the method of QG2-MME (Equation [10.5.3]) if **S** uses the QG coding of Equations [10.2.3]-[10.2.5], and becomes GBLUP-MME using the RQG2-MME method (Equation 12.1.13] if **S** is replaced with $\mathbf{Z}_\alpha = \mathbf{ZT}_\alpha$ (Equation 12.1.15].

## 12.6 Comparison between QG and GBLUP Methods

The QG methods and the GBLUP methods based on the RQG models have the same prediction accuracy due to the identical genomic values and genomic variance-covariance matrix.

The choice of the $k_i$ formula does not affect the accuracy of genomic prediction but affects the interpretation and application of the genetic variance and genomic relationships.

Although the QG and GBLUP methods have the same prediction accuracy, GBLUP using genomic relationships has two major advantages over the QG methods without using genomic relationship matrices.

First, the use of genomic relationships provides a genomic version of the traditional theory and methods of BLUP that uses pedigree relationships, and this genomic version can utilize a wealth of BLUP-based theoretical results and computational strategies.

Second, the additive genetic variance under the GBLUP methods ( $\sigma^2_\alpha=k_i\sigma^2_{\alpha o}$ , Equation [12.1.7]) can be used for estimating genomic heritability for three reasons:

1) $\sigma^2_\alpha=k_i\sigma^2_{\alpha o}$ is a common variance to all individuals,
2) $\sigma^2_\alpha=k_i\sigma^2_{\alpha o}$ is the total additive variance of all SNPs or the average of the additive variances of all individuals,
3) $\sigma^2_\alpha=k_i\sigma^2_{\alpha o}$ is invariant to the number of SNPs if the SNPs provide a sufficient coverage of the genome.

In contrast, the QG model does not have a method to estimate additive genetic variance for calculating genomic heritabilities for the following reasons:

1) The additive genetic variance under the QG model ( $\sigma^2_{\alpha o}$ , Equations [10.1.5] and [10.1.9]) is an inverse function of the number of SNPs. As the number of SNPs increases or decreases, the value of each element in $\mathbf{W}_\alpha\mathbf{W}_\alpha'$ changes in the same direction and the $\sigma^2_{\alpha o}$ estimate changes in the opposite direction, i.e., as the number of SNPs increases or decreases, $\sigma^2_{\alpha o}$ decreases or increases. Consequently, the $\sigma^2_{\alpha o}$ estimate does not have a unique interpretation and cannot be used for estimating genomic heritability.
2) $\sigma^2_{\alpha o}$ estimates the additive variance of one SNP and the numerical value of $\sigma^2_{\alpha o}$ generally is a too small to represent the total genetic additive variance of all SNPs.
3) The QG model does not have another additive variance common to all individuals for estimating heritability. These comparisons will be further illustrated in the next chapter.

In summary, the GBLUP methods have advantages over the QG methods for being the genomic version of BLUP and for providing methods of estimating genomic heritability, whereas the QG methods can be considered as preparations for the GBLUP methods and have no advantage over GBLUP, although the QG and GBLUP methods have the same prediction accuracy. The QG and GBLUP methods differ by a constant and have the same computing efficiency.

# CHAPTER 13: GENOMIC RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION

Variance component estimation provides estimates of variance components for calculating GBLUP and and for estimating heritability. Genomic restricted maximum likelihood estimation (GREML) is the genomic version of REML using pedigree relationships.

## 13.1 GREML Formulations

**GREML-CE**

GREML-CE using the RQG1-CE model has the same formulations as REML using pedigree relationships except some definition changes due to the use SNPs. The iterative formulations for GREML-CE are:

$$\sigma_{\alpha}^{2(i+1)} = \frac{\sigma_{\alpha}^{2(i)}\mathbf{y}\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}_g\mathbf{Z}'\mathbf{P}^{(i)}\mathbf{y}}{\text{tr}(\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}_g\mathbf{Z}')} \quad [13.1.1]$$

$$\sigma_{e}^{2(i+1)} = \frac{\sigma_{e}^{2(i)}\mathbf{y}\mathbf{P}^{(i)}\mathbf{P}^{(i)}\mathbf{y}}{\text{tr}(\mathbf{P}^{(i)})} \quad [13.1.2]$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$ (Equation [2.2.8]), and $\mathbf{A}_g$ = genomic additive relationship matrix (Equation 11.1.1] or [11.1.3]).

**GREML-MME**

The iterative formulations for GREML-MME method based on the RQG2-MME model are:

$$\sigma_{\alpha}^{2(i+1)} = \frac{\hat{\boldsymbol{\alpha}}^{(i)}{}'\hat{\boldsymbol{\alpha}}^{(i)}}{m - \text{tr}(\mathbf{C}^{\alpha\alpha(i)})\lambda_{\alpha}^{(i)}} \quad [13.1.3]$$

$$\sigma_{e}^{2(i+1)} = \frac{\hat{\mathbf{e}}^{(i)}{}'\hat{\mathbf{e}}^{(i)}}{N - [r - \text{tr}(\mathbf{C}^{\alpha\alpha(i)})\lambda_{\alpha}^{(i)}]} \quad [13.1.4]$$

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}_{\alpha}\hat{\boldsymbol{\alpha}} \quad [13.1.5]$$

$$\mathbf{C}^{\alpha\alpha} = (\mathbf{Z}_{\alpha}'\mathbf{M}\mathbf{Z}_{\alpha} + \lambda_{\alpha}\mathbf{I}_m)^{-1} \quad [13.1.6]$$

$$\mathbf{M} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' \quad [13.1.7]$$

The two sets of GREML formulations have identicial results but different computing properie. GREML-CE is more efficient when the number of individuals is smaller than the number of SNPs, whereas GREML-MME is more efficient when the number of SNPs is smaller than the number of individuals.

## 13. 2 SNP Heritability

With the variance components from GREML, genomic heritability can be estimated to evaluate the contributions of SNPs to the phenotypic variance, and to study the sizes of individual SNPs relative to each other.

**Total SNP heritability and each SNP heritability**

Genomic additive heritability measures the fraction of the phenotypic variance explained by additive effects of all SNP and is estimated by:

$$h_\alpha^2 = \sigma_\alpha^2 / \sigma_y^2 \qquad [13.2.1]$$

where $\sigma_y^2 = \sigma_\alpha^2 + \sigma_e^2$ = phenotypic variance. For the $i^{th}$ SNP marker, additive heritability or heritability in the narrow sense ($h_{\alpha i}^2$) can be estimated as:

$$h_{\alpha i}^2 = \sigma_{\alpha i}^2 / \sigma_y^2 = (\hat{\alpha}_i^2 / \sum_{i=1}^m \hat{\alpha}_i^2) h_\alpha^2 = (\hat{\alpha}_i^2 / \hat{\boldsymbol{\alpha}}'\hat{\boldsymbol{\alpha}}) h_\alpha^2 \qquad [13.2.2]$$

Therefore, the heritability of all SNPs can be expressed as the summation of all heritabilities of individual SNPs:

$$h_\alpha^2 = \sum_{i=1}^m h_{\alpha i}^2 \qquad [13.2.3]$$

Based on Equations [13.2.1]-[13.2.3], the SNP heritability of any part or several regions of the genome can be estimated as a sum of the corresponding single-SNP heritabilities. However, the interpretation for any subset of the total heritability is not straightforward.

**Dependency of the Size of each SNP heritability and Number of SNPs**

A problem of genomic heritability estimation is the dependence of the size of the heritability estimate for a single SNP on the number of SNPs in the model: the heritability of any single SNP decreases as the number of SNPs increases or vice versa. Assuming the SNPs provide a sufficient coverage of the genome, this dependency can be formulated as:

$$\hat{\alpha}_{ri} = \hat{\alpha}_i / \sqrt{r} \qquad [13.2.4]$$

$$h_{\alpha ri}^2 = (\hat{\alpha}_{ri}^2 / \sum_{i=1}^{rm} \hat{\alpha}_{ri}^2) h_\alpha^2 = (\hat{\alpha}_i^2 / r) / (\sum_{i=1}^{rm} \hat{\alpha}_i^2 / r) h_\alpha^2 = h_{\alpha i}^2 / r \qquad [13.2.5]$$

$$h_\alpha^2 = \sum_{i=1}^{rm} h_{\alpha ri}^2 = \sum_{i=1}^{m} h_{\alpha i}^2 \qquad [13.2.6]$$

where $\hat{\alpha}_{ri}$ = additive GBLUP of the $i^{th}$ SNP from the mixed model with m SNPs repeated r times, $\hat{\alpha}_i$ = additive GBLUP of the $i^{th}$ SNP from the model with m SNPs.

Equation [13.2.4] states that the additive effect of the $i^{th}$ SNP becomes $\sqrt{r}$ times smaller than the SNP effect from any of the r sets if the SNP data is repeated r times.

Equation [13.2.5] states that the SNP heritability becomes r rimes smaller than the SNP effect from any of the r sets if the SNP data is repeated r times.

Equation [13.2.6] states that the total SNP heritability is invariant to the number of the SNPs, and remain unchanged even when the SNPs are repeated r times.

**Numerical example of dependency of the size of each SNP heritability and number of SNPs**

Numerical results showed that the results of Equations [13.2.4]-[13.2.6] approximately hold for real SNP data that were not repeats of any subset of SNPs (Table 10.10.1).

The total additive heritability of all was $\hat{h}^2_{\alpha 3} = 0.368$ using 41,108 SNPs, and was $\hat{h}^2_{\alpha 1} = 0.360$ using half of the SNPs (20,554 SNPs selected as every other SNPs of the 41,108 SNPs), showing that the total additive heritability remained relatively unchanged for the whole set or half set of SNPs. This was consistent with Equation [13.2.6] that the total SNP heritability is invariant to the number of SNPs.

In contract, the average heritability of single SNPs using 20,554 SNPs ( $\overline{h}^2_{\alpha 2}$ ) without the other half set of SNPs in the model was nearly twice as large as the average heritability using 41,108 SNPs ( $\overline{h}^2_{\alpha 1} / \overline{h}^2_{\alpha 3}$ =1.955). The average heritability of single SNPs of 20,554 SNPs ( $\overline{h}^2_{\alpha 2}$ ) with the other half set of SNPs in the model was nearly the same as the average SNP heritability of all 41,108 SNPs ( $\overline{h}^2_{\alpha 2} / \overline{h}^2_{\alpha 3}$ =0.998). These results were consistent with Equation [13.2.4] that the SNP heritability becomes r rimes smaller than the SNP effect from any of the r sets if the SNP data is repeated r times.

**Table 13.2.1.** Estimates of SNP additive heritabilities of teat number using 41,108 autosome SNPs and 20,554 SNPs selected as every other SNPs of the 41,108 SNPs. (Tan et al., 2017)

| SNP set | Average $\hat{h}^2_{k\alpha}$ per SNP | ratio | Total $\hat{h}^2_{\alpha i}$ |
|---|---|---|---|
| 20,554 SNPs | $\overline{h}^2_{\alpha 1} = 1.75(10^{-5})$ | $\overline{h}^2_{\alpha 1} / \overline{h}^2_{\alpha 2}$ =1.960 | $\hat{h}^2_{\alpha 1} = 0.360$ |
| 20,554 SNPs from all 41,108 SNPs in the mixed model | $\overline{h}^2_{\alpha 2} = 8.93(10^{-6})$ | $\overline{h}^2_{\alpha 2} / \overline{h}^2_{\alpha 3}$ =0.998 | $\hat{h}^2_{\alpha 2} = 0.184$ |
| 41,108 SNPs | $\overline{h}^2_{\alpha 3} = 8.95(10^{-6})$ | $\overline{h}^2_{\alpha 1} / \overline{h}^2_{\alpha 3}$ =1.955 | $\hat{h}^2_{\alpha 3} = 0.368$ |

These numerical results along with the theoretical results of Equations [13.2.4]-[13.2.6] have two conclusions. First, single SNP heritability or the heritability of a subset of SNPs cannot be interpreted as the fraction of phenotypic variance explained by the SNP or SNP set. Second, the total additive SNP heritability is the fraction of the phenotypic variance explained by all SNPs irrespective of the number of SNPs in the model.

## Methods to address the dependency between the number of SNPs and the size of SNP heritability estimates

Two methods based on the comparison between heritability estimates from the full model with all SNPs and a reduced model, but neither method provides a perfect solution to interpret the heritability of a single SNP or a subset of SNPs.

The first method to define the reduced model is to drop the target SNP from the reduced model. Then the difference between heritability estimates of the SNP from the full model and this reduced model is the contribution of the SNP to the total heritability of all SNPs. The problem of this method is the underestimated heritability for the target SNP that is not in the reduced model, because the effect of this excluded SNP could be explained by other SNPs in linkage disequilibrium with the excluded SNP. Numerical results supported this conclusion.

The second method removes the effect of the target SNP by fitting the target SNP as fixed effect in the reduced model. The difference between heritability estimates of the SNP from the full model and this reduced model is the contribution of the target SNP to the total heritability of all SNPs. This method was termed as 'partial heritability' and was virtually unaffected by the number of SNP. The problem of partial heritability is the potential overestimate of the SNP contribution to heritability, because the SNP fitted as fixed effect could have removed some effects of other SNPs due to linkage disequilibrium with the target SNP. Overall, the first method tends to provide a lower bound and the second method tends to provide an upper bound of the SNP contribution to the total SNP heritability of all SNPs.

## Interpretation of SNP effects and heritability estimates

The dependency between the number of SNPs in the mixed model and the sizes of SNP heritability estimates does not affect the ratio of two SNP heritabilities. Therefore, the SNP heritability estimates should be interpreted relative to each other. For example, commonly used Manhattan plots can be used for identifying large SNP heritabilities.

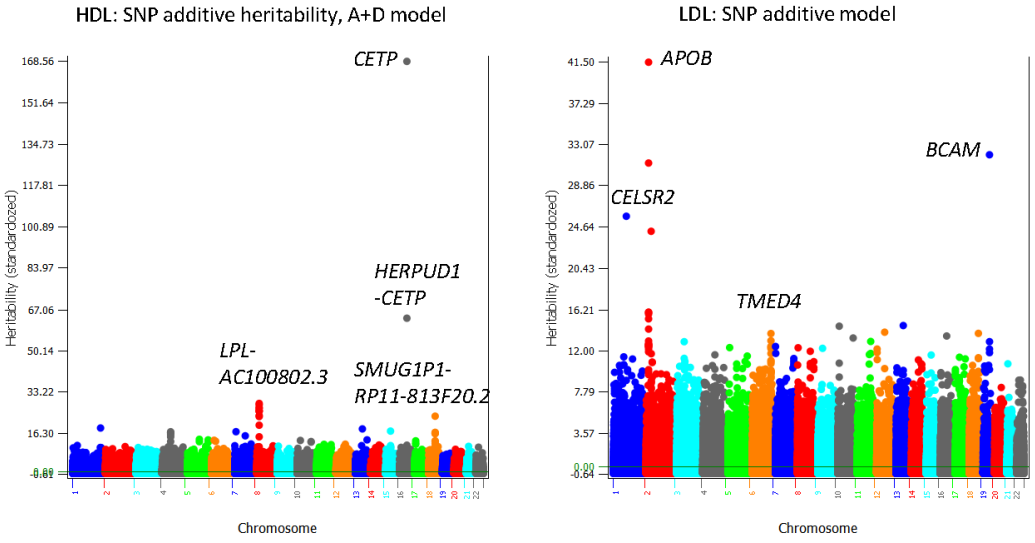



**Figure 13.2.1** Heritability estimates of 380,705 SNP markers for high density lipoproteins (HDL) and low density lipoproteins (LDL) (Liang et al., 2020).

## 13.3 SAS Programs for GREML-CE and GREML-MME

The GREML_CE.SAS shows the step by step implementation of GBLUP with reliability and GREML under the RQG1-CE model, and the GREML_CE.SAS shows the step by step implementation of GBLUP with reliability and GREML under the RQG2-MME model.

The GREML_CE.SAS program can run both the RQG1-CE model and the QG1-CE model with a few changes in the definition of genomic additive relationship matrix, and can use both Definition I and Definition 2 of the genomic additive relationship matrix.

Similarly, the GREML_MME.SAS program can run both the RQG2-MME model and the QG2-MME model with a changes in the definition of of the model matrix of additive effects, and can use both Definition I and Definition 2 of the genomic additive relationship matrix.

**Exercise 13.3.1**

Use GREML_CE.SAS to verify the relationship between the additive variance under the RQG models and the additive variance under the QG model:

$$\sigma^2_\alpha = \text{Var}(\alpha_j) = \text{Var}(\sqrt{k_i}\alpha_{oj}) = k_i\sigma^2_{\alpha o} = \text{genomic additive variance} \qquad [12.1.7]$$

where $\sigma^2_\alpha$ measures the total additive heritability of all SNPs or the average additive heritability of all individuals under the RQG model, and $\hat{\sigma}^2_{\alpha o}$ is the common additive variance to all SNPs under the QG model. This verification is done in two runds:

1) Run GREML_CE.SAS using genomic relationship matrices:

   AG = AG_1;

The GREML estimate of $\sigma^2_\alpha$ is $\sigma^2_\alpha = 0.3688034$ at iteration 12 with tolerance $<10^{-8}$, and $k_1 = 4.48$ .

2) Run GREML_CE.SAS using $\mathbf{W}_\alpha\mathbf{W}_\alpha'$ as the genomic relationship matrices:

   AG = WW_A;

The GREML estimate of $\sigma^2_{\alpha o}$ is $\hat{\sigma}^2_{\alpha o}$ = 0.0823222 at iteration 10 with tolerance $<10^{-8}$. The GBLUP and its reliability for additive values remained unchanged from the GREML_CE program using genomic relationships.

Combining results from the above two steps verifies Equation [12.1.7]:

$$\hat{\sigma}^2_\alpha = k_1\hat{\sigma}^2_{\alpha o} = (4.48)(0.0823222) = 0.3688034 .$$

**Exercise 13.3.2**

Use GREML_MME.SAS to verify the relationship between the additive variance under the RQG models and the additive variance under the QG model as described by Equation [12.1.7]. This verification is done in two runds:

1) Run GREML_MME.SAS using $\mathbf{T}_\alpha$ matrix:
   TA = TA_1;

The GREML estimate of $\sigma^2_\alpha$ is $\sigma^2_\alpha = 0.3688034$ at iteration 12 with tolerance $<10^{-8}$, and $k_1 = 4.48$
.

2) Run GREML_CE.SAS using $\mathbf{W}_\alpha$ :
   TA = WA;

Combining results from the above two steps verifies Equation [12.1.7].

**Exercise 13.3.3**

This exercise is the comparison between GREML_CE and GREML_MME. The two programs can have 6 runs with identical **G** matrix, GBLUP of additive values and reliability estimates :
:

- QG1-CE model: $\hat{\sigma}^2_{\alpha o} = 0.0823222$
  - $k_1 = 4.48$, $\hat{\sigma}^2_\alpha = 0.3688034$; $k_2 = 5.36$, $\hat{\sigma}^2_\alpha = 0.4412469$

**G**

| | | | | |
|---|---|---|---|---|
| 0.3984394 | -0.260138 | -0.079029 | 0.0526862 | -0.111958 |
| -0.260138 | 0.5630837 | -0.079029 | -0.111958 | -0.111958 |
| -0.079029 | -0.079029 | 0.4313682 | -0.095494 | -0.177816 |
| 0.0526862 | -0.111958 | -0.095494 | 0.2831883 | -0.128423 |
| -0.111958 | -0.111958 | -0.177816 | -0.128423 | 0.5301548 |

| GBLUP_A | REL_A |
|---|---|
| -0.26904 | 0.7274088 |
| 0.5077794 | 0.7996113 |
| -0.685917 | 0.6773303 |
| 0.382873 | 0.5469923 |
| 0.0643047 | 0.0048438 |

- RQG1-CE model, $k_1 = 4.48$, $\hat{\sigma}^2_\alpha = 0.3688034$

**G**

| | | | | |
|---|---|---|---|---|
| 0.3984394 | -0.260138 | -0.079029 | 0.0526862 | -0.111958 |
| -0.260138 | 0.5630837 | -0.079029 | -0.111958 | -0.111958 |
| -0.079029 | -0.079029 | 0.4313682 | -0.095494 | -0.177816 |
| 0.0526862 | -0.111958 | -0.095494 | 0.2831883 | -0.128423 |
| -0.111958 | -0.111958 | -0.177816 | -0.128423 | 0.5301549 |

| GBLUP_A | REL_A |
|---|---|
| -0.26904 | 0.7274088 |
| 0.5077794 | 0.7996113 |
| -0.685917 | 0.6773303 |
| 0.382873 | 0.5469923 |
| 0.0643047 | 0.0048438 |

- RQG1-CE model, $k_2 = 5.36$, $\hat{\sigma}^2_\alpha = 0.4412469$

**G**

| | | | | |
|---|---|---|---|---|
| 0.3984394 | -0.260138 | -0.079029 | 0.0526862 | -0.111958 |
| -0.260138 | 0.5630837 | -0.079029 | -0.111958 | -0.111958 |
| -0.079029 | -0.079029 | 0.4313682 | -0.095494 | -0.177816 |
| 0.0526862 | -0.111958 | -0.095494 | 0.2831883 | -0.128423 |
| -0.111958 | -0.111958 | -0.177816 | -0.128423 | 0.5301549 |

| GBLUP_A | REL_A |
|---|---|
| -0.26904 | 0.7274088 |
| 0.5077794 | 0.7996113 |
| -0.685917 | 0.6773303 |
| 0.382873 | 0.5469923 |
| 0.0643047 | 0.0048438 |

- QG2-MME model: $\hat{\sigma}^2_{\alpha o} = 0.0823222$
  - $k_1 = 4.48$, $\hat{\sigma}^2_{\alpha} = 0.3688034$; $k_2 = 5.36$, $\hat{\sigma}^2_{\alpha} = 0.4412469$

**G**

| | | | | |
|---|---|---|---|---|
| 0.3984394 | -0.260138 | -0.079029 | 0.0526862 | -0.111958 |
| -0.260138 | 0.5630837 | -0.079029 | -0.111958 | -0.111958 |
| -0.079029 | -0.079029 | 0.4313682 | -0.095494 | -0.177816 |
| 0.0526862 | -0.111958 | -0.095494 | 0.2831883 | -0.128423 |
| -0.111958 | -0.111958 | -0.177816 | -0.128423 | 0.5301548 |

| GBLUP_A | REL_A |
|---|---|
| -0.26904 | 0.7274088 |
| 0.5077794 | 0.7996113 |
| -0.685917 | 0.6773303 |
| 0.382873 | 0.5469923 |
| 0.0643047 | 0.0048438 |

- RQG2-MME model, $k_1 = 4.48$, $\hat{\sigma}^2_{\alpha} = 0.3688034$

**G**

| | | | | |
|---|---|---|---|---|
| 0.3984394 | -0.260138 | -0.079029 | 0.0526862 | -0.111958 |
| -0.260138 | 0.5630837 | -0.079029 | -0.111958 | -0.111958 |
| -0.079029 | -0.079029 | 0.4313682 | -0.095494 | -0.177816 |
| 0.0526862 | -0.111958 | -0.095494 | 0.2831883 | -0.128423 |
| -0.111958 | -0.111958 | -0.177816 | -0.128423 | 0.5301549 |

| GBLUP_A | REL_A |
|---|---|
| -0.26904 | 0.7274088 |
| 0.5077794 | 0.7996113 |
| -0.685917 | 0.6773303 |
| 0.382873 | 0.5469923 |
| 0.0643047 | 0.0048438 |

- RQG2-MME model, $k_2 = 5.36$, $\hat{\sigma}^2_\alpha = 0.4412469$

**G**

| | | | | |
|---|---|---|---|---|
| 0.3984394 | -0.260138 | -0.079029 | 0.0526862 | -0.111958 |
| -0.260138 | 0.5630837 | -0.079029 | -0.111958 | -0.111958 |
| -0.079029 | -0.079029 | 0.4313682 | -0.095494 | -0.177816 |
| 0.0526862 | -0.111958 | -0.095494 | 0.2831883 | -0.128423 |
| -0.111958 | -0.111958 | -0.177816 | -0.128423 | 0.5301549 |

| GBLUP_A | REL_A |
|---|---|
| -0.26904 | 0.7274088 |
| 0.5077794 | 0.7996113 |
| -0.685917 | 0.6773303 |
| 0.382873 | 0.5469923 |
| 0.0643047 | 0.0048438 |

