Notes on the paper "Reconciling the analysis of IBD and IBS in complex trait studies" and other digressions

Daniel Sorensen

July 25, 2013

Contents

Below are some notes and reflections concerning the important paper by Powell et al. (2010).

The first point is related to the traditional way of calculating probabilities of IBD at a given locus or gametic relationships (between the two gametes of individual i, that gives rise to its inbreeding coefficient F_i, or between gametes of individuals i and j, that results in the gametic relationship between the gametes of these individuals, F_{ij}) from a known pedigree spanning several generations. The classical model assumes the existence of a base population, of size N say, consisting of $2N$ different alleles at a locus. IBD probabilities are estimated with reference to this base population.

Data on genetic markers can be used to estimate probabilities of IBD and often there is no obvious base population. The authors suggest that the base population can be arbitrary, and one choice could be the current generation involving the typed individuals. In this way, the gene frequencies of the observed markers refer to this cohort consisting of the typed individuals. The authors remind us that probabilities of IBD can be interpreted as correlations of uniting gametes (Wright, 1922). By defining the base population as the population of typed individuals, some values of F may be negative. However this has an interpretation in terms of the correlation coefficient; it means that the particular individual is less homozygous than the average.

1 Inbreeding as a correlation of uniting gametes

Let

$$X_i = \begin{cases} 0 & \text{if } G_i = A_2A_2 \\ 1 & \text{if } G_i = A_1A_2 \\ 2 & \text{if } G_i = A_1A_1 \end{cases}$$

and

$$X_{ij} = \begin{cases} 0 & \text{if } j = 1 \text{ and } G_{i1} = A_1, \text{ with probability } (1-p) \\ 1 & \text{if } j = 2 \text{ and } G_{i2} = A_2, \text{ with probability } p, \end{cases}$$

where G_{ij} is the jth allele $(j = 1, 2)$ of individual i. Then,

$$\begin{aligned} X_i &= X_{i1} + X_{i2}, \\ E(X_i) &= 2p, \\ Var(X_i) &= 2p(1-p)(1+F_i). \end{aligned} \tag{1}$$

For individual i, let

$$Z = \begin{cases} 0 & \text{if } X_{i1} \not\equiv X_{i2}, \text{ with probability } (1-F_i) \\ 1 & \text{if } X_{i1} \equiv X_{i2}, \text{ with probability } F_i. \end{cases}$$

The covariance between uniting gametes in individual i is defined as

$$\begin{aligned} Cov(X_{i1}, X_{i2}) &= E[Cov(X_{i1}, X_{i2})|Z] + Cov[E(X_{i1}|Z), E(X_{i2}|Z)] \\ &= (Cov(X_{i1}, X_{i2})|Z = 0)(1-F_i) + (Cov(X_{i1}, X_{i2})|Z = 1)F_i + 0 \\ &= 0 + Var(X_{ij})F_i = p(1-p)F_i, \end{aligned} \tag{2}$$

and the correlation between uniting gametes in individual i is

$$\frac{Cov\left(X_{i1},X_{i2}\right)}{\sqrt{Var\left(X_{i1}\right)Var\left(X_{i1}\right)}}=\frac{p\left(1-p\right)F_{i}}{p\left(1-p\right)}=F_{i},$$

the inbreeding coefficient of individual i. If instead one computes the correlation between two arbitrary gametes k and l, from individuals i and j,

$$\frac{Cov\left(X_{ik},X_{jl}\right)}{\sqrt{Var\left(X_{ik}\right)Var\left(X_{jl}\right)}}=\frac{p\left(1-p\right)F_{kl}}{p\left(1-p\right)}=F_{kl},$$

the gametic relationship between gametes k and l.

2 Estimating relatedness from SNPs

Using the concept of correlation between uniting gametes, one can estimate the gametic relationship between gametes 1 and 2, using

$$\begin{aligned}\widehat{F}_{12} &= \frac{Cov\left(X_{i1},X_{j2}\right)}{\sqrt{Var\left(X_{i1}\right)Var\left(X_{j2}\right)}}\\ &= \frac{\left(X_{i1}-E\left(X_{i1}\right)\right)\left(X_{j2}-E\left(X_{j2}\right)\right)}{p\left(1-p\right)}\\ &= \frac{\left(X_{i1}-p\right)\left(X_{j2}-p\right)}{p\left(1-p\right)}. \qquad (3)\end{aligned}$$

We note that the term $\left(X_{i1}-E\left(X_{i1}\right)\right)\left(X_{j2}-E\left(X_{j2}\right)\right)$ is strictly not equal to $Cov\left(X_{i1},X_{j2}\right)=E\left[\left(X_{i1}-E\left(X_{i1}\right)\right)\left(X_{j2}-E\left(X_{j2}\right)\right)\right]$. It is best interpreted as a sample covariance with known means. A similar observation holds for several of the expressions below.

2.1 Expected value of $\widehat{F}_{12}$

The expected value of $\widehat{F}_{12}$ (given the gene frequency p) is obtained as follows.

$$\begin{aligned}E\left[\left(X_{i1}-p\right)\left(X_{j2}-p\right)\right] &= E\left(X_{i1},X_{j2}\right)-pE\left(X_{i1}\right)-pE\left(X_{j2}\right)+p^{2}\\ &= E\left(X_{i1},X_{j2}\right)-p^{2}. \qquad (4)\end{aligned}$$

The first term on the right hand side is

$$\begin{aligned}E\left(X_{i1},X_{j2}\right) &= Cov\left(X_{i1},X_{j2}\right)+E\left(X_{i1}\right)E\left(X_{j2}\right)\\ &= Cov\left(X_{i1},X_{j2}\right)+p^{2}.\end{aligned}$$

Therefore, substituting in (4),

$$\begin{aligned}E\left[\left(X_{i1}-p\right)\left(X_{j2}-p\right)\right] &= Cov\left(X_{i1},X_{j2}\right)\\ &= p\left(1-p\right)F_{12}\end{aligned}$$

from (2). Therefore,

$$E\left(\widehat{F}_{12}\right) = \frac{p\left(1-p\right)F_{12}}{p\left(1-p\right)} = F_{12}.$$

2.2 Variance of $\widehat{F}_{12}$

The variance of $\widehat{F}_{12}$ (given gene frequency, and assuming $F_{12} = 0$) is obtained as follows.

$$Var\left(\widehat{F}_{12}\right) = \frac{Var\left[\left(X_{i1}-p\right)\left(X_{j2}-p\right)\right]}{p^2\left(1-p\right)^2}.$$

The numerator is

$$\begin{gathered} Var\left[\left(X_{i1}-p\right)\left(X_{j2}-p\right)\right] = Var\left[\left(X_{i1}X_{j2}\right) - pX_{i1} - pX_{j2} + p^2\right] \\ = Var\left(X_{i1}X_{j2}\right) + 2p^2 Var\left(X_{i1}\right) - 2Cov\left(X_{i1}X_{j2}, pX_{i1}\right) - 2Cov\left(X_{i1}X_{j2}, pX_{i2}\right) \\ +2Cov\left(pX_{i1}, pX_{j2}\right). \end{gathered} \tag{5}$$

The expressions for each term are as follows. Since $F_{12} = 0$, X_{i1} and X_{j2} are independent. Then,

$$\begin{aligned} Var\left(X_{i1}X_{j2}\right) &= E\left(X_{i1}^2\right)E\left(X_{j2}^2\right) - \left[E\left(X_{i1}\right)\right]^2\left[E\left(X_{i2}\right)\right]^2 \\ &= \left[p\left(1-p\right)+p^2\right]^2 - p^4 \\ &= p^2\left(1-p^2\right). \end{aligned}$$

$$2p^2 Var\left(X_{i1}\right) = 2p^3\left(1-p\right).$$

$$\begin{aligned} 2Cov\left(X_{i1}X_{j2}, pX_{i1}\right) &= 2\left[E\left(X_{i1}^2 X_{j2}p\right) - E\left(X_{i1}X_{j2}\right)E\left(X_{i1}p\right)\right] \\ &= 2\left[E\left(X_{i1}^2\right)E\left(X_{j2}p\right) - E\left(X_{i1}\right)E\left(X_{i2}\right)E\left(X_{i1}p\right)\right] \\ &= 2\left[\left(p\left(1-p\right)+p^2\right)p^2\right] - p^4 \\ &= 2p^3\left(1-p\right). \end{aligned}$$

$$2Cov\left(X_{i1}X_{j2}, pX_{i2}\right) = 2p^3\left(1-p\right).$$

$$2Cov\left(pX_{i1}, pX_{j2}\right) = 0.$$

Collecting terms,

$$\begin{aligned} Var\left[\left(X_{i1}-p\right)\left(X_{j2}-p\right)\right] &= p^2\left(1-p^2\right) + 2p^3\left(1-p\right) - 4p^3\left(1-p\right) \\ &= p^2\left(1-p^2\right) - 2p^3\left(1-p\right) \\ &= \left[p\left(1-p\right)\right]^2. \end{aligned} \tag{6}$$

Therefore,

$$Var\left(\widehat{F}_{12}\right) = \frac{Var\left[\left(X_{i1}-p\right)\left(X_{j2}-p\right)\right]}{p^2\left(1-p\right)^2} = 1.$$

