Covariance between relatives: A reminder

Daniel Sorensen*

April 11, 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; in the diagonal of the 4×4 table)
- $N_{ij} = 1$ (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 (dividing by the number of alleles in i, 2 for diploids)

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

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. With random mating and no inbreeding the expected additive genetic relationship a_{ij} is also the correlation between the additive genetic values of individuals i and j.

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 (centred allele content of the genotype) 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.

- $N_{ij} = 1$,

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

  the gametic variance.

- $N_{ij} = 2$,

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

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

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

Then, marginally with respect to N_{ij},

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

where the last line uses (1).

1.1.1 An alternative derivation

The traditional derivation of the covariance between relatives uses the concept of identity by descent (IBD). Two genes are IBD if they are biochemical replicates produced without mutation from a common ancestral gene. The probability that genes A_x and A_y at locus A are IBD is denoted $\Pr(A_x = A_y)$. If genes A_x and A_y at locus A belong in the same individual Z, $\Pr(A_x = A_y) = F_Z$, the inbreeding coefficient of individual Z.

The coefficient of parentage between i and j is the probability that a gene drawn at random from a particular locus in i is IBD with a gene drawn at random from the same locus in j. The probability of drawing a paternal or maternal gene from individual i (or from individual j) is 1/2. Therefore the probability of drawing any of the four possible combinations of maternal or paternal genes from i and j is equal to 1/4, the product of these independent evants. If the two individuals i and j have genotypes at locus k, $A_{ikm}A_{ikp}$ and $A_{jkm}A_{jkp}$, where m and p stand for the maternally and paternally inherited gametes, then the coefficient of parentage between i and j is

$$\begin{aligned}\Theta_{ij} &= \frac{1}{4}(\Pr(A_{ikm} = A_{jkm}) + \Pr(A_{ikm} = A_{jkp}) + \Pr(A_{ikp} = A_{jkm}) + \Pr(A_{ikp} = A_{jkp})) \\ &= \frac{1}{4}(\Theta_{ikm,jkm} + \Theta_{ikm,jkp} + \Theta_{ikp,jkm} + \Theta_{ikp,jkp}). \end{aligned} \tag{3}$$

The expected additive genetic relationship a_{ij} between individuals i and j is twice the coefficient of parentage:

$$a_{ij} = 2\Theta_{ij}. \tag{4}$$

New notation is introduced that will be useful for the next section. Let z^*_{ik} denote the allele content of individual i at locus k that can take values $z^*_{ik} = 0, 1, 2$. The allele content is the result of independent contributions from the two gametes inherited by i:

$$z^*_{ik} = z^*_{ikm} + z^*_{ikp},$$

where each gametic contribution $z^*_{ikx} = 0, 1$, $x = m, p$, is a binary random variable with expected value $\mathrm{E}(z^*_{ikx}) = \Pr(z^*_{ikx} = 1) = p_k$ and variance $\mathrm{Var}(z^*_{ikx}) = p_k(1 - p_k)$ (not to confuse the expected value p_k with the subscript p indicating a gamete from paternal origin). From now on the gametic contributions z^*_{ikx} are centred, so that $z_{ikx} = z^*_{ikx} - p_k$, and therefore $\mathrm{E}(z_{ikx}) = 0$ and $\mathrm{Var}(z_{ikx}) = p_k(1 - p_k)$.

The additive genetic value of individual i at locus k is

$$\alpha_k\, z_{ik} = \alpha_k(z_{ikm} + z_{ikp})$$

and the additive genetic variance contributed by locus k in the large population maintained by random mating (ensuring that z_{ikm} and z_{ikp} are independent) is

$$\mathrm{Var}(\alpha_k\, z_{ik}|\alpha_k) = \alpha_k^2\, 2p_k(1-p_k). \tag{5}$$

Consider two individuals i and j with additive genetic values $\alpha_k\, z_{ik}$ and $\alpha_k\, z_{jk}$. The covariance between the additive genetic values of i and j is

$$\mathrm{Cov}(\alpha_k\, z_{ik}, \alpha_k\, z_{jk}|\alpha_k) = \alpha_k^2\ \mathrm{Cov}(z_{ik}, z_{jk}). \tag{6}$$