## GREML_CE.SAS

```
/* ANSC8141: GREML_CE.SAS */
/* RQG1-CE MODEL: GREML, GBLUP, RELIABILITY */
/* 5 INDIVIDUALS, 10 MARKERS */
/* ORIGINAL VERSION OCTOBER 2012  */

PROC IML;
*RESET PRINT ;

VA = 2; VE = 7;
NB = 3; NU1=5;
NREC = 8;
N_IND =5;
NA0=NB+1;  ND0 = NB+NU1+1;
NA = NB+NU1; NC = NA;
Y = {1.7, 1.2, 1.3, 2.1, 2.3, 3.1, 4.2, 4.3};
X1 = J(8,1,1);
DX = {1,1,1,1,1,2,2,2}; * MISSING PHENOTYPE FOR IND 5;
X2 = DESIGN(DX);        * MISSING PHENOTYPE FOR IND 5;
X = X1||X2;
DZ = {1,1,1,2,2,3,4,4};
Z1 = DESIGN(DZ);
Z0 = J(8,1,0);          * MISSING PHENOTYPE FOR IND 5;
Z = Z1||Z0;             * MISSING PHENOTYPE FOR IND 5;
XX = X`*X;
XZ = X`*Z;
ZZ = Z`*Z;
XY = X`*Y;
ZY = Z`*Y;
C1 = XX||XZ||XZ;
RHS = XY//ZY//ZY;

/* TABLE 6.3: 10 MARKERS, 5 INDIVIDUALS */
/* LOC1:  0-0-1-0-1 */
/* LOC2:  2-0-1-1-1 */
/* LOC3:  0-0-1-0-1 */
/* LOC4:  1-1-2-0-1 */
/* LOC5:  0-2-0-1-1 */
/* LOC6:  2-0-1-1-0 */
/* LOC7:  0-2-2-0-0 */
/* LOC8:  1-0-2-2-0 */
/* LOC9:  1-2-2-2-0 */
/* LOC10: 2-2-1-1-0 */

/* CALCULATION OF ALLELE FREQUENCIES */
P1_II = 3/5;
```

```
P1_IJ = 2/5;
P1_JJ = 0/5;
P1_I = P1_II +0.5*P1_IJ;
P1_J = P1_JJ + 0.5*P1_IJ;

P2_II = 1/5;
P2_IJ = 3/5;
P2_JJ = 1/5;
P2_I = P2_II +0.5*P2_IJ;
P2_J = P2_JJ + 0.5*P2_IJ;

P3_II = 3/5;
P3_IJ = 2/5;
P3_JJ = 0/5;
P3_I = P3_II +0.5*P3_IJ;
P3_J = P3_JJ + 0.5*P3_IJ;

P4_II = 1/5;
P4_IJ = 3/5;
P4_JJ = 1/5;
P4_I = P4_II +0.5*P4_IJ;
P4_J = P4_JJ + 0.5*P4_IJ;

P5_II = 2/5;
P5_IJ = 2/5;
P5_JJ = 1/5;
P5_I = P5_II +0.5*P5_IJ;
P5_J = P5_JJ + 0.5*P5_IJ;

P6_II = 2/5;
P6_IJ = 2/5;
P6_JJ = 1/5;
P6_I = P6_II +0.5*P6_IJ;
P6_J = P6_JJ + 0.5*P6_IJ;

P7_II = 3/5;
P7_IJ = 0/5;
P7_JJ = 2/5;
P7_I = P7_II +0.5*P7_IJ;
P7_J = P7_JJ + 0.5*P7_IJ;

P8_II = 2/5;
P8_IJ = 1/5;
P8_JJ = 2/5;
P8_I = P8_II +0.5*P8_IJ;
P8_J = P8_JJ + 0.5*P8_IJ;

P9_II = 1/5;
```

```
P9_IJ = 1/5;
P9_JJ = 3/5;
P9_I = P9_II +0.5*P9_IJ;
P9_J = P9_JJ + 0.5*P9_IJ;

P10_II = 1/5;
P10_IJ = 2/5;
P10_JJ = 2/5;
P10_I = P10_II + 0.5*P10_IJ;
P10_J = P10_JJ + 0.5*P10_IJ;

/* 2PQ */
HET1 = 2*P1_I*P1_J;
HET2 = 2*P2_I*P2_J;
HET3 = 2*P3_I*P3_J;
HET4 = 2*P4_I*P4_J;
HET5 = 2*P5_I*P5_J;
HET6 = 2*P6_I*P6_J;
HET7 = 2*P7_I*P7_J;
HET8 = 2*P8_I*P8_J;
HET9 = 2*P9_I*P9_J;
HET10 = 2*P10_I*P10_J;
/* SUM OF 2PQ */
SUM_H = HET1+HET2+HET3+HET4+HET5+HET6+HET7+HET8+HET9+HET10;

/* ADDITIVE SNP CODING */
WA1_0 = 2*P1_J;
WA1_1 = P1_J - P1_I;
WA1_2 = -2*P1_I;

WA2_0 = 2*P2_J;
WA2_1 = P2_J - P2_I;
WA2_2 = -2*P2_I;

WA3_0 = 2*P3_J;
WA3_1 = P3_J - P3_I;
WA3_2 = -2*P3_I;

WA4_0 = 2*P4_J;
WA4_1 = P4_J - P4_I;
WA4_2 = -2*P4_I;

WA5_0 = 2*P5_J;
WA5_1 = P5_J - P5_I;
WA5_2 = -2*P5_I;

WA6_0 = 2*P6_J;
WA6_1 = P6_J - P6_I;
```

```
WA6_2 = -2*P6_I;

WA7_0 = 2*P7_J;
WA7_1 = P7_J - P7_I;
WA7_2 = -2*P7_I;

WA8_0 = 2*P8_J;
WA8_1 = P8_J - P8_I;
WA8_2 = -2*P8_I;

WA9_0 = 2*P9_J;
WA9_1 = P9_J - P9_I;
WA9_2 = -2*P9_I;

WA10_0 = 2*P10_J;
WA10_1 = P10_J - P10_I;
WA10_2 = -2*P10_I;

ZA1 =  WA1_0//WA1_0//WA1_1//WA1_0//WA1_1;
ZA2 =  WA2_2//WA2_0//WA2_1//WA2_1//WA2_1;
ZA3 =  WA3_0//WA3_0//WA3_1//WA3_0//WA3_1;
ZA4 =  WA4_1//WA4_1//WA4_2//WA4_0//WA4_1;
ZA5 =  WA5_0//WA5_2//WA5_0//WA5_1//WA5_1;
ZA6 =  WA6_2//WA6_0//WA6_1//WA6_1//WA6_0;
ZA7 =  WA7_0//WA7_2//WA7_2//WA7_0//WA7_0;
ZA8 =  WA8_1//WA8_0//WA8_2//WA8_2//WA8_0;
ZA9 =  WA9_1//WA9_2//WA9_2//WA9_2//WA9_0;
ZA10 = WA10_2//WA10_2//WA10_1//WA10_1//WA10_0;
WA = ZA1||ZA2||ZA3||ZA4||ZA5||ZA6||ZA7||ZA8||ZA9||ZA10;

IN = I(NREC);
WW_A = WA*WA`;
K1=SUM_H;
K2 = TRACE(WW_A)/N_IND;
TA_1 = WA/SQRT(K1);
TA_2 = WA/SQRT(K2);

*AG_1=WW_A;
AG_1 = TA_1*TA_1`;
*AG_2 = TA_2*TA_2`;

AG = AG_1;
*AG = AG_2;

PRINT K1 K2;

TOL=.00000001;
K=0;          DIFA=1000;             DIFE=1000;
```

```
*----- ITERATION STARTS ---------------------------------------;
START;
DO WHILE((DIFA>TOL | DIFE>TOL) & K <100);
ZAZ = Z*AG*Z`;
V = ZAZ*VA + IN*VE;
IV = INV(V);
XIVX = X`*IV*X;
IXIVX = GINV(XIVX);
XIVY = X`*IV*Y;
P = IV - IV*X*IXIVX*X`*IV;
YY_A = Y`*P*ZAZ*P*Y;
YY_E = Y`*P*P*Y;
TR_A = TRACE(P*ZAZ);
TR_E = TRACE(P);
   REMLA = YY_A*VA/TR_A;
   REMLE = YY_E*VE/TR_E;
DIFA=ABS(REMLA - VA);    DIFE=ABS(REMLE - VE);
VA=REMLA;      VE=REMLE;         K=K+1;
PRINT K VA VE DIFA DIFE;
*---- END OF ITERATIONS -------------------------------------;
END;

*VAK1=K1*VA;
*VAK2=K2*VA;
*PRINT VAK1 VAK2 K1 K2;

V = ZAZ*VA + IN*VE;
IV = INV(V);
XIVX = X`*IV*X;
IXIVX = GINV(XIVX);
P = IV - IV*X*IXIVX*X`*IV;
GBLUP_A = VA*AG*Z`*P*Y;
ZPZ = Z`*P*Z;
VHAT_A = DIAG(VECDIAG(VA*AG*ZPZ*AG`*VA));
G_A = AG*VA;
GA_DIAG = DIAG(VECDIAG(G_A));
IGA_DIAG = INV(GA_DIAG);
REL_A = VECDIAG(VHAT_A*IGA_DIAG);
PRINT GBLUP_A REL_A;
FINISH;
RUN;
```

## GREML_MME.SAS

```
/* ANSC8141: GREML_MME.SAS */
/* RQG2-MME MODEL: GREML, GBLUP, RELIABILITY */
/* 5 INDIVIDUALS, 10 MARKERS */
/* ORIGINAL VERSION SEPT/1/2012  */
PROC IML;
*RESET PRINT ;

VA = 2; VE = 7;
N_IND =5;
NMARKERS=10;
NU1=10;
NB = 3;
NREC = 8;
NA0=NB+1;
NA = NB+NU1;
NC = NA;

Y = {1.7, 1.2, 1.3, 2.1, 2.3, 3.1, 4.2, 4.3};
X1 = J(8,1,1);
DX = {1,1,1,1,1,2,2,2}; * MISSING PHENOTYPE FOR IND 5;
X2 = DESIGN(DX);        * MISSING PHENOTYPE FOR IND 5;
X = X1||X2;
DZ = {1,1,1,2,2,3,4,4};
Z1 = DESIGN(DZ);
Z0 = J(8,1,0);          * MISSING PHENOTYPE FOR IND 5;
Z = Z1||Z0;             * MISSING PHENOTYPE FOR IND 5;
XX = X`*X;
XZ = X`*Z;
ZZ = Z`*Z;
XY = X`*Y;
ZY = Z`*Y;


/* TABLE 6.3: 10 MARKERS, 5 INDIVIDUALS */
/* LOC1:  0-0-1-0-1 */
/* LOC2:  2-0-1-1-1 */
/* LOC3:  0-0-1-0-1 */
/* LOC4:  1-1-2-0-1 */
/* LOC5:  0-2-0-1-1 */
/* LOC6:  2-0-1-1-0 */
/* LOC7:  0-2-2-0-0 */
/* LOC8:  1-0-2-2-0 */
/* LOC9:  1-2-2-2-0 */
/* LOC10: 2-2-1-1-0 */

/* CALCULATION OF ALLELE FREQUENCIES */
```

```
P1_II = 3/5;
P1_IJ = 2/5;
P1_JJ = 0/5;
P1_I = P1_II +0.5*P1_IJ;
P1_J = P1_JJ + 0.5*P1_IJ;

P2_II = 1/5;
P2_IJ = 3/5;
P2_JJ = 1/5;
P2_I = P2_II +0.5*P2_IJ;
P2_J = P2_JJ + 0.5*P2_IJ;

P3_II = 3/5;
P3_IJ = 2/5;
P3_JJ = 0/5;
P3_I = P3_II +0.5*P3_IJ;
P3_J = P3_JJ + 0.5*P3_IJ;

P4_II = 1/5;
P4_IJ = 3/5;
P4_JJ = 1/5;
P4_I = P4_II +0.5*P4_IJ;
P4_J = P4_JJ + 0.5*P4_IJ;

P5_II = 2/5;
P5_IJ = 2/5;
P5_JJ = 1/5;
P5_I = P5_II +0.5*P5_IJ;
P5_J = P5_JJ + 0.5*P5_IJ;

P6_II = 2/5;
P6_IJ = 2/5;
P6_JJ = 1/5;
P6_I = P6_II +0.5*P6_IJ;
P6_J = P6_JJ + 0.5*P6_IJ;

P7_II = 3/5;
P7_IJ = 0/5;
P7_JJ = 2/5;
P7_I = P7_II +0.5*P7_IJ;
P7_J = P7_JJ + 0.5*P7_IJ;

P8_II = 2/5;
P8_IJ = 1/5;
P8_JJ = 2/5;
P8_I = P8_II +0.5*P8_IJ;
P8_J = P8_JJ + 0.5*P8_IJ;
```

```
P9_II = 1/5;
P9_IJ = 1/5;
P9_JJ = 3/5;
P9_I = P9_II +0.5*P9_IJ;
P9_J = P9_JJ + 0.5*P9_IJ;

P10_II = 1/5;
P10_IJ = 2/5;
P10_JJ = 2/5;
P10_I = P10_II + 0.5*P10_IJ;
P10_J = P10_JJ + 0.5*P10_IJ;

/* 2PQ */
HET1 = 2*P1_I*P1_J;
HET2 = 2*P2_I*P2_J;
HET3 = 2*P3_I*P3_J;
HET4 = 2*P4_I*P4_J;
HET5 = 2*P5_I*P5_J;
HET6 = 2*P6_I*P6_J;
HET7 = 2*P7_I*P7_J;
HET8 = 2*P8_I*P8_J;
HET9 = 2*P9_I*P9_J;
HET10 = 2*P10_I*P10_J;
/* SUM OF 2PQ */
SUM_H = HET1+HET2+HET3+HET4+HET5+HET6+HET7+HET8+HET9+HET10;

/* ADDITIVE SNP CODING */
WA1_0 = 2*P1_J;
WA1_1 = P1_J - P1_I;
WA1_2 = -2*P1_I;

WA2_0 = 2*P2_J;
WA2_1 = P2_J - P2_I;
WA2_2 = -2*P2_I;

WA3_0 = 2*P3_J;
WA3_1 = P3_J - P3_I;
WA3_2 = -2*P3_I;

WA4_0 = 2*P4_J;
WA4_1 = P4_J - P4_I;
WA4_2 = -2*P4_I;

WA5_0 = 2*P5_J;
WA5_1 = P5_J - P5_I;
WA5_2 = -2*P5_I;

WA6_0 = 2*P6_J;
```

```
WA6_1 = P6_J - P6_I;
WA6_2 = -2*P6_I;

WA7_0 = 2*P7_J;
WA7_1 = P7_J - P7_I;
WA7_2 = -2*P7_I;

WA8_0 = 2*P8_J;
WA8_1 = P8_J - P8_I;
WA8_2 = -2*P8_I;

WA9_0 = 2*P9_J;
WA9_1 = P9_J - P9_I;
WA9_2 = -2*P9_I;

WA10_0 = 2*P10_J;
WA10_1 = P10_J - P10_I;
WA10_2 = -2*P10_I;

ZA1 =  WA1_0//WA1_0//WA1_1//WA1_0//WA1_1;
ZA2 =  WA2_2//WA2_0//WA2_1//WA2_1//WA2_1;
ZA3 =  WA3_0//WA3_0//WA3_1//WA3_0//WA3_1;
ZA4 =  WA4_1//WA4_1//WA4_2//WA4_0//WA4_1;
ZA5 =  WA5_0//WA5_2//WA5_0//WA5_1//WA5_1;
ZA6 =  WA6_2//WA6_0//WA6_1//WA6_1//WA6_0;
ZA7 =  WA7_0//WA7_2//WA7_2//WA7_0//WA7_0;
ZA8 =  WA8_1//WA8_0//WA8_2//WA8_2//WA8_0;
ZA9 =  WA9_1//WA9_2//WA9_2//WA9_2//WA9_0;
ZA10 = WA10_2//WA10_2//WA10_1//WA10_1//WA10_0;
WA = ZA1||ZA2||ZA3||ZA4||ZA5||ZA6||ZA7||ZA8||ZA9||ZA10;

IN = I(NREC);
WW_A = WA*WA`;
K1=SUM_H;
K2 = TRACE(WW_A)/N_IND;
TA_1 = WA/SQRT(K1);
TA_2 = WA/SQRT(K2);

*TA = WA;
TA = TA_1;

PRINT K1 K2;
XY = X`*Y;
Z1 = Z*TA;
XZ1 = X`*Z1;
ZZ1 = Z1`*Z1;
Z1Y = Z1`*Y;
RHS = XY//Z1Y;
```

```
C1 = XX||XZ1;

IM = I(NMARKERS);
IN = I(NREC);

TOL=.00000001;
K=0;          DIFA=1000;          DIFE=1000;
*----- iteration starts ---------------------------------------;
start;
DO WHILE((DIFA>TOL | DIFE>TOL) & K <100);

RATIO_A = VE/VA;
CAA = ZZ1 + IM*RATIO_A;

C2 = XZ1`||CAA;
C = C1//C2;
IC = GINV(C);
PEV_AA = IC(|NA0:NA,NA0:NA|)*RATIO_A;
SOL = IC*RHS;
bhat=sol(|1:nb,* |);
Ahat=sol(|nA0:nA,*|);

TRACE_A = TRACE(PEV_AA);
DENOM_A = NU1 - TRACE_A;
AAHAT = AHAT`*AHAT;

ehat = y - x*bhat - Z1*Ahat;
EEHAT = EHAT`*EHAT;
C_RANK = TRACE(C*IC);
DENOM_E = C_RANK - TRACE_A;

   REMLA = AAHAT/DENOM_A;
   REMLE = EEHAT/(NREC - DENOM_E);

difa=abs(remla - va);      dife=abs(remle - ve);
va=remla;      ve=remle;          k=k+1;
print k va ve difa dife;
*---- end of iterations -------------------------------------;
end;

*VA_K1=K1*VA;
*VA_K2=K2*VA;
*PRINT VA_K1 VA_K2 K1 K2;

RATIO_A = VE/VA;
CAA = ZZ1 + IM*RATIO_A;
C2 = XZ1`||CAA;
C = C1//C2;
```

```
IC = GINV(C);
SOL = IC*RHS;
Ahat=sol(|nA0:nA,*|);
GBLUP_A = TA*AHAT;

PEV = IC(|NA0:NC,NA0:NC|);
PEV_AA = IC(|NA0:NA,NA0:NA|)*VE;
JM = J(NU1,1,1);
IM = I(NU1);

VHAT_ALFA = IM*VA - PEV_AA;
VHAT_A = TA*VHAT_ALFA*TA`;
G_A = TA*TA`*VA;
GA_DIAG = DIAG(VECDIAG(G_A));
IGA_DIAG = INV(GA_DIAG);
REL_A = VECDIAG(VHAT_A*IGA_DIAG);
PRINT GBLUP_A REL_A;
FINISH;
RUN;
```

## 13.4 GBLUP and GREML Exercises Using GVCBLUP

The purpose of these exercises is to gain hands-on experience for real data analysis of GBLUP and GREML using the GVCBLUP package, which is available at:

https://animalgene.umn.edu/gvcblup

Download the Windows version, GVCBLUP_Win.zip file. This zip file contains 3 folders: Test_data, win32, win64, and a pdf file of user manual.

The win32 and win64 folders each contains three compiled programs: greml_ce that implements the RQG1-CE methods, greml_qm that implements the RQG2-MME methods, and gcorrmx for calculating genomic relationships, along with a .dll file.

The Test_data has 2 data folders, p1k_m3k and p3k_m1k, along with a read_me file.

**Creation of working directories**

Copy all files in win32 or win64 folder to p1k_m3k and p3k_m1k folders so that the executable programs and the input data are in the same directory.

**Review of input files**

1) phenotypic data and SNP data,
2) parameter file,
3) 'readme' file and user manual.

**Running greml_ce and greml_qm programs in p3k_m1k folder**

1) greml_ce and greml_qm have identical results of GBLUP, reliability, variance components, and heritability estimates,
2) greml_ce and greml_qm have identical results of variance components at every iteration,
3) greml_ce and greml_qm required the same number of iterations,
4) greml_ce and greml_qm required different computing times.

**Running greml_ce and greml_qm programs in p1k_m3k folder**

1) greml_ce and greml_qm have identical results of GBLUP, reliability, variance components, and heritability estimates,
2) greml_ce and greml_qm have identical results of variance components at every iteration,
3) greml_ce and greml_qm required the same number of iterations,
4) greml_ce and greml_qm required different computing times.

**Comparison of greml_ce and greml_qm**

1) greml_ce and greml_qm have identical results for both p3k_m1k and p1k_m3k datasets,
2) greml_ce is faster than greml_qm for p1k_m3k,
3) greml_qm is faster than greml_ce for p3k_m1k.

**Comparison of EM-REML and AI-REML**

1) in the p3k_m1k folder, run greml_ce and greml_qm using EM-REML,
2) compare the number of iterations of EM-REML with that of AI-REML.

**Calculation of the accuracy of predicting phenotypic values**

This exrcise calculates the prediction accuracy of one validation population as an example of k-fold validations.

**Calculation of genomic relationship matrices**

This exrcise calculates genomic additive relationship matrix using Definitions I, II, IV and V.

# CHAPTER 14: GENOMIC PREDICTION USING HAPLOTYPES

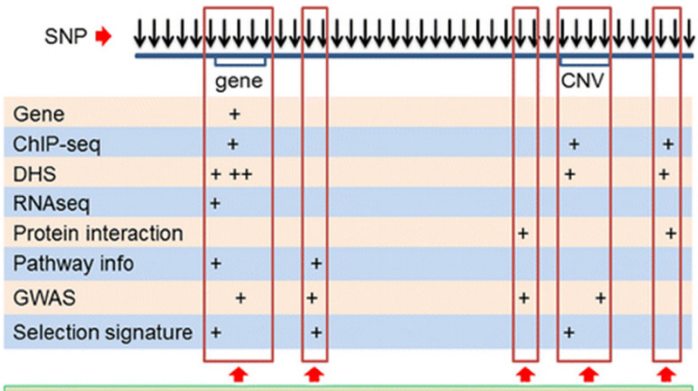
## Motivation: local high-order epistasis effects

- **Haplotype analysis opens many possibilities to use:**
  - functional genomic information
    - Gene, Chip-seq, RNA expression, GWAS results
  - Structural information
    - Linkage disequilibrium (LD blocks)
    - Fixed number of SNPs per haplotype block
    - Fixed distance per haplotype block
- **Examples of haplotype genomic prediction**
  - Wheat, maize, soybean, pig, sheep, cattle, chicken, human
  - Degree of successes varies: negative to substantial accuracy increases
  - Much fewer studies than those of single-locus genomic prediction
  - Computing aspect can be challenging
    - Difficult to exhaust all possible haplotype configurations

University of Minnesota

1

## Functional genomic information for haplotype blocks

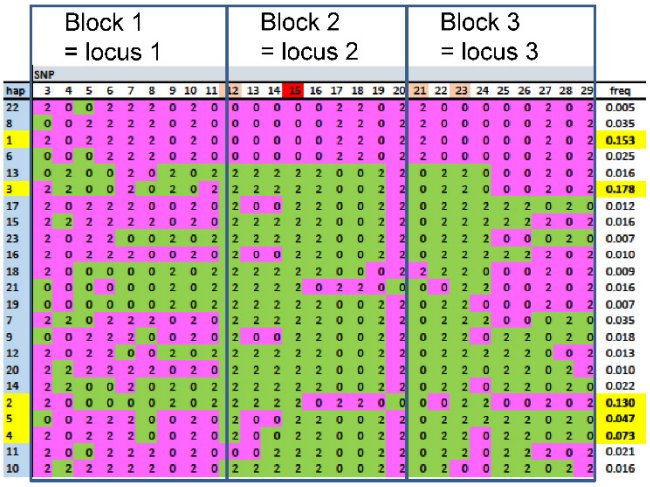



2

University of Minnesota

2

## Haplotype block = 'locus', haplotype = 'allele'




**Each SNP has 2 alleles: 0 and 2**

**27 SNPs in 3 blocks**

**9 SNPs per block**

**Each block as a locus**

**locus 1:**
**18 alleles (haplotypes)**
**locus 2:**
**7 alleles (haplotypes)**
**locus 3:**
**14 alleles (haplotypes)**

University of Minnesota

3

## Notations: one haplotype as one 'allele'

| Allele, frequency | $A_1$, $p_1$ | $A_2$, $p_2$ | $A_3$, $p_3$ | $A_4$, $p_4$ |
|---|---|---|---|---|
| $A_1$, $p_1$ | $A_1A_1$<br>$P_{11}$<br>$g_{11}$ | $A_1A_2$<br>$P_{12}$<br>$g_{12}$ | $A_1A_3$<br>$P_{13}$<br>$g_{13}$ | $A_1A_4$<br>$P_{14}$<br>$g_{14}$ |
| $A_2$, $p_2$ | | $A_2A_2$<br>$P_{22}$<br>$g_{22}$ | $A_2A_3$<br>$P_{23}$<br>$g_{23}$ | $A_2A_4$<br>$P_{24}$<br>$g_{24}$ |
| $A_3$, $p_3$ | | | $A_3A_3$<br>$P_{33}$<br>$g_{33}$ | $A_3A_4$<br>$P_{34}$<br>$g_{34}$ |
| $A_4$, $p_4$ | | | | $A_4A_4$<br>$P_{44}$<br>$g_{44}$ |

$A_i$ = allele i

$p_i$ = allele frequency

$A_iA_j$ = genotype ij

$P_{ij}$ = genotypic frequency

$g_{ij}$ = genotypic value

University of Minnesota

4

## Frequency and mean

- **allelic array of h alleles**

  $\sum_{i=1}^{h} p_i A_i, \quad A_i$ = allel i, $p_i$ = allele frequency of $A_i$

- **genotypic array of h(h+1)/2 genotypes**

  $\sum_{i=1}^{h}\sum_{j=1}^{h} P_{ij}A_iA_j$, $P_{ij}$ = genotypic frequency of $A_iA_j$ genotype

- **Hardy-Weinberg equilibrium (HWE):**

  $\sum_{i=1}^{h}\sum_{j=1}^{h} P_{ij}A_iA_j = (\sum_{i=1}^{h} p_i A_i)^2$

- **allele frequency calculated from genotypic frequencies**

  $p_i = P_{ii} + \frac{1}{2}\sum_{j\neq i}^{h} P_{ij}$

- **multi-allelic mean**

  $\mu_i = [2P_{ii}g_{ii} + \sum_{j\neq i}^{h} P_{ij}g_{ij}]/[2P_{ii} + \sum_{j\neq i}^{h} P_{ij}] = \sum_{j=1}^{h} p_j g_{ij}$

- **population mean**

  $\mu = \sum_{i=1}^{h}\sum_{j=1}^{h} P_{ij}g_{ij} = \sum_{i=1}^{h} p_i^2 g_{ii} + 2\sum_{i=1}^{h-1}\sum_{j=i+1}^{h} p_i p_j g_{ij} = \sum_{k=1}^{h} p_k \mu_k$

University of Minnesota

5

## Quantitative genetics definitions of effect and value

- **allelic effect:** $a_i = \mu_i - \mu = \sum_{j\neq i}^{h} p_j(\mu_i - \mu_j) = \sum_{j\neq i}^{h} p_j \alpha_{ij}$
- **additive effect:**

  $\alpha_{ij} = a_i - a_j = \mu_i - \mu_j = \sum_{k=1}^{h} p_k(g_{ik} - g_{jk}) = -\alpha_{ji}$

- **h(h-1)/2 additive effects are possible for h alleles**
- **h-1 independent additive effects, e.g.,**

  $\alpha_{1j} = a_1 - a_j = \mu_1 - \mu_j, \quad j = 2,\ldots,h$

- **additive value:** $a_{ij} = a_i + a_j$
- **dominance effect :** $\delta_{ij} = g_{ij} - \frac{1}{2}(g_{ii} + g_{jj})$
- **dominance value** (dominance deviation)**:** $d_{ij} = g_{ij} - \mu - a_{ij}$

University of Minnesota

6

## Multi-allelic genetic values and variances

- **additive (breeding) values**

$$a_{ij} = a_i + a_j = -(1-2p_i)\alpha_{1i} - (1-2p_j)\alpha_{1j} + 2\sum_{k\neq ij}^{h} p_k \alpha_{1k}$$

$$a_{ii} = 2a_i = -2(1-2p_i)\alpha_{1i} + 2\sum_{k\neq ij}^{h} p_k \alpha_{1k}$$

- **dominance values (deviations)**

$$d_{ij} = g_{ij} - \mu - a_i - a_j = [1 - p_i(1-p_j) - p_j(1-p_i)]\delta_{ij} - (1-2p_i)\sum_{k\neq i,j}^{h} p_k \delta_{ik} - (1-2p_j)\sum_{f\neq i,j}^{h} p_f \delta_{jf} + 2\sum_{k\neq i,j}^{h-1} p_k \sum_{f=k+1}^{h} p_f \delta_{kf}$$

$$d_{ii} = g_{ii} - \mu - 2a_i = -2(1-p_i)\sum_{k\neq i}^{h} p_k \delta_{ik} + 2\sum_{k\neq i}^{h-1} p_k \sum_{f=k+1}^{h} p_f \delta_{kf}$$

- **genotypic values and variance**

$$g_{ij} = \mu + a_{ij} + d_{ij} \qquad \sigma_g^2 = \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j g_{ij}^2 - \mu^2$$

- **additive and dominance variances**

$$\sigma_a^2 = \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j a_{ij}^2 \qquad \sigma_d^2 = \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j d_{ij}^2$$

University of Minnesota

7

## Multi-allelic partition of genotypic values

$$g_{ij} = \mu + a_{ij} + d_{ij} = \mu + \sum_{k=2}^{h} w_{\alpha}^{ij,k}\alpha_{1k} + \sum_{k=1}^{h-1}\sum_{f=k+1}^{h} w_{\delta}^{ij,kf}\delta_{kf}$$

$w_{\alpha}^{ij,k} = 2p_k$ for i,j≠k ( $a_{ij}$ and $\alpha_{1k}$ do not share allele k)

$w_{\alpha}^{ij,k} = -(1-2p_k)$ for i≠j but i=k or j=k ( $a_{ij}$ and $\alpha_{1k}$ share allele k, i≠j)

$w_{\alpha}^{ij,k} = -2(1-p_k)$ for i=j=k ( $a_{ij}$ and $\alpha_{1k}$ share allele k, i=j)

$w_{\delta}^{ij,kf} = 1 - p_i(1-p_j) - p_j(1-p_i)$ for ij=kf ( $d_{ij}$ and $\delta_{kf}$ share 2 alleles)

$w_{\delta}^{ij,kf} = -p_k(1-2p_i)$ for i≠j and i=f ( $d_{ij}$ and $\delta_{kf}$ share allele f, i≠j)

$w_{\delta}^{ij,kf} = -p_f(1-2p_j)$ for i≠j and j=k ( $d_{ij}$ and $\delta_{kf}$ share allele k, i≠j)

$w_{\delta}^{ij,kf} = -2p_k(1-p_i)$ for i=j and i=f ( $d_{ij}$ and $\delta_{kf}$ share allele f, i=j)

$w_{\delta}^{ij,kf} = 2p_k p_f$ for i,j≠k,f ( $d_{ij}$ and $\delta_{kf}$ share no allele, i=j or i≠j)