3 Estimating relationships between diploid individuals

Expression (3) can readily be generalised to compute the coefficient of relationship between diploid individuals. At a given locus, for individual i, the random variable X_i defines the number of copies of the A_1 allele (known as gene content in the genetics literature). It can also be interpreted as a label for the genotype. The coefficient of relationship between individuals i and j is defined as

$$\begin{aligned} \widehat{A}_{ij} &= \widehat{F}_{ij} = \frac{Cov\left(X_i, X_j\right)}{\sqrt{Var\left(X_i\right) Var\left(X_j\right)}} \\ &= \frac{\left(X_i - 2p\right)\left(X_j - 2p\right)}{2p\left(1-p\right)}. \end{aligned} \tag{7}$$

the coefficient of relationship takes the form of a "sampling" correlation with known mean (gene frequencies) between the genotypes (represented by the random variable X, that can take values $0, 1$ or 2). At this point we note that

$$\begin{aligned} Var\left(X_i\right) &= Var\left(X_{i1} + X_{i2}\right) \\ &= Var\left(X_{i1}\right) + Var\left(X_{i2}\right) + 2Cov\left(X_{i1}, X_{i2}\right) \\ &= p\left(1-p\right) + p\left(1-p\right) + 2p\left(1-p\right)F_i \\ &= 2p\left(1-p\right)\left(1+F_i\right), \end{aligned}$$

so the term $\sqrt{\left(1+F_i\right)\left(1+F_j\right)}$ is ignored in the denominator of (7).

3.1 Expected value of $\widehat{A}_{ij}$

Given p the expected value is

$$\begin{aligned} E\left[\left(X_i - 2p\right)\left(X_j - 2p\right)\right] &= E\left(X_i, X_j\right) - 2pE\left(X_i\right) - 2pE\left(X_j\right) + 4p^2 \\ &= E\left(X_i, X_j\right) - 4p^2. \end{aligned}$$

The expectation of the first term is

$$\begin{aligned} E\left(X_i, X_j\right) &= Cov\left(X_i, X_j\right) + E\left(X_i\right)E\left(X_j\right) \\ &= Cov\left(X_i, X_j\right) + 4p^2. \end{aligned}$$

Substituting above gives

$$E\left[\left(X_i - 2p\right)\left(X_j - 2p\right)\right] = Cov\left(X_i, X_j\right),$$

as it should. The covariance can be written as

$$\begin{aligned} Cov\left(X_i, X_j\right) &= Cov\left(X_{i1} + X_{i2}, X_{j1} + X_{j2}\right) \\ &= 4Cov\left(X_{i1}, X_{j1}\right). \end{aligned}$$

The covariance term takes the form

$$\begin{aligned} Cov\left(X_{i1}, X_{j1}\right) &= E\left[Cov\left(X_{i1}, X_{j1}\right)|Z\right] + Cov\left[E\left(X_{i1}|Z\right), E\left(X_{j1}|Z\right)\right] \\ &= \left(Cov\left(X_{i1}, X_{j1}\right)|Z=0\right)\Pr\left(X_{i1} \not\equiv X_{j1}\right) + \left(Cov\left(X_{i1}, X_{j1}\right)|Z=1\right)\Pr\left(X_{i1} \equiv X_{j1}\right) + 0 \\ &= Var\left(X_{i1}\right)\Pr\left(X_{i1} \equiv X_{j1}\right) \\ &= p\left(1-p\right)\Pr\left(X_{i1} \equiv X_{j1}\right). \end{aligned}$$

The coefficient of coancestry between individuals i and j is defined as

$$\varphi_{ij} = \frac{1}{4}\left[\Pr\left(X_{i1} \equiv X_{j1}\right) + \Pr\left(X_{i1} \equiv X_{j2}\right) + \Pr\left(X_{i2} \equiv X_{j1}\right) + \Pr\left(X_{i2} \equiv X_{j2}\right)\right].$$

Then,

$$\begin{aligned} E\left[\left(X_i - 2p\right)\left(X_j - 2p\right)\right] &= Cov\left(X_i, X_j\right) \\ &= 4Cov\left(X_{i1}, X_{j1}\right) \\ &= 4p\left(1-p\right)\frac{1}{4}\varphi_{ij} \\ &= 2p\left(1-p\right)F_{ij}, \end{aligned} \tag{8}$$

where $F_{ij} = A_{ij} = 2\varphi_{ij}$ is the correlation between additive genetic values of related individuals, or additive genetic relationship. Therefore,

$$E\left(\widehat{A}_{ij}\right) = \frac{2p\left(1-p\right)F_{ij}}{2p\left(1-p\right)} = A_{ij} = F_{ij}, \tag{9}$$

provided the term $\sqrt{\left(1+F_i\right)\left(1+F_j\right)}$ is ignored in the denominator of (7).

An important observation is that the coefficient of relationship obtained using marker information provides an "actual or true" relationship. This relationship differs from the pedigree based relationship (even when the number of markers tends to "infinity"), which is based on expected relationships. The expected relationship between full-sibs is 0.5. This expectation is taken over replicated full-sib individuals. But given two full-sibs, their actual relationship will vary around the expected value of 0.5, and the markers disclose this deviation.

3.2 The special case when $i = j$

When $i = j$ (the same individual),

$$\begin{aligned} E\left(\widehat{A}_i\right) &= \frac{E\left(X_i - 2p\right)^2}{2p\left(1-p\right)} \\ &= \frac{E\left(X_i^2\right) - 4p^2}{2p\left(1-p\right)} \\ &= \frac{Var\left(X_i\right) + \left[E\left(X_i\right)\right]^2 - 4p^2}{2p\left(1-p\right)} \\ &= \frac{Var\left(X_i\right)}{2p\left(1-p\right)} \\ &= \frac{Var\left(X_{i1} + X_{i2}\right)}{2p\left(1-p\right)} \\ &= \frac{2p\left(1-p\right)\left(1+F_i\right)}{2p\left(1-p\right)} \\ &= \left(1+F_i\right). \end{aligned}$$

3.3 The variance of $\widehat{A}_{ij}$

Assume $A_{ij} = 0$ and $F_i = F_j = 0$. To compute this variance, first notice that

$$\begin{aligned} \left(X_i - 2p\right)\left(X_j - 2p\right) = \left[\left(X_{i1} - p\right) + \left(X_{i2} - p\right)\right]\left[\left(X_{j1} - p\right) + \left(X_{j2} - p\right)\right] \\ = \left(X_{i1} - p\right)\left(X_{j1} - p\right) + \left(X_{i1} - p\right)\left(X_{j2} - p\right) + \left(X_{i2} - p\right)\left(X_{j1} - p\right) + \left(X_{i2} - p\right)\left(X_{j2} - p\right). \end{aligned}$$

Using (6), each of the 4 terms has variance $p^2\left(1-p\right)^2$. All covariance terms are of the form

$$\begin{aligned} Cov\left[\left(X_{i1} - p\right)\left(X_{j1} - p\right), \left(X_{i1} - p\right)\left(X_{j2} - p\right)\right] = E\left[\left(X_{i1} - p\right)^2\left(X_{j1} - p\right)\left(X_{j2} - p\right)\right] \\ -E\left[\left(X_{i1} - p\right)\left(X_{j1} - p\right)\right]E\left[\left(X_{i1} - p\right)\left(X_{j2} - p\right)\right] \\ = E\left(X_{i1} - p\right)^2 E\left(X_{j1} - p\right)E\left(X_{j2} - p\right) - \left[E\left(X_{i1} - p\right)\right]^2 E\left(X_{j1} - p\right)E\left(X_{j2} - p\right) \\ = E\left(X_{j1} - p\right)E\left(X_{j2} - p\right)Var\left(X_{i1}\right) = 0. \end{aligned}$$

Therefore,

$$Var\left(\widehat{A}_{ij}\right) = \frac{4p^2\left(1-p\right)^2}{4p^2\left(1-p\right)^2} = 1.$$

4 A genomic model

Consider the model for the record of individual j

$$y_j = \mu + W_1 b_1 + W_2 b_2 + e_j, \tag{10}$$

where the SNP effects are

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I\sigma_b^2\right],$$

and

$$W_i = \frac{X_i - 2p_i}{\sqrt{2p_i\left(1-p_i\right)}}, \quad i = 1, 2 \tag{11}$$

which results in $E\left(W_i\right) = 0$ and $Var\left(W_i\right) = 1$. Let g_M denote the genomic value (captured by the SNPs), equal to

$$\begin{aligned} g_M &= W_1 b_1 + W_2 b_2 \\ &= Wb. \end{aligned}$$

Then, conditionally on W,

$$\begin{aligned} Var\left(g_M|W_1, W_2\right) &= WVar\left(b\right)W' \\ &= WW'\sigma_b^2 \\ &= W_1^2\sigma_b^2 + W_2^2\sigma_b^2. \end{aligned} \tag{12}$$

and the unconditional variance is

$$Var\left(g_M\right) = \sigma_{g_M}^2 = 2\sigma_b^2 \tag{13}$$

known as the genomic variance, equal to the variance per SNP times the number of SNPs. Notice that (12) does not include a contribution due to disequilibrium between marker loci 1 and 2.