The covariance term is

$$\mathrm{Cov}(z_{ik}, z_{jk}) = \mathrm{Cov}(z_{ikm} + z_{ikp}, z_{jkm} + z_{jkp}). \tag{7}$$

There are four terms contributing to this covariance and in view of the centring each is of the form

$$\mathrm{Cov}(z_{ikm}, z_{jkm}) = \mathrm{E}(z_{ikm}\, z_{jkm}). \tag{8}$$

Let W be a binary random variable that takes the value 1 if A_{ikx} is IBD with A_{jkx}, $x = m, p$, and 0 otherwise. Then

$$\begin{aligned}
\mathrm{E}(z_{ikm}\, z_{jkm}) &= \mathrm{E}_w(E(z_{ikm}\, z_{jkm}|W)) \\
&= \mathrm{E}(z_{ikm}\, z_{jkm}|W=1)\,\Pr(W=1) + \mathrm{E}(z_{ikm}\, z_{jkm}|W=0)\,\Pr(W=0) \\
&= \mathrm{E}(z_{ikm}\, z_{jkm}|W=1)\,\Pr(W=1).
\end{aligned} \tag{9}$$

The second term drops out because if A_{ikm} and A_{jkm} are not IBD the two alleles are independent, $\mathrm{E}(z_{ikm}\, z_{jkm}|W=0) = \mathrm{E}(z_{ikm}|W=0)\,\mathrm{E}(z_{jkm}|W=0) = \mathrm{E}(z_{ikm})\,\mathrm{E}(z_{jkm}) = 0$. If A_{ikm} and A_{jkm} are IBD they are the same allele and $\mathrm{E}(z_{ikm}\, z_{jkm}|W=1) = \mathrm{E}(z_{ikm}^2) = p_k(1-p_k)$. On the basis of these results, expression (7) is

$$\mathrm{Cov}(z_{ik}, z_{jk}) = 4\Theta_{ij}\, p_k(1-p_k),$$

where Θ_{ij} is defined in (3). From (6) the additive genetic covariance between i and j is

$$\begin{aligned}
\mathrm{Cov}(\alpha_k\, z_{ik}, \alpha_k\, z_{jk}|\alpha_k) &= 4\Theta_{ij}\, \alpha_k^2\, p_k(1-p_k) \\
&= 2a_{ij}\, \alpha_k^2\, p_k(1-p_k).
\end{aligned} \tag{10}$$

The equality in the second line follows from (4).

1.2 Covariance involving different loci

Let $\tilde{\Theta}_{ikm,jlm}$ denote the probability that an allele drawn from locus k in the maternal gamete of individual i and an allele drawn from locus l in the maternal gamete of individual j are copies of genes that originate from the gamete of a common ancestor. More generally, the property that two alleles from different loci taken from two individuals i and j are copies of genes that originate from the gamete of a common ancestor is known as *equivalence by descent*, $\tilde{\Theta}_{ij}$ (EBD, Weir and Cockerham, 1974).

The centred allele contents of individuals i and j at loci k and l, respectively are

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

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

$$\mathrm{Cov}(z_{ik}, z_{jl}) = \mathrm{Cov}(z_{ikm} + z_{ikp}, z_{jlm} + z_{jlp}). \tag{11}$$

Let the binary random variable W take the value 1, if a randomly drawn allele from i at locus k and an allele from j at locus l are EBD. There are 4 terms contributing to (11) that have the following form

$$\begin{aligned} \mathrm{Cov}(z_{ikm}, z_{jlm}) &= \mathrm{E}(z_{ikm}\, z_{jlm}) \\ &= \mathrm{E}_W[\mathrm{E}(z_{ikm}\, z_{jlm}|W)] \\ &= \mathrm{E}(z_{ikm}\, z_{jlm}|W=1)\Pr(W=1) + \mathrm{E}(z_{ikm}\, z_{jlm}|W=0)\Pr(W=0) \\ &= D_{kl}\tilde{\Theta}_{ikm,jlm}, \end{aligned} \tag{12}$$