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{a} + \mathbf{d} = \mathbf{1}\mu + \mathbf{W}_{\alpha}\boldsymbol{\alpha}_{ho} + \mathbf{W}_{\delta}\boldsymbol{\delta}_{ho}$$

University of Minnesota

8

## Numerical example confirming multi-allelic partition

| Haplotype | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Frequency | 0.4 | 0.3 | 0.2 | 0.1 |

| Haplotype | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | $g_{11}$=25 | $g_{12}$=18 | $g_{13}$=15 | $g_{14}$=10 |
| 2 | | $g_{22}$=30 | $g_{23}$=33 | $g_{24}$=40 |
| 3 | | | $g_{33}$=17 | $g_{34}$=12 |
| 4 | | | | $g_{44}$=35 |

$$\sigma_g^2 = \text{genotypic variance} = \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j g_{ij}^2 - \mu^2 = 71.0419$$

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 \, ?$$

$$\sigma_a^2 + \sigma_d^2 = 71.0419?$$

University of Minnesota

9

## Estimated additive effects and values

- **additive effects:** $\boldsymbol{\alpha}_{ho}' = [-7.4 \quad -1.1 \quad -2.5]'$
- **additive values and variance**

$$\mathbf{a}_h = \begin{bmatrix} a_{11} \\ a_{22} \\ a_{33} \\ a_{44} \\ a_{12} \\ a_{13} \\ a_{14} \\ a_{23} \\ a_{24} \\ a_{34} \end{bmatrix} = \begin{bmatrix} 2p_2 & 2p_3 & 2p_4 \\ -2(1-p_2) & 2p_3 & 2p_4 \\ 2p_2 & -2(1-p_3) & 2p_4 \\ 2p_2 & 2p_3 & -2(1-p_4) \\ -(1-2p_2) & 2p_3 & 2p_4 \\ 2p_2 & -(1-2p_3) & 2p_4 \\ 2p_2 & 2p_3 & -(1-2p_4) \\ -(1-2p_2) & -(1-2p_3) & 2p_4 \\ -(1-2p_2) & 2p_3 & -(1-2p_4) \\ 2p_2 & -(1-2p_3) & -(1-2p_4) \end{bmatrix} \begin{bmatrix} \alpha_{12} \\ \alpha_{13} \\ \alpha_{14} \end{bmatrix} = \begin{bmatrix} 0.6 & 0.4 & 0.2 \\ -1.4 & 0.4 & 0.2 \\ 0.6 & -1.6 & 0.2 \\ 0.6 & 0.4 & -1.8 \\ -0.4 & 0.4 & 0.2 \\ 0.6 & -0.6 & 0.2 \\ 0.6 & 0.4 & -0.8 \\ -0.4 & -0.6 & 0.2 \\ -0.4 & 0.4 & -0.8 \\ 0.6 & -0.6 & -0.8 \end{bmatrix} \begin{bmatrix} -7.4 \\ -1.1 \\ -2.5 \end{bmatrix} = \begin{bmatrix} -5.38 \\ 9.42 \\ -3.18 \\ -0.38 \\ 2.02 \\ -4.28 \\ -2.88 \\ 3.12 \\ 4.52 \\ -1.78 \end{bmatrix}$$

$$\sigma_a^2 = \sum_{i=1}^{h}\sum_{j=1}^{h} p_i p_j a_{ij}^2 = 20.1178$$

University of Minnesota

10

## Estimated dominance effects and values

- **dominance effects:** $\boldsymbol{\delta}_{ho}' = [-9.5 \quad -6 \quad -20 \quad 9.5 \quad 7.5 \quad -14]'$
- **dominance values and variance**

$$\mathbf{d}_h = \begin{bmatrix} d_{11} \\ d_{22} \\ d_{33} \\ d_{44} \\ d_{12} \\ d_{13} \\ d_{11} \\ d_{23} \\ d_{24} \\ d_{31} \end{bmatrix} = \begin{bmatrix} -2p_2(1-p_1) & -2p_3(1-p_1) & -2p_4(1-p_1) & 2p_2p_3 & 2p_2p_4 & 2p_3p_4 \\ -2p_1(1-p_2) & 2p_1p_3 & 2p_1p_4 & -2p_3(1-p_2) & -2p_4(1-p_2) & 2p_3p_4 \\ 2p_1p_2 & -2p_1(1-p_3) & 2p_1p_4 & -2p_2(1-p_3) & 2p_2p_4 & -2p_4(1-p_3) \\ 2p_1p_2 & 2p_1p_3 & -2p_1(1-p_4) & 2p_2p_3 & -2p_2(1-p_4) & -2p_3(1-p_4) \\ w_\delta^{12,12} & -p_3(1-2p_1) & -p_4(1-2p_1) & -p_3(1-2p_2) & -p_4(1-2p_2) & 2p_3p_4 \\ -p_2(1-2p_1) & w_\delta^{13,13} & -p_4(1-2p_1) & -p_2(1-2p_3) & 2p_2p_4 & -p_4(1-2p_3) \\ -p_2(1-2p_1) & -p_3(1-2p_1) & w_\delta^{14,14} & 2p_2p_3 & -p_2(1-2p_4) & -p_3(1-2p_4) \\ -p_1(1-2p_2) & -p_1(1-2p_3) & 2p_1p_4 & w_\delta^{23,23} & -p_4(1-2p_2) & -p_4(1-2p_3) \\ -p_1(1-2p_2) & 2p_1p_3 & -p_1(1-2p_4) & -p_3(1-2p_2) & w_\delta^{24,24} & -p_3(1-2p_4) \\ 2p_1p_2 & -p_1(1-2p_3) & -p_1(1-2p_4) & -p_2(1-2p_3) & -p_2(1-2p_4) & w_\delta^{34,34} \end{bmatrix} \begin{bmatrix} \delta_{12} \\ \delta_{13} \\ \delta_{14} \\ \delta_{23} \\ \delta_{24} \\ \delta_{34} \end{bmatrix} = \begin{bmatrix} 8.29 \\ -1.51 \\ -1.91 \\ 13.29 \\ -6.11 \\ -2.81 \\ -9.21 \\ 7.79 \\ 13.39 \\ -8.31 \end{bmatrix}$$

$$\sigma_d^2 = \sum_{i=1}^{h} \sum_{j=1}^{h} p_i p_j d_{ij}^2 = 50.9241$$

University of Minnesota

11

## Genotypic values and variance based on partitioning

- **genotypic values from genetic partition**

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{a}_h + \mathbf{d}_h = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} (22.09) + \begin{bmatrix} -5.38 \\ 9.42 \\ -3.18 \\ -0.38 \\ 2.02 \\ -4.28 \\ -2.88 \\ 3.12 \\ 4.52 \\ -1.78 \end{bmatrix} + \begin{bmatrix} 8.29 \\ -1.51 \\ -1.91 \\ 13.29 \\ -6.11 \\ -2.81 \\ -9.21 \\ 7.79 \\ 13.39 \\ -8.31 \end{bmatrix} = \begin{bmatrix} 25 \\ 30 \\ 17 \\ 35 \\ 18 \\ 15 \\ 10 \\ 33 \\ 40 \\ 12 \end{bmatrix} = \begin{bmatrix} g_{11} \\ g_{22} \\ g_{33} \\ g_{44} \\ g_{12} \\ g_{13} \\ g_{14} \\ g_{23} \\ g_{24} \\ g_{34} \end{bmatrix}$$

- **genotypic variance from genetic partition**

$$\sigma_g^2 = \sigma_a^2 + \sigma_d^2 = 20.1178 + 50.9241 = 71.0419$$

University of Minnesota

12

## Additive multi-allelic haplotype mixed model

- **Mixed model with haplotype effects**

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZW}_{\alpha h}\boldsymbol{\alpha}_{ho} + \mathbf{e} = \mathbf{Xb} + \mathbf{ZT}_{\alpha h}\boldsymbol{\alpha}_{h} + \mathbf{e} = \mathbf{Xb} + \mathbf{Za}_{h} + \mathbf{e}$$

$$\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \mathbf{ZG}_{\alpha h}\mathbf{Z}' + \sigma_e^2\mathbf{I}_N = \mathbf{Z}(\sigma_{\alpha h}^2\mathbf{A}_h)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N$$

$$\mathbf{T}_{\alpha h} = \mathbf{W}_{\alpha h} / k_{\alpha h}^{1/2} \qquad \boldsymbol{\alpha}_h = k_{\alpha h}^{1/2}\boldsymbol{\alpha}_{ho}$$

$k_{\alpha h}$ = average of diagonal elements of $\mathbf{W}_{\alpha h}\mathbf{W}_{\alpha h}'$

- **Haplotype genomic relationship matrix**

$$\mathbf{A}_h = \mathbf{T}_{\alpha h}\mathbf{T}_{\alpha h}' = \mathbf{W}_{\alpha h}\mathbf{W}_{\alpha h}' / k_{\alpha h}$$

- **GBLUP and GREML**

$$\hat{\mathbf{a}}_h = \sigma_{\alpha h}^2\mathbf{A}_h\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) \qquad \hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$\sigma_{\alpha h}^{2\,(k+1)} = \sigma_{\alpha h}^{2\,(k)}\mathbf{y}\mathbf{P}^{(k)}\mathbf{Z}\mathbf{A}_h\mathbf{Z}'\mathbf{P}^{(k)}\mathbf{y} / \mathrm{tr}(\mathbf{P}^{(k)}\mathbf{Z}\mathbf{A}_h\mathbf{Z}')$$

$$\sigma_e^{2(k+1)} = \sigma_e^{2(k)}\mathbf{y}\mathbf{P}^{(k)}\mathbf{P}^{(k)}\mathbf{y} / \mathrm{tr}(\mathbf{P}^{(k)})$$

University of Minnesota

13

## Seven models of GVCHAP (Prakapenka et al., 2020)

**4 haplotype models**:

- Model 1: a + d + h
- Model 2: a + h
- Model 3: d + h
- Model 4: h

a = SNP additive, d = SNP dominance, h = haplotype additive

**3 SNP models:**

- Model 5: a + d
- Model 6: a
- Model 7: d

**Model selection:**

- Heritability > a required value such as 1%
- The effect type increases prediction accuracy

University of Minnesota

14

## Accuracy and heritability of haplotype genomic prediction

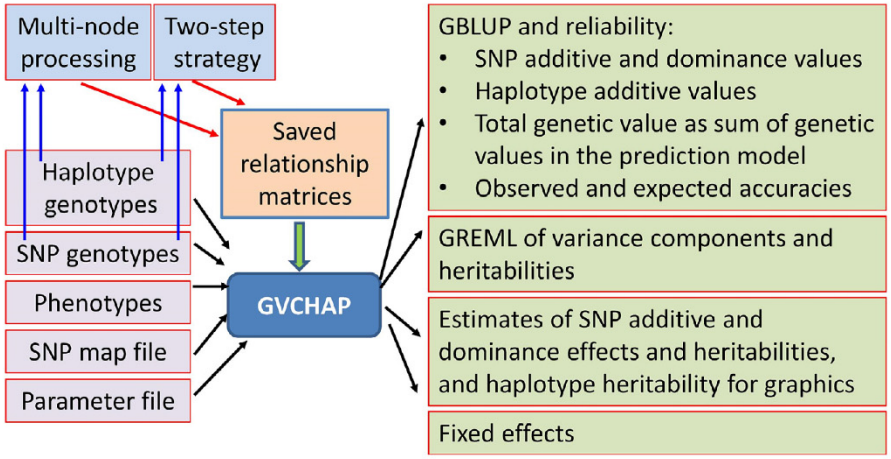
- **Cross-validations to evaluate prediction accuracy**
  - Correlation between haplotype GBLUP and phenotypic values in validation populations ($R_h$)
  - Correlation between single-locus GBLUP and phenotypic values in validation populations ($R_s$)
  - Accuracy increase due to haplotypes = $(R_h - R_s)/R_s$
- **Haplotype epistasis heritability** (Liang et al., 2020)

$\hat{h}^2_E = \hat{h}^2_g - \hat{h}^2_s = \hat{h}^2_{\alpha h} - \hat{h}^2_{\alpha 1}$ for haplotype-only and additive-only models

University of Minnesota

15

## GVCHAP input and output files (Prakapenka et al., 2020)

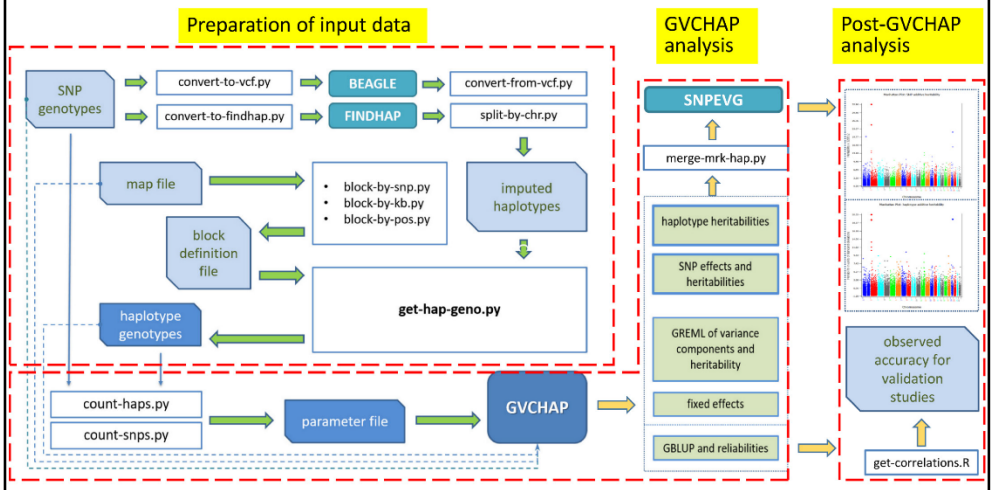



16

University of Minnesota

16

## GVCHAP computing pipeline (Prakapenka et al., 2020)




University of Minnesota

17

## The FHS example of haplotype genomic prediction

(Liang et al., 2020. Frontiers in Genetics, 1:588907)

**Framingham Hear Study (FHS) data**

- 3657-7564 individuals for 7 traits
  - ✓ high density lipoproteins (**HDL**), low density lipoproteins (**LDL**),
  - ✓ total cholesterol (**TC**), triglycerides (**TG**), height (**HTo**),
  - ✓ weight (**WT**), and body mass index (**BMIo**)
- 8 SNP sets: 40,941-380,705 SNPs
  - ✓ Effect of SNP density on haplotype accuracy
  - ✓ Effect of MAF (0.05 vs 0.10) on haplotype accuracy

**Structural genomic information for haplotype blocks**

- Fixed number of SNPs per block, fixed distance per block

**Functional genomic information for haplotype blocks**

- Codding genes
- Noncoding genes
- Chip-seq sites

University of Minnesota

18

## Haplotype statistics using structural information

**TABLE 1** | Statistics of haplotype blocks defined by fixed chromosome distance (380K, MAF = 0.05).

| Distance (Kb) | 5 | 50 | 100 | 250 | 500 |
|---|---|---|---|---|---|
| Total number of haplotypes | 1,123,531 | 4,868,466 | 9,601,998 | 21,315,586 | 26,992,496 |
| Number of blocks | 92,993 | 47,701 | 25,774 | 10,596 | 5,339 |
| Average number of haplotypes per block | 12.08 | 102.06 | 372.55 | 2,011.66 | 5,055.72 |
| Minimum SNPs per block | 2 | 2 | 2 | 2 | 2 |
| Maximum SNPs per block | 16 | 50 | 75 | 148 | 260 |
| Average number of SNPs per block | 2.71 | 7.92 | 14.75 | 35.93 | 71.3 |

**TABLE 2** | Statistics of haplotype blocks defined by fixed number of SNPs (380K, MAF = 0.05).

| Number of SNPs per block | 2 | 12 | 22 | 30 | 50 |
|---|---|---|---|---|---|
| Total number of haplotypes | 1,472,716 | 6,676,664 | 13,241,867 | 18,101,920 | 26,127,332 |
| Number of blocks | 190,357 | 31,734 | 17,317 | 12,699 | 7,624 |
| Average number of haplotypes/block | 7.74 | 210.39 | 764.67 | 1,425.46 | 3,426.98 |
| Minimum distance in block (Kb) | 0.01 | 2.76 | 10.99 | 22.08 | 50.41 |
| Maximum distance in block (Kb) | 4,548.44 | 29,754.01 | 29,774.74 | 25,970.14 | 29,898.03 |
| Average distance per block (Kb) | 7.26 | 80.71 | 153.27 | 210.87 | 358.19 |

University of Minnesota

19

## Haplotype statistics using functional information

**TABLE 3** | Statistics of haplotype blocks defined by gene boundaries and ChIP-seq sites (380K, MAF = 0.05).

| | Autosome genes | Coding genes | Noncoding genes | ChIP-seq |
|---|---|---|---|---|
| Total number of haplotypes | 7,419,624 | 5,571,918 | 1,946,912 | 13,368,940 |
| Number of blocks | 18,080 | 12,676 | 10,111 | 21,474 |
| Average number of haplotypes per block | 410.38 | 439.56 | 192.55 | 622.56 |
| Minimum SNPs per block | 2 | 2 | 2 | 2 |
| Maximum SNPs per block | 87 | 87 | 64 | 104 |
| Average number of SNPs per block | 13.49 | 13.95 | 8.56 | 17.63 |
| Minimum distance in block (Kb) | 1.14 | 1.14 | 1.78 | 4.04 |
| Maximum distance in block (Kb) | 150.0 | 150.0 | 150.0 | 150.0 |
| Average distance per block (Kb) | 90.13 | 95.34 | 49.82 | 116.85 |
| Autosome coverage (Mb) | 1557.23 | 1158.16 | 463.33 | 2423.24 |
| % of autosomes | 48.53 | 36.08 | 14.44 | 75.51 |
| 4-Kb extended coverage (Mb) | 1629.55 | 1208.49 | 503.77 | 2509.14 |
| % of autosomes by 4-Kb extended coverage | 50.78 | 37.66 | 15.70 | 78.19 |

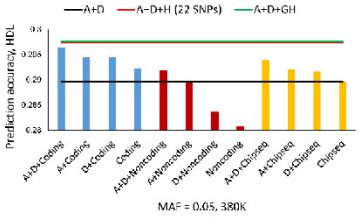
University of Minnesota

20

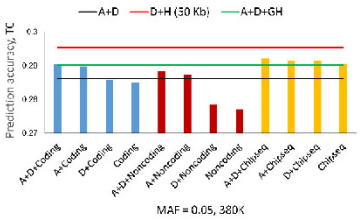
# Haplotypes of coding and non-coding genes affect phenotypes

## 1. Prediction accuracies

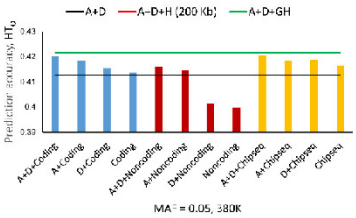
**a. High density lipoproteins**




**b. Total cholesterol**

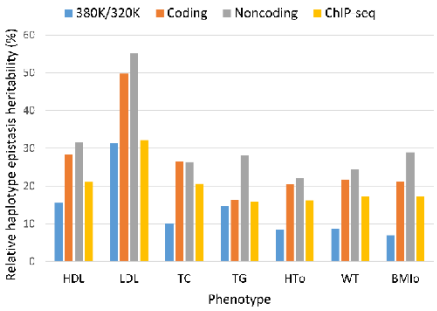



**c. Height**

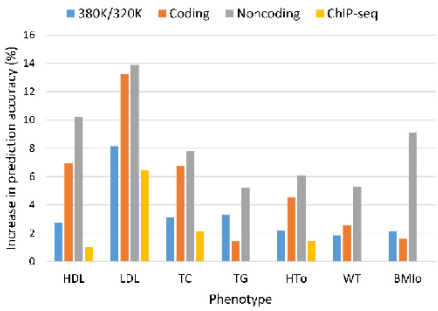



## 2. Heritability and accuracy increases

**d. Relative haplotype epistasis heritability**

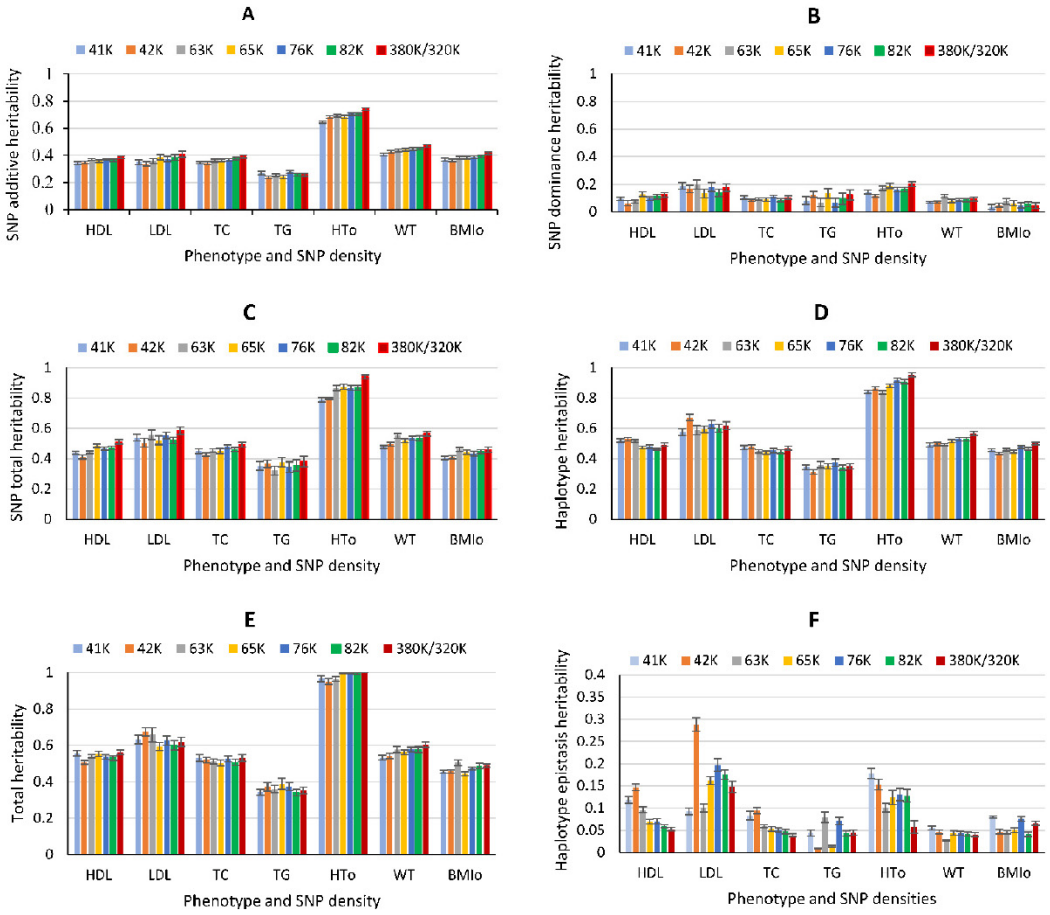



**e. Accuracy increase due to haplotypes**



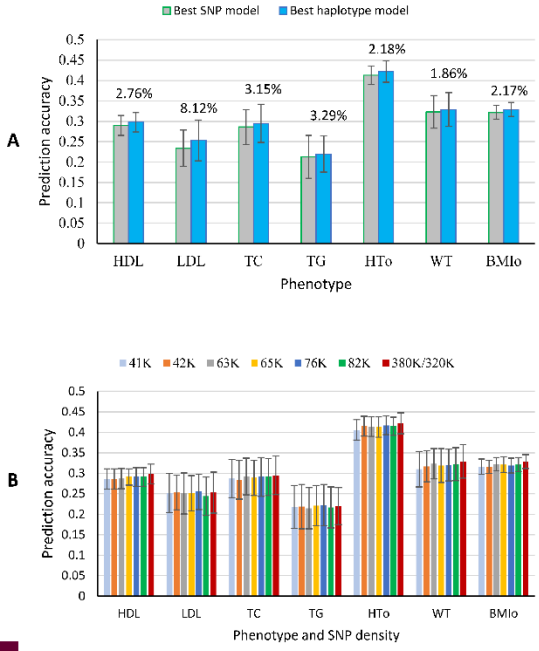

University of Minnesota

# Heritability estimates




University of Minnesota

## Accuracies of best prediction models




University of Minnesota

23

## Conclusion

- **Haplotype genomic prediction provides a method for genomic prediction to utilize and investigate:**
  - **Local high-order epistasis effects**
  - **Functional genomic information**
  - **Structural genomic information**
  - **Haplotypes from 41-380 K SNP sets had similar accuracies**
- **Numerous possible haplotype configurations are a computing challenge**

University of Minnesota

24

# CHAPTER 15: GENOMIC PREDICTION USING EPISTASIS EFFECTS

## 15.1 Two-locus Quantitative genetics (QG) Model with Additive and Dominance Effects

This section describes the QG-model of SNP additive and dominance effects of two loci that will be used for defining the epitasis models. Assuming two SNPs with alleles *A* and *a* at SNP 1 and *B* and *b* at SNP 2, the two-locus model with SNP additive and dominance effects for the nine two-locus genotypic values can be expressed as:

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{d}_1 + \mathbf{d}_2 = \mathbf{1}\mu + \mathbf{w}_{\alpha 1}\alpha_{1o} + \mathbf{w}_{\alpha 2}\alpha_{2o} + \mathbf{w}_{\delta 1}\delta_{1o} + \mathbf{w}_{\delta 2}\delta_{2o} \quad [15.1.1]$$

$$= \begin{bmatrix} g_{AABB} \\ g_{AABb} \\ g_{AAbb} \\ g_{AaBB} \\ g_{AaBb} \\ g_{Aabb} \\ g_{aaBB} \\ g_{aaBb} \\ g_{aabb} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} w_{\alpha 1}^{11} \\ w_{\alpha 1}^{11} \\ w_{\alpha 1}^{11} \\ w_{\alpha 1}^{12} \\ w_{\alpha 1}^{12} \\ w_{\alpha 1}^{12} \\ w_{\alpha 1}^{22} \\ w_{\alpha 1}^{22} \\ w_{\alpha 1}^{22} \end{bmatrix} \alpha_{1o} + \begin{bmatrix} w_{\alpha 2}^{11} \\ w_{\alpha 2}^{12} \\ w_{\alpha 2}^{22} \\ w_{\alpha 2}^{11} \\ w_{\alpha 2}^{12} \\ w_{\alpha 2}^{22} \\ w_{\alpha 2}^{11} \\ w_{\alpha 2}^{12} \\ w_{\alpha 2}^{22} \end{bmatrix} \alpha_{2o} + \begin{bmatrix} w_{\delta 1}^{11} \\ w_{\delta 1}^{11} \\ w_{\delta 1}^{11} \\ w_{\delta 1}^{12} \\ w_{\delta 1}^{12} \\ w_{\delta 1}^{12} \\ w_{\delta 1}^{22} \\ w_{\delta 1}^{22} \\ w_{\delta 1}^{22} \end{bmatrix} \delta_{1o} + \begin{bmatrix} w_{\delta 2}^{11} \\ w_{\delta 2}^{12} \\ w_{\delta 2}^{22} \\ w_{\delta 2}^{11} \\ w_{\delta 2}^{12} \\ w_{\delta 2}^{22} \\ w_{\delta 2}^{11} \\ w_{\delta 2}^{12} \\ w_{\delta 2}^{22} \end{bmatrix} \delta_{2o}$$

where

$\mathbf{g} = 9 \times 1$ column vector of genotypic values [15.1.2]

$\mu$ = common mean [15.1.3]

$\alpha_{io}$ = additive effect of SNP i ($i = 1, 2$) [15.1.4]

$\mathbf{w}_{\alpha i} = 9 \times 1$ model matrix of $\alpha_{io}$ [15.1.5]

$\delta_{io}$ = dominance effect of SNP i ($i = 1, 2$) [15.1.6]

$\mathbf{w}_{\delta 1} = 9 \times 1$ model matrix of $\delta_{io}$ [15.1.7]

$\mathbf{a}_i = \mathbf{w}_{\alpha i}\alpha_{io} = 9 \times 1$ column vector of additive values of SNP i ($i = 1, 2$) [15.1.8]

$\mathbf{d}_i = \mathbf{w}_{\delta 1}\delta_{1o} = 9 \times 1$ column vector of dominance values of SNP i ($i = 1, 2$) [15.1.9]

The additive codings of the genotypes of the $i^{th}$ SNP in $\mathbf{w}_{\alpha i}$ are $w_{\alpha i}^{11} = 2q_i$, $w_{\alpha i}^{12} = q_i - p_i$, $w_{\alpha i}^{22} = -2p_i$, and the dominance codings in $\mathbf{w}_{\delta 1}$ are $w_{\delta i}^{11} = -2q_i^2$, $w_{\delta i}^{12} = 2p_i q_i$, and $w_{\delta i}^{22} = -2p_i^2$ for $A_1A_1$, $A_1A_2$ and $A_2A_2$ genotypes of the $i^{th}$ SNP respectively, with $p_i$ = allele frequency of $A_1$ and $q_i$ = allele frequency of $A_2$. Equation (15.1.1) is the foundation for pairwise epistasis effects as the interaction between $\alpha_{1o}$ and $\alpha_{2o}$, $\alpha_{1o}$ and $\delta_{2o}$, $\delta_{1o}$ and $\alpha_{1o}$, and $\delta_{1o}$ and $\delta_{2o}$ under the assumption of linkage equilibrium (LE) that allows simplified genomic epistasis relationship matrices based on SNP additive and dominance relationship matrices without creating the epistasis

model matrices. Subscript 'o' denotes a genetic effect form the original quantitative genetics model and later will be removed in the raparameterized and equivalent model resulting from the use of genomic relationship matrices.