On the other hand, if (10) had been parameterised as

$$\begin{aligned} y_j &= \mu + X_1 b_1 + X_2 b_2 + e_j \\ &= \mu + b_1 X_1 + b_2 X_2 + e_j \\ &= \mu + b'X + e_j \\ &= \mu + g_M + e_j, \end{aligned}$$

and one conditions on the b's instead,

$$\begin{aligned} Var\left(g_M|b_1, b_2\right) &= Var\left(bX|b\right) \\ &= bVar\left(X\right)b' \\ &= b'\begin{pmatrix} Var\left(X_1\right) & Cov\left(X_1X_2\right) \\ Cov\left(X_1X_2\right) & Var\left(X_2\right) \end{pmatrix} b \\ &= \begin{pmatrix} b_1 & b_2 \end{pmatrix}\begin{pmatrix} 2p_1\left(1-p_1\right) & 2D_{12} \\ 2D_{12} & p_2\left(1-p_2\right) \end{pmatrix}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\ &= 2b_1^2 p_1\left(1-p_1\right) + 2b_2^2 p_2\left(1-p_2\right) + 4b_1 b_2 D_{12}, \end{aligned}$$

where $2D_{12} = Cov\left(X_1, X_2\right)$. Here the $X's$ are treated as random and the inference is conditional on the b's, and this incorporates the contribution due to disequilibrium. Unconditionally with respect to the b's, the variance is of course equal to (13), because the b's are a priori uncorrelated.

4.1 The covariance between two individuals

Consider the records of two individuals

$$\begin{aligned} y_i &= \mu + W_{1i}b_1 + W_{2i}b_2 + e_i, \\ y_j &= \mu + W_{1j}b_1 + W_{2j}b_2 + e_j. \end{aligned}$$

The covariance between their genomic breeding values is

$$\begin{aligned} Cov\left(g_{Mi}, g_{Mj}|W_1, W_2\right) &= Cov\left(W_{1i}b_1 + W_{2i}b_2, W_{1j}b_1 + W_{2j}b_2\right) \\ &= W_{1i}W_{1j}\sigma_b^2 + W_{2i}W_{2j}\sigma_b^2 \end{aligned}$$

and the other terms vanish because $Cov\left(b_1, b_2\right) = 0$. The unconditional covariance is

$$\begin{aligned} Cov\left(g_{Mi}, g_{Mj}\right) &= Cov\left[E\left(g_{Mi}|W_1, W_2\right), E\left(g_{Mj}|W_1, W_2\right)\right] + E\left[Cov\left(g_{Mi}, g_{Mj}|W_1, W_2\right)\right] \\ &= E\left[Cov\left(g_{Mi}, g_{Mj}|W_1, W_2\right)\right] \\ &= \sigma_b^2 E\left[W_{1i}W_{1j} + W_{2i}W_{2j}\right] \\ &= \sigma_b^2\left(A_{ij,1} + A_{ij,2}\right) \\ &= 2\sigma_b^2 A_{ij} \\ &= \sigma_{g_M}^2 A_{ij}, \end{aligned}$$

where $A_{ij} = \frac{1}{2}\left(A_{ij,1} + A_{ij,2}\right)$, the average (over markers) of the additive genetic relationship between individuals i and j.

In general, if the model in matrix notation is

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\mu + \mathbf{W}\mathbf{b} + \mathbf{e} \\ &= \mathbf{1}\mu + \mathbf{g_M} + \mathbf{e}, \end{aligned}$$

with

$$\mathbf{b} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_b^2\right)$$

then

$$\begin{aligned} Var\left(\mathbf{g_M}\right) &= \mathbf{W}\mathbf{W}'\sigma_b^2 \\ &= \frac{1}{m_G}\mathbf{W}\mathbf{W}'\sigma_{g_M}^2. \end{aligned}$$

The term $\frac{1}{m_G}\mathbf{W}\mathbf{W}'$ is the average (over SNPs) *realised* additive genetic relationship among individuals and

$$\sigma_{g_M}^2 = m_G\sigma_b^2 \tag{14}$$

is the sum of the variances of all (m_G) SNPs.

It is instructive to make quite explicit the structure of $\mathbf{W}\mathbf{W}'$. For two individuals and two markers it takes the form

$$\begin{aligned} \mathbf{W}\mathbf{W}' &= \begin{bmatrix} W_{1i} & W_{2i} \\ W_{1j} & W_{2j} \end{bmatrix}\begin{bmatrix} W_{1i} & W_{1j} \\ W_{2i} & W_{2j} \end{bmatrix} \\ &= \begin{bmatrix} W_{1i}W_{1i} + W_{2i}W_{2i} & W_{1i}W_{1j} + W_{2i}W_{2j} \\ W_{1j}W_{1i} + W_{2j}W_{2i} & W_{1j}W_{1j} + W_{2j}W_{2j} \end{bmatrix}. \end{aligned}$$

This shows that disequilibrium among loci cannot be accounted for when the inference is conditional on W.

On the other hand, if the model is parameterised as

$$\begin{aligned} y_i &= \mu + b_1 X_{1i} + b_2 X_{2i} + e_i \\ y_j &= \mu + b_1 X_{1j} + b_2 X_{2j} + e_j. \end{aligned}$$

In matrix notation,

$$y = 1\mu + bX + e,$$

where

$$b = \begin{bmatrix} b_1 & b_2 & 0 & 0 \\ 0 & 0 & b_1 & b_2 \end{bmatrix}$$

and

$$X' = [X_{1i}, X_{2i}, X_{1j}, X_{2j}].$$

and the $X's$ are treated as random with variance-covariance structure

$$\begin{bmatrix} Var(X_{1i}) & Cov(X_{1i}, X_{2i}) & Cov(X_{1i}, X_{1j}) & Cov(X_{1i}, X_{2j}) \\ & Var(X_{2i}) & Cov(X_{2i}, X_{1j}) & Cov(X_{2i}, X_{2j}) \\ & & Var(X_{1j}) & Cov(X_{1j}, X_{2j}) \\ & & & Var(X_{2j}) \end{bmatrix},$$

then the 2×2 off-diagonal block defines covariances between the two individuals. These covariances include terms involving the same locus, like $Cov(X_{1i}, X_{1j})$ and $Cov(X_{2i}, X_{2j})$, and terms involving different loci, like $Cov(X_{1i}, X_{2j})$ and $Cov(X_{2i}, X_{1j})$. The within individual terms include inbreeding contributions in the diagonals, and disequlibrium contributions in the off-diagonals.

The covariance between i and j takes the form

$$\begin{aligned} & Cov(X_{1i}b_1 + X_{2i}b_2, X_{1j}b_1 + X_{2j}b_2 | b_1, b_2) \\ = \; & b_1^2 Cov(X_{1i}, X_{1j}) + b_2^2 Cov(X_{2i}, X_{2j}) + b_1 b_2 Cov(X_{1i}, X_{2j}) + b_2 b_1 Cov(X_{2i}, X_{1j}). \end{aligned}$$

From (8), the within locus covariances between individuals i and j are

$$Cov(X_{ki}, X_{kj}) = 2p_k(1 - p_k) A_{ij,k}, \quad k = 1, 2,$$

where $A_{ij,k}$ is the additive genetic relationship between individuals i and j at locus k computed using the SNP markers. The between loci covariances between individuals i and j involve each 4 terms, each one of the form

$$D_{1i,2j} = \Pr(X_{11i} = 1, X_{21j} = 1) - \Pr(X_{11i} = 1) \Pr(X_{21j} = 1).$$

The conditional covariance between the genomic breeding values of individuals i and j is then

$$\begin{aligned} Cov(X_{1i}b_1 + X_{2i}b_2, X_{1j}b_1 + X_{2j}b_2 | b_1, b_2) &= 2b_1^2 p_1(1 - p_1) A_{ij,1} + 2b_2^2 p_2(1 - p_2) A_{ij,2} \\ &\quad + 4b_1 b_2 D_{1i,2j} + 4b_2 b_1 D_{1j,2i}. \end{aligned}$$

5 Marker effects are functions of disequilibria

5.1 Example 1: Marker effects fixed and genotypes random

Consider the task of predicting an unobserved genotypic value for a quantitative trait, $g = X_a a$, based on observed genotypic information from two markers X_1, X_2. Here, X_a is a label for the unobserved genotype at the causal locus, X_1 and X_2 are labels for the observed marker genotypes, a is the effect of the causal locus. Assume that the variance-covariance structure of $[X_a, X_1, X_2]$ is

$$\begin{bmatrix} \sigma_{aa} & \sigma_{a1} & \sigma_{a2} \\ \sigma_{a1} & \sigma_{11} & \sigma_{12} \\ \sigma_{a2} & \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

