Covariance between relatives: A reminder

Daniel Sorensen*

January 26, 2023

1 Covariance between relatives

The covariance between relatives for a purely additive genetic model with two loci in LD is briefly sketched out. The development involves covariance terms between individuals at the same locus, and covariance terms between individuals at different loci. The term for covariances at the same locus is derived first.

1.1 Covariance at a single locus

An example motivates the general case. Imagine a locus denoted A. The genotype of a father is A_1A_2 and of a mother A_3A_4. Consider two offspring from these parents, and the possible number of alleles shared identical by descent (IBD) between the two. There are 16 possible genotype combinations for the two offspring genotypes (arranged in a 4×4 table, where the columns are the possible genotypes for offspring 1, and the rows the possible genotypes for offspring 2). The number of alleles shared IBD between the two offspring i and j, N_{ij}, can take the following values

- $N_{ij} = 2$ (4 cases out of 16)
- $N_{ij} = 1$ (8 cases out of 16)
- $N_{ij} = 0$ (4 cases out of 16)

Therefore

$$\begin{aligned}
\mathrm{E}(N_{ij}) &= 0\,\Pr(N_{ij}=0) + 1\,\Pr(N_{ij}=1) + 2\,\Pr(N_{ij}=2) \\
&= 1\frac{1}{2} + 2\frac{1}{4} = 1
\end{aligned}$$

*Center for Quantitative Genetics and Genomics, Aarhus University, C F Møllers Alle 3, bygning 1130, 8000 Aarhus Denmark

and the expected proportion of alleles shared IBD is

$$\frac{\mathrm{E}(N_{ij})}{2} = a_{ij} \tag{1}$$

where a_{ij} is also known as the expected additive genetic relationship between i and j, which is the element in the ith row and jth column of the additive genetic relationship matrix A. In the present example the expected proportion is $a_{ij} = 0.5$, the expected number is 1, but the two full-sibs can share 0, 1 or 2 alleles IBD, with probabilities 1/4, 1/2 and 1/4, respectively.

Denote the additive genetic value, or breeding value of individual j

$$g_j = \alpha z_j$$

where α is the additive genetic effect for a locus (or additive effect of a gene substitution), and z_j is the centred genotypic code for the locus. Due to the centring of z

$$\mathrm{E}(g_j|\alpha) = \alpha\,\mathrm{E}(z_j) = 0.$$

The additive genetic variance in the population contributed by the locus is

$$V_g = \mathrm{E}\big(g_j^2|\alpha\big) = \alpha^2\,\mathrm{Var}(z_j).$$

Consider the covariance between offspring i and j, conditional on N_{ij}. There are three possible outcomes

- $N_{ij} = 0$,

  $$\begin{aligned}\mathrm{Cov}(g_i, g_j|N_{ij} = 0) &= \mathrm{E}(g_i\,g_j|N_{ij} = 0) - \mathrm{E}(g_i|N_{ij} = 0)\,\mathrm{E}(g_j|N_{ij} = 0)\\ &= \mathrm{E}(g_i|N_{ij} = 0)\,\mathrm{E}(g_j|N_{ij} = 0) - E(g_i|N_{ij} = 0)\,\mathrm{E}(g_j|N_{ij} = 0) = 0,\end{aligned}$$

  because if individuals do not share alleles IBD, the $g's$ are independent.

- $N_{ij} = 1$,

  $$\mathrm{Cov}(g_i, g_j|N_{ij} = 1) = \frac{1}{2}Vg,$$

  the gametic variance.

- $N_{ij} = 2$,

  $$\mathrm{Cov}(g_i, g_j|N_{ij} = 2) = Vg,$$

  the additive genetic variance at the locus. These three cases can be written compactly as

  $$\mathrm{Cov}(g_i, g_j|N_{ij}) = \frac{N_{ij}}{2}Vg, \quad N_{ij} = 0, 1, 2. \tag{2}$$

Then, marginally with respect to N_{ij},

$$\begin{aligned}
\mathrm{Cov}(g_i, g_j) &= \mathrm{E}[\mathrm{Cov}(g_i, g_j|N_{ij})] + \mathrm{Cov}[\mathrm{E}(g_i|N_{ij})\,\mathrm{E}(g_j|N_{ij})] \\
&= \mathrm{E}[\mathrm{Cov}(g_i, g_j|N_{ij})] \\
&= \frac{\mathrm{E}(N_{ij})}{2} Vg \\
&= a_{ij} Vg \qquad (3)
\end{aligned}$$

where the last line uses (1).