## 15.2 Pairwise Epistasis Effects and Values for Two Loci

The epistasis effects defined by Cockerham method can cover high-order epistasis effects. However, the biological significance of high-order epistasis is unknown. This study only considers second- and third-order epistasis. Based on the single-SNP model of two loci defined by Equation (15.1.1) and the epistasis effects defined by Cockerham (Cockerham, 1954), the model matrices of pairwise epistasis effects for n individuals are expressed as the Hadamard products between the model matrices of the additive and dominance effects:

$$\mathbf{E}_2 = \mathbf{w}_{\alpha 1} \# \mathbf{w}_{\alpha 2}(\alpha\alpha)_o + \mathbf{w}_{\alpha 1} \# \mathbf{w}_{\delta 2}(\alpha\delta)_o + \mathbf{w}_{\alpha 2} \# \mathbf{w}_{\delta 1}(\delta\alpha)_o + \mathbf{w}_{\delta 1} \# \mathbf{w}_{\delta 2}(\delta\delta)_o \quad [15.2.1]$$

where '#' indicates the Hadamard product, and

$(\alpha\alpha)_o$ = additive × additive (A×A) epistasis effect [15.2.2]
= allele × allele interaction effect

$(\alpha\delta)_o$ = additive × dominance (A×D) epistasis effects [15.2.3]
= allele × genotype interaction effect

$(\delta\alpha)_o$ = dominance × additive (D×A) epistasis effect [15.2.4]
= genotype × allele interaction effect

$(\delta\delta)_o$ = dominance × dominance (D×D) epistasis effect [15.2.5]

$\mathbf{w}_{\alpha 1} \# \mathbf{w}_{\alpha 2}$ = coefficients of $(\alpha\alpha)_o$ [15.2.6]

$\mathbf{w}_{\alpha 1} \# \mathbf{w}_{\delta 2}$ = coefficients of $(\alpha\delta)_o$ [15.2.7]

$\mathbf{w}_{\alpha 2} \# \mathbf{w}_{\delta 1}$ = coefficients of $(\delta\alpha)_o$ [15.2.8]

$\mathbf{w}_{\delta 1} \# \mathbf{w}_{\delta 2}$ = coefficients of $(\delta\delta)_o$ [15.2.9]

## 15.3 Pairwise Epistasis Effects and Values for Multiple Loci

For m SNPs and n individuals, the total epistasis values of the genome as the summation of all pairwise epistasis values defined by Equations [15.2.1]- [15.2.9] are:

$$\begin{aligned}
\mathbf{E}_2 &= \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(\mathbf{w}_{\alpha 1}^{i} \# \mathbf{w}_{\alpha 2}^{j})(\alpha\alpha)_o^{ij} + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(\mathbf{w}_{\alpha 1}^{i} \# \mathbf{w}_{\delta 2}^{j})(\alpha\delta)_o^{ij} \\
&+ \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(\mathbf{w}_{\delta 1}^{i} \# \mathbf{w}_{\alpha 2}^{j})(\delta\alpha)_o^{ij} + \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(\mathbf{w}_{\delta 1}^{i} \# \mathbf{w}_{\delta 2}^{j})(\delta\delta)_o^{ij} \\
&= \mathbf{W}_{\alpha\alpha}(\boldsymbol{\alpha\alpha})_o + [\mathbf{W}_{\alpha\delta}(\boldsymbol{\alpha\delta})_o + \mathbf{W}_{\delta\alpha}(\boldsymbol{\delta\alpha})_o] + \mathbf{W}_{\delta\delta}(\boldsymbol{\delta\delta})_o \quad [15.3.1] \\
&= \mathbf{W}_{\alpha\alpha}(\boldsymbol{\alpha\alpha})_o + \mathbf{W}_{\alpha\delta}^{(2)}(\boldsymbol{\alpha\delta})_o^{(2)} + \mathbf{W}_{\delta\delta}(\boldsymbol{\delta\delta})_o \\
&= \mathbf{aa} + \mathbf{ad} + \mathbf{da} + \mathbf{dd} = \mathbf{aa} + (\mathbf{ad})^{(2)} + \mathbf{dd}
\end{aligned}$$

where

$(\boldsymbol{\alpha\alpha})_o$ = $\binom{m}{2}$×1 column vector of A×A epistasis effects [15.3.2]

$= m(m-1)/2$ = allele × allele interaction effects

$\mathbf{W}_{\alpha\alpha} = n\times\binom{m}{2}$ model matrix of $(\boldsymbol{\alpha\alpha})_o$ [15.3.3]

$(\boldsymbol{\alpha\delta})_o = \binom{m}{2}\times 1$ column vector of A×D epistasis effects [15.3.4]

= allele × genotype interaction effects

$\mathbf{W}_{\alpha\delta} = n\times\binom{m}{2}$ model matrix of $(\boldsymbol{\alpha\delta})_o$ [15.3.5]

$(\boldsymbol{\delta\alpha})_o = \binom{m}{2}\times 1$ column vector of D×A epistasis effects [15.3.6]

= genotype × allele interaction effects,

$\mathbf{W}_{\delta\alpha} = n\times\binom{m}{2}$ model matrix of $(\boldsymbol{\delta\alpha})_o$ [15.3.7]

$(\boldsymbol{\delta\delta})_o = \binom{m}{2}\times 1$ column vector of D×D epistasis effects [15.3.8]

= genotype × genotype interaction effects

$\mathbf{W}_{\delta\delta} = n\times\binom{m}{2}$ model matrix of $(\boldsymbol{\delta\delta})_o$ [15.3.9]

$\mathbf{aa} = \mathbf{W}_{\alpha\alpha}(\boldsymbol{\alpha\alpha})_o = n\times 1$ column vector of genomic A×A epistasis values [15.3.10]

$\mathbf{ad} = \mathbf{W}_{\alpha\delta}(\boldsymbol{\alpha\delta})_o = n\times 1$ column vector of genomic A×D epistasis values [15.3.11]

$\mathbf{da} = \mathbf{W}_{\delta\alpha}(\boldsymbol{\delta\alpha})_o = n\times 1$ column vector of genomic D×A epistasis values [15.3.12]

$\mathbf{dd} = \mathbf{W}_{\delta\delta}(\boldsymbol{\delta\delta})_o = n\times 1$ column vector of genomic D×D epistasis values [15.3.13]

$(\boldsymbol{\alpha\delta})_o^{(2)} = [(\boldsymbol{\alpha\delta})_o', (\boldsymbol{\delta\alpha})_o']'$

$= 2\binom{m}{2}\times 1$ column vector of A×D and D×A epistasis effects [15.3.14]

$\mathbf{W}_{\alpha\delta}^{(2)} = (\mathbf{W}_{\alpha\delta}, \mathbf{W}_{\delta\alpha}) = n\times[2\binom{m}{2}]$ model matrix of $(\boldsymbol{\alpha\delta})_o^{(2)}$ [15.3.15]

$(\mathbf{ad})^{(2)} = \mathbf{W}_{\alpha\delta}^{(2)}(\boldsymbol{\alpha\delta})_o^{(2)}$

$= n\times 1$ column vector of genomic A×D and D×A epistasis values [15.3.16]

Combining the epistasis model of Equation [15.31] with the signle-locus model of Equation [1.4.1], the QG model of multiple loci for n individuals are:

$$\begin{aligned}\mathbf{g} &= \mathbf{1}\mu + \mathbf{W}_\alpha\boldsymbol{\alpha}_o + \mathbf{W}_\delta\boldsymbol{\delta}_o + \mathbf{W}_{\alpha\alpha}(\boldsymbol{\alpha\alpha})_o + \mathbf{W}_{\alpha\delta}^{(2)}(\boldsymbol{\alpha\delta})_o^{(2)} + \mathbf{W}_{\delta\delta}(\boldsymbol{\delta\delta})_o \\ &= \mathbf{1}\mu + \mathbf{a} + \mathbf{d} + \mathbf{aa} + (\mathbf{ad})^{(2)} + \mathbf{dd}\end{aligned} \quad [15.3.17]$$

The variance-covaraince matrix is:

$$\begin{aligned}\mathrm{var}(\mathbf{g}) = \mathbf{G} &= \mathrm{var}(\mathbf{g}) + \mathrm{var}(\mathbf{g}) + \mathrm{var}(\mathbf{g}) + \mathrm{var}(\mathbf{g}) + \mathrm{var}(\mathbf{g}) \\ &= \mathbf{G}_a + \mathbf{G}_d + \mathbf{G}_{aa} + \mathbf{G}_{ad} + \mathbf{G}_{\delta\delta} \\ &= \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}_\alpha' + \sigma^2_{\delta o}\mathbf{W}_\delta\mathbf{W}_\delta' + \sigma^2_{\alpha\alpha o}\mathbf{W}_{\alpha\alpha}\mathbf{W}_{\alpha\alpha}' + \sigma^2_{\alpha\delta o}\mathbf{W}^{(2)}_{\alpha\delta}\mathbf{W}^{(2)}_{\alpha\delta}' + \sigma^2_{\delta\delta o}\mathbf{W}_{\delta\delta}\mathbf{W}_{\delta\delta}'\end{aligned} \quad [15.3.18]$$

The QG model of Equations [15.3.1]-[15.3.18] provides an exact genetic understanding of the notations in the model. With this understanding, the notations of the QG model can be simplied using the multifactorial notations.

## 15.4 Multifactorial Notations for QG Model with Pairwise Epistasis Effects

Using multifactorial notations, the QG model of Equation [15.3.17] can be expressed as:

$$\mathbf{g} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{W}_i \boldsymbol{\tau}_{io} = \mu\mathbf{I} + \sum_{i=1}^{f} \mathbf{u}_i \quad [15.4.1]$$

where $\boldsymbol{\tau}_{io}$ = genetic effects of the $i^{th}$ effect type from the original QG model, $\mathbf{W}_i$ = model matrix of $\boldsymbol{\tau}_{io}$, $\mathbf{u}_i = \mathbf{W}_i \boldsymbol{\tau}_{io}$ = genetic values of the $i^{th}$ effect type, f = number pf genetic factors in the model (= 5 for Equation [15.3.17]).

The variance-covariance matrix of Equation [15.3.18] in multifactorial notations is:

$$\mathbf{G} = \sum_{i=1}^{f} \sigma^2_{io} \mathbf{W}_i \mathbf{W}_i' \quad [15.4.2]$$

## 15.5 Genomic Epistasis Relationship Matrices

### General formulations

Genomic relationship matrices are needed for calculating GBLUP and GREML. Based on the general multifactorial model of Equation [15.4.1], the genomic relationship matrices are:

$$\mathbf{S}_i = (\mathbf{W}_i \mathbf{W}_i') / K_i \text{ , } i=1,\ldots 5 \quad [15.5.1]$$

$$K_i = tr(\mathbf{W}_i \mathbf{W}_i')/n \text{ , } i=1,\ldots 5 \quad [15.5.2]$$

where $\mathbf{S}_i$ = genomic epistasis relationship matrix of the $i^{th}$ effect type, and $K_i$ = the average of the diagonal elements of $\mathbf{W}_i\mathbf{W}_i'$, noting that $K_i$ of Equation [15.5.2] may differ from $k_i$ of Equation [11.1.2] ot [11.1.4] for additive relationship matrix.

Equations [15.5.1] and [15.5.2] use the original model matrices. For epistasis effects, the model matrices ($\mathbf{W}_i$, i=3, 4, 5) are difficult or impossible to compute. Two methods are available for calculating Equations Equations [15.5.1] and [15.5.2] using the model matrices of additive and dominance effects ($\mathbf{W}_i$, i=1, 2) without creating the epistasis model matrices.

The computing difficulty due to creating epistasis model matrices can be removed by computing the approximate genomic epistasis relationship matrices (AGERM) as the genomic version of Henderson's Hadamard products between additive and dominance relationship matrices (Su et al., 2012; Muñoz et al., 2014; Vitezica et al., 2017), or by the exact genomic epistasis relationship matrices (EGERM). AGERM contains intra-locus epistasis that should not exist (Martini et al., 2020) and EGERM removes intra-locus epistasis from AGERM based on products between SNP genomic additive and dominance relationship matrices (Jiang and Reif, 2020; Martini et al., 2020). Although EGERM is theoretically more appealing than AGERM for being exact, the difference between these two methods in prediction accuracy and heritability estimates was nonexistent for a Holstein dataset with 78,964 SNPs (Liang et al., 2022) and was negligible for a swine dataset with 52,842 SNPs (Da et al., 2022), but EGERM required 21 times as much computing time as required by AGERM for the Holstein dataset, and required about 9 times as

much computing time as required by AGERM for the swine dataset in this article. The lack of difference between these two methods and the computing inefficiency of EGERM should favor AGERM for its mathematical simplicity and computing efficiency at least for datasets with 50,000 or more SNPs.

**Approximate genomic epistasis relationship matrices (AGERM)**

For pairwise epistasis effects, AGERM as the genomic version of Henderson's Hadamard products of pedigree additive and dominance relationship matrices (Henderson, 1985) (Equations [1.8.10]) are:

$$\mathbf{S}_3 = \mathbf{S}_{\alpha\alpha} = \mathbf{S}_1 \# \mathbf{S}_1 = \mathbf{A} \# \mathbf{A} = \text{A}\times\text{A relationship matrix} \quad [15.5.3]$$

$$\mathbf{S}_4 = \mathbf{S}_{\alpha\delta} = \mathbf{S}_1 \# \mathbf{S}_2 = \mathbf{A} \# \mathbf{D} = \text{A}\times\text{D relationship matrix} \quad [15.5.4]$$

$$\mathbf{S}_5 = \mathbf{S}_{\delta\delta} = \mathbf{S}_2 \# \mathbf{S}_2 = \mathbf{D} \# \mathbf{D} = \text{D}\times\text{D relationship matrix} \quad [15.5.5]$$

where

$$\mathbf{S}_1 = \mathbf{S}_\alpha = \mathbf{A} = \mathbf{W}_\alpha \mathbf{W}_\alpha' / (2\sum_{i=1}^{m} p_i q_i) = \mathbf{W}_1 \mathbf{W}_1' / (2\sum_{i=1}^{m} p_i q_i) \quad [15.5.6]$$

$$\mathbf{S}_2 = \mathbf{S}_\delta = \mathbf{D} = \mathbf{W}_\delta \mathbf{W}_\delta' / (4\sum_{i=1}^{m} p_i^2 q_i^2) = \mathbf{W}_2 \mathbf{W}_2' / (4\sum_{i=1}^{m} p_i^2 q_i^2) \quad [15.5.7]$$

The AGERM for pairwise epistasis effects of Equations [15.5.3]-[15.5.5] are based on the additive and dominance model matrices of additive and dominance effects of Equations [15.5.6]-[15.5.7] and avoids using any epistasis model matrices.

**Exact genomic epistasis relationship matrices (EGERM)**

The EGERM forlulations based on SNP additive and dominance model matrices (Jiang and Reif, 2020) are:

$$\mathbf{W}_3 \mathbf{W}_3' = \mathbf{W}_{\alpha\alpha} \mathbf{W}_{\alpha\alpha}' = \frac{1}{2}[(\mathbf{W}_\alpha \mathbf{W}_\alpha') \# (\mathbf{W}_\alpha \mathbf{W}_\alpha') - (\mathbf{W}_\alpha \# \mathbf{W}_\alpha)(\mathbf{W}_\alpha \# \mathbf{W}_\alpha)'] \quad [15.5.8]$$

$$\begin{aligned} \mathbf{W}_4 \mathbf{W}_4' &= \mathbf{W}_{\alpha\delta}^{(2)} \mathbf{W}_{\alpha\delta}^{(2)\prime} = \mathbf{W}_{\alpha\delta} \mathbf{W}_{\alpha\delta}' + \mathbf{W}_{\delta\alpha} \mathbf{W}_{\delta\alpha}' \\ &= \{[(\mathbf{W}_\alpha \mathbf{W}_\alpha') \# (\mathbf{W}_\delta \mathbf{W}_\delta') - (\mathbf{W}_\alpha \# \mathbf{W}_\delta)(\mathbf{W}_\alpha \# \mathbf{W}_\delta)'] \\ &\quad + [(\mathbf{W}_\alpha \mathbf{W}_\alpha') \# (\mathbf{W}_\delta \mathbf{W}_\delta') - (\mathbf{W}_\delta \# \mathbf{W}_\alpha)(\mathbf{W}_\delta \# \mathbf{W}_\alpha \#)']\} / 2 \\ &= [(\mathbf{W}_\alpha \mathbf{W}_\alpha') \# (\mathbf{W}_\delta \mathbf{W}_\delta') - (\mathbf{W}_\alpha \# \mathbf{W}_\delta)(\mathbf{W}_\alpha \# \mathbf{W}_\delta)'] \end{aligned} \quad [15.5.9]$$

$$\mathbf{W}_5 \mathbf{W}_5' = \mathbf{W}_{\delta\delta} \mathbf{W}_{\delta\delta}' = \frac{1}{2}[(\mathbf{W}_\delta \mathbf{W}_\delta') \# (\mathbf{W}_\delta \mathbf{W}_\delta') - (\mathbf{W}_\delta \# \mathbf{W}_\delta)(\mathbf{W}_\delta \# \mathbf{W}_\delta)'] \quad [15.5.10]$$

In Equation [15.5.9], $\mathbf{W}_\alpha \# \mathbf{W}_\delta = \mathbf{W}_\delta \# \mathbf{W}_\alpha$. For notation simplicity, the EGERM use the definitions of Equations [15.5.1] and [15.5.2]:

$$\mathbf{S}_3 = (\mathbf{W}_3 \mathbf{W}_3') / \text{K}_3 \quad [15.5.11]$$

$$\mathbf{S}_4 = (\mathbf{W}_4 \mathbf{W}_4') / \text{K}_4 \quad [15.5.12]$$

$$\mathbf{S}_5 = (\mathbf{W}_5\mathbf{W}_5')/K_5 \quad [15.5.13]$$

$$K_3 = tr(\mathbf{W}_3\mathbf{W}_3')/n \quad [15.5.14]$$
$$K_4 = tr(\mathbf{W}_4\mathbf{W}_4')/n \quad [15.5.15]$$
$$K_5 = tr(\mathbf{W}_5\mathbf{W}_5')/n \quad [15.5.16]$$

To be consistent with the $K_i$ notation for EGERM, the SNP additive and dominance relationship matrices also use definitions of Equations [15.5.1] and [15.5.2]:

$$\mathbf{S}_1 = (\mathbf{W}_1\mathbf{W}_1')/K_1 \quad [15.5.17]$$
$$\mathbf{S}_2 = (\mathbf{W}_2\mathbf{W}_2')/K_2 \quad [15.5.18]$$

$$K_1 = tr(\mathbf{W}_1\mathbf{W}_1')/n \quad [15.5.19]$$
$$K_2 = tr(\mathbf{W}_2\mathbf{W}_2')/n \quad [15.5.20]$$

With the understanding of the epistasis genomic relationship matrices, formulations of GBLUP and GREML can be developed.

## 15.6 Reparameterized MF Model, GBLUP, GREM

**Reparameterized multifactorial model**

The use of genomic relationship matrices for genomic prediction results in a reparameterized QG model. Using the multifactorial notations of the QG model, the reparameterized multifactorial (RMF) model is:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{g} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{Z}\textstyle\sum_{i=1}^{f}\mathbf{T}_i\boldsymbol{\tau}_i + \mathbf{e} \\ &= \mathbf{X}\mathbf{b} + \mathbf{Z}\textstyle\sum_{i=1}^{f}\mathbf{u}_i + \mathbf{e}\end{aligned} \quad [15.6.1]$$

$$\begin{aligned}\mathbf{V} &= \mathbf{Z}\mathbf{G}\mathbf{Z}' + \sigma_e^2\mathbf{I}_N = \mathbf{Z}(\textstyle\sum_{i=1}^{f}\mathbf{G}_i)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N \\ &= \mathbf{Z}(\textstyle\sum_{i=1}^{f}\sigma_i^2\mathbf{T}_i\mathbf{T}_i')\mathbf{Z}' + \sigma_e^2\mathbf{I}_N = \mathbf{Z}(\textstyle\sum_{i=1}^{f}\sigma_i^2\mathbf{S}_i)\mathbf{Z}' + \sigma_e^2\mathbf{I}_N\end{aligned} \quad [15.6.2]$$

where $\mathbf{y}$ = N×1 column vector of phenotypic observations, $\mathbf{Z}$ = N×n incidence matrix allocating phenotypic observations to each individual = identity matrix for one observation per individual (N = n), N = number of observations, n = number of individuals, $\mathbf{b}$ = c×1 column vector of fixed effects such as heard-year-season in dairy cattle, c = number of fixed effects, $\mathbf{X}$ = N×c model matrix of $\mathbf{b}$, $\mathbf{e}$ = N×1 column vector of random residuals, $\sigma_e^2$ = residual variance, and

$$\boldsymbol{\tau}_i = \sqrt{K_i}\,\boldsymbol{\tau}_{io} = \text{the genetic effects of the } i^{th} \text{ effect type} \quad [15.6.3]$$
$$\mathbf{T}_i = \mathbf{W}_i/\sqrt{K_i} = \text{model matrix of the genetic effects of the } i^{th} \text{ effect type} \quad [15.6.4]$$

$\mathbf{u}_i = \mathbf{T}_i\boldsymbol{\tau}_i$ = the genetic values of the $i^{th}$ effect type [15.6.5]

$\sigma_i^2 = K_i\sigma_{io}^2$ = common variance of the genetic effects of the $i^{th}$ effect type [15.6.6]

$\mathbf{G}_i = \sigma_i^2\mathbf{S}_i$ = variance-covariance matrix of the genetic values of the $i^{th}$ effect type [15.6.7]

$\mathbf{G} = \sum_{i=1}^{f}\mathbf{G}_i$ = variance-covariance matrix of all genetic values [15.6.8]

The phenotypic values (**y**) are assumed to follow a normal distribution with mean **Xb** and variance-covariance matrix of **V**. The methods described below for genomic estimation and prediction are based on the CE method because the MME method is computationally unfeasible for the epistasis model. Equations [15.6.3]-[15.6.7] are the reprameterization of the QG model in multifactorial notations of Equations [15.4.1] and [15.4.2].

Table 15.6.1 summarizes the exact correspondence of the notations between the QG model of Equations [15.3.17]-[15.3.18], the multifactorial (MF) model of Equations [15.4.1]-[15.4.2], and the reparameterized multifactorial (RMF) model of Equations [15.5.11]-[15.5.20] and [15.6.3]-[15.6.6].

## GBLUP and reliability

Based on the multifactorial genetic model of Equations 16 and 17, the GBLUP of the genetic values of the $i^{th}$ effect type ($\hat{\mathbf{u}}_i$) and the best linear unbiased estimator (BLUE) or generalized least squares (GLS) estimator of fixed effect ($\hat{\mathbf{b}}$) are:

$$\hat{\mathbf{u}}_i = \sigma_i^2\mathbf{S}_i\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \sigma_i^2\mathbf{S}_i\mathbf{Z}'\mathbf{P}\mathbf{y}, \quad i = 1,...,f \qquad [15.6.9]$$

where $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}$. The GBLUP of total genetic values of the n individuals is the summation of all types of genetic values:

$$\hat{\mathbf{g}} = \sum_{i=1}^{f}\hat{\mathbf{u}}_i \qquad [15.6.10]$$

The reliability of the GBLUP of the total genetic value (Equation 20) of the $j^{th}$ individual is:

$$R_{gj}^2 = [\mathbf{G}(\mathbf{Z}'\mathbf{P}\mathbf{Z}')\mathbf{G}']_{jj} / \theta_{jj} \qquad [15.6.11]$$

where $\mathbf{G} = \sum_{i=1}^{f}\mathbf{G}_i = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_i$ (Equation [15.6.8]), $\theta_{jj} = \sum_{i=1}^{f}G_i^{jj} = \sum_{i=1}^{f}\sigma_i^2\mathbf{S}_i^{jj}$, and subscript or superscript jj denotes the $j^{th}$ diagonal element. The reliability formula for any or a combination of genetic values can be readily derived from Equation 21, e.g., the reliability of $\hat{\mathbf{u}}_3$ (GBLUP of haplotype additive values) is obtained from Equation [15.6.11] by deleting all terms except $\mathbf{G}_3(\mathbf{Z}'\mathbf{P}\mathbf{Z}')\mathbf{G}_3'$ in the numerator and $\sigma_3^2\mathbf{S}_3^{jj}$ in the denominator, with changes in the **V** and **P** matrices accordingly.

## GREML and heritabilities

The EM-REML iterative algorithm for the multifactorial model of Equations [15.6.1] and [15.6.2] is:

$$\sigma_i^{2(j+1)} = \sigma_i^{2(j)}\mathbf{y}\mathbf{P}^{(j)}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'\mathbf{P}^{(j)}\mathbf{y} / \mathrm{tr}(\mathbf{P}^{(j)}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'),\ i=1,...,f \qquad [15.6.12]$$

$$\sigma_e^{2(j+1)} = \sigma_e^{2(j)}\mathbf{y}\mathbf{P}^{(k)}\mathbf{P}^{(j)}\mathbf{y} / \mathrm{tr}(\mathbf{P}^{(j)}) \qquad [15.6.13]$$

where j = iteration number. The estimate of the genomic heritability for each type of genetic effects ($h_i^2$) and the total heritability of all types of genetic effects ($H^2$) are:

$$h_i^2 = \sigma_i^2/\sigma_y^2 \qquad i=1,...,f \qquad [15.6.14]$$

$$H^2 = \sum_{i=1}^{f} h_i^2 \qquad [15.6.15]$$

where $\sigma_y^2 = \sum_{i=1}^{f}\sigma_i^2 + \sigma_e^2$ = phenotypic variance.

The heritability estimates of Equation [15.6.14] can be used for model selection by removing effect types with heritability estimates below a user determined threshold value from the prediction model. The total heritability of Equation [15.6.15] provides an estimate of the total genetic contribution to the phenotypic variance.