Then, given a, the best linear predictor of $X_a a$ takes the form

$$\widehat{g} = g(X_1, X_2) = aE(X_a) + b_1 (X_1 - E(X_1)) + b_2 (X_2 - E(X_2)), \tag{15}$$

where, labeling $\widehat{g}_i = b_i (X_i - E(X_i))$,

$$b_1 = a\frac{\sqrt{\sigma_{aa}}}{\sqrt{\sigma_{11}}}\left(\frac{r_{g,\widehat{g}_1} - r_{g,\widehat{g}_2} r_{\widehat{g}_1,\widehat{g}_2}}{1 - r^2_{\widehat{g}_1,\widehat{g}_2}}\right), \tag{16a}$$

$$b_2 = a\frac{\sqrt{\sigma_{aa}}}{\sqrt{\sigma_{22}}}\left(\frac{r_{g,\widehat{g}_2} - r_{g,\widehat{g}_1} r_{\widehat{g}_1,\widehat{g}_2}}{1 - r^2_{\widehat{g}_1,\widehat{g}_2}}\right) \tag{16b}$$

are the marker effects. In these expressions

$$\sigma_{ii} = 2p_i (1 - p_i), \quad i = a, 1, 2$$

and the correlation between the genotypic value at the causal locus, g, and $g_{M_i} = X_i b_i$, the genomic value of marker i take the form

$$\begin{aligned} r_{g,g_{M_i}} &= \frac{Cov(X_a a, X_i b_i)}{\sqrt{Var(X_a a)\, Var(X_i b_i)}} = \frac{Cov(X_a, X_i)}{\sqrt{Var(X_a)\, Var(X_i)}} \\ &= \frac{\sigma_{ai}}{\sqrt{\sigma_{aa}\sigma_{ii}}} \\ &= \frac{D_{ai}}{\sqrt{p_a(1-p_a)\, p_i(1-p_i)}} \quad i = 1, 2 \end{aligned} \tag{17}$$

which is a correlation of gene frequencies involving the causal locus and the marker. The term $Cov(X_a, X_i) = D_{ai}$ is the linkage disequilibrium (covariance of gene frequencies) involving the causal locus and the marker. Similarly, for the two marker loci,

$$\begin{aligned} r_{12} &= \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)\, Var(X_2)}} \\ &= \frac{D_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}. \end{aligned}$$

In the model, the predictor of the genetic value g is $\widehat{g}$. The correlation between these two quantities is

$$\begin{aligned}\rho_{g,\widehat{g}} &= \frac{Cov\left(g,\widehat{g}\right)}{\sqrt{Var\left(g\right)Var\left(\widehat{g}\right)}} \\ &= \frac{Cov\left(X_a a, X_1 b_1 + X_2 b_2\right)}{\sqrt{Var\left(X_a a\right)Var\left(X_1 b_1 + X_2 b_2\right)}} \\ &= \frac{ab_1 Cov\left(X_a X_1\right) + ab_2 Cov\left(X_a X_2\right)}{\sqrt{a^2 Var\left(X_a\right)\left[b_1^2 Var\left(X_1\right) + b_2^2 Var\left(X_2\right) + 2b_1 b_2 Cov\left(X_1, X_2\right)\right]}}. \end{aligned} \tag{18}$$

In this approach, the $X's$ are treated as random and the partial regression coefficients b_1 and b_2 are parameters, assumed known in the context of best linear prediction. The point is that marker effects b_1 and b_2 in expressions (16) are functions of linkage disequilibria involving the causal locus and the markers. Also, unconditionally with respect to a, marker effects are correlated.

5.2 Example 2: Marker effects random and genotypes fixed

Assume that the model for the datum y (with mean 0) is

$$\mathbf{y} = W_a a + W_b b + e \tag{19}$$

where the $W's$, defined in (11), are labels for the causal locus, whose effect is a, and for the marker locus, whose effect is b. Assume that

$$E\left(a, b, e\right) = \left(0, 0, 0\right)$$

and

$$Var\left(a, b, e\right) = diag\left(\sigma_a^2, \sigma_b^2, \sigma_e^2\right).$$

Write (19) as

$$y = Wu + e \tag{20}$$

where

$$W = \left(W_a, W_b\right), \quad u' = \left(a, b\right).$$

Then the posterior mean of u is the solution to

$$\left(W'W + D^{-1}\right)\widehat{u} = Z'y, \tag{21}$$

where

$$D^{-1} = \begin{bmatrix} \frac{\sigma_e^2}{\sigma_a^2} & 0 \\ 0 & \frac{\sigma_e^2}{\sigma_b^2} \end{bmatrix}.$$

The solution is

$$\begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix} = \frac{1}{\frac{\sigma_e^2}{\sigma_a^2}\frac{\sigma_e^2}{\sigma_b^2}W_a W_a W_b W_b - \left(W_a W_b\right)^2} \begin{bmatrix} W_b W_b \frac{\sigma_e^2}{\sigma_b^2} & -W_a W_b \\ -W_a W_b & W_a W_a \frac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} W_a \\ W_b \end{bmatrix} y. \tag{22}$$

The posterior mean of the marker effect is

$$\widehat{b} = \frac{W_a W_b \left(\frac{\sigma_e^2}{\sigma_a^2} - 1\right) W_b}{\frac{\sigma_e^2}{\sigma_a^2}\frac{\sigma_e^2}{\sigma_b^2} W_a W_a W_b W_b - \left(W_a W_b\right)^2} y$$

which indicates that it is a function of the disequilibrium between the marker and the causal locus ($W_a W_b$).

6 Computing $\rho_{Y\widehat{Y}}$ with limited information

Here we address how imperfect knowledge of marker effects and ignorance about the true genetic model affect $\rho_{Y\widehat{Y}}$ in a cross-validation scenario.

We first review briefly some basic results of best linear prediction. Consider two random variables Y and X of finite non-zero variance. A linear function $k + \beta X$ predicts Y with mean squared error $E\left(y - k + \beta X\right)^2$. This is minimised with $k = E\left(Y\right) - \beta E\left(X\right)$ and $\beta = Cov\left(Y, X\right) / Var\left(X\right)$. The best linear predictor of Y is

$$\widehat{Y} = E\left(Y\right) + \beta\left(X - E\left(X\right)\right) \tag{23}$$

with

$$\begin{aligned} E\left(\widehat{Y}\right) &= E_X\left[E\left(\widehat{Y}|X\right)\right] \\ &= E\left(Y\right). \end{aligned}$$

One may write

$$Y = \widehat{Y} + \left(Y - \widehat{Y}\right) = \widehat{Y} + e \tag{24}$$

where e, the prediction error, is uncorrelated with $\widehat{Y}$. Then it is easily verified that

$$\begin{aligned} E\left(e\right) &= E\left(Y - \widehat{Y}\right) = E\left(Y\right) - E\left(\widehat{Y}\right) = 0, \\ Var\left(\widehat{Y}\right) &= \frac{\left[Cov\left(Y, X\right)\right]^2}{Var\left(X\right)} = \beta^2 Var\left(X\right) = \rho_{YX}^2 Var\left(Y\right), \\ Cov\left(Y, \widehat{Y}\right) &= Var\left(\widehat{Y}\right), \text{ and therefore, } \beta_{Y,\widehat{Y}} = 1, \\ Var\left(e\right) &= Var\left(Y\right) - \frac{\left[Cov\left(Y, X\right)\right]^2}{Var\left(X\right)} = \left(1 - \rho_{YX}^2\right) Var\left(Y\right), \\ Cov\left(\widehat{Y}, e\right) &= 0, \\ \rho_{YX} &= \rho_{Y\widehat{Y}}, \\ \rho_{Y\widehat{Y}} &= \sqrt{\frac{Var\left(\widehat{Y}\right)}{Var\left(Y\right)}}. \end{aligned}$$

6.1 Predicting unobserved genetic value using observed marker information

In this section it is assumed that marker effects (labelled as b) are known parameters. Consider first the case of predicting $X_a a = g$ with one marker W_1 genotype only. Here X_a is a label for the genotype at the unobserved causal locus, a the substitution effect at the causal locus, and W_1 is a label for the genotype at the observed marker. Then, given a, the best linear predictor of g is

$$\widehat{g} = aE(X_a) + b_1(W_1 - E(W_1)).$$

The decomposition $g = \widehat{g} + e$ results in

$$Var(g) = \rho^2_{g\widehat{g}} Var(g) + \left(1 - \rho^2_{g\widehat{g}}\right) Var(g).$$

The proportion of variance of the causal locus explained by the marker X_1, given a, is

$$\rho^2_{g\widehat{g}} Var(g) = 2a^2 p_a (1 - p_a) r^2_{a1} \tag{25}$$

where $Var(X_a a|a) = 2a^2 p_a (1 - p_a)$ and

$$\begin{aligned} \rho_{g\widehat{g}} &= r_{a1} = \frac{Cov(X_a a, b_1 W_1)}{\sqrt{Var(X_a a|a)\, Var(b_1 W_1|b_1)}} \\ &= \frac{D_{a1}}{\sqrt{p_a(1-p_a)\, p_1(1-p_1)}}. \end{aligned}$$

where $Var(b_1 W_1|b_1) = 2b_1^2 p_1(1 - p_1)$. Expression (25) shows that if the causal locus is rare, the proportion of variance explained by the marker is very small. As an example, if the frequency of the causal locus is 0.1, that of the SNP is 0.4, and assuming that $D_{a1} = -0.1 \times 0.4 = -0.04$, then $r^2_{a1} = 0.074$, and the variance explained by the marker is $0.013a^2$. Even when $r^2_{a1} = 1$, if the frequency of the causal locus is 0.1, the variance explained by the marker is $0.18a^2$.