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 in the gametes of the common ancestor's generation. The equality in the first line holds because terms like $\mathrm{E}(z_{ikm})$ are equal to zero. The second term in the third line vanishes when $W = 0$, because if the alleles are not EBD, they originate from different independent gametes from the common ancestor. Therefore, $\mathrm{E}(z_{ikm}|W=0) = \mathrm{E}(z_{ikm}) = 0$ and $\mathrm{E}(z_{ikm}\, z_{jlm}|W=0) = \mathrm{E}(z_{ikm})\,\mathrm{E}(z_{jlm}) = 0$. Summing over all 4 terms yields

$$\begin{aligned} \mathrm{Cov}(z_{ik}, z_{jl}) &= D_{kl}\Big(\tilde{\Theta}_{ikm,jlm} + \tilde{\Theta}_{ikm,jlp} + \tilde{\Theta}_{ikp,jlm} + \tilde{\Theta}_{ikp,jlp}\Big) \\ &= 2\tilde{a}_{ij}D_{kl}, \end{aligned} \tag{13}$$

where $\tilde{a}_{ij}$ is the expected additive genetic relationship between i and j, since

$$\tilde{a}_{ij} = 2\tilde{\Theta}_{ij} = \frac{1}{2}\Big(\tilde{\Theta}_{ikm,jlm} + \tilde{\Theta}_{ikm,jlp} + \tilde{\Theta}_{ikp,jlm} + \tilde{\Theta}_{ikp,jlp}\Big).$$

The covariance between additive genetic values of locus k of individual i and locus l of individual j is

$$\mathrm{Cov}(\alpha_k z_{ik}, \alpha_l z_{jl}|\alpha_k, \alpha_l) = 2\tilde{a}_{ij}\,\alpha_k\,\alpha_l\,D_{kl}, \tag{14}$$

and the contribution to the covariances between additive genetic values of individuals i and j from different loci, including the 8 terms associated with loci k and l in i and loci k and l in j is

$$\alpha_k\,\alpha_l[\mathrm{Cov}(z_{ik}, z_{jl}|\alpha_k, \alpha_l) + \mathrm{Cov}(z_{il}, z_{jk}|\alpha_k, \alpha_l)] = 4\tilde{a}_{ij}\,\alpha_k\,\alpha_l\,D_{kl}. \tag{15}$$

In (13), (14) and (15) $\tilde{a}_{ij}$ is used to distinguish it from a_{ij} in (2). The latter involves the probability of IBD for alleles of the same locus, whereas the former considers the probability of EBD of alleles from different loci. The example below illustrates that the probability of transmission of two alleles drawn from a common ancestor to produce two independent gametes is the same regardless of whether the alleles belong in the same or in different loci. In this case there is no need to use different notation for $\tilde{a}_{ij}$ and a_{ij} (see also Lynch and Walsh (1998), page 151).

1.2.1 Example

Imagine a sire that at two linked loci A and B has genotype $A_pA_m//B_pB_m$, where m and p stand for the maternal and the paternal haplotype of the sire. Thus, the maternal haplotype carries alleles A_mB_m and the paternal haplotype A_pB_p. This sire produces four possible gametes; two non-recombinant types with probabilities

$$\begin{aligned}\Pr(A_pB_p) &= \frac{1}{2}(1-c),\\ \Pr(A_mB_m) &= \frac{1}{2}(1-c)\end{aligned}$$

and two recombinant types

$$\begin{aligned}\Pr(A_pB_m) &= \frac{1}{2}c,\\ \Pr(A_mB_p) &= \frac{1}{2}c,\end{aligned}$$

where $c \in \left[0, \frac{1}{2}\right]$ is the probability of recombination between loci A and B. The marginal probability of transmission of any of the two alleles at each locus is $\frac{1}{2}$.