## AI GREML for reparameterized multifactorial model

The AI-REML iterative algorithm for the RMF model of Equations [15.6.1] and [15.6.2] is:

$$\boldsymbol{\theta}^{(j+1)} = \boldsymbol{\theta}^{(j)} + \left(\mathbf{AI}^{(j)}\right)^{-1}\boldsymbol{\Delta}^{(j)} \qquad [15.6.16]$$

where $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, ..., \sigma_f^2, \sigma_{f+1}^2)'$ = $(f+1)\times 1$ column vector of variance-covariance components, $\sigma_{f+1}^2 = \sigma_e^2$ = residual variance, $\boldsymbol{\Delta} = (\Delta_1, \Delta_2, ..., \Delta_f, \Delta_{f+1})'$ = $(f+1)\times 1$ column vector of the partial derivatives of the log residual likelihood function with respect to each variance component, and j = iteration number. A typical term in $\boldsymbol{\Delta}$ ($\Delta_i$) and a typical term in **AI** ($AI_{ik}$) are:

$$\begin{aligned}\Delta_i &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\right) + \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\mathbf{y} \\ &= -\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{Z}\mathbf{S}_i\mathbf{Z}') + \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}\mathbf{S}_i\mathbf{Z}'\mathbf{P}\mathbf{y}, \quad i=1,...,f+1\end{aligned} \qquad [15.6.17]$$

$$\begin{aligned}AI_{ik} &= \frac{1}{2}\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_k^2}\mathbf{P}\mathbf{y} \\ &= \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{Z}'\mathbf{S}_i\mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{S}_k\mathbf{Z}'\mathbf{P}\mathbf{y}, \quad i, k=1,...,f+1\end{aligned} \qquad [15.6.18]$$

where $\mathbf{S}_{f+1} = \mathbf{I}_N$. As discussed in Chapter 8, EM-type and AI-REML can be used jointly: EM-type converges but can be very slow, whereas AI-REML is fast but may fail.

**Table 15.6.1** Notations of the quantitative genetics (QG) model, multifactorial (MF) model, and reparameterized MF (RMF) model for genomic prediction.

| | Additive | Dominance | A×A | A×D | D×D |
|---|---|---|---|---|---|
| | Quantitative genetics (QG) model | | | | |
| Effect | $\boldsymbol{\alpha}_o$ | $\boldsymbol{\delta}_o$ | $(\boldsymbol{\alpha\alpha})_o$ | $(\boldsymbol{\alpha\delta})_o^{(2)}$ | $(\boldsymbol{\delta\delta})_o$ |
| Model matrix | $\mathbf{W}_\alpha$ | $\mathbf{W}_\delta$ | $\mathbf{W}_{\alpha\alpha}$ | $\mathbf{W}_{\alpha\delta}^{(2)}$ | $\mathbf{W}_{\delta\delta}$ |
| Value | $\mathbf{a} = \mathbf{W}_\alpha\boldsymbol{\alpha}_o$ | $\mathbf{d} = \mathbf{W}_\delta\boldsymbol{\delta}_o$ | $\mathbf{aa} = \mathbf{W}_{\alpha\alpha}(\boldsymbol{\alpha\alpha})_o$ | $(\mathbf{ad})^{(2)} = \mathbf{W}_{\alpha\delta}^{(2)}(\boldsymbol{\alpha\delta})_o^{(2)}$ | $\mathbf{dd} = \mathbf{W}_{\delta\delta}(\boldsymbol{\delta\delta})_o$ |
| var(effects) | $\sigma^2_{\alpha o}\,\mathbf{I}_n$ | $\sigma^2_{\alpha o}\,\mathbf{I}_n$ | $\sigma^2_{\alpha\alpha o}\,\mathbf{I}_n$ | $\sigma^2_{\alpha\delta o}\,\mathbf{I}_n$ | $\sigma^2_{\delta\delta o}\,\mathbf{I}_n$ |
| var(values) | $\mathbf{G}_a = \sigma^2_{\alpha o}\mathbf{W}_\alpha\mathbf{W}_\alpha'$ | $\mathbf{G}_d = \sigma^2_{\delta o}\mathbf{W}_\delta\mathbf{W}_\delta'$ | $\mathbf{G}_{aa} = \sigma^2_{\alpha\alpha o}\mathbf{W}_{\alpha\alpha}\mathbf{W}_{\alpha\alpha}'$ | $\mathbf{G}_{ad} = \sigma^2_{\alpha\delta o}\mathbf{W}^{(2)}_{\alpha\delta}\mathbf{W}^{(2)\prime}_{\alpha\delta}$ | $\mathbf{G}_{\delta\delta} = \sigma^2_{\delta\delta o}\mathbf{W}_{\delta\delta}\mathbf{W}_{\delta\delta}'$ |
| | Multifactorial (MF) model | | | | |
| Effect | $\boldsymbol{\tau}_{1o}$ | $\boldsymbol{\tau}_{2o}$ | $\boldsymbol{\tau}_{3o}$ | $\boldsymbol{\tau}_{4o}$ | $\boldsymbol{\tau}_{5o}$ |
| Model matrix | $\mathbf{W}_1$ | $\mathbf{W}_2$ | $\mathbf{W}_3$ | $\mathbf{W}_4$ | $\mathbf{W}_5$ |
| Value | $\mathbf{u}_1 = \mathbf{W}_1\boldsymbol{\tau}_{1o}$ | $\mathbf{u}_2 = \mathbf{W}_2\boldsymbol{\tau}_{2o}$ | $\mathbf{u}_3 = \mathbf{W}_3\boldsymbol{\tau}_{3o}$ | $\mathbf{u}_4 = \mathbf{W}_4\boldsymbol{\tau}_{4o}$ | $\mathbf{u}_5 = \mathbf{W}_5\boldsymbol{\tau}_{5o}$ |
| var(effects) | $\sigma^2_{1o}\,\mathbf{I}_n$ | $\sigma^2_{2o}\,\mathbf{I}_n$ | $\sigma^2_{3o}\,\mathbf{I}_n$ | $\sigma^2_{4o}\,\mathbf{I}_n$ | $\sigma^2_{5o}\,\mathbf{I}_n$ |
| var(values) | $\mathbf{G}_1 = \sigma^2_{1o}\mathbf{W}_1\mathbf{W}_1'$ | $\mathbf{G}_2 = \sigma^2_{2o}\mathbf{W}_2\mathbf{W}_2'$ | $\mathbf{G}_3 = \sigma^2_{3o}\mathbf{W}_3\mathbf{W}_3'$ | $\mathbf{G}_4 = \sigma^2_{4o}\mathbf{W}_4\mathbf{W}_4'$ | $\mathbf{G}_5 = \sigma^2_{5o}\mathbf{W}_5\mathbf{W}_5'$ |
| | Reparameterized multifactorial (RMF) model | | | | |
| Effect | $\boldsymbol{\tau}_1 = \sqrt{K_1}\boldsymbol{\tau}_{1o}$ | $\boldsymbol{\tau}_2 = \sqrt{K_2}\boldsymbol{\tau}_{2o}$ | $\boldsymbol{\tau}_3 = \sqrt{K_3}\boldsymbol{\tau}_{3o}$ | $\boldsymbol{\tau}_4 = \sqrt{K_4}\boldsymbol{\tau}_{4o}$ | $\boldsymbol{\tau}_5 = \sqrt{K_5}\boldsymbol{\tau}_{5o}$ |
| Model matrix | $\mathbf{T}_1 = \mathbf{W}_1/\sqrt{K_1}$ | $\mathbf{T}_2 = \mathbf{W}_2/\sqrt{K_2}$ | $\mathbf{T}_3 = \mathbf{W}_3/\sqrt{K_3}$ | $\mathbf{T}_4 = \mathbf{W}_4/\sqrt{K_4}$ | $\mathbf{T}_5 = \mathbf{W}_5/\sqrt{K_5}$ |
| Value | $\mathbf{u}_1 = \mathbf{T}_1\boldsymbol{\tau}_1$ | $\mathbf{u}_2 = \mathbf{T}_2\boldsymbol{\tau}_2$ | $\mathbf{u}_3 = \mathbf{T}_3\boldsymbol{\tau}_3$ | $\mathbf{u}_4 = \mathbf{T}_4\boldsymbol{\tau}_4$ | $\mathbf{u}_5 = \mathbf{T}_5\boldsymbol{\tau}_5$ |
| var(effect) | $\sigma^2_1 = K_1\sigma^2_{1o}$ | $\sigma^2_2 = K_2\sigma^2_{2o}$ | $\sigma^2_3 = K_3\sigma^2_{3o}$ | $\sigma^2_4 = K_4\sigma^2_{4o}$ | $\sigma^2_5 = K_5\sigma^2_{5o}$ |
| var(effects) | $\sigma^2_1\,\mathbf{I}_n$ | $\sigma^2_2\,\mathbf{I}_n$ | $\sigma^2_3\,\mathbf{I}_n$ | $\sigma^2_4\,\mathbf{I}_n$ | $\sigma^2_5\,\mathbf{I}_n$ |
| Genomic relationships | $\mathbf{S}_1 = \mathbf{T}_1\mathbf{T}_1'$ | $\mathbf{S}_2 = \mathbf{T}_2\mathbf{T}_2'$ | $\mathbf{S}_3 = \mathbf{T}_3\mathbf{T}_3'$ | $\mathbf{S}_4 = \mathbf{T}_4\mathbf{T}_4'$ | $\mathbf{S}_5 = \mathbf{T}_5\mathbf{T}_5'$ |
| K value | $K_1 = \text{tr}(\mathbf{W}_1\mathbf{W}_1')/n$ | $K_2 = \text{tr}(\mathbf{W}_2\mathbf{W}_2')/n$ | $K_3 = \text{tr}(\mathbf{W}_3\mathbf{W}_3')/n$ | $K_4 = \text{tr}(\mathbf{W}_4\mathbf{W}_4')/n$ | $K_5 = \text{tr}(\mathbf{W}_5\mathbf{W}_5')/n$ |
| var(values) | $\mathbf{G}_1 = \sigma^2_1\mathbf{S}_1$ | $\mathbf{G}_2 = \sigma^2_2\mathbf{S}_2$ | $\mathbf{G}_3 = \sigma^2_3\mathbf{S}_3$ | $\mathbf{G}_4 = \sigma^2_4\mathbf{S}_4$ | $\mathbf{G}_5 = \sigma^2_5\mathbf{S}_5$ |

## 15.7 Model Selection

The epistasis model of Equations [15.6.1]-[15.62.] may have a large number of effect types. For the model with epistasis effects upto the fourth order, 14 effect types are possible (Table 15.7.1). However, some effect types may not contribute to the phenotypic variance and the prediction accuracy. Therefore, model selection is necessary to include the effect types that contribute to the phenotypic variance and prediction accuracy in the final prediction model.

**Table 15.7.1** Heritability estimates of the full model with SNP and epistasis effects up to the fourth-order (Liang et al., 2023).

| | HDL | LDL | TC | TG | HTo | WT | BMIo |
|---|---|---|---|---|---|---|---|
| A | **0.241** | **0.284** | **0.331** | **0.221** | **0.648** | **0.424** | **0.320** |
| D | **0.055** | **0.034** | **0.085** | **0.079** | **0.134** | **0.071** | **0.012** |
| A×A | **0.356** | **0.423** | **0.145** | **0.135** | **0.192** | **0.106** | **0.219** |
| A×D | **0.111** | 0.000 | 0.003 | 0.000 | 0.000 | **0.057** | **0.126** |
| D×D | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | **0.015** | 0.008 |
| A×A×A | 0.002 | 0.000 | 0.004 | 0.000 | 0.000 | **0.017** | 0.000 |
| A×A×D | 0.002 | 0.000 | 0.001 | 0.000 | 0.000 | **0.014** | 0.000 |
| A×D×D | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.008 | 0.000 |
| D×D×D | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 |
| A×A×A×A | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.004 | 0.000 |
| A×A×A×D | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.003 | 0.000 |
| A×A×D×D | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.002 | 0.000 |
| A×D×D×D | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| D×D×D×D | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |
| Total heritability | 0.768 | 0.742 | 0.573 | 0.437 | 0.974 | 0.728 | 0.686 |
| Initial model | A+D+ AA+AD | A+D+AA | A+D+AA | A+D+AA | A+D+AA | A+D+AA +AD+DD +AAA+AAD | A+D+ AA+AD |
| Final model selected by 10-fold validations | A+D+ AA+AD | A+D+AA | A+D+AA | A+D | A+D+AA | A+D+AD | A+AD |

A is additive effect. D is dominance effect. A×A, A×D and D×D are second-order (pairwise) epistasis effects. A×A×A, A×A×D, A×D×D and D×D×D are third-order epistasis effects. A×A×A×A, A×A×A×D, A×A×D×D, A×D×D×D and D×D×D×D are fourth-order epistasis effects. Entries in bold are heritability estimates greater than 0.01 for the initial prediction models. HDL is the normality transformed high density lipoproteins. LDL is the normality transformed low density lipoproteins. TC is the normality transformed total cholesterol. TG is the normality transformed triglycerides. WT is the normality transformed weight (WT). HTo is the original phenotypic observations without normality transformation of height, BMIo is the original phenotypic observations without normality transformation of body mass index.

Model selection may include three steps:

1) Initial selection of epistasis models was to exclude effect types with little or no contribution to the phenotypic variance from further evaluation for prediction accuracy using heritability estimation,
2) Final model selection to include effect types that contribute to prediction accuracy using validation studies,
3) Select the model with the smallest number of effect types among several models with the same prediction accuracy.

Table 15.7.1 is an example of initial selection of epistasis models requiring a minimal heritability estimate of 1% for any effect type to be included in the initial epistasis model. With this requirement, four traits (only had A×A effects in the initial epistasis models, two traits also had A×D in addition to A×A, and one trait (WT) had the most complex initial epistasis model.

Ten-fold validations showed that the A+D+AD model was the final epistasis model for WT, where A×D, A×A×A and A×A×D had no contribution to prediction accuracy (Table 15.7.2).

The ten-fold validations also showed that A×A did not contribute to the prediction accuracy for WT, and resulted a 2.35% accuracy decrease for TG. The final models with the highest prediction accuracies and the smallest numbers of effect types identified by the ten-fold validation studies are listed at the bottom of Table 15.7.1.

**Table 15.7.2** Prediction accuracy of alternative models for weight (WT) from 10-fold validations.

| Prediction model | Accuracy of predicting phenotypic values | Accuracy increase over 'A+D' model (%) |
|---|---|---|
| A+D | 0.323 | 0 |
| A | 0.322 | -0.31 |
| A+D+AA | 0.324 | 0.31 |
| A+D+AA+AD | 0.325 | 0.62 |
| A+D+AA+DD | 0.325 | 0.62 |
| A+D+AA+AD+DD | 0.325 | 0.62 |
| **A+D+AD** | 0.325 | 0.62 |
| A+D+DD | 0.325 | 0.62 |
| A+D+AA+AD+DD+AAA+AAD | 0.325 | 0.62 |

The model in bold face is the best prediction model with the highest prediction accuracy and smallest number of effect types.

# CHAPTER 16: GENOME-WIDE ASSOCIATION STUDY WITH CORRECTIONS TO REDUCE SIGNIFICANT EFFECTS BY CHANCE

Genome-wide association study (GWAS) detects chromosome locations of significant SNP effects affecting a phenotype using SNPs covering the entire genome and has been a widely used approach for the discovery of genetic variants affecting phenotypes. A GWAS typically uses a population with phenotypic observations and SNP genotypes, and conducts statistical tests for significant associations between SNPs and the phenotypic observations.

A major issue of GWAS is the correction of the phenotypic values for sample stratification, which results in false positive effects due to different population structures rather than true genetic effects associated with the phenotypic values. More broadly, stratification correction could be considered a general approach to reduce false significant effects by chance. Many methods of GWAS analysis are available. This chapter describes two types of sample stratification correction: using principlal component analysis (PCA) or multidimensional scaling (MDS), and using relationships among individuals.

## 16.1 GWAS with Stratification Correction Using MDS method

The PCA method uses eigenvactors whereas the MDS method uses 'dimensions' as covariables in the statistical model. However, the PCA and MDS methods have a high correlation about 99%. The following uses the MDS method implemented by PLINK (Purcell et al., 2007) as an example of using PCA or MDS for sample stratification correction.

The statistical model for the PLINK analysis can be expressed as:

$$\mathbf{y} = \mathbf{X}_b\mathbf{b} + \mathbf{X}_1\mathbf{b}_1 + \alpha\mathbf{x} + \mathbf{e} \quad [16.1.1]$$

$$E(\mathbf{y}) = \mathbf{X}_b\mathbf{b} + \mathbf{X}_1\mathbf{b}_1 + \alpha\mathbf{x} \quad [16.1.2]$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \sigma_e^2\mathbf{I} \quad [16.1.3]$$

where

$\mathbf{y}$ = column vector of phenotypic deviation,
$\mathbf{b}$ = fixed nongenetic effects,
$\mathbf{X}_b$ = model matrix of $\mathbf{b}$,
$\mathbf{b}_1$ = fixed effects of the MDS dimensions,
$\mathbf{X}_1$ = matrix of the MDS dimension(s) calculated by PLINK from the matrix of identity by state,
$\alpha$ = SNP additive effect,
$\mathbf{x}$ = column vector as model matrix of $\alpha$ created by PLINK,
$\mathbf{e}$ = random residuals,
$\sigma_e^2$ = residual variance.

For the model of Equations [16.1.1]-[16.1.3], the statistical significance of the additive effect ($\alpha$) is tested one at a time until all SNPs are tested. The number of dimensions to be included in the model of [16.1.1] can be determined by the changes in SNP effects and genomic inflation factor: the number of dimensions is enough when the statistical significance and the genomic inflation factor stabilizes (Figure 16.1.1).

factor: the number of dimensions is enough when the statistical significance and the genomic inflation factor stabilizes (Figure 16.1.1).

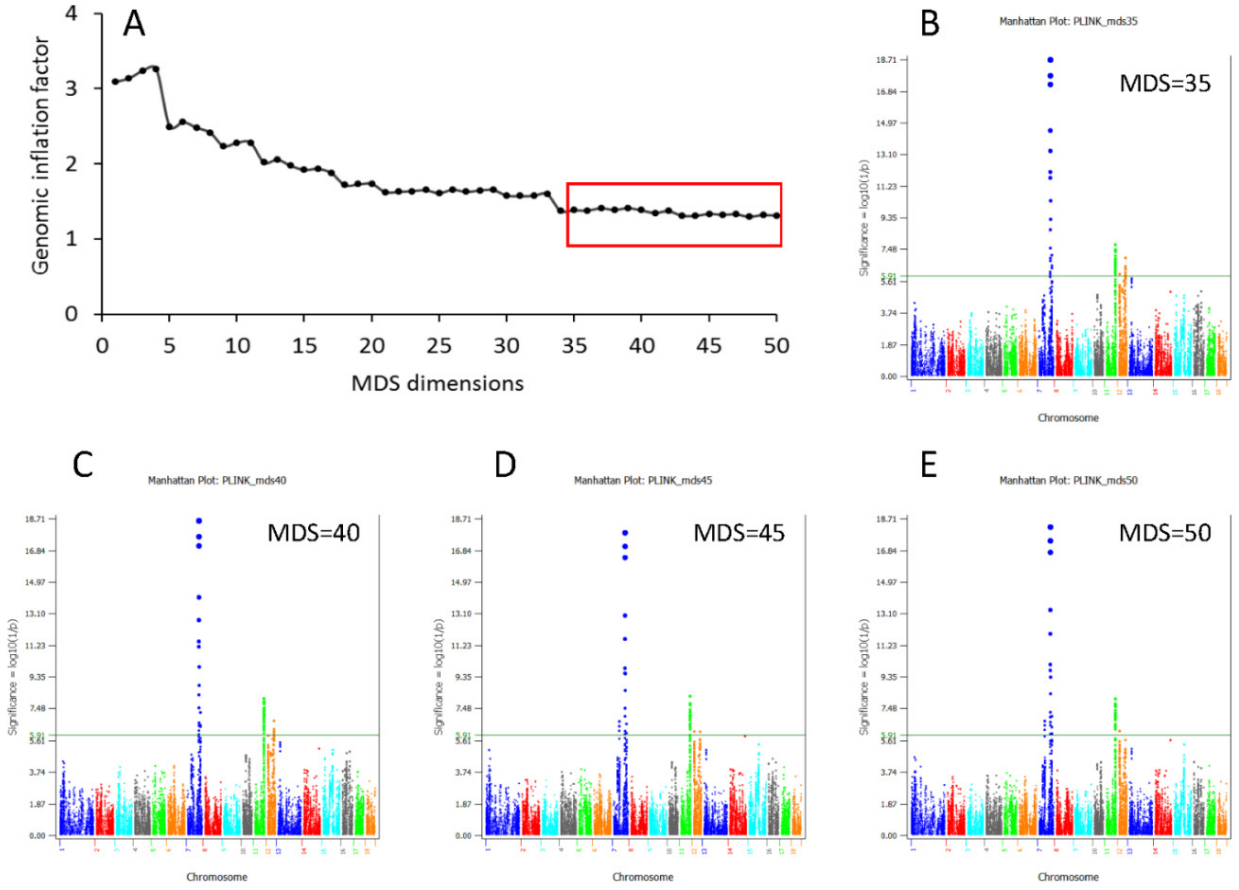



**Figure 16.1.1** Changes in genomic inflation factor and patterns of Manhattan plots of SNP significance due to increased numbers of the multidimensional scaling (MDS) dimensions as fixed effects in the statistical model for stratification correction (Tan et al., 2017). **A:** Genomic inflation factor remained relative unchanged as the number of MDS dimensions increased beyond the first 35 MDS dimensions. **B-E:** GWAS significance PLINK using 35-50 MDS dimensions showing that the significance patterns were virtually unchanged with the exception of Chr12 that had decreasing significance with increased number of MDS dimensions. All p-values in the figures were in log(1/p) scale. The horizontal green line indicates the statistical significance with the Bonferroni correction assuming a 5% genome-wide type-I error.

## 16.2 GWAS with Stratification Correction Using Mixed Models with Relationship Matrices

The mixed model for GWAS with a relationship matrix can be expressed as:

$$\mathbf{y} = \mathbf{X}_b\mathbf{b} + \mathbf{X}_g\mathbf{g} + \mathbf{Z}\mathbf{a} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e} \quad [16.2.1]$$

$$E(\mathbf{y}) = \mathbf{X}\mathbf{b} \quad [16.2.2]$$

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} = \sigma_a^2\,\mathbf{Z}\mathbf{A}\mathbf{Z}' + \sigma_e^2\,\mathbf{I} \quad [16.2.3]$$

$\mathbf{g} = (g_{11}, g_{12}, g_{22})'$ = column vector of genotypic values of the three SNP genotypes $A_1A_1$, $A_1A_2$ and $A_2A_2$;

$\mathbf{X}_g$ = model matrix of **g**;
**b** = (μ, **g**')',
**X** = (**I**, $\mathbf{X}_g$);
**a** = column vector of additive polygenic values;
**Z** = model matrix of **a**;
**e** = random residuals;
$\sigma_a^2$ = additive variance;
**A** = additive relationship matrix; [16.2.4]
$\sigma_e^2$ = residual variance.

The significance of additive and dominance ffects can be tested using a t-test:

$$t_j = \frac{|L_j|}{\sqrt{var(L_j)}} = \frac{|\mathbf{s}_j \hat{\mathbf{g}}|}{v\sqrt{\mathbf{s}_j (\mathbf{X}'\mathbf{X})^-_{gg} \mathbf{s}_j'}},\ j = a,d \qquad [16.2.5]$$

where

$L_j$ = additive or dominance contrast;
$\sqrt{var(L_j)}$ = standard deviation of the additive or dominance contrast;
$\mathbf{s}_a$ = additive contrast coefficients = $(P_{11}/p_1,\ 0.5P_{12}(p_2 - p_1)/(p_1p_2),\ -P_{22}/p_2)$;
$\mathbf{s}_d$ = dominance contrast coefficients = (−0.5, 1, −0.5);
$v^2 = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})/(n - k)$ = estimated residual variance;
$\hat{\mathbf{g}}$ = column vector of the estimated SNP genotypic effects of $g_{11}$, $g_{12}$, and $g_{22}$;
$(\mathbf{X}'\mathbf{X})^-_{gg}$ = submatrix of $(\mathbf{X}'\mathbf{X})^-$ corresponding to $\hat{\mathbf{g}}$;
$p_1$ = frequency of $A_1$ allele;
$p_2$ = frequency of $A_2$ allele;
$P_{11}$ = frequency of $A_1A_1$ genotype,
$P_{12}$ = frequency of $A_1A_2$ genotype,
$P_{22}$ = frequency of $A_2A_2$ genotype,
n = number of observations,
k = rank of **X**.

The formula of $\mathbf{s}_a$ defined above allows Hardy-Weinberg disequilibrium, and simplifies to $(p_1,\ p_2 - p_1,\ -p_2)$ under Hardy-Weinberg equilibrium. The t-test of Equation 5 for additive effects accounts for variations associated with allele frequencies. Generally, the statistical significance of

additive effects represented by the t-value of Equation [16.2.5] decreases as the allele frequencies deviate further away from equal allele frequencies.

## 16.3 Equivalence between Using Relationship Matrices and Removing Genetic Values from the Phenotypic Values

The **A** matrix of Equation [16.2.4] can be a matrix of IBS or IBD, genomic additive relationships or pedigree additive relationships. This section shows that the use of additive relationships is equivalent to removing the additive values from the phenotypic values. Using genomic additive relationships is equivalent to removing genomic additive values (GBLUP), and using pedigree additive relationships is equivalent to removing pedigree additive values (BLUP). The use of IBS or IBD matrix of SNPs is expected to have similar effects as using genomic additive relationship matrix. The fixed effects including the SNP genotypic values can be estimated by the GLS estimator or BLUE from the mixed model equations under the assumption of $\mathbf{R} = \sigma_e^2\mathbf{I}$:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad [16.3.1]$$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{a}}) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{y} - \mathbf{Z}\hat{\mathbf{a}}) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}_{*} \qquad [16.3.2]$$

where $\mathbf{y}_{*} = \mathbf{y} - \mathbf{Z}\hat{\mathbf{a}}$, and where $\hat{\mathbf{a}}$ = the best linear unbiased prediction (BLUP) of **a**. Equation [16.3.1] is the GLS solution and Equation [16.3.2] is the MME solution of BLUE. These two equations yield identical results.

Equations [16.3.1] and [16.3.2] have two important messages.

First, the GLS solution of Equation [16.3.1] can be calculated as the least squares (LS) solution of Equation [16.3.2] if $\hat{\mathbf{a}}$ is known, noting that the LS version of BLUE given by Equation [16.3.2] is computationally efficient relative to the GLS of Equation [16.3.1] requiring the **V** inverse, or the joint MME solutions of $\hat{\mathbf{b}}$ and $\hat{\mathbf{a}}$ requiring the **A** inverse. This result is the basis the approximate GLS (AGLS) method for GWAS in Holstein cattle with $\hat{\mathbf{a}}$ replaced by the esimates of **a** ($\tilde{\mathbf{a}}$):

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{a}}) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}^{*} \qquad [16.3.3]$$

where $\mathbf{y}^{*} = \mathbf{y} - \mathbf{Z}\tilde{\mathbf{a}}$, and where $\tilde{\mathbf{a}}$ = column vector of 2(PTA), PTA = predicted transmission ability from routine genetic evaluation using pedigree relationships. Equation [16.3.3] achieves the benefit of sample stratification correction from mixed models without the computing difficulty of inverting **V** or **A**.

Second, the GLS solution or BLUE of Equation [16.3.1] in fact removes $\hat{\mathbf{a}}$ from the phenotypic observations as shown by the equivalence between Equations [16.3.1] and [16.3.2] even though Equation [16.3.1] does not show the removal of $\hat{\mathbf{a}}$ explicitly. The $\hat{\mathbf{a}}$ is the GBLUP if genomic relationships are used, and is BLUP is pedigree relationships are used.

# CHAPTER 17: LITERATURE DISCUSSION

This chapter is in-class discussion of published articles on two topics: single-step genomic evaluation, and the n-m confounding.