With many markers, the decomposition (24) still holds, but the expression for the (multiple) correlation between g and $\widehat{g}$ is less transparent. However the result $\rho_{g\widehat{g}} = [Var(\widehat{g})/Var(g)]^{1/2}$ holds for any number of markers.

We consider now a model discussed in Goddard (2009), that assumes that each marker genotype is correlated with one causal genotype, but that pairs of markers and causal genotypes, are independent. The model assumes that there is an equal number of markers and causal genotypes. For simplicity we consider here 2 pairs, where each pair consists of a marker genotype and a causal genotype. The causal genotypic value of an individual is

$$g = X_1 a_1 + X_2 a_2$$

where X_i is a label (random variable) for the genotype at the ith causal locus, and a_i is the effect of the ith causal genotype. The conditional variance is

$$Var\left(g|a_1,a_2\right)=a_1^2\sigma_{X_1}^2+a_2^2\sigma_{X_2}^2$$

where $Var\left(X_i\right)=\sigma_{X_i}^2$. The best linear predictor of g is

$$\widehat{g}=E\left(g\right)+b_1\left(W_1-E\left(W_1\right)\right)+b_2\left(W_2-E\left(W_2\right)\right) \tag{26}$$

where W_i is a label (random variable) for the genotype at the ith marker locus, and b_i is the effect of the ith marker genotype. The derivation of (26) uses standard best linear prediction theory and is sketched at the end of this subsection. In (26),

$$\begin{aligned} b_i &= \frac{Cov\left(g,W_i|a_1,a_2\right)}{Var\left(W_i\right)} \\ &= a_i\frac{\sigma_{ii}}{\sigma_{W_i}^2}, \quad i=1,2, \end{aligned} \tag{27}$$

where $\sigma_{ii}=Cov\left(X_i,W_i\right)$, and $\sigma_{W_i}^2=Var\left(W_i\right)$, $i=1,2$. Then

$$\begin{aligned} Var\left(\widehat{g}\right) &= a_1^2\frac{\left(\sigma_{11}\right)^2}{\sigma_{W_1}^2}+a_2^2\frac{\left(\sigma_{22}\right)^2}{\sigma_{W_2}^2} \\ &= a_1^2\sigma_{X_1}^2r_{11}^2+a_2^2\sigma_{X_2}^2r_{22}^2 \end{aligned} \tag{28}$$

where r_{ii}^2 is the (squared) coefficient of linkage disequilibrium between the ith pair of causal and marker genotypes. The covariance is

$$Cov\left(g,\widehat{g}\right)=Var\left(\widehat{g}\right),$$

and the squared correlation

$$\begin{aligned} \rho_{g,\widehat{g}}^2 &= \frac{Var\left(\widehat{g}\right)}{Var\left(g\right)} \\ &= \frac{a_1^2\sigma_{X_1}^2r_{11}^2+a_2^2\sigma_{X_2}^2r_{22}^2}{a_1^2\sigma_{X_1}^2+a_2^2\sigma_{X_2}^2} \end{aligned} \tag{29}$$

The denominator is the total (causal) genetic variance of the trait. Therefore $\rho_{g,\widehat{g}}^2=1$ when the marker genotype is the causal genotype. (Compare the above with (18), derived under a different genetic model). Similarly,

$$\begin{aligned} \rho_{y,\widehat{y}}^2 &= \frac{Var\left(\widehat{y}\right)}{Var\left(y\right)} \\ &= \frac{a_1^2\sigma_{X_1}^2r_{11}^2+a_2^2\sigma_{X_2}^2r_{22}^2}{a_1^2\sigma_{X_1}^2+a_2^2\sigma_{X_2}^2+\sigma_e^2}\frac{Var\left(g\right)}{Var\left(g\right)} \\ &= \rho_{g,\widehat{g}}^2h^2 \\ &= h_G^2, \end{aligned} \tag{30}$$

the genomic heritability.

NOTE

The derivation of (26) is as follows.

6.2 The genomic variance

Typically one computes genomic values for each individual. Given the above model, with m independent pairs of marker-causal locus genotypes, the genomic values are defined as

$$\begin{aligned} g &= \sum_{i=1}^{m} b_i W_i \\ &= \sum_{i=1}^{m} a_i \frac{\sigma_{ii}}{\sigma^2_{W_i}} W_i \end{aligned}$$

where $a_i r_{ii} \frac{\sigma_{X_i}}{\sigma_{W_i}} = b_i$, the regression of the causal locus genotypic value on marker genotype, given the substitution effect of the causal locus a_i. Then the genomic variance is

$$\begin{aligned} Var\left(g|b\right) &= \sum_{i=1}^{m} b_i^2 Var\left(W_i\right) \\ &= \sum_{i=1}^{m} a_i^2 \sigma^2_{X_i} r^2_{ii}. \end{aligned} \tag{31}$$

When the markers are causal, $r^2_{ii} = 1$ and $Var\left(g|b\right) = \sum_{i=1}^{m} a_i^2 \sigma^2_{X_i}$ is the additive genetic variance of the trait, assuming linkage equilibrium. This expression provides one explanation for the so-called *missing heritability*.

6.3 Cross-validation

The following problem is considered below. Assume data $Y_i^t = bX_i + e_i, \quad i = 1, 2, \ldots, N$ from which an estimate $\widehat{b}$ of b is obtained. These are referred to as the *training data* Y_i^t. The genomic value is now defined as

$$g_M = bX_i.$$

An independent (given X) data (the *test data*) labelled $Y_i, \quad i = 1, 2, \ldots$, are to be predicted using $\widehat{Y}_i = \widehat{g}_M = \widehat{b}X_i$ and the first objective is to obtain an expression for the (marginal, with respect to $\widehat{b}$ and X) correlation between g_M and $\widehat{Y}_i$, $\rho_{g_M,\widehat{Y}}$. Later we consider the correlation between Y_i and $\widehat{Y}_i$, $\rho_{Y,\widehat{Y}}$ and show that (Goddard, 2009)

$$\rho_{Y,\widehat{Y}} = \rho_{Y,g_M}\, \rho_{g_M,\widehat{Y}} \tag{32}$$

provided that Y is conditionally independent of $\widehat{Y}$, given g_M. The correlation between a datum from the test data and its genomic value is of the form

$$\rho_{Y,g_M} = \frac{Cov\left(Y, g_M\right)}{\sqrt{Var\left(Y\right) Var\left(g_M\right)}}$$

and

$$Cov\left(Y, g_M\right) = Cov\left(g, g_M\right) \tag{33}$$

where g is the true, unobserved genetic value. Here we are expressing the observed phenotype in the validating data as

$$Y_i = g_i + \epsilon_i$$

where g_i is the unobserved genetic vlaue of the causal genotype of individual i. As shown above in (29), $Cov\left(g, g_M\right)$ is a function of the degree of linkage disequilibrium between the markers and unobserved causal factors and of the effect of the causal locus on the trait

Two modelling scenarios are considered. In the first one, SNP effects are treated as unknown, fixed parameters, and are estimated using least squares. In the second scenario, SNP effects are random variables and are predicted using BLUP. Before doing this we show the derivation of (32).

6.4 Factorising the correlation between Y and $\widehat{Y}$ *a la* Goddard

Assume that $\left(Y, \widehat{Y}, g_M\right)$ are multivariate normally distributed with covariance structure

$$Var\left(Y, \widehat{Y}, g_M\right) = \begin{bmatrix} \sigma^2_Y & \sigma_{Y,\widehat{Y}} & \sigma_{Y,g_M} \\ \sigma_{Y,\widehat{Y}} & \sigma^2_{\widehat{Y}} & \sigma_{\widehat{Y},g_M} \\ \sigma_{Y,g_M} & \sigma_{\widehat{Y},g_M} & \sigma^2_{g_M} \end{bmatrix}.$$

Then

$$\begin{aligned} Var\left(Y, \widehat{Y}|g_M\right) &= \begin{bmatrix} \sigma^2_Y & \sigma_{Y,\widehat{Y}} \\ \sigma_{Y,\widehat{Y}} & \sigma^2_{\widehat{Y}} \end{bmatrix} - \begin{bmatrix} \sigma_{Y,g_M} \\ \sigma_{\widehat{Y},g_M} \end{bmatrix} \begin{bmatrix} \sigma_{Y,g_M} & \sigma_{\widehat{Y},g_M} \end{bmatrix} \frac{1}{\sigma^2_{g_M}} \\ &= \begin{bmatrix} \sigma^2_Y & \sigma_{Y,\widehat{Y}} \\ \sigma_{Y,\widehat{Y}} & \sigma^2_{\widehat{Y}} \end{bmatrix} - \frac{1}{\sigma^2_{g_M}} \begin{bmatrix} \left(\sigma_{Y,g_M}\right)^2 & \sigma_{Y,g_M}\sigma_{\widehat{Y},g_M} \\ \sigma_{Y,g_M}\sigma_{\widehat{Y},g_M} & \left(\sigma_{\widehat{Y},g_M}\right)^2 \end{bmatrix}. \end{aligned}$$

