Theory and Methods for Genomic Prediction

November 11, 2013

Contents

1 Basic Concepts

1.1 Crossingover and Recombination

An odd number of crossovers between two loci results in a recombination between them. Because crossing over takes place at random, the probability of recombination is higher for loci that are farther apart than for loci that are closer to each other. This provides the basis for genetic linkage analysis, where recombination rates between loci are used to order genes on chromosomes. For example, if the recombination rate between locus A and B is $r_{AB} = 0.1$, between B and C is $r_{BC} = 0.1$, and between A and C is $r_{AC} = 0.19$, we can arrange the loci in the order ABC. Note that $r_{AC} < r_{AB} + r_{BC}$. This is because recombinations between A and B and between B and C result in an even number of crossovers between A and C with no recombination between A and C.

1.2 Interference

Interference is the lack of independence in recombinations at different intervals on a chromosome. Consider three loci ordered as ABC. If recombination in the A-B interval is independent from recombination in the B-C interval, the probability of a double recombinant, denoted by g_{11}, is

$$g_{11} = r_{AB}r_{BC}$$

where r_{ij} is the probability of a recombination between loci i and j. If recombinations in the two intervals are not independent, the above probability is give by

$$g_{11} = cr_{AB}r_{BC} \tag{1}$$

where c is called the coefficient of coincidence. Interference is quantified as $I = 1 - c$. Thus, under independence, $c = 1$ and $I = 0$.

1.3 Map Distance

The map distance x between two loci, in Morgan units, is defined as the expected number of crossovers between them. Unlike recombination rates, map distances are additive.

Map distances between loci provide a convenient set of parameters for models used in linkage analysis. Consider the gametes produced by a parent heterozygous at each of k loci. Each such gamete corresponds to a recombination event that can be indexed by a $k - 1 \times 1$ vector $\boldsymbol{\epsilon_i}$ where element j of $\boldsymbol{\epsilon_i}$ is 1 if there was a recombination between locus j and $j + 1$ or is 0 otherwise. Thus, in linkage analysis with k loci, there are 2^{k-1} recombination events that need to be modeled. The probability of each of these recombination events can be treated as a parameter. Taking into consideration that these probabilities sum to one, this approach would give rise to $2^{k-1} - 1$ parameters that need to be estimated. However, using the relationship between map distance and recombination rate, probabilities of the 2^{k-1} recombination events can be computed from the $k-1$ map distances between adjacent loci. Then, these $k-1$ map distances become the parameters for linkage analysis. The relationship between map distance and recombination rate is discussed next.

1.4 Map Functions

Map functions provide a transformation from map distance to recombination rate. Two approaches have been used to derive map functions. In the first, a probability model is assumed for the number of crossovers in an interval of length x. Then, recombination rate is calculated as the probability of an odd number of crossovers in the