## 17.1 Single-step Genomic Evaluation

Single-step genomic evaluation uses pedigree and genomic relationships jointly.

## 17.2 The Issue of n-m Confounding

This course use the terminology of 'n-m confounding' to refer to the result that the maximum number of individuals that contribute to the prediction accuracy is the number of SNPs, and that the selection of different training individuals in a population with the number of individuals greater than the number of SNPs does not affect the prediction accuracy.

**HANDOUT**
**ANSC 8141, Fall 2023**



1

## BP, BLP, BLUP (Chapter 5 of Henderson)

- **BP = best predictor**
  - Maximize correlation between BP and the true value
  - E(w/y) as predictor
  - Unbiased
  - Requires distribution knowledge
- **BLP = best linear predictor**
  - Same formula as BP
  - Selection index is BLP
  - Does not require distribution knowledge
  - Does not have method to estimate fixed effects
- **BLUP = best linear unbiased prediction**
  - Provides a method to estimate fixed effects
  - Minimizes prediction error variance



2

## Chapter 5: Prediction Error Variance and Reliability

$$\begin{aligned}\text{cov}(\hat{\mathbf{a}},\hat{\mathbf{b}}') &= \text{cov}\{\mathbf{GZ'Py},[(\mathbf{X'V^{-1}X})^{-}\mathbf{X'V^{-1}y}]'\}\\ &= \text{cov}\{\mathbf{GZ'Py},\mathbf{y'V^{-1}X}(\mathbf{X'V^{-1}X})^{-}]\}\\ &= \mathbf{GZ'PVar(y)V^{-1}X}(\mathbf{X'V^{-1}X})^{-}\\ &= \mathbf{GZ'PX}(\mathbf{X'V}^{-1}\mathbf{X})^{-}]\\ &= 0, \qquad \text{by } \mathbf{PX}=0\end{aligned}$$

$$\begin{aligned}\text{Var}(\hat{\mathbf{a}}) &= \mathbf{GZ'PVar(y)PZG})\\ &= \mathbf{GZ'PVPZG})\\ &= \mathbf{GZ'PZG}, \quad \text{by } \mathbf{PVP}=\mathbf{P}\end{aligned}$$

$$\begin{aligned}\text{cov}(\mathbf{a},\hat{\mathbf{b}}') &= \text{cov}\{\mathbf{a}, [(\mathbf{X'V}^{-1}\mathbf{X})^{-}\mathbf{X'V}^{-1}\mathbf{y}]'\}\\ &= \text{cov}(\mathbf{a},\mathbf{y'V^{-1}X'}(\mathbf{X'V^{-1}X})^{-}]\}\\ &= \mathbf{GZ'V}^{-1}\mathbf{X'}(\mathbf{X'V}^{-1}\mathbf{X})\end{aligned}$$

$$\begin{aligned}\text{PEV} &= \text{var}(\hat{\mathbf{a}}-\mathbf{a}) = \text{var}(\mathbf{a}) - \text{var}(\hat{\mathbf{a}})\\ &= \mathbf{G} - \mathbf{GZ'PZG}\end{aligned}$$

$$\begin{aligned}\text{var}(\hat{\mathbf{a}}) &= \mathbf{GZ'PZG}\\ &= \text{var}(\mathbf{a}) - \text{PEV} = \mathbf{G} - \text{PEV}\\ &= \mathbf{G} - \mathbf{C}^{aa} \qquad \text{for original MME}\\ &= \mathbf{G} - \mathbf{C}^{aa}\sigma_e^2 \qquad \text{for simplified MME}\end{aligned}$$

$$R_i^2 = \text{var}(\hat{\mathbf{a}})_{ii} / \text{var}(\mathbf{a})_{ii}$$

$$R_i^2 = \sigma_a^2(\mathbf{AZ'PZA})_{ii}/A_{ii}$$

$$R_i^2 = 1 - \text{PEV}_{ii}/(A_{ii}\sigma_a^2)$$

ANSC 8141, University of Minnesota

3

$$R_{ai}^2 = \sigma_\alpha^2\left(\mathbf{A}_g\mathbf{Z'PZA}_g\right)_{ii}/A_{ii} = \left(\mathbf{G}_a\mathbf{Z'PZG}_a\right)_{ii}/(\sigma_\alpha^2 A_{ii})$$

GZPZG 5 rows 5 cols (numeric)

| | | | | |
|---|---|---|---|---|
| **0.0165841** | -0.020267 | -0.001852 | 0.0053616 | 0.0001737 |
| -0.020267 | **0.0257147** | -0.001523 | -0.004067 | 0.0001428 |
| -0.001852 | -0.001523 | **0.0153558** | -0.01054 | -0.00144 |
| 0.0053616 | -0.004067 | -0.01054 | **0.0082575** | 0.0009882 |
| 0.0001737 | 0.0001428 | -0.00144 | 0.0009882 | **0.000135** |

G_A 5 rows 5 cols (numeric)

| | | | | |
|---|---|---|---|---|
| **0.0760894** | -0.049678 | -0.015092 | 0.0100614 | -0.02138 |
| -0.049678 | **0.1075313** | -0.015092 | -0.02138 | -0.02138 |
| -0.015092 | -0.015092 | **0.0823777** | -0.018236 | -0.033957 |
| 0.0100614 | -0.02138 | -0.018236 | **0.05408** | -0.024525 |
| -0.02138 | -0.02138 | -0.033957 | -0.024525 | **0.1012429** |

ANSC 8141, University of Minnesota

4

$$R^2_{ai} = \sigma^2_\alpha \left(\mathbf{A}_g \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{A}_g\right)_{ii} / A_{ii} = \left(\mathbf{G}_a \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_a\right)_{ii} / \left(\sigma^2_\alpha A_{ii}\right)$$

REL_A 5 rows 1 col (numeric)

| | | |
|---|---|---|
| 0.2179556 | = | 0.0165841/0.0760894 |
| 0.2391367 | | 0.0257147/0.1075313 |
| 0.1864077 | | 0.0153558/0.0823777 |
| 0.1526908 | | 0.0082575/0.05408 |
| 0.0013331 | | 0.000135/0.1012429 |

ANSC 8141, University of Minnesota

5

**Reliability for all individuals, with/without observations** (Da et al., 2014)

- CE method for predicting genetic values

$$R^2_{ai} = \sigma^2_\alpha \left(\mathbf{A}_g \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{A}_g\right)_{ii} / a_{ii}$$

$$R^2_{di} = \sigma^2_\delta \left(\mathbf{D}_g \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{D}_g\right)_{ii} / d_{ii}$$

$$R^2_{gi} = \left(\mathbf{G}_\alpha \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\alpha + \mathbf{G}_\alpha \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\delta + \mathbf{G}_\delta \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\alpha + \mathbf{G}_\delta \mathbf{Z}'\mathbf{P}\mathbf{Z}\mathbf{G}_\delta\right)_{ii} / \left(a_{ii}\sigma^2_\alpha + d_{ii}\sigma^2_\delta\right)$$

- MME method for predicting genetic values

$$R^2_{ai} = 1 - \lambda_\alpha \left(\mathbf{T}_\alpha \mathbf{C}^{\alpha\alpha} \mathbf{T}_\alpha'\right)_{ii} / a_{ii}$$

$$R^2_{di} = 1 - \lambda_\delta \left(\mathbf{T}_\delta \mathbf{C}^{\delta\delta} \mathbf{T}_\delta'\right)_{ii} / d_{ii}$$

$$R^2_{gi} = 1 - \sigma^2_e \left(\mathbf{T}_\alpha \mathbf{C}^{\alpha\alpha} \mathbf{T}_\alpha' + \mathbf{T}_\alpha \mathbf{C}^{\alpha\delta} \mathbf{T}_\delta' + \mathbf{T}_\delta \mathbf{C}^{\delta\alpha} \mathbf{T}_\alpha' + \mathbf{T}_\delta \mathbf{C}^{\delta\delta} \mathbf{T}_\delta'\right)_{ii} / \left(a_{ii}\sigma^2_\alpha + d_{ii}\sigma^2_\delta\right)$$

ANSC 8141, University of Minnesota

6

## Reliability for individuals without observations

- CE method for predicting genetic values

$$R^2_{ai} = \sigma^2_\alpha \left(\mathbf{A}_{01}\mathbf{Z}'\mathbf{PZA}_{10}\right)_{ii} / a_{ii}$$

$$R^2_{di} = \sigma^2_\delta \left(\mathbf{D}_{01}\mathbf{Z}'\mathbf{PZD}_{10}\right)_{ii} / d_{ii}$$

$$R^2_{gi} = \frac{\left(\sigma^4_\alpha \mathbf{A}_{01}\mathbf{Z}'\mathbf{PZA}_{10} + \sigma^2_\alpha \mathbf{A}_{01}\mathbf{Z}'\mathbf{PZD}_{10}\sigma^2_\delta + \sigma^2_\delta \mathbf{D}_{01}\mathbf{Z}'\mathbf{PZA}_{10}\sigma^2_\alpha + \sigma^4_\delta \mathbf{D}_{01}\mathbf{Z}'\mathbf{PZD}_{10}\right)_{ii}}{a_{ii}\sigma^2_\alpha + d_{ii}\sigma^2_\delta}$$

- MME method for predicting genetic values

$$R^2_{ai} = 1 - \lambda_\alpha \left(\mathbf{T}_{\alpha 0}\mathbf{C}^{\alpha\alpha}\mathbf{T}_{\alpha 0}'\right)_{ii} / a_{ii}$$

$$R^2_{di} = 1 - \lambda_\delta \left(\mathbf{T}_{\delta 0}\mathbf{C}^{\delta\delta}\mathbf{T}_{\delta 0}'\right)_{ii} / d_{ii}$$

$$R^2_{gi} = 1 - \sigma^2_e \left(\mathbf{T}_{\alpha 0}\mathbf{C}^{\alpha\alpha}\mathbf{T}_{\alpha 0}' + \mathbf{T}_{\alpha 0}\mathbf{C}^{\alpha\delta}\mathbf{T}_{\delta 0}' + \mathbf{T}_{\delta 0}\mathbf{C}^{\delta\alpha}\mathbf{T}_{\alpha 0}' + \mathbf{T}_{\delta 0}\mathbf{C}^{\delta\delta}\mathbf{T}_{\delta 0}'\right)_{ii} / \left(a_{ii}\sigma^2_\alpha + d_{ii}\sigma^2_\delta\right)$$



7

## Chapter 6: Maximum likelihood estimation

$$L(\mathbf{V};\mathbf{y}) \propto -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})$$

$$\frac{\partial L(\mathbf{V};\mathbf{y})}{\partial \sigma^2_j} = -\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma^2_j}\right) + (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma^2_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb}) = 0$$

$$\frac{\partial(\log|\mathbf{A}|)}{\partial x} = \mathrm{tr}\left[\mathbf{A}^{-1}\frac{\partial(\mathbf{A})}{\partial x}\right] \qquad \frac{\partial(\mathbf{A}^{-1})}{\partial x} = -\mathbf{A}^{-1}\frac{\partial(\mathbf{A})}{\partial x}\mathbf{A}^{-1}$$

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma^2_j}\right) = (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma^2_j}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})$$

$$\mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{V}^*_j\right) = (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}\mathbf{V}^*_j\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})$$



8

## Chapter 7. Conversion from CE to MME (1 of 3)

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})\mathbf{C}^{-}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1}$$

$$\mathrm{tr}(\mathbf{P}) = \mathrm{tr}[\mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})\mathbf{C}^{-}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1}] = \mathrm{N}/\sigma_e^2 - \mathrm{tr}[\mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})\mathbf{C}^{-}]$$

$$\mathrm{tr}[\mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1}(\mathbf{X},\mathbf{Z})\mathbf{C}^{-}] = \mathrm{tr}(\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\end{bmatrix}\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix}^{-})/\sigma_e^2$$

$$= \mathrm{tr}[(\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix} - \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{bmatrix})\begin{bmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba}\\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{bmatrix}]/\sigma_e^2$$

$$= \mathrm{tr}[\begin{bmatrix}\mathbf{C}_{bb} & \mathbf{C}_{ba}\\ \mathbf{C}_{ab} & \mathbf{C}_{aa}\end{bmatrix}\begin{bmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba}\\ \mathbf{C}b & \mathbf{C}^{aa}\end{bmatrix} - \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba}\\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{bmatrix}]/\sigma_e^2$$

$$= [\mathrm{r} - \mathrm{tr}(\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{C}^{bb} & \mathbf{C}^{ba}\\ \mathbf{C}^{ab} & \mathbf{C}^{aa}\end{bmatrix})]/\sigma_e^2 = [\mathrm{r} - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]/\sigma_e^2, \quad \mathrm{r} = \text{rank of } \mathbf{C}$$

$$\mathrm{tr}(\mathbf{P}) = \{\mathrm{N} - [\mathrm{r} - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]\}/\sigma_e^2$$

ANSC 8141, University of Minnesota

9

## Chapter 7. Conversion from CE to MME (2 of 3)

$$\mathrm{tr}(\mathbf{PZAZ}') = \mathrm{tr}(\mathbf{PZGZ}')/\sigma_a^2 = \mathrm{tr}[\mathbf{P}(\mathbf{ZGZ}' + \mathbf{R} - \mathbf{R})/\sigma_a^2$$
$$= [\mathrm{tr}(\mathbf{PV}) - \mathrm{tr}(\mathbf{PR})]/\sigma_a^2$$

$$\mathrm{tr}(\mathbf{PV}) = \mathrm{tr}\{[(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{V}\}$$
$$= \mathrm{tr}[(\mathbf{I}_N - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{X}'] = \mathrm{N} - \mathrm{tr}[\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}]$$
$$= \mathrm{N} - \mathrm{r}_x, \quad \mathrm{r}_x = \text{rank of } \mathbf{X}$$

$$\mathrm{tr}(\mathbf{PR}) = \mathrm{tr}(\mathbf{P})\sigma_e^2 = \{\mathrm{N} - [\mathrm{r}_x - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]\}$$

$$\mathrm{tr}(\mathbf{PZAZ}') = [\mathrm{tr}(\mathbf{PV}) - \mathrm{tr}(\mathbf{PR})]/\sigma_a^2 = \{(\mathrm{N} - \mathrm{r}_x) - \{\mathrm{N} - [\mathrm{r} - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]\}\}/\sigma_a^2$$
$$= \{-\mathrm{r}_x + [\mathrm{r} - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]\}/\sigma_a^2 = [\mathrm{n} - \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]/\sigma_a^2$$

ANSC 8141, University of Minnesota

10

## Chapter 7. Conversion from CE to MME (3 of 3)

$$\mathbf{y}'\mathbf{P}\mathbf{V}_a^*\mathbf{P}\mathbf{y} = \text{tr}(\mathbf{y}'\mathbf{P}\mathbf{Z}\mathbf{G}\mathbf{G}^{-1}\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{y})/\sigma_a^2 = \text{tr}(\hat{\mathbf{a}}'\mathbf{G}^{-1}\hat{\mathbf{a}})/\sigma_a^2$$

$$\mathbf{y}'\mathbf{P}\mathbf{V}_e^*\mathbf{P}\mathbf{y} = \text{tr}[(\mathbf{y}'\mathbf{P}\mathbf{R})\mathbf{R}^{-1}\mathbf{R}^{-1}(\mathbf{R}\mathbf{P}\mathbf{y})]/(\sigma_e^2)^2 = \text{tr}(\hat{\mathbf{e}}'\hat{\mathbf{e}})/(\sigma_e^2)^2$$

$$\mathbf{y}'\mathbf{P}\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{P}\mathbf{y} = \text{tr}(\mathbf{P}\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{P}) \longrightarrow \text{tr}(\hat{\mathbf{a}}'\mathbf{G}^{-1}\hat{\mathbf{a}})/\sigma_a^2 = [n - \text{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]/\sigma_a^2$$

$$\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y} = \text{tr}(\mathbf{P}) \longrightarrow \text{tr}(\hat{\mathbf{e}}'\hat{\mathbf{e}})/(\sigma_e^2)^2 = N - [r - \text{tr}(\mathbf{G}^{-1}\mathbf{C}^{aa})]\}/\sigma_e^2$$

$$\frac{\partial L}{\partial \sigma_j^2} = -\frac{1}{2}\text{tr}(\mathbf{P}\mathbf{V}_j^*) + \frac{1}{2}\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{y}$$



11

## Chapters 7 & 13: REML (left) and GREML (right)

$$\sigma_a^{2(i+1)} = \frac{\sigma_a^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}\mathbf{Z}'\mathbf{P}^{(i)}\mathbf{y}}{\text{tr}(\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}\mathbf{Z}')} \longleftrightarrow \sigma_\alpha^{2(i+1)} = \frac{\sigma_\alpha^{2(i)}\mathbf{y}\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}_g\mathbf{Z}'\mathbf{P}^{(i)}\mathbf{y}}{\text{tr}(\mathbf{P}^{(i)}\mathbf{Z}\mathbf{A}_g\mathbf{Z}')}$$

$$\sigma_e^{2(i-1)} = \frac{\sigma_e^{2(i)}\mathbf{y}'\mathbf{P}^{(i)}\mathbf{P}^{(i)}\mathbf{y}}{\text{tr}(\mathbf{P}^{(i)})} \longleftrightarrow \sigma_e^{2(i+1)} = \frac{\sigma_e^{2(i)}\mathbf{y}\mathbf{P}^{(i)}\mathbf{P}^{(i)}\mathbf{y}}{\text{tr}(\mathbf{P}^{(i)})}$$

$$\sigma_a^{2(i+1)} = \frac{\hat{\mathbf{a}}^{(i)}{}'\mathbf{A}^{-1}\hat{\mathbf{a}}^{(i)}}{[n - \text{tr}(\mathbf{A}^{-1}\mathbf{C}^{aa(i)})\lambda^{(i)}]} \longleftrightarrow \sigma_\alpha^{2(i+1)} = \frac{\hat{\boldsymbol{\alpha}}^{(i)}{}'\hat{\boldsymbol{\alpha}}^{(i)}}{m - \text{tr}(\mathbf{C}^{\alpha\alpha(i)})\lambda_\alpha^{(i)}}$$

$$\sigma_e^{2(i+1)} = \frac{\hat{\mathbf{e}}^{(i)}{}'\hat{\mathbf{e}}^{(i)}}{N - [r - \text{tr}(\mathbf{A}^{-1}\mathbf{C}^{aa(i)})\lambda^{(i)}]} \longleftrightarrow \sigma_e^{2(i+1)} = \frac{\hat{\mathbf{e}}^{(i)}{}'\hat{\mathbf{e}}^{(i)}}{N - [r - \text{tr}(\mathbf{C}^{\alpha\alpha(i)})\lambda_\alpha^{(i)}]}$$



12

## Chapter 8: Comparison between Newton-Raphson, Scoring, and AI-REML

Newton-Raphson: $-\mathbf{H} = -\frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*) + \frac{1}{2}\mathbf{yPV}_j^*\mathbf{PV}_k^*\mathbf{Py}$

cancel

Scoring: $\mathbf{I}(\theta) = \mathrm{E}(-\mathbf{H}) = \frac{1}{2}\mathrm{tr}(\mathbf{PV}_j^*\mathbf{PV}_k^*)$

AI-REML: $\mathbf{AI} = \frac{1}{2}[\mathbf{I}(\theta) - \mathbf{H}] = \frac{1}{2}[-\mathrm{E}(\mathbf{H}) - \mathbf{H}] = \frac{1}{2}\mathbf{yPV}_j^*\mathbf{PV}_k^*\mathbf{Py}$

ANSC 8141, University of Minnesota

13

## AI-REML for CE method with additive and dominance SNP effects

(Similar to Lee and van der Werf (2006) with definition and notation changes)

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \left(\mathbf{AI}^{(i)}\right)^{-1}\boldsymbol{\Delta}^{(i)} \qquad \boldsymbol{\theta} = \left[\sigma_\alpha^2 \quad \sigma_\delta^2 \quad \sigma_e^2\right]'$$

$$\mathbf{AI} = \frac{1}{2}\begin{pmatrix} \mathbf{y'PV}_\alpha^*\mathbf{PV}_\alpha^*\mathbf{Py} & \mathbf{y'PV}_\alpha^*\mathbf{PV}_\delta^*\mathbf{Py} & \mathbf{y'PV}_\alpha^*\mathbf{PPy} \\ \mathbf{y'PV}_\delta^*\mathbf{PV}_\alpha^*\mathbf{Py} & \mathbf{y'PV}_\delta^*\mathbf{PV}_\delta^*\mathbf{Py} & \mathbf{y'PV}_\delta^*\mathbf{PPy} \\ \mathbf{y'PPV}_\alpha^*\mathbf{Py} & \mathbf{y'PPV}_\delta^*\mathbf{Py} & \mathbf{y'PPPy} \end{pmatrix}$$

$$\boldsymbol{\Delta} = -\frac{1}{2}\begin{bmatrix} \mathrm{tr}(\mathbf{PV}_\alpha^*) - \mathbf{y'PV}_\alpha^*\mathbf{Py} \\ \mathrm{tr}(\mathbf{PV}_\delta^*) - \mathbf{y'PV}_\delta^*\mathbf{Py} \\ \mathrm{tr}(\mathbf{P}) - \mathbf{y'PPy} \end{bmatrix}$$

$$\mathbf{V}_\delta^* = \mathbf{ZD}_g\mathbf{Z}' \qquad \mathbf{V}_\alpha^* = \mathbf{ZA}_g\mathbf{Z}'$$

ANSC 8141, University of Minnesota

14

## AI-REML for MME method with additive and dominance SNP effects (Da et al., 2014)

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \left(\mathbf{AI}^{(i)}\right)^{-1}\boldsymbol{\Delta}^{(i)} \qquad \boldsymbol{\theta} = \left[\sigma_\alpha^2 \quad \sigma_\delta^2 \quad \sigma_e^2\right]'$$

$$\mathbf{AI} = \frac{1}{2}\begin{pmatrix} \hat{\boldsymbol{\alpha}}'\mathbf{B}_1\mathbf{Z}_1\hat{\boldsymbol{\alpha}}/(\sigma_\alpha^2)^3 & \hat{\boldsymbol{\alpha}}'\mathbf{B}_1\mathbf{Z}_2\hat{\boldsymbol{\delta}}/[(\sigma_\alpha^2)^2\sigma_\delta^2] & \hat{\boldsymbol{\alpha}}'\mathbf{B}_1\hat{\mathbf{e}}/[(\sigma_\alpha^2)^2\sigma_e^2] \\ \hat{\boldsymbol{\alpha}}'\mathbf{B}_1\mathbf{Z}_2\hat{\boldsymbol{\delta}}/[(\sigma_\alpha^2)^2\sigma_\delta^2] & \hat{\boldsymbol{\delta}}'\mathbf{B}_2\mathbf{Z}_2\hat{\boldsymbol{\delta}}/(\sigma_\delta^2)^3 & \hat{\boldsymbol{\delta}}'\mathbf{B}_2\hat{\mathbf{e}}/[(\sigma_\delta^2)^2\sigma_e^2] \\ \hat{\boldsymbol{\alpha}}'\mathbf{B}_1\hat{\mathbf{e}}/[(\sigma_\alpha^2)^2\sigma_e^2] & \hat{\boldsymbol{\delta}}'\mathbf{B}_2\hat{\mathbf{e}}/[(\sigma_\delta^2)^2\sigma_e^2] & \hat{\mathbf{e}}'\mathbf{B}_3\hat{\mathbf{e}}/(\sigma_e^2)^3 \end{pmatrix}$$

$$\mathbf{B}_1 = \mathbf{C}^{\alpha\alpha}\mathbf{Z}_1'\mathbf{M} + \mathbf{C}^{\alpha\delta}\mathbf{Z}_2'\mathbf{M} = \sigma_\alpha^2\mathbf{Z}_1'\mathbf{P} = \mathbf{Z}_1'\mathbf{B}_3/\sigma_e^2$$
$$\mathbf{B}_2 = \mathbf{C}^{\alpha\delta}\mathbf{Z}_1'\mathbf{M} + \mathbf{C}^{\delta\delta}\mathbf{Z}_2'\mathbf{M} = \sigma_\alpha^2\mathbf{Z}_2'\mathbf{P} = \mathbf{Z}_2'\mathbf{B}_3/\sigma_e^2$$
$$\mathbf{B}_3 = \mathbf{M} - \mathbf{M}\mathbf{Z}_u\mathbf{H}^{-1}\mathbf{Z}_u'\mathbf{M} = \sigma_e^2\mathbf{P}$$

$$\mathbf{Z}_1 = \mathbf{Z}\mathbf{T}_\alpha \qquad \mathbf{M} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$$
$$\mathbf{Z}_2 = \mathbf{Z}\mathbf{T}_\delta \qquad \begin{pmatrix}\mathbf{C}^{\alpha\alpha} & \mathbf{C}^{\alpha\delta}\\ \mathbf{C}^{\delta\alpha} & \mathbf{C}^{\delta\delta}\end{pmatrix} = \begin{pmatrix}\mathbf{Z}_1'\mathbf{M}\mathbf{Z}_1 + \mathbf{I}_m\lambda_\alpha & \mathbf{Z}_1'\mathbf{M}\mathbf{Z}_2\\ \mathbf{Z}_2'\mathbf{M}\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{M}\mathbf{Z}_2 + \mathbf{I}_m\lambda_\delta\end{pmatrix}^{-1}$$
$$\mathbf{Z}_u = (\mathbf{Z}_1, \mathbf{Z}_2)$$

ANSC 8141, University of Minnesota

15

## Chapter 8: Proof of $E(\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}\mathbf{y}) = \mathrm{tr}(\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*)$

$$E(\mathbf{x}'\mathbf{A}\mathbf{x}) = E[\mathrm{tr}(\mathbf{x}'\mathbf{A}\mathbf{x})] = E[\mathrm{tr}(\mathbf{A}\mathbf{x}\mathbf{x}')] = \mathrm{tr}[\mathbf{A}E(\mathbf{x}\mathbf{x}')] = \mathrm{tr}\{\mathbf{A}[\mathrm{var}(\mathbf{x}) + E(\mathbf{x})E(\mathbf{x})']\} \quad \text{A1.6}$$

$$E(\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}\mathbf{y}) = \mathrm{tr}\{\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*[\mathrm{Var}(\mathbf{P}\mathbf{y}) + E(\mathbf{P}\mathbf{y})E(\mathbf{P}\mathbf{y})']\}$$
$$= \mathrm{tr}(\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}\mathbf{V}\mathbf{P}) + \mathrm{Var}(\mathbf{P}\mathbf{y}) = \mathrm{tr}(\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}) + \mathbf{0}$$
$$= \mathrm{tr}(\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*), \qquad \mathbf{P}\mathbf{V}\mathbf{P} = \mathbf{P}, \quad \mathbf{P}\mathbf{X} = \mathbf{0}$$

$$E(\mathbf{P}\mathbf{y}) = E[\mathbf{P}(\mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{e})] = E(\mathbf{P}\mathbf{X}\mathbf{b}) + E(\mathbf{P}\mathbf{Z}\mathbf{a}) + E(\mathbf{P}\mathbf{e})$$
$$= \mathbf{P}\mathbf{X}\mathbf{b} + \mathbf{P}\mathbf{Z}E(\mathbf{a}) + \mathbf{P}E(\mathbf{e}) = \mathbf{0}, \quad \mathbf{P}\mathbf{X} = \mathbf{0}$$

$$E(\mathbf{y}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}\mathbf{y}) = \mathrm{tr}\{\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}[\mathrm{Var}(\mathbf{y}) + E(\mathbf{y})E(\mathbf{y})']\}$$
$$= \mathrm{tr}(\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}\mathbf{V}) + \mathrm{tr}(\mathbf{b}'\mathbf{X}'\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*\mathbf{P}\mathbf{X}\mathbf{b}) = \mathrm{tr}(\mathbf{P}\mathbf{V}_j^*\mathbf{P}\mathbf{V}_k^*)$$