The off-diagonal term is

$$Cov\left(Y, \widehat{Y}|g_M\right) = \sigma_{Y,\widehat{Y}} - \frac{1}{\sigma^2_{g_M}}\sigma_{Y,g_M}\sigma_{\widehat{Y},g_M}.$$

The diagonal terms are

$$\begin{aligned} Var\left(Y|g_M\right) &= \sigma^2_Y - \frac{\left(\sigma_{Y,g_M}\right)^2}{\sigma^2_{g_M}} = \sigma^2_Y\left(1 - \rho^2_{Y,g_M}\right), \\ Var\left(\widehat{Y}|g_M\right) &= \sigma^2_{\widehat{Y}}\left(1 - \rho^2_{\widehat{Y},g_M}\right). \end{aligned}$$

The conditional correlation is then

$$\begin{aligned} \rho_{Y,\widehat{Y}|g_M} &= \frac{\sigma_{Y,\widehat{Y}}}{\sigma_Y\sigma_{\widehat{Y}}\sqrt{\left(1 - \rho^2_{Y,g_M}\right)\left(1 - \rho^2_{\widehat{Y},g_M}\right)}} - \frac{\frac{1}{\sigma^2_{g_M}}\sigma_{Y,g_M}\sigma_{\widehat{Y},g_M}}{\sigma_Y\sigma_{\widehat{Y}}\sqrt{\left(1 - \rho^2_{Y,g_M}\right)\left(1 - \rho^2_{\widehat{Y},g_M}\right)}} \\ &= \frac{\rho_{Y,\widehat{Y}}}{\sqrt{\left(1 - \rho^2_{Y,g_M}\right)\left(1 - \rho^2_{\widehat{Y},g_M}\right)}} - \frac{\rho_{Y,g_M}\rho_{\widehat{Y},g_M}}{\sqrt{\left(1 - \rho^2_{Y,g_M}\right)\left(1 - \rho^2_{\widehat{Y},g_M}\right)}}. \end{aligned} \tag{34}$$

If

$$p\left(Y,\widehat{Y}|g_M\right)=p\left(Y|g_M\right)p\left(\widehat{Y}|g_M\right)$$

then $\rho_{Y,\widehat{Y}|g_M}=0$ and (34) shows that

$$\rho_{Y,\widehat{Y}}=\rho_{Y,g_M}\rho_{\widehat{Y},g_M}.$$

6.5 Marker effects as fixed parameters - least squares estimation

6.5.1 The case of 1 marker

To address this question, consider first the single marker model

$$Y_i=bX_i+e_i,\quad i=1,2,\ldots,N, \tag{35}$$

with $Var\left(Y_i|X_i\right)=\sigma_e^2$ and where N is the number of individuals. In (35), $X_i\sim(0,2p\left(1-p\right))$; thus, in contrast with previous notation, X_i represents now a centered (but not scaled) random variable that labels the marker genotype and b is treated as a known parameter. The genomic value is

$$g_M=bX_i \tag{36}$$

and

$$Var\left(g_M\right)=\sigma_{g_M}^2=b^2 2p\left(1-p\right)$$

is the genomic variance.

The variance of Y_i given b, is

$$Var\left(Y_i|b\right)=\sigma_{g_M}^2+\sigma_e^2.$$

Assume that b is estimated by least squares using test data Y^t, and that the predictor is (we use here $\widehat{Y}$ for the predictor, not to be confused with (23))

$$\widehat{Y}_i=\widehat{b}X_i, \tag{37}$$

where $\widehat{b}$ is the least squares estimator which has the distribution

$$\widehat{b}\sim N\left(b,\frac{\sigma_e^2}{N2p\left(1-p\right)}\right) \tag{38}$$

and N is the size of the training data. The denominator arises by writing $Var\left(\widehat{b}|X\right)=\frac{\sigma_e^2}{\sum_{i=1}^N\left(X_i-\overline{X}\right)^2}$, and expressing $\sum_{i=1}^N\left(X_i-\overline{X}\right)^2=N\widehat{Var}\left(X_i\right)=N2p\left(1-p\right)$, assuming p is known.

To obtain an expression for $\rho_{g_M,\widehat{Y}}=\frac{Cov\left(g_M,\widehat{Y}\right)}{\sqrt{Var(g_M)Var\left(\widehat{Y}\right)}}$ we proceed as follows. Firstly,

$$\begin{aligned}Cov\left(g_M,\widehat{Y}|X\right) &= Cov\left(bX,\widehat{b}X|X\right)\\ &= X^2Cov\left(b,\widehat{b}\right)=0\end{aligned}$$

because b is a parameter. The marginal covariance is

$$\begin{aligned} Cov\left(g_M, \widehat{Y}\right) &= E\left[Cov\left(g_M, \widehat{Y}|X\right)\right] + Cov\left[E\left(g_M|X\right), E\left(\widehat{Y}|X\right)\right] \\ &= 0 + Cov_X\left[bX, bX\right] \\ &= b^2 Var\left(X\right) = \sigma^2_{g_M}. \end{aligned} \tag{39}$$

NOTE: $E\left(\widehat{Y}|X\right) = E\left(\widehat{b}X|X\right) = XE\left(\widehat{b}\right) = bX$.

The same result is obtained (of course) if instead one conditions on $\widehat{b}$. Thus

$$\begin{aligned} Cov\left(g_M, \widehat{Y}|\widehat{b}\right) &= Cov\left(bX, \widehat{b}X|\widehat{b}\right) \\ &= \widehat{b}\widehat{b}Var\left(X\right) \end{aligned}$$

and

$$\begin{aligned} Cov\left(g_M, \widehat{Y}\right) &= E\left[Cov\left(g_M, \widehat{Y}|\widehat{b}\right)\right] + Cov\left[E\left(g_M|\widehat{b}\right), E\left(\widehat{Y}|\widehat{b}\right)\right] \\ &= E\left[\widehat{b}\widehat{b}Var\left(X\right)\right] \\ &= b^2 Var\left(X\right) = \sigma^2_{g_M}. \end{aligned}$$

and the second term on the right hand side of the first line is zero because $E\left(\widehat{Y}|\widehat{b}\right) = E\left(\widehat{b}X|\widehat{b}\right) = \widehat{b}E\left(X\right) = 0$.

Also,

$$Var\left(\widehat{Y}|\widehat{b}\right) = \widehat{b}^2 Var\left(X\right).$$

Notice that if b is known without error, $Var\left(\widehat{Y}|\widehat{b}\right) = Var\left(\widehat{Y}|b\right) = b^2 Var\left(X\right) = \sigma^2_{g_M}$ and

$$\begin{aligned} \rho^2_{g_M\widehat{Y}} &= \frac{\left(b^2 Var\left(X\right)\right)^2}{Var\left(X\right) b^2 Var\left(g_M\right)} \\ &= \frac{b^2 Var\left(X\right)}{Var\left(g_M\right)} = 1 \end{aligned}$$

,

When b is estimated from the data, the unconditional variance is

$$\begin{aligned} Var\left(\widehat{Y}\right) &= E\left[Var\left(\widehat{Y}|\widehat{b}\right)\right] + Var\left[E\left(\widehat{Y}|\widehat{b}\right)\right] \\ &= E\left[Var\left(\widehat{Y}|\widehat{b}\right)\right] \\ &= Var\left(X\right) E\left(\widehat{b}^2\right) \\ &= Var\left(X\right)\left(Var\left(\widehat{b}\right) + b^2\right) \\ &= \sigma^2_{g_M} + \frac{\sigma^2_e}{N}. \end{aligned}$$

Then,

$$\begin{aligned}\rho^2_{g_M,\widehat{Y}} &= \frac{\left(\sigma^2_{g_M}\right)^2}{\sigma^2_{g_M}\left(\sigma^2_{g_M}+\frac{\sigma^2_e}{N}\right)} \\ &= \frac{\sigma^2_{g_M}}{\sigma^2_{g_M}+\frac{\sigma^2_e}{N}} \\ &= \frac{g^2}{\frac{(1-g^2)}{N}+g^2}, \end{aligned} \tag{40}$$

where $1-g^2=\sigma^2_e/Var\,(Y)$. When N is large, $\rho^2_{g_M,\widehat{Y}}\approx 1$.

6.5.2 The case of m_G markers

The extended model assumes that m_G marker genotypes are available. The linear structure for a datum is

$$Y_{ij}=\sum_{j=1}^{m_G} b_j X_{ij}+e_i; \quad j=1,2,\ldots,m_G; \quad i=1,2,\ldots,N,$$

where as before, the $b's$ are treated as fixed but unknown paprameters. The genomic value is

$$g_{M_i}=\sum_{j=1}^{m_G} b_j X_{ij}.$$

The variance of Y_{ij} given all the $b's$ is

$$\begin{aligned} Var\,(Y_{ij}|b) &= \sum_{j=1}^{m_G} b_j^2 Var\,(X_{ij})+\sum_{i<j} 2b_i b_j Cov(X_i,X_j)+\sigma^2_e \\ &= \sigma^2_{g_M}+\sigma^2_e=\sigma^2. \end{aligned}$$

At this point, we note that here with the $b's$ treated as fixed effects,

$$\sigma^2_{g_M}=\sum_{j=1}^{m_G} b_j^2 Var\,(X_{ij})+\sum_{i<j} 2b_i b_j Cov(X_i,X_j) \tag{41}$$

the genomic variance has a diffferent interpretation than in (13). In the latter, the $b's$ are random, and the variance is marginalised with respect to b and W. Here it is only marginalised over the $X's$.