The sire mates with two randomly chosen females and produces one offspring from each mating, half-sibs X and Y.

- Consider locus A only. Offspring X inherits genes A_p or A_m from its father and genes N_p or N_m from its mother. Offspring Y inherits genes A_p or A_m from its father and genes M_p or M_m from its mother. Of the 16 possible pair of genes that can be drawn from X and Y, in two cases the pair of genes are IBD since they originate from the common father. These cases are the events defined by drawing A_p from X and A_p from Y, and A_m from X and A_m from Y. Therefore $\Theta_{XY} = 1/8$, since each of the 16 events occurs with equal probability.

- Consider two loci and the event defined by randomly choosing a gene from locus 1 in X and a gene from locus 2 in Y. Importantly, these are two independent events. At locus 1, X can inherit genes A_p or A_m from its father and genes M_p or M_m from its mother. At locus 2, Y can inherit genes B_p or B_m from its father, and genes N_p or N_m from its mother. The random draw of a gene from locus 1 in X and a gene from locus 2 in Y defines 16 possible events each with probability 1/16. In 2 out of the 16 events, the genes drawn from X and Y are copies of the haplotype of their common father: A_mB_m and A_pB_p. Therefore $\widetilde{\Theta}_{XY} = \Theta_{XY} = 1/8$. (The same argument holds when instead a gene from locus 2 is drawn from X and a gene from locus 1 is drawn from Y. These events occur each with probability 1/2).

The general result is that when additive genetic relationships are computed tracing related individuals to their most recent common ancestor, the probability of EBD for genes at different loci in different gametes is the same as the probability of IBD for alleles at the same locus (Lynch and Walsh, 1998).

The covariance between two relatives i and j in an additive genetic model has a contribution from covariances between additive genetic values at the same locus, given by (10) and a contribution from covariances between additive genetic values at different loci, given by (15). The general expression for a model of additive gene action within and between q loci is

$$\begin{aligned}\mathrm{Cov}(\alpha z_i, \alpha z_j|\alpha) &= 2a_{ij}\sum_{k=1}^{q}\alpha_k^2 p_k(1-p_k) + 4a_{ij}\sum_{k=1}^{q-1}\sum_{l=k+1}^{q}\alpha_k\alpha_l D_{kl} \\ &= a_{ij}\left[2\sum_{k=1}^{q}\alpha_k^2 p_k(1-p_k) + 4\sum_{k=1}^{q-1}\sum_{l=k+1}^{q}\alpha_k\alpha_l D_{kl}\right], \qquad (16)\end{aligned}$$

where a_{ij} is the expected additive genetic relationship between i and j. The term in square brackets is the additive genetic variance that has a component due to single loci and a component from pairs of correlated loci (linkage disequilibrium).

Under the present model of random mating, in the absence of inbreeding, the coefficient of correlation between the additive genetic values of i and j is

$$\mathrm{Corr}(\alpha z_i, \alpha z_j|\alpha) = a_{ij}.$$

1.3 Remarks

The covariance between relatives in multiloci systems is part of a subject that is not easily accessible. An exact general treatment involving only pairs of loci constitutes a formidable challenge leading to unwieldy expressions, as shown by Weir and Cockerham (1977). 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 Weir et al. (1980).

References

Lynch, M. and B. Walsh (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates.

Weir, B. and C. C. Cockerham (1974). Behavior of pairs of loci in finite monoecious populations. *Theoretical Population Biology 6*, 323–354.

Weir, B., C. C. Cockerham, and J. Reynolds (1980). The effects of linkage and linkage disequilibrium on the covariance of noninbred relatives. *Heredity 45*, 351–359.

Weir, B. S. and C. C. Cockerham (1977). Two-locus theory in quantitative genetics. In E. Pollak, O. Kempthorne, and T. B. Bailey (Eds.), *Proceedings of the International Conference on Quantitative Genetics*, pp. 247–269. The Iowa State University Press, Ames, Iowa.