ANSC 8141, University of Minnesota

16

## Content of GVCBLUP

- **GREML_ce** for m>q, m = # of SNPs, q = # of individuals
- **GREML_qm (mme)** for q>m
- GREML_ce and GREML_qm have identical results
  - GBLUP, reliability
  - GREML estimates of variance components
  - Heritability of all SNPs with standard deviation
  - Heritability of each SNP for Manhattan plots
  - SNP effect of each SNP for Manhattan plots
  - Can use any of Definitions I, II, IV, V for genomic relationship
  - Additive and dominance effects
- **GCORRMX**
  - Genomic relationships, Definitions I, II, IV, V
  - Genomic correlations, Definitions III, VI
- **GVCeasy as graphical interface**

ANSC 8141, University of Minnesota

17

## Goals of GVCBLUP exercise

1. **Computing speed of CE and MME versions**
   1) GREML_ce for p1k_m3k, p3k_m1k
   2) GREML_qm for p1k_m3k, p3k_m1k
   3) Additive and dominance
   4) Training vs validation
2. **Comparison of different relationship matrix**
   1) GREML_ce for p1k_m3k, additive only
   2) Definitions I, II, IV, V
3. **Input and output files**
4. **Manhattan plots of SNP heritabilities and effects**

ANSC 8141, University of Minnesota

18

- **Comparison of computing time by GREML_ce and GREML_qm**
  - 2366 individuals (q = 2366)
  - 45878 SNPs (m = 45878)
  - desktop computer with 64 Gb memory and 8 cores (16 threats)
- **Computing strategy makes a major difference in feasibility**

| | GREML_CE | GREML_QM | QM/CE ratio |
|---|---|---|---|
| **Time per iteration: A (additive only)** | <1 sec<br>**V**: 2366x2366 | ~680 sec<br>MME: 45878x45878 | >680<br>19.4 |
| **Time per iteration: A+D** | 1 sec<br>**V**: 2366x2366 | ~9000 sec<br>MME: 91756x91756 | ~9000<br>38.8 |

ANSC 8141, University of Minnesota

19

Approximate generalized least squares (AGLS) method
https://www.frontiersin.org/articles/10.3389/fgene.2019.00412/full

$$\mathbf{X'R^{-1}X\hat{b}+X'R^{-1}Z\hat{a}=X'R^{-1}y} \quad \mathbf{R}=\sigma_e^2\mathbf{I} \quad \mathbf{R}^{-1}=(1/\sigma_e^2)\mathbf{I} \quad \mathbf{X'X\hat{b}+X'Z\hat{a}=X'y}$$

$$\hat{\mathbf{b}}=(\mathbf{X'X})^{-}(\mathbf{X'y}-\mathbf{X'Z\hat{a}})=(\mathbf{X'X})^{-}\mathbf{X'}(\mathbf{y}-\mathbf{Z\hat{a}})$$

$$=(\mathbf{X'X})^{-}\mathbf{X'y}_{*}, \qquad \mathbf{y}_{*}=\mathbf{y}-\mathbf{Z\hat{a}}$$

$$=(\mathbf{X'V^{-1}X})^{-}\mathbf{X'V^{-1}y}$$

$$=\text{BLUE or GLS estimator of } \mathbf{b}$$

$$\hat{\mathbf{b}}=(\mathbf{X'X})^{-}\mathbf{X'y}, \qquad \hat{\mathbf{b}}=(\mathbf{X'X})^{-}\mathbf{X'y}$$

$$=\text{LS estimator of } \mathbf{b}$$

$$\hat{\mathbf{b}}=(\mathbf{X'X})^{-}\mathbf{X'y}^{*}, \qquad \mathbf{y}^{*}=\mathbf{y}-\mathbf{Z\tilde{a}}$$

$$=\text{AGLS estimator of } \mathbf{b}$$

$$\tilde{\mathbf{a}}=\text{additive values from routine genetic evaluation}$$

ANSC 8141, University of Minnesota

20

# APPENDIX 1: MATRIX ALGEBRA

## Generalized Inverse

$\mathbf{A}^-$ is a generalized inverse if:

$\mathbf{AA}^-\mathbf{A} = \mathbf{A}$ and $\mathbf{A}^-\mathbf{AA}^- = \mathbf{A}^-$ [A1.1]

## Trace

Definition

$\mathbf{A}$ = n × n square matrix

$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$ [A1.2]

Useful property

$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, $\text{tr}(\mathbf{ABCD}) = \text{tr}(\mathbf{BCDA}) = \text{tr}(\mathbf{CDAB}) == \text{tr}(\mathbf{DABC})$ [A1.3]

$\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$ [A1.4]

## Quadratic Forms

Definition

$\mathbf{A}$ = n × n square matrix, $\mathbf{x}$ = n × 1 column vector

$y = \mathbf{x}'\mathbf{Ax} = \sum_{j=1}^{n}\sum_{i=1}^{n} x_i x_j a_{ij}$ [A1.5]

Example 1:

$$\mathbf{x}' = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix},\ \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$y = \mathbf{x}'\mathbf{Ax} = x_1^2 a_{11} + x_2^2 a_{22} + x_3^2 a_{33} + x_1 x_2 (a_{12} + a_{21}) + x_1 x_3 (a_{13} + a_{31}) + x_2 x_3 (a_{23} + a_{32})$$

If $\mathbf{A}$ is symmetric with $a_{ij} = a_{ji}$,

$$y = \mathbf{x}'\mathbf{Ax} = x_1^2 a_{11} + x_2^2 a_{22} + x_3^2 a_{33} + 2x_1 x_2 a_{12} + 2x_1 x_3 a_{13} + 2x_2 x_3 a_{23}$$

Expectation of quadratic form

$E(\mathbf{x}'\mathbf{Ax}) = E[\text{tr}(\mathbf{Axx}')] = \text{tr}[\mathbf{A}E(\mathbf{xx}')] = \text{tr}\{\mathbf{A}[\text{var}(\mathbf{x}) + E(\mathbf{x})E(\mathbf{x})']\}$ [A1.6]

Example: if $\text{var}(\mathbf{y}) = \mathbf{V}$, $E(\mathbf{y}) = \mathbf{Xb}$,

$E(\mathbf{y}'\mathbf{Qy}) = \text{tr}\{\mathbf{Q}[\text{var}(\mathbf{y}) + E(\mathbf{y})E(\mathbf{y})']\} = \text{tr}\{\mathbf{Q}[\mathbf{V} + (\mathbf{Xb})(\mathbf{Xb})']\}$

## Matrix Derivatives

### Definition of matrix derivatives

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} f(x)_1 \\ f(x)_2 \\ f(x)_3 \\ f(x)_4 \end{bmatrix}, \quad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \frac{\partial y_3}{\partial x} \\ \frac{\partial y_4}{\partial x} \end{bmatrix}, \quad \frac{\partial \mathbf{y}'}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \frac{\partial y_3}{\partial x} & \frac{\partial y_4}{\partial x} \end{bmatrix} \qquad [A1.7]$$

### Formulations

$\mathbf{y} = x\mathbf{a}$ = n × 1 column vector, $\mathbf{a}$ = n × 1 column vector, x = scalar

$$\frac{\partial y}{\partial x} = \frac{\partial (x\mathbf{a})}{\partial x} = \mathbf{a} \qquad [A1.8]$$

$\mathbf{y} = x\mathbf{a}'$ = 1 × n row vector, $\mathbf{a}'$ = 1 × n row vector, x = scalar

$$\frac{\partial y}{\partial x} = \frac{\partial (x\mathbf{a}')}{\partial x} = \mathbf{a}' \qquad [A1.9]$$

$y = \mathbf{a}'\mathbf{x}$ = 1 × 1 scalar, $\mathbf{a}'$ = 1 × n row vector, $\mathbf{x}$ = n × 1 column vector

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial (\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a} \qquad [A1.10]$$

$\mathbf{y} = \mathbf{A}\mathbf{x}$ = n × 1 column vector, $\mathbf{A}$ = n × m matrix, $\mathbf{x}$ = m × 1 column vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial (\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{A}\mathbf{x})'}{\partial \mathbf{x}} = \mathbf{A}' \qquad [A1.11]$$

$\mathbf{y} = \mathbf{x}'\mathbf{A}$ = 1 × n row vector, $\mathbf{A}$ = m × n matrix, $\mathbf{x}'$ = 1 × m row vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x}'\mathbf{A})}{\partial \mathbf{x}} = \mathbf{A} \qquad [A1.12]$$

$y = \mathbf{x}'\mathbf{A}\mathbf{x}$ = 1 × 1 quadratic form, $\mathbf{A}$ = n × n matrix, $\mathbf{x}$ = n × 1 column vector

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x} \qquad [A1.13]$$

$\mathbf{A}^{-1}$ = inverse of $\mathbf{A}$ = f(x)

$$\frac{\partial(\mathbf{A}^{-1})}{\partial x} = -\mathbf{A}^{-1}\frac{\partial(\mathbf{A})}{\partial x}\mathbf{A}^{-1} \qquad [A1.15]$$

y = log|**A**| = logarithm of determinant = f(x)

$$\frac{\partial(\log|\mathbf{A}|)}{\partial x} = \mathrm{tr}[\mathbf{A}^{-1}\frac{\partial(\mathbf{A})}{\partial x}] \qquad [A1.16]$$

**F** = **f**($x_j$) = p × q matrix, **G** = **g**($x_j$) = q × r matrix, j=1,…,m

$$\frac{\partial \mathbf{FG}}{\partial x_j} = \mathbf{F}\frac{\partial(\mathbf{G})}{\partial x_j} + \frac{\partial(\mathbf{F})}{\partial x_j}\mathbf{G} \qquad \text{(Harville, 1997)} \qquad [A1.17]$$

**F** = **f**($x_j$) = p × p matrix

$$\frac{\partial[\mathrm{tr}(\mathbf{F})]}{\partial x_j} = \mathrm{tr}\frac{\partial \mathbf{F}}{\partial x_j} \qquad [A1.18]$$

# APPENDIX 2: GENETIC PARTITION

(This is Chapter 11 of ANSC 5200, Fall 2022)
(A similar version was published as Text S1, Da et al., 2014, PLoS One 9(1):e87666)

## 11. GENETIC MODELING OF QUANTITATIVE TRAIT LOCI: SINGLE-LOCUS MODEL

### Summary of main points

● A genotypic value of a bi-allelic locus is partitioned into the sum of the population mean, additive value and dominance value
● The additive value is the sum of two allelic effects
● Each allelic effect is a deviation of the allelic mean from the population mean
● Additive effect is the difference between the two allelic effects
● The dominance value (deviation) is the deviation of the genotypic value from the population mean and the additive value
● The dominance effect is the difference between the heterozygous dominance value and the average of the two homozygous dominance values, and is also the difference between heterozygous genotypic value and the average of the two homozygous genotypic values
● The genotypic variance is partitioned into additive and dominance variances

● An additive value can be expressed as a function of the additive effect, providing a quantitative genetics model for candidate gene testing, genome-wide association study (GWAS), and genomic prediction
● A dominance value can be expressed as a function of the dominance effect, providing a quantitative genetics model for candidate gene testing, genome-wide association study (GWAS), and genomic prediction
● The quantitative genetics model in this chapter is used for genomic prediction of quantitative traits

### References

Falconer DS, Mackay TFC: Introduction to Quantitative Genetics, 4 edn. Harlow, Essex, UK, Longmans Green; 1996. Chapters 7 and 8.
Lynch M, Walsh B: Genetics and analysis of quantitative traits, Sinauer Associates, Inc., 1998. Chapter 4.

### Fundamental assumption for quantitative trait

● The phenotype of a quantitative trait is affected by genetic and environment factors

### Two general models of a quantitative trait

1) $y = G + E + e$
2) $y = G + E + I_{GE} + e$

y = phenotypic observation of the quantitative trait
G = genotypic value of the gene(s) affecting the quantitative trait
E = environmental value of the quantitative trait
$I_{GE}$ = the phenotypic value attributable to the gene x environment interaction
e = random residual

- Modeling G is the task of genetic modeling of quantitative trait loci (QTL)
- This chapter considers y = G + E + e only
- Genetic partition is the foundation of quantitative genetics
- Models from genetic partition have been used for detection of genetic effects and genomic prediction

## 11.1 Partition of genotypic values

- Linear partitioning of genotypic values is the primary genetic modeling of quantitative traits
- A bi-allelic locus is assumed to affect the quantitative trait
- One locus with two alleles, $A_1$ and $A_2$
- Allele frequencies: p($A_1$) = p, and p($A_2$) = q

- Under Hardy-Weinberg equilibrium (HWE) assumption, the genotypic array satisfies

$$p^2 A_1A_1 + 2pqA_1A_2 + q^2 A_2A_2 = (pA_1 + qA_2)^2$$

### Population mean of genotypic values and average effect of a gene

**Table 11.1.1** Calculation of population mean and average effect

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Number of individuals | $N_{11}$ | $N_{12}$ | $N_{22}$ |
| Genotypic frequency: general expression | $P_{11} = N_{11}/N$ | $P_{12} = N_{12}/N$ | $P_{22} = N_{22}/N$ |
| Genotypic frequency under HWE | $p^2$ | $2pq$ | $q^2$ |
| Number of $A_1$ | 2 | 1 | 0 |
| Number of $A_2$ | 0 | 1 | 2 |
| Genotypic value: general definition | $g_{11}$ | $g_{12}$ | $g_{22}$ |
| Genotypic value: Falconer's definition | a | d | −a |

- $N = N_{11} + N_{12} + N_{22}$

$\mu$ = population mean value of a gene
= sum of (genotypic frequency × genotypic value)

$$= P_{11} g_{11} + P_{12} g_{12} + P_{22} g_{22} \quad \text{in general} \qquad (11.1.1)$$

$$= p^2 g_{11} + 2pq g_{12} + q^2 g_{22} \quad \text{under HWE} \qquad (11.1.2)$$

### Average effect of an allele

average effect of an allele = the mean of genotypic values of all genotypes containing the allele, measured as deviation from population mean.

Allelic mean of $A_1$ allele

$$\mu_1 = \frac{2N_{11}g_{11} + N_{12}g_{12}}{2N_{11} + N_{12}} = \frac{N_{11}g_{11} + ½N_{12}g_{12}}{N_{11} + ½N_{12}}$$

$$= \frac{\frac{N_{11}}{N}g_{11} + ½\frac{N_{12}}{N}g_{12}}{\frac{N_{11}}{N} + ½\frac{N_{12}}{N}} = \frac{P_{11}g_{11} + ½P_{12}g_{12}}{P_{11} + ½P_{12}} = \frac{P_{11}g_{11} + ½P_{12}g_{12}}{p}$$

$$= \frac{P_{11}}{p}g_{11} + ½\frac{P_{12}}{p}g_{12} = P_{11.1}g_{11} + ½P_{12.2}g_{12} \quad \text{in general} \qquad (11.1.3)$$

$$= \frac{p^2}{p}g_{11} + ½\frac{2pq}{p}g_{12} = pg_{11} + qg_{12} \text{ under HWE} \qquad (11.1.4)$$

- $P_{11.1} = P_{11}/p$ = conditional probability of $A_1A_1$ genotype among all genotypes with $A_1$ allele
- $P_{12.1} = P_{12}/p$ = conditional probability of $A_1A_2$ genotype among all genotypes with $A_1$ allele

Average effect of $A_1$ allele

$$a_1 = \mu_1 - \mu \text{ = average effect of } A_1 \text{ allele} \qquad (11.1.5)$$

$$= (p\,g_{11} + q\,g_{12}) - (p^2 g_{11} + 2pq\,g_{12} + q^2 g_{22})$$

$$= p(1-p)\ g_{11} + q(1-2p)\,g_{12} - q^2 g_{22}$$

$$= pq\,g_{11} + q(q-p)\,g_{12} - q^2 g_{22}$$

$$= q[p(g_{11} - g_{12}) + q(g_{12} - g_{22})] \qquad (11.1.6)$$

Allelic mean of $A_2$ allele

$$\mu_2 = \frac{N_{12}g_{11} + 2N_{22}g_{12}}{N_{12} + N_{22}} = \frac{N_{12}g_{11} + ½N_{22}g_{12}}{N_{12} + ½N_{22}}$$

$$= \frac{½\frac{N_{12}}{N}g_{12} + \frac{N_{22}}{N}g_{22}}{½\frac{N_{21}}{N} + \frac{N_{22}}{N}} = \frac{½P_{12}g_{12} + P_{22}g_{22}}{½P_{12} + P_{22}} = \frac{½P_{12}g_{12} + P_{22}g_{22}}{q}$$

$$= ½\frac{P_{12}}{q}g_{12} + \frac{P_{22}}{q}g_{12} = P_{12.2}g_{12} + ½P_{22.2}g_{22} \text{ in general} \qquad (11.1.7)$$

$$= ½\frac{2pq}{q}g_{12} + \frac{q^2}{q}g_{22} = pg_{12} + qg_{22} \text{ under HWE} \qquad (11.1.8)$$

- $P_{12.2} = P_{12}/q$ = conditional probability of $A_1A_2$ genotype among all genotypes with $A_2$ allele
- $P_{22.2} = P_{22}/q$ = conditional probability of $A_2A_2$ genotype among all genotypes with $A_2$ allele

Average effect of $A_2$

$$a_2 = \mu_2 - \mu \text{ = average effect of } A_2 \text{ allele} \quad (11.1.9)$$

$$= (p\,g_{12} + q\,g_{22}) - (p^2 g_{11} + 2pq\,g_{12} + q^2 g_{22})$$

$$= -p^2 g_{11} + p(p-q)\,g_{12} + pq\,g_{22}$$

$$= -p[p(g_{11} - g_{12}) + q(g_{12} - g_{22})] \quad (11.1.10)$$

- Average effects sum to zero, i.e., $p\,a_1 + q\,a_2 = 0$
- Average effect of gene is also called "additive effect of gene", meaning that one gene adds one effect to the genotypic value

**Average effect of gene substitution (α)**

= the difference between the average effect of $A_1$ and the average effect of $A_2$

$$\alpha = a_1 - a_2 = \mu_1 - \mu_2 \quad (11.1.11)$$

$$= p(g_{11} - g_{12}) + q(g_{12} - g_{22}) \quad (11.1.12)$$

$$= p(g_{11}) + (q-p)\,g_{12} - q\,g_{22} \quad (11.1.13)$$

$$= (g_{11} - g_{22})/2 \quad \text{if } p = q$$

$$a_1 = q\alpha, \quad \text{by equations (11.1.6) and (11.1.12)} \quad (11.1.14)$$

$$a_2 = -p\alpha, \quad \text{by equations (11.1.10) and (11.1.12)} \quad (11.1.15)$$

- $\alpha = p(g_{11}) + (q-p)\,g_{12} - q\,g_{22}$ is the additive contrast of genotypic values
- $\alpha$, $a_1$, $a_2$ are measures of additive effect (allelic effect)

**Additive (breeding) value**

- Breeding value of a genotype is the additive effect of the genotype and is a sum of average allele effects of the genotype, the summation being made over the pair of alleles at each locus and over all loci.

$$a_{ij} = a_i + a_j \quad (11.1.16)$$

**Table 11.1.2** Additive (breeding) value

| Genotype | Additive (breeding) value |
|---|---|
| $A_1A_1$ | $a_{11} = a_1 + a_1 = 2\,a_1 = 2q\alpha$ |
| $A_1A_2$ | $a_{12} = a_1 + a_2 = (q-p)\alpha$ |
| $A_2A_2$ | $a_{22} = a_2 + a_2 = 2\,a_2 = -2p\alpha$ |

Note that additive vales sum to zero, i.e., $p^2 a_{11} + 2pq\,a_{12} + q^2 a_{22} = 0$ .

**Dominance value (deviation)**

Dominance value (deviation) is the deviation of a genotypic value from its mean and additive

value, i.e.,

$$d_{ij} = g_{ij} - \mu - a_{ij} = t_{ij} - a_{ij} \tag{11.1.17}$$
$$t_{ij} = g_{ij} - \mu,\ t_{11} = g_{11} - \mu,\ t_{12} = g_{12} - \mu,\ t_{22} = g_{22} - \mu$$

**Table 11.1.3** Dominance value (deviation)

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Corrected genotypic value | $t_{11}$ | $t_{12}$ | $t_{22}$ |
| Additive (breeding) value | $a_{11} = 2a_1$<br>$= 2q\alpha$ | $a_{12} = a_1 + a_2$<br>$= (q-p)\alpha$ | $a_{22} = 2a_2$<br>$= -2p\alpha$ |
| Dominance value (deviation) | $d_{11} = t_{11} - a_{11}$<br>$= -2q^2\delta$ | $d_{12} = t_{12} - a_{12}$<br>$= 2pq\delta$ | $d_{22} = t_{22} - a_{22}$<br>$= -2p^2\delta$ |

$$\begin{aligned} d_{11} &= t_{11} - a_{11} = g_{11} - \mu - 2a_1 \\ &= g_{11} - \mu - 2(\mu_1 - \mu) \\ &= g_{11} - 2\mu_1 + \mu \\ &= g_{11} - 2(p\,g_{11} + q\,g_{12}) + (p^2 g_{11} + 2pq\,g_{12} + q^2 g_{22}) \\ &= (1-2p)\ g_{11} - 2q\,g_{12} + (p^2 g_{11} + 2pq\,g_{12} + q^2 g_{22}) \\ &= q^2(g_{11} + g_{22} - 2g_{12}) \\ &= -2q^2[g_{12} - \tfrac{1}{2}(g_{11} + g_{22})] \\ &= -2q^2\delta \end{aligned} \tag{11.1.18}$$

$$\begin{aligned} d_{12} &= t_{12} - a_{12} = g_{12} - \mu - (a_1 + a_2) \\ &= g_{12} - \mu_1 - \mu_2 + \mu \\ &= g_{12} - (p\,g_{11} + q\,g_{12}) - (p\,g_{12} + q\,g_{22}) \\ &\quad + (p^2 g_{11} + 2pq\,g_{12} + q^2 g_{22}) \\ &= -pq(g_{11} + g_{22}) + 2pq\,g_{12} \\ &= 2pq[g_{12} - \tfrac{1}{2}(g_{11} + g_{22})] \\ &= 2pq\delta \end{aligned} \tag{11.1.19}$$

$$\begin{aligned} d_{22} &= t_{22} - a_{22} = g_{22} - \mu - 2a_2 \\ &= g_{22} - 2(p\,g_{12} + q\,g_{22}) + (p^2 g_{11} + 2pq\,g_{12} + q^2 g_{22}) \\ &= (1 - 2q + q^2)g_{22} - 2p(1-q)\ g_{12} + p^2 g_{11} \\ &= (p - pq)\ g_{22} - 2p^2 g_{12} + p^2 g_{11} = p^2 g_{22} - 2p^2 g_{12} + p^2 g_{11} \\ &= p^2(g_{11} + g_{22} - 2g_{12}) \\ &= -2p^2[g_{12} - \tfrac{1}{2}(g_{11} + g_{22})] \\ &= -2p^2\delta \end{aligned} \tag{11.1.20}$$

**Dominance effect (δ)**

$$\delta = g_{12} - \tfrac{1}{2}(g_{11} + g_{22}) = d_{12} - \tfrac{1}{2}(d_{11} + d_{22}) \quad (11.1.21)$$

= dominance contrast of genotypic values
= dominance effect = contrast of dominance deviations

● Dominance deviations sum to zero, i.e., $p^2 d_{11} + 2pq\, d_{12} + q^2 d_{22} = 0$ .

**Summary of partition of genotypic values**

$$g_{ij} = \mu + a_i + a_j + d_{ij} \quad (11.1.22)$$

$$= \mu + a_{ij} + d_{ij} \quad (11.1.23)$$

**Interpretation of additive and dominance effects**

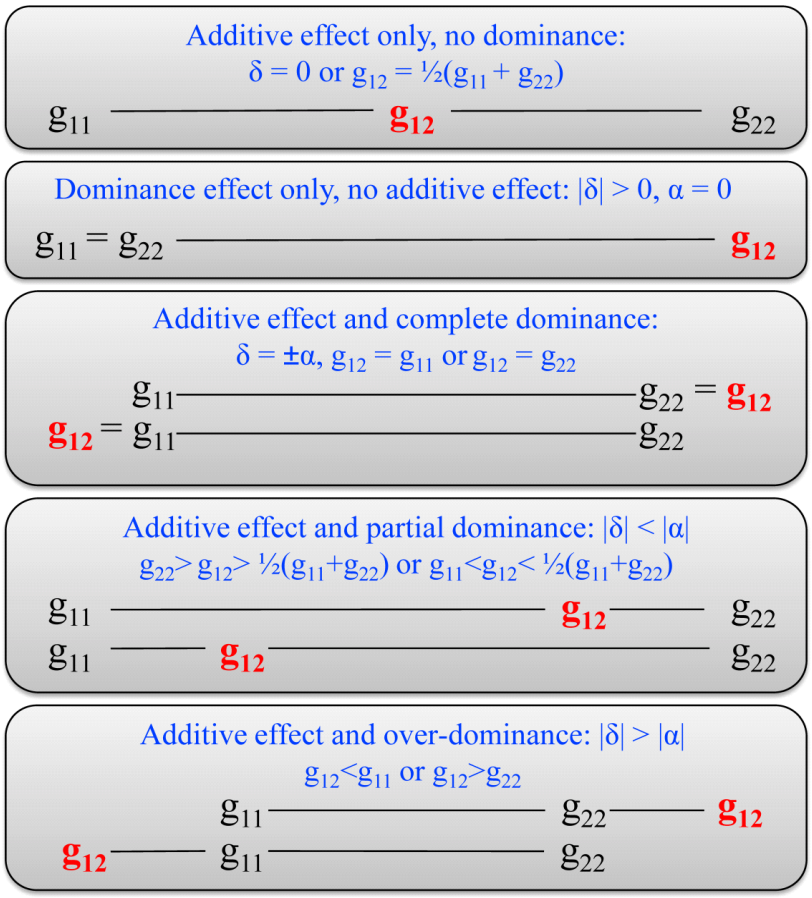



**Figure 11.1.1** Quantitative genetics interpretation of additive and dominance effects. (Assuming equal allele frequency for additive effect)

**11.2 Partition of genotypic values in matrix notations**

● Quantitative genetics model from genetic partition

**Model 1**

$$g_{ij} = \mu + a_{ij} + d_{ij} \quad (11.2.1)$$

or

$$g_{11} = \mu + (2q\alpha) - (2q^2\delta) \quad (11.2.2)$$

$$g_{12} = \mu + [(q - p)\alpha] + (2pq\delta) \quad (11.2.3)$$

$$g_{22} = \mu + (-2p\alpha) + (-2p^2\delta) \quad (11.2.4)$$

In matrix notations,

$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} + \begin{bmatrix} a_{11} \\ a_{12} \\ a_{22} \end{bmatrix} + \begin{bmatrix} d_{11} \\ d_{12} \\ d_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} + \begin{bmatrix} 2q\alpha \\ (q-p)\alpha \\ -2p\alpha \end{bmatrix} + \begin{bmatrix} -2q^2\delta \\ 2pq\delta \\ -2p^2\delta \end{bmatrix} = \begin{bmatrix} 1 & 2q & -2q^2 \\ 1 & q-p & 2pq \\ 1 & -2p & -2p^2 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \delta \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 2q \\ (q-p) \\ -2p \end{bmatrix} \alpha + \begin{bmatrix} -2q^2 \\ 2pq \\ -2p^2 \end{bmatrix} \delta$$

$$\mathbf{g} = \mathbf{1}\mu + \mathbf{w}_\alpha\alpha + \mathbf{w}_\delta\delta = \mathbf{I}\mu + \mathbf{a} + \mathbf{d} \quad (11.2.5)$$