The predictor is

$$\widehat{Y}_{ij}=\sum_{j=1}^{m_G}\widehat{b}_j X_{ij}.$$

Now

$$\begin{aligned} Cov\left(g_{M_i}, \widehat{Y}_i|\widehat{b}\right) &= Cov\left(\sum_{j=1}^{m_G} b_j X_{ij}, \sum_{j=1}^{m_G} \widehat{b}_j X_{ij}|\widehat{b}\right) \\ &= \sum_{j=1}^{m_G} \widehat{b}_j b_j Var\left(X_j\right) + \sum_{k,j} b_k \widehat{b}_j Cov\left(X_k, X_j\right) \end{aligned}$$

and unconditionally with respect to $\widehat{b}$,

$$\begin{aligned} Cov\left(g_{M_i}, \widehat{Y}_i\right) &= \sum_{j=1}^{m_G} b_j^2 Var\left(X_j\right) + \sum_{k<j} 2 b_k b_j Cov(X_k, X_j) \\ &= \sigma^2_{g_M}, \end{aligned} \tag{42}$$

the total variance due to the markers. Similarly,

$$\begin{aligned} Var\left(\widehat{Y}|\widehat{b}\right) &= Var\left(\textstyle\sum_{j=1}^{m_G} \widehat{b}_j X_j\right) \\ &= \textstyle\sum_{j=1}^{m_G} \widehat{b}_j^2 Var\left(X_j\right) + \sum_{i<j} 2\widehat{b}_i \widehat{b}_j Cov\left(X_i, X_j\right), \end{aligned}$$

and the marginal variance is

$$\begin{aligned} Var\left(\widehat{Y}\right) &= E\left[Var\left(\textstyle\sum_j \widehat{b}_j X_j\right)\right] + Var\left[E\left(\widehat{Y}|\widehat{b}\right)\right] \\ &= \textstyle\sum_{j=1}^{m_G} Var\left(X_j\right) E\left(\widehat{b}_j^2\right) + \sum_{i<j} 2Cov\left(X_i, X_j\right) E\left(\widehat{b}_i \widehat{b}_j\right) \\ &= \textstyle\sum_{j=1}^{m_G} Var\left(X_j\right)\left(Var\left(\widehat{b}_j\right) + b_j^2\right) + \sum_{i<j} 2Cov\left(X_i, X_j\right)\left(Cov\left(\widehat{b}_i, \widehat{b}_j\right) + b_i b_j\right) \\ &= \sigma^2_{g_M} + \textstyle\sum_j Var\left(X_j\right) Var\left(\widehat{b}_j\right) + \sum_{i<j} 2Cov\left(X_i, X_j\right) Cov\left(\widehat{b}_i, \widehat{b}_j\right). \end{aligned} \tag{43}$$

The second term in the first line is zero because $E\left(\widehat{Y}|\widehat{b}\right) = 0$. Also in the 4th line

$$\begin{aligned} \sigma^2_{g_M} &= \textstyle\sum_{j=1}^{m_G} Var\left(X_j\right) b_j^2 + \sum_{i<j} 2Cov\left(X_i, X_j\right) b_i b_j \\ &= \textstyle\sum_{j=1}^{m_G} 2p_j(1-p_j) b_j^2 + \sum_{i<j} 2D_{ij} b_i b_j \end{aligned}$$

which differs from the definition given before in (13) as mentioned above.

As an approximation, we assume

$$Var\left(\widehat{b}_j\right) = \frac{\sigma_e^2}{2Np_j\left(1-p_j\right)}$$

and then

$$\begin{aligned} \textstyle\sum_{j=1}^{m_G} Var\left(X_j\right) Var\left(\widehat{b}_j\right) &= \textstyle\sum_{j=1}^{m_G} 2p_j\left(1-p_j\right) \dfrac{\sigma_e^2}{2Np_j\left(1-p_j\right)} \\ &= \frac{m_G \sigma_e^2}{N}. \end{aligned}$$

We shall also ignore the third term in (43), which is equivalent to assuming that the m_G markers represent an *effective number of markers* whose correlation is zero. Therefore, ignoring the third term $\sum_{i<j} 2Cov\left(X_i, X_j\right) Cov\left(\widehat{b}_i, \widehat{b}_j\right)$

$$\begin{aligned} Var\left(\widehat{Y}\right) &\approx \sigma^2_{g_M} + \textstyle\sum_j Var\left(X_j\right) Var\left(\widehat{b}_j\right) \\ &= \sigma^2_{g_M} + \frac{m_G \sigma^2_e}{N}, \end{aligned}$$

and, approximately

$$\begin{aligned} \rho^2_{g_M,\widehat{Y}} &\approx \frac{\left[Cov\left(g_{m_i}, \widehat{Y}_i\right)\right]^2}{Var\left(g_{M_i}\right) Var\left(Var\widehat{Y}_i\right)} \\ &= \frac{\left(\sigma^2_{g_M}\right)^2}{\sigma^2_{g_M}\left(\sigma^2_{g_M} + \frac{m_G\sigma^2_e}{N}\right)} \\ &= \frac{\sigma^2_{g_M}}{\frac{m_G\sigma^2_e}{N} + \sigma^2_{g_M}} \\ &= \frac{g^2}{\frac{m_G(1-g^2)}{N} + g^2}. \end{aligned} \tag{44}$$

This expression reduces to (40) when $m_G = 1$. The point is that depending on m_G/N, the squared correlation between observed and predicted values may be significantly smaller than the proportion of variance due to the markers, g^2.

Incidentally, (44) produces

$$m_G = \frac{Ng^2\left(1 - \rho^2_{g_M,\widehat{Y}}\right)}{\rho^2_{g_M,\widehat{Y}}\left(1 - g^2\right)}. \tag{45}$$

6.6 Marker effects as random variables - BLUP

When W and b are treated as random, the genomic variance conditional on the observed genotypes is

$$\begin{aligned} Var\left(w_i'b|w_i\right) &= Var\left(g_{Mi}|w_i\right) \\ &= w_i'w_i\sigma^2_b \\ &= \textstyle\sum_{i=1}^{m_G} w_i^2\sigma^2_b, \end{aligned}$$

where w_i' is the ith row (individual) of $W = \{w_{ij}\}$, $j = 1, 2, \ldots, m_G$ and m_G is the observed number of marker genotypes. Unconditionally,

$$Var\left(g_M\right) = \sigma^2_{g_M} = m_G\sigma^2_b$$

as indicated in (14) and different from (41). Given $\sigma^2_{g_M}$, the larger m_G the smaller the variance of the SNP effect σ^2_b. The genomic heritability or the proportion of the total variance explained by the SNP genotypes is

$$g^2 = \frac{\sigma^2_{g_M}}{\sigma^2},$$

where $\sigma^2 = \sigma^2_{g_M} + \sigma^2_e$ is the marginal variance of Y.

As above an expression for the squared correlation between observed data g_M and predicted data $\widehat{Y}$, $\rho^2_{g_M,\widehat{Y}}$ is derived.

The model is

$$y = Wb + e, \tag{46}$$

y is $(N \times 1)$, W is $(N \times m_G)$ and b is $(m_G \times 1)$. Further,

$$\begin{aligned} y|W, b &\sim N\left(Wb, I\sigma^2_e\right), \\ b &\sim N\left(0, I\sigma^2_b\right) \end{aligned}$$

and the variance components σ^2_e, σ^2_b are assumed "known". Although inferences are conditional on $W = \{w_{ij}\}$, since the elements of W are centered and scaled as in (11),

$$\begin{aligned} w_{ij} &\sim (0, 1), \\ Cov\left(w_{ij}, w_{ik}\right) &= E\left(w_{ij}, w_{ik}\right) \\ &= r_{jk} = D_{jk} \end{aligned}$$

where D_{jk} is the linkage disequilibrium parameter between SNP j and k.