1.2 Covariance involving different loci

Let Θ_{ij} denote the coefficient of coancestry, equal to the probability that two gametes, one from individual i and the other from individual j, are IBD. Consider two loci, k and l. At locus k individual i has genotype z_{ik}, coded as $(0, 1, 2)$ and at locus l, individual j has genotype z_{jl}, also originally coded as $(0, 1, 2)$. Let $z_{ikm} = 0, 1$ and $z_{ikp} = 0, 1$, be the binary random variables representing the maternal and paternal contributions to z_{ik} with similar coding for z_{jl}. Then

$$\begin{aligned}
z_{ik} &= z_{ikm} + z_{ikp}, \\
z_{jl} &= z_{jlm} + z_{jlp.}
\end{aligned}$$

Assume that these binary random variables are centred so that $E(z_{ikm})$, say, is equal to zero, which renders z_{ik} and z_{jl} also centred.

The covariance between z_{ik} and z_{jl} is

$$Cov(z_{ik}, z_{jl}) = Cov(z_{ikm} + z_{ikp}, z_{jlm} + z_{jlp}). \qquad (4)$$

Let the binary random variable W take the value 1, if a randomly drawn gamete from i is IBD with a randomly drawn gamete from j, and zero otherwise. Then $\Pr(W = 1) = \Theta_{ij}$. There are 4 terms contributing to (4) and all have the following form:

$$\begin{aligned}
Cov(z_{ikm}, z_{jlm}) &= E(z_{ikm}\, z_{jlm}) - E(z_{ikm})E(z_{jlm}) \\
&= E_W[E(z_{ikm}, z_{jlm}|W)] \\
&= E(z_{ikm}, z_{jlm}|W = 1) \Pr(W = 1) + E(z_{ikm}, z_{jlm}|W = 0) \Pr(W = 0) \\
&= D_{kl}\Theta_{ij}, \qquad (5)
\end{aligned}$$

where D_{kl}, the linkage disequilibrium parameter between loci k and l, is here the covariance between the maternal allele at locus k, and the maternal allele at locus l, and Θ_{ij} is the probability that the gametes drawn are IBD. The second term in the third line vanishes when $W = 0$, because if the gametes are not IBD, they are independent and $E(z_{ikm}\, z_{jlm}|W) = E(z_{ikm})E(z_{jlm}) = 0$. Summing over all 4 terms yields

$$\begin{aligned}
Cov(z_{ik}, z_{jl}) &= 4D_{kl}\Theta_{ij} \\
&= 2a_{ij}D_{kl}, \qquad (6)
\end{aligned}$$

where a_{ij} is the expected additive genetic relationship between i and j, since $a_{ij} = 2\Theta_{ij}$. The covariance between additive genetic values of individuals i and j is

$$Cov(\alpha_k z_{ik}, \alpha_l z_{jl} | \alpha_k, \alpha_l) = 2a_{ij}\, \alpha_k\, \alpha_l\, D_{kl}. \tag{7}$$

1.3 A remark

At this moment it is useful to reflect on the sampling scheme that gives rise to the covariance structure given by (6). Gametes are assumed to be randomly drawn from a pool of gametes where the Mendelian lottery has taken place. The sampling process does not account for changes of the disequilibrium parameter due to recombination as new cohorts are produced, or for changes due to selective forces over generations. The disequilibrium parameter D_{kl}, involving loci k and l, is assumed to be the same across individuals. The sampling scheme assumed is an approximation that provides the basis for a static description of the population and leads to a model best suited for pedigrees involving cohorts born within a generation, as found often in human pedigrees. The mathematics of this approximation is equivalent to that of a single locus model, with two individuals and two alleles randomly sampled, one from each individual. Then, setting $k = l$ and replacing D_{kl} in expression (5) by $p_k(1 - p_k)$, (7) reduces to $2a_{ij}\alpha_k^2 p_k(1 - p_k)$, the same as (3). An exact general treatment involving pairs of loci constitutes a formidable challenge leading to unwieldy expressions, as shown by **?**. The curious reader may wish to glance with awe at formula (6) for the genetic variance in their article, that is almost two pages long! Results assuming lack of inbreeding, epistasis and assortative mating, but accounting for dominance, linkage, and for the dynamics of the linkage disequilibrium parameter over generations, lead to simpler expressions and are given by **?**.

References