● Equation 11.2.5 is used for GWAS and genomic prediction

● Quantitative genetics model with equal allele frequency: p = q = ½

$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 2q \\ (q-p) \\ -2p \end{bmatrix} \alpha + \begin{bmatrix} -2q^2 \\ 2pq \\ -2p^2 \end{bmatrix} \delta = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \alpha + \begin{bmatrix} -½ \\ ½ \\ -½ \end{bmatrix} \delta \quad (11.2.6)$$

**11.3 Different forms of quantitative genetics models**

● Two different genetic models, Model 2 and Model 3 as special cases of Model 1 assuming equal allele frequencies
● Models 1-3 have been used for detecting genetic effects
● Models 1 and 3 have been used for genomic prediction

**Model 2: 1-0-(−1) additive coding and 0-1-0 dominance coding**

$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \delta, \quad (11.3.1)$$

**Model 2 is a reparameterized model of Model 1 assuming equal allele frequencies**

- Equal QTL allele frequencies: p = q = ½
- adding −½δ and ½δ to the right-hand-side of Equations 11.2.6 lads to:

$$g_{11} = \mu + \alpha - ½\delta = (\mu - ½\delta) + \alpha = \mu^* + \alpha \quad (11.3.2)$$

$$g_{12} = \mu + ½\delta = (\mu - ½\delta) + \delta = \mu^* + \delta \quad (11.3.3)$$

$$g_{22} = (\mu - ½\delta) - \alpha = \mu^* - \alpha \quad (11.3.4)$$

$$\mu^* = \mu - ½\delta \quad (11.3.5)$$

or
$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu^* + \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \delta \quad (11.3.6)$$

**<u>Model 3</u>: 2-1-0 additive coding and 0-1-0 dominance coding**

- One copy of an allele has one additive effect
- Only heterozygous genotype has dominance effect
- Example: $A_1A_1$ has 2 copies of $A_1$, $A_1A_2$ has 1 copy of $A_1$, $A_2A_2$ has 0 copy of $A_1$
- 2-1-0 additive coding,

$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \delta \quad (11.3.7)$$

**Model 2 is a reparameterized model of Model 1 assuming equal allele frequencies**

- adding −α and α to the right-hand-side of Equations 11.3.3-11.3.5 leads to:

$$g_{11} = \mu^* - \alpha + \alpha + \alpha = (\mu - \alpha - ½\delta) + 2\alpha = \mu^{**} + 2\alpha \quad (11.3.8)$$

$$g_{12} = \mu^* - \alpha + \alpha + \delta = (\mu - \alpha - ½\delta) + \alpha + \delta \quad = \mu^{**} + \alpha + \delta \quad (11.3.9)$$

$$g_{22} = \mu^* - \alpha + \alpha - \alpha = (\mu - \alpha - ½\delta) = \mu^{**} \quad (11.3.10)$$

$$\mu^{**} = \mu - \alpha - ½\delta \quad (11.3.11)$$

or
$$\begin{bmatrix} g_{11} \\ g_{12} \\ g_{22} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu^{**} + \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \delta \quad (11.3.12)$$

## 11.4 Partition of genotypic variance

**Additive variance**
Additive genetic variance is the variance of additive values or breeding values.

**Table 11.4.1** Calculation of additive variance under HWE

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Genotypic freq | $p^2$ | $2pq$ | $q^2$ |
| Additive (breeding) value | $a_{11} = 2q\alpha$ | $a_{12} = (q-p)\alpha$ | $a_{22} = -2p\alpha$ |
| (Breeding value)$^2$ | $4q^2\alpha^2$ | $(q-p)^2\alpha^2$ | $4p^2\alpha^2$ |
| Freq × (breeding value)$^2$ | $4p^2q^2\alpha^2$ | $2pq(q-p)^2\alpha^2$ | $4p^2q^2\alpha^2$ |

$\sigma_a^2$ = sum of frequency x (additive value)$^2$ (11.4.1)
(mean of additive values = 0)

$$\begin{aligned}\sigma_a^2 &= 4p^2q^2\alpha^2 + 2pq(q-p)^2\alpha^2 + 4p^2q^2\alpha^2 \\ &= 2pq\alpha^2[2pq + (q-p)^2 + 2pq] \\ &= 2pq\alpha^2(2pq + q^2 - 2pq + p^2 + 2pq) \\ &= 2pq\alpha^2(p+q)^2 \\ &= 2pq\alpha^2 \end{aligned} \quad (11.4.2)$$

**Dominance variance**
Dominance variance is the variance of dominance values or deviations.

**Table 11.4.2** Calculation of dominance variance under HWE

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Genotypic freq | $p^2$ | $2pq$ | $q^2$ |
| Dominance value (deviation) | $-2q^2\delta$ | $2pq\delta$ | $-2p^2\delta$ |
| (Dominance value)$^2$ | $4q^4\delta^2$ | $4q^2p^2\delta^2$ | $4p^4\delta^2$ |
| Freq × (Dominance value)$^2$ | $4p^2q^4\delta^2$ | $8p^3q^3d\delta^2$ | $4p^4q^2\delta^2$ |

$\sigma_d^2$ = sum of frequency × (dominance value)$^2$ (11.4.3)
(mean of dominance values = 0)

$$\begin{aligned} &= 4p^2q^4\delta^2 + 8p^3q^3\delta^2 + 4p^4q^2\delta^2 \\ &= 4p^2q^2\delta^2(q^2 + 2pq + p^2) \\ &= 4p^2q^2\delta^2 \end{aligned} \quad (11.4.4)$$

**Covariance between additive and dominance valuess**

**Table 11.4.3** Calculation of covariance between additive and dominance values

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Genotypic frequency | $p^2$ | $2pq$ | $q^2$ |
| A (additive value) | $2q\alpha$ | $(q-p)\alpha$ | $-2p\alpha$ |
| D (dominance value) | $-2q^2\delta$ | $2pq\delta$ | $-2p^2\delta$ |
| A x D | $-4q^3\alpha\delta$ | $2pq(q-p)\alpha\delta$ | $4p^3\alpha\delta$ |
| Freq x (A x D) | $-4p^2q^3\alpha\delta$ | $4p^2q^2(q-p)\alpha\delta$ | $4p^3q^2\alpha\delta$ |

$$\begin{aligned}\text{cov(A, D)} &= \text{sum of frequency} \times (\text{A x D})^2 \\ &\quad (\text{mean of A} = 0, \text{ and mean of D} = 0) \\ &= -4p^2q^3\alpha\delta + 4p^2q^2(q-p)\alpha\delta + 4p^3q^2\alpha\delta \\ &= -4p^2q^2\alpha\delta(-q+q-p+p) \\ &= 0\end{aligned}$$

**Partition of genetic variance**

$$\begin{aligned}\sigma_g^2 &= \sigma_a^2 + \sigma_d^2 + 2\text{cov(A, D)} \\ &= \sigma_a^2 + \sigma_d^2 = 2pq\alpha^2 + 4p^2q^2\delta^2 \end{aligned} \quad (11.4.5)$$

**Direct calculation of $\sigma_g^2$ as confirmation of partition of genetic variance**

$$g_{11} - \mu = 2q(\alpha - q\delta)$$
$$g_{12} - \mu = (q-p)\alpha + 2pq\delta$$
$$g_{22} - \mu = -2p(\alpha + p\delta)$$

**Table 11.4.5** Calculation of genotypic variance

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Genotypic frequency | $p^2$ | $2pq$ | $q^2$ |
| Genotypic deviation | $2q(\alpha-q\delta)$ | $(q-p)\alpha+2pq\delta$ | $-2p(\alpha+p\delta)$ |
| (Genotypic deviation)$^2$ | $4q^2(\alpha-q\delta)^2$ | $[(q-p)\alpha+2pq\delta]^2$ | $4p^2(\alpha+p\delta)^2$ |
| Freq x (genotypic deviation)$^2$ | $4p^2q^2(\alpha-q\delta)^2$ | $2pq[(q-p)\alpha+2pq\delta]^2$ | $4p^2q^2(\alpha+p\delta)^2$ |

$$\sigma_g^2 = 4p^2q^2(\alpha - q\delta)^2 + 2pq[(q-p)\alpha+2pq\delta]^2 + 4p^2q^2(\alpha+p\delta)^2 \quad (11.4.6)$$

$$= 4p^2q^2(\alpha^2 - 2q\alpha\delta + q^2\delta^2) + 2pq[(q-p)^2\alpha^2 + 4pq(q-p)\alpha\delta + 4p^2q^2\delta^2] + 4p^2q^2(\alpha^2 + 2p\alpha\delta + p^2\delta^2)$$

$$= \{4p^2q^2\alpha^2 + 2pq(q-p)^2\alpha^2 + 4p^2q^2\alpha^2\} \quad (= \sigma_a^2)$$
$$+ \{4p^2q^2(-2q\alpha\delta) + 2pq[4pq(q-p)\alpha\delta] + 4p^2q^2(2p\alpha\delta)\} \quad (= \text{cov}(A, D) = 0)$$
$$+ \{4p^2q^2(q^2\delta^2) + 2pq(4p^2q^2\delta^2) + 4p^2q^2(p^2\delta^2)\} \quad (= \sigma_d^2)$$
$$= \sigma_a^2 + \sigma_d^2$$

**Partition of phenotypic variance**

$$y = G + e = \mu + a + d + e$$
$$\sigma_y^2 = \text{var}(y) = \text{var}(G) + \text{var}(e) = \sigma_a^2 + \sigma_d^2 + \sigma_e^2 \quad (11.4.7)$$

**Heritability**

- ratio of a genetic variance to the phenotypic variance
- phenotypic variance explained by the genetic variance

$$h_a^2 = \sigma_a^2/\sigma_y^2 = \text{additive heritability} = \text{heritability in the narrow sense} \quad (11.4.8)$$

$$h_d^2 = \sigma_d^2/\sigma_y^2 = \text{dominance heritability} \quad (11.4.9)$$

$$h_g^2 = \sigma_g^2/\sigma_y^2 = \text{heritability in the broad sense} \quad (11.4.10)$$

# APPENDIX 3: HWE, HWD, LE, LD, IBS, IBD

(These are abraviated sections of ANSC 5200, Fall 2022)

## 2.3 Hardy Weinberg Equilibrium (HWE) and Disequilibrium (HWD)

Hardy-Weinberg equilibrium (HWE)

Definition: The population has constant gene and genotypic frequencies
Conditions:

- Random mating: "any individual has an equal chance of mating with any other individual in the population" (Falconer and Mackay, 1996, Introduction to Quantitative Genetics)
- No systematic forces changing allele and genotypic frequencies

Expected genotypic frequencies of a biallelic locus under Hardy-Weinberg equilibrium (HWE)

- Mathematical definition of HWE: $P_{ii} = p_i^2$, $P_{ij} = 2p_i p_j$, i,j = 1,2

**Table 2.2.1** Genotypic frequencies of a biallelic locus under Hardy-Weinberg equilibrium.

| Genotype | $A_1A_1$ | $A_1A_2$ | $A_2A_2$ |
|---|---|---|---|
| Number of individuals | $N_{11}$ | $N_{12}$ | $N_{22}$ |
| Genotypic frequency in general | $P_{11} = N_{11}/N$ | $P_{12} = N_{12}/N$ | $P_{22} = N_{22}/N$ |
| Genotypic frequency under HWE | $P_{11} = p_1^2$ | $P_{12} = 2p_1p_2$ | $P_{22} = p_2^2$ |

$N = N_{11} + N_{12} + N_{22}$ = total number of individuals

- Relationship between allele and genotypic frequencies under HWE:
$(p_1 + p_2)^2 = p_1^2 + 2p_1p_2 + p_2^2$

Hardy-Weinberg disequilibrium (HWD)

- Definition of HWD: $P_{ii} \neq p_i^2$, $P_{ij} \neq 2p_i p_j$, i,j = 1,2

- Genotypic frequency as a function of the HWE frequency and HWD parameter.

An observed genotypic frequency can be expressed a function of the expected genotypic frequency under HWE plus a deviation from HWE:

$$P_{11} = p_1^2 + \Delta_{11} \quad (2.2.1)$$

$$P_{12} = 2p_1p_2 + \Delta_{12} \quad (2.2.2)$$

$$P_{22} = p_2^2 + \Delta_{22} \quad (2.2.3)$$

$\Delta_{ij}$ = deviation from the expected genotypic frequency under HWE

Using $p_1 = P_{11} + 0.5P_{12}$ of Equation (2.1.1), Equations (2.2.1-2.2.3) reduce to:

$$P_{11} = p_1^2 + \Delta \quad (2.2.4)$$
$$P_{12} = 2p_1p_2 - 2\Delta \quad (2.2.5)$$
$$P_{22} = p_2^2 + \Delta \quad (2.2.6)$$

$\Delta$ = HWD parameter.

Proof:

$$\begin{aligned} P_{11} + 0.5P_{12} &= p_1 \\ &= p_1^2 + \Delta_{11} + 0.5(2p_1p_2 + \Delta_{12}) \\ &= p_1^2 + p_1p_2 + \Delta_{11} + 0.5\Delta_{12} \\ &= p_1(p_1 + p_2) + \Delta_{11} + 0.5\Delta_{12} \\ &= p_1 + \Delta_{11} + 0.5\Delta_{12} \end{aligned}$$

$\Delta_{11} + 0.5\Delta_{12} = 0$, $\Delta_{11} = -0.5\Delta_{12}$

$$P_{22} + 0.5P_{12} = p_2 = p_2 + \Delta_{11} + 0.5\Delta_{12}$$

$\Delta_{22} + 0.5\Delta_{12} = 0$, $\Delta_{22} = -0.5\Delta_{12} = \Delta_{11} = \Delta$

$\Delta_{12} = -2\Delta_{11} = -2\Delta_{22} = -2\Delta$

- Test of HWD: can be used for screening genome-wide SNP markers

From (2.2.4)-(2.2.6),

$$\Delta = P_{11} - p_1^2 = P_{22} - p_2^2 = (2p_1p_2 - P_{12})/2 \quad (2.2.7)$$

$$\chi^2 = \frac{N\Delta^2}{p_1^2p_2^2}, \quad \text{d.f.} = 1 \quad (2.2.8)$$

$H_0$: HWE, $H_1$: HWD

Accept $H_0$ and reject $H_1$ if $\chi^2 < \chi^2(\text{d.f.}, \alpha)$

Reject $H_0$ and accept $H_1$ if $\chi^2 \geq \chi^2(\text{d.f.}, \alpha)$

$\chi^2(\text{d.f.}, \alpha)$ = threshold $\chi^2$ value to declare significance

d.f. = degrees of freedom

$\alpha$ = significance level = type-I error

## 8.1 Gametic disequilibrium

### Linkage disequilibrium between two loci

Linkage disequilibrium (LD)
- Non-random association between alleles, between allele and genotype, and between genotypes
- LD is the foundation for genome-wide association study

Linkage equilibrium (LE)
- Association between two linked loci is the same as two unlinked loci
- Relationship between gametic array and genotypic array is defined by that for two independent loci

Gametic and genotypic array for two loci

(Recall: gametic and genotypic arrays for one locus, Equations 2.4.1-2.4.3)
- $A_iB_j$ = haplotype with alleles $A_i$ and $B_j$ from a parent, i = 1, ..., n, j = 1,...,m
- $P_{ij}$ = probability of $A_iB_j$
- Gametic array in the population: $P_{11}A_1B_1 + ... + P_{nm}A_nB_m$
- Genotypic array in the population

$$(P_{11}A_1B_1 + ... + P_{nm}A_nB_m)^2$$
$$= P_{11}^2(A_1B_1/A_1B_1) + ... + P_{nm}^2(A_nB_m/A_nB_m)$$
$$+ 2P_{11}P_{12}(A_1B_1/A_1B_2) + ... + 2P_{(n-1)m}P_{nm}[A_{(n-1)}B_m/A_nB_m] \quad (8.1.1)$$

- Two loci are independent if:
  a) the two loci are unlinked, or
  b) the two loci are in LE

- Gametic array for two independent loci or two loci in LE

$$P_{11}A_1B_1 + ... + P_{nm}A_nB_m = (p_1A_1 + ... + p_nA_n)(q_1B_1 + ... + q_mB_m) \quad (8.1.2)$$

- Genotypic array for two independent loci or two loci in LE

$$(P_{11}A_1B_1 + ... + P_{nm}A_nB_m)^2 = (p_1A_1 + ... + p_nA_n)^2(q_1B_1 + ... + q_mB_m)^2 \quad (8.1.3)$$

$$\left(\sum_{i=1}^{n} p_iA_i\right)^2 = \left(\sum_{i=1}^{n} p_i^2A_iA_i\right) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_ip_jA_iA_j$$
$$\left(\sum_{i=1}^{m} p_iB_i\right)^2 = \left(\sum_{i=1}^{m} p_i^2B_iB_i\right) + 2\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} p_ip_jB_iB_j$$

$p_i$ = allele frequency of $A_i$, i = 1,...,n ; $q_j$ = allele frequency of $B_j$, j = 1,...,m

**LD analysis of haplotype data**

- Two bi-allelic loci assumed
- Allele frequencies

  $p_A$ = prob{A}= frequency of allele A of locus 1
  $p_a$ = prob{a} = frequency of allele A of locus 1
  $q_B$ = prob{B} = frequency of allele B of locus 2
  $q_b$ = prob{b} = frequency of allele b of locus 2

- Population gametic frequency under LE

  $Q_{AB} = p_A q_B$, $Q_{Ab} = p_A q_b$, $Q_{aB} = p_a q_B$, $Q_{ab} = p_a q_b$

- Population haplotype (gametic) frequency under LD

  $P_{AB}$ = prob{AB} = $p_A q_B + D_{AB}$, $P_{Ab}$ = prob{Ab} = $p_A q_b + D_{Ab}$
  $P_{aB}$ = prob{aB} = $p_a q_B + D_{aB}$, $P_{ab}$ = prob{ab} = $p_a q_b + D_{ab}$

  $D_{ij}$ = deviation of gametic frequency from linkage equilibrium (i = A,a; j = B,b)

  $p_A = P_{AB} + P_{Ab} = p_A q_B + D_{AB} + p_A q_b + D_{Ab} = p_A + D_{AB} + D_{Ab}$
  $D_{AB} + D_{Ab} = 0$,
  $D_{AB} = -D_{Ab}$

  $P_{AB} = p_A q_B + D$, $P_{Ab} = p_A q_b - D$
  $P_{aB} = p_a q_B - D$, $P_{ab} = p_a q_b + D$

  D = LD parameter = $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$

<u>Estimation of LD</u>

$P_{AB}P_{ab} = (p_A q_B + D)(p_a q_b + D) = p_A q_B\, p_a q_b + D p_a q_b + D p_A q_B + D^2$
$P_{Ab}P_{aB} = (p_A q_b - D)(p_a q_B - D) = p_A q_b\, p_a q_B - D p_a q_B - D p_A q_b + D^2$

$\therefore D = P_{AB}P_{ab} - P_{Ab}P_{aB}$ (8.1.4)

Statistical test of LD

**Table 8.1.1** Linkage disequilibrium test for haplotype data

| | B | b | |
|---|---|---|---|
| A | $n_{11}$<br>$P_{AB} = n_{11} / n$ | $n_{12}$<br>$P_{Ab} = n_{12} / n$ | $n_{\bullet 1} = n_{11} + n_{12}$<br>$p_A = P_{AB} + P_{Ab} = n_{1\cdot} / n$ |
| a | $n_{21}$<br>$P_{aB} = n_{21} / n$ | $n_{22}$<br>$P_{ab} = n_{22} / n$ | $n_{\bullet 2} = n_{21} + n_{22}$<br>$p_a = P_{aB} + P_{ab} = n_{2\cdot} / n$ |
| | $n_{1\bullet} = n_{11} + n_{12}$<br>$q_B = P_{AB} + P_{aB} = n_{1}/n$ | $n_{2\bullet} = n_{12} + n_{22}$<br>$q_b = P_{Ab} + P_{ab} = n_{\cdot 2}/n$ | n<br>1 |

$$\chi^2 = n\rho^2, \qquad \text{d.f.} = 1 \tag{8.1.5}$$

$$\rho = \frac{\frac{n_{11}n_{22} - n_{12}n_{21}}{n^2}}{\sqrt{n_{1\bullet}n_{2\bullet}n_{\bullet 1}n_{\bullet 2}\frac{1}{n^4}}} = \frac{D}{\sqrt{p_A p_a q_B q_b}} \tag{8.1.6}$$

## 9.1 Identity By Descent and Identity By State

Identity by descent (IBD): two copies of the same allele originated from the same ancestor
Identity by state (IBS): two copies of the same allele with unknown ancestral origin

- IBS = upper bound of IBD

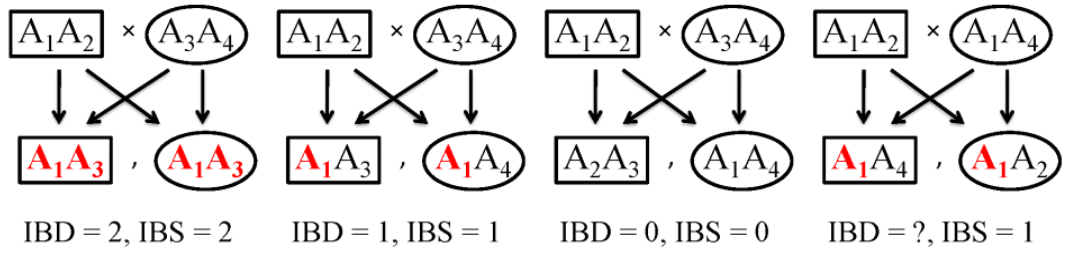



**Figure 9.1.1.** Examples of identity by descent (IBD) and identity by state (IBS).

## REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93(2):743-752.

Bian, C., D. Prakapenka, C. Tan, R. Yang, D. Zhu, X. Guo, D. Liu, G. Cai, Y. Li, and Z. Liang. 2021. Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. Genetics Selection Evolution 53(1):1-19.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol 42(2):1-8.

Cockerham, C. C. 1954. An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39(6):859.

Da, Y., C. Wang, S. Wang, and G. Hu. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PloS One 9(1):e87666.

Da, Y. 2015. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. BMC genetics 16(1):144.

Da, Y., Z. Liang, and D. Prakapenka. 2022. Multifactorial methods integrating haplotype and epistasis effects for genomic estimation and prediction of quantitative traits. Frontiers in Genetics 13:922369doi: doi: 10.3389/fgene.2022.922369

Draper, N. R., and H. Smith. 1998. Applied regression analysis. John Wiley & Sons.
Endelman, J. B., and J.-L. Jannink. 2012. Shrinkage estimation of the realized relationship matrix. G3: Genes, Genomes, Genetics 2(11):1405-1413.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. 4 ed. Longmans Green, Harlow, Essex, UK.

Fernando, R. L., H. Cheng, and D. J. Garrick. 2016. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. Genetics Selection Evolution 48(1):1-12.

Fisher, R. A. 1918. The Correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52(02):399-433.

Gilmour, A. R., R. Thompson, and B. R. Cullis. 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics:1440-1450.

Hartley, H. O., and J. N. Rao. 1967. Maximum-likelihood estimation for the mixed analysis of variance model. Biometrika 54(1-2):93-108.

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72(358):320-338.

Harville, D. A. 1997. Matrix algebra from a statistician's perspective. Springer-Verlag New York, Inc.

Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome 53(11):876-883.

Hazel, L. N. 1943. The genetic basis for constructing selection indexes. Genetics 28(6):476-490.

Henderson, C. 1975. Rapid method for computing the inverse of a relationship matrix. Journal of dairy science 58(11):1727-1730.

Henderson, C. 1977. Best linear unbiased prediction of breeding values not in the model for records. Journal of dairy science 60(5):783-787.
Henderson, C. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Ontario.
Henderson, C. 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of animal science 60(1):111-117.
Hoerl, A. E. 1960. Application of ridge regression analysis to regression problems. Chemical Engineering Progress 58:54-59.
Jiang, J., L. Ma, D. Prakapenka, P. M. VanRaden, J. B. Cole, and Y. Da. 2019. A large-scale genome-wide association study in US Holstein cattle. Frontiers in genetics 10:412.
Jiang, Y., and J. C. Reif. 2020. Efficient algorithms for calculating epistatic genomic relationship matrices. Genetics 216(3):651-669.
Johnson, D., and R. Thompson. 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. Journal of dairy science 78(2):449-456.
Kempthorne, O. 1954. The correlation between relatives in a random mating population. Proceedings of the Royal Society of London. Series B-Biological Sciences 143(910):103-113.
Liang, Z., C. Tan, D. Prakapenka, L. Ma, and Y. Da. 2020. Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. Frontiers in Genetics 11(1461):588907. doi: 10.3389/fgene.2020.588907
Liang, Z., D. Prakapenka, K. L. Parker Gaddis, M. J. VandeHaar, K. A. Weigel, R. J. Tempelman, J. E. Koltes, J. E. P. Santos, H. M. White, and F. Peñagaricano. 2022. Impact of epistasis effects on the accuracy of predicting phenotypic values of residual feed intake in U. S Holstein cows. Frontiers in Genetics:2930.
Liang, Z., D. Prakapenka, and Y. Da. 2023. Comparison of the accuracy of epistasis and haplotype models for genomic prediction of seven human phenotypes. Biomolecules 13(10):1478.
Lee, S. H., and J. H. van der Werf. 2006. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genetics Selection Evolution 38(1):1-19.
Legarra, A., C. Robert-Granié, E. Manfredi, and J.-M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180(1):611-618.
Ma, L., H. B. Runesha, D. Dvorkin, J. Garbe, and Y. Da. 2008. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies. BMC bioinformatics 9(1):315.
Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. Multivariate analysis. Academic press.
Meuwissen, T., B. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819-1829.
Misztal, I. 2015. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics: 115.182089.
Muñoz, P. R., M. F. Resende Jr, S. A. Gezan, M. D. V. Resende, G. de Los Campos, M. Kirst, D. Huber, and G. F. Peter. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198(4):1759-1768.

Patterson, H. D., and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58(3):545-554.

Pollak, E. J. 1984. Transformation for systematically missing data to facilitate multiple trait evaluations. Page 55 in Genetic Research. Report to Eastern AI Coop., Inc., Dept. of Anim. Sci., Cornell Univ., Ithaca, NY.

Prakapenka, D., C. Wang, Z. Liang, C. Bian, C. Tan, and Y. Da. 2020. GVCHAP: a computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. Frontiers in Genetics 11:282. doi: 10.3389/fgene.2020.00282

Prakapenka, D., Z. Liang, J. Jiang, L. Ma, and Y. Da. 2021. A Large-scale genome-wide association study of epistasis effects of production traits and daughter pregnancy rate in US Holstein cattle. Genes 12(7):1089.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, and M. J. Daly. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81(3):559-575.

Searle, S. R. 1982. Matrix algebra useful for statistics. John Wiley & Sons, New York, New York.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance components. John Wiley & Sons.

Smith, H. F. 1936. A discriminant function for plant selection. Annals of Eugenics 7(3):240-250.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund. 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PloS One 7(9):e45293.

Tan, C., Z. Wu, J. Ren, Z. Huang, D. Liu, X. He, D. Prakapenka, R. Zhang, N. Li, and Y. Da. 2017. Genome-wide association study and accuracy of genomic prediction for teat number in Duroc pigs using genotyping-by-sequencing. Genetics Selection Evolution 49(1):35.

Tier, B. 1990. Computing inbreeding coefficients quickly. Genetics Selection Evolution 22(4):419.

VanRaden, P. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414-4423.

Vitezica, Z. G., A. Reverter, W. Herring, and A. Legarra. 2018. Dominance and epistatic genetic variances for litter size in pigs using genomic models. Genetics Selection Evolution 50(1):1-8.

Wang, C., and Y. Da. 2014. Quantitative genetics model as the unifying model for defining genomic relationship and inbreeding coefficient. PLoS ONE 9: e114484

Wang, C., D. Prakapenka, S. Wang, S. Pulugurta, H. B. Runesha, and Y. Da. 2014. GVCBLUP: a computer package for genomic prediction and variance component estimation of additive and dominance effects. BMC Bioinformatics 15(1):270.