BLUP of b is

$$\widehat{b} = E\left(b|y\right) = \left(W'W + Ik\right)^{-1} W'y \tag{47}$$

with $k = \sigma^2_e/\sigma^2_b$. The prediction error variance is

$$\begin{aligned} Var\left(\widehat{b} - b\right) &= Var\left(b|y\right) \quad \text{(not a function of } y\text{)} \\ &= E\left[Var\left(b|y\right)\right] \\ &= \sigma^2_e\left(W'W + Ik\right)^{-1}. \end{aligned} \tag{48}$$

In general,

$$\begin{aligned} Var\left(b\right) &= E\left[Var\left(b|y\right)\right] + Var\left[E\left(b|y\right)\right] \\ &= Var\left(\widehat{b} - b\right) + Var\left(\widehat{b}\right). \end{aligned} \tag{49}$$

As shown below, $Var\left(\widehat{b}_i - b_i\right) \approx \left(N + k\right)^{-1}\sigma^2_e$ and

$$Var\left(\widehat{b}_i\right) \approx \sigma^2_b - \frac{\sigma^2_e}{N + k}. \tag{50}$$

To obtain an expression for $\rho^2_{g_M,\widehat{Y}}$ we need $Var\left(g_M\right) = \sigma_{g_M}$, $Var\left(\widehat{Y}\right)$ and $Cov\left(g_M, \widehat{Y}\right)$. The marginal (with respect to b and W) variance of the data is

$$\begin{aligned} Var\left(Y\right) &= m_G\sigma_b^2 + \sigma_e^2 \\ &= \sigma^2_{g_M} + \sigma_e^2 = \sigma^2. \end{aligned}$$

The predicted value of Y is

$$\widehat{Y}_i = \textstyle\sum_{j=1}^{m_G} w_{ij}\widehat{b}_j.$$

The conditional expectation is

$$E\left(\widehat{Y}_i|\widehat{b}\right) = \textstyle\sum_{j=1}^{m_G} \widehat{b}_j E\left(w_{ij}\right) = 0$$

and the conditional variance

$$Var\left(\widehat{Y}_i|\widehat{b}\right) = \textstyle\sum_{j=1}^{m_G} \widehat{b}_j^2 + \sum_{k<l} 2\widehat{b}_k\widehat{b}_l D_{kl}.$$

Unconditionally,

$$\begin{aligned} Var\left(\widehat{Y}_i\right) &= E\left[Var\left(\widehat{Y}_i|\widehat{b}\right)\right] \\ &= \textstyle\sum_{j=1}^{m_G} Var\left(\widehat{b}_j\right) + 2\sum_{k<l} r_{kl} Cov\left(\widehat{b}_k, \widehat{b}_l\right) \\ &= \textstyle\sum_{j=1}^{m_G} Var\left(\widehat{b}_j\right) + 2\sum_{k<l} \\ &\approx \textstyle\sum_{j=1}^{m_G}\left[\sigma_b^2 - \dfrac{\sigma_e^2}{N+k}\right] \quad \text{ignoring the second term above} \\ &= m_G\sigma_b^2 - \dfrac{m_G\sigma_e^2}{N+k} \\ &= \sigma^2_{g_M} - \dfrac{m_G\left(1-g^2\right)\sigma^2}{N + \frac{(1-g^2)\sigma^2}{\sigma_b^2}} \\ &= \sigma^2_{g_M} - \dfrac{\left(1-g^2\right)\sigma^2}{\frac{N}{m_G} + \frac{(1-g^2)\sigma^2}{m_G\sigma_b^2}} \\ &= \sigma^2_{g_M} - \dfrac{\left(1-g^2\right)\sigma^2}{\frac{N}{m_G} + \frac{(1-g^2)}{g^2}} \\ &= \sigma^2_{g_M}\dfrac{g^2}{g^2 + \left(1-g^2\right)\frac{m_G}{N}}. \end{aligned} \tag{51}$$

The squared correlation is

$$\begin{aligned}\rho^2_{g_M,\widehat{Y}} &= \frac{Cov\left(g_M,\widehat{Y}\right)^2}{Var\left(g_M\right)Var\left(\widehat{Y}\right)} \\ &= \frac{Var\left(\widehat{Y}\right)}{Var\left(g_M\right)} \\ &= \frac{g^2}{g^2+\left(1-g^2\right)\frac{m_G}{N}} \end{aligned} \tag{52}$$

because $Cov\left(g_M,\widehat{Y}\right)=Var\left(\widehat{Y}\right)$. Notice that (52) is equal to (44), which seems a strange result despite the different definitions of $\sigma^2_{g_M}$.

NOTE 1

$$\begin{aligned} Cov\left(\widehat{Y}_i, g'_{Mi}|w_i\right) &= Cov\left(w_i'b, \widehat{b}'w_i|w_i\right) \\ &= Cov\left[w_i'\widehat{b}, \left(\widehat{b}-\left(\widehat{b}-b\right)\right)'w_i\right] \\ &= w_i'\left[Var\left(\widehat{b}\right)-Cov\left(\widehat{b}, \left(\widehat{b}-b\right)'\right)\right]w_i \\ &= w_i'Var\left(\widehat{b}\right)w_i \\ &= Var\left(\widehat{Y}_i|w_i\right). \end{aligned}$$

NOTE 2 - For two records and three SNPs, (46) is

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + e$$

and

$$W'W+Ik = \begin{bmatrix} w_{11}w_{11}+w_{21}w_{21}+k & w_{11}w_{12}+w_{21}w_{22} & w_{11}w_{13}+w_{21}w_{23} \\ w_{11}w_{12}+w_{21}w_{22} & w_{12}w_{12}+w_{22}w_{22}+k & w_{12}w_{13}+w_{22}w_{23} \\ w_{11}w_{13}+w_{21}w_{23} & w_{12}w_{13}+w_{22}w_{23} & w_{13}w_{13}+w_{22}w_{23}+k \end{bmatrix}. \tag{53}$$

The jth diagonal element has the form

$$\sum_{i=1}^{N} w_{ij}^2 + k \tag{54}$$

where the subscript i refers to individual and j to the SNP. Since $E\left(w_{ij}^2\right) = 1$ and $E\left(w_{ij}\right) = 0$, $Var\left(w_{ij}\right) = 1$ and we write (54) as $NVar\left(w_{ij}\right) + k = N + k$. The off-diagonal terms are function of disequilibria (covariances) involving different SNP genotypes. Parameter k is typically a very large number, because the variance per SNP σ_b^2 is very small relative to σ_e^2. The diagonal element of the inverse of (53) will be approximated by $(N+k)^{-1}$ and therefore,

$$Var\left(\widehat{b} - b\right) \approx I\left(N+k\right)^{-1}\sigma_e^2. \tag{55}$$

6.7 Influence of degree of relationship on predictive ability

We now consider how the degree of relationship between the training and validating data affect the predictive ability of the genomic model. To formalise this, we consider

$$Cov\left(\widehat{Y}, g_M'|W\right).$$

Firstly, BLUP of b is computed using the training data $Y_t = W_t b + e$, where as before

$$\begin{aligned} b &\sim N\left(0, I\sigma_b^2\right) \\ e &\sim N\left(0, I\sigma_e^2\right). \end{aligned}$$

Then

$$\begin{aligned} \widehat{b} &= C_t^{-1}W_t'Y_t \\ &= C_t^{-1}W_t'W_t b + C_t^{-1}W_t'e, \end{aligned}$$

where

$$C_t = W_t'W_t + Ik, \quad k = \sigma_e^2/\sigma_b^2.$$

In these expressions, W_t of order $N \times m$, has elements $W_{t,ij}$, $i = 1, 2, \ldots, m$ labelling markers and $j = 1, 2, \ldots, N$ labelling individuals in the training data. The predicted phenotype in the validating data is

$$\begin{aligned} \widehat{Y} &= W_v\widehat{b} \\ &= W_v C_t^{-1}W_t'W_t b + W_v C_t^{-1}W_t'e. \end{aligned}$$

The unobserved genomic effect in the validating data is

$$g_M = W_v b$$

and

$$\begin{aligned} Cov\left(\widehat{Y}, g_M'|W\right) &= Cov\left(W_v C_t^{-1}W_t'W_t b + W_v C_t^{-1}W_t'e, b'W_v'|W\right) \\ &= W_v C_t^{-1}W_t'W_t W_v'\sigma_b^2 \end{aligned} \tag{56}$$

because $Cov\,(b, e') = 0$. The covariance is a function of the degree of genomic relationship between the training and the validating datasets, via the terms $W_t W_v'$. For the ith phenotypic value, $Cov\left(\widehat{Y}_i, g_{M_i}|W\right) = W_{v,i} C_t^{-1} W_t' W_t W_{v,i}' \sigma_b^2$, where $W_{v,i}$ is the ith row of matrix W_v. In this case

$$W_t W_{v,i}' = \begin{bmatrix} W_{t,11}W_{v,1i} + W_{t,21}W_{v,2i} + \cdots + W_{t,m1}W_{v,mi} \\ W_{t,12}W_{v,1i} + W_{t,22}W_{v,2i} + \cdots + W_{t,m2}W_{v,mi} \\ \vdots \\ W_{t,1N}W_{v,1i} + W_{t,2N}W_{v,2i} + \cdots + W_{t,mN}W_{v,mi} \end{bmatrix}$$

an $N \times 1$ column vector where the jth row is the sum over the m markers of terms of the form $W_{t,kj}W_{v,ki}$. This term corresponds to individual j in the training data and indvidual i in the validating data. The elements in this vector describe the degree of marker captured relationship between individual i in the validating data, and all the N individuals in the training data. Therefore, dividing row j say of $W_t W_{v,i}'$ by the number of markers m, yields an estimate of the average (over markers) relationship between individual j in the training data and i in the validating data.

References

Goddard, M. E. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica 136*, 245–252.

Powell, J. E., P. M. Visscher, and M. E. Goddard (2010). Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews, Genetics 11*, 800–805.

Wright, S. (1922). Coefficients of inbreeding and relationship. *American Naturalist 56*, 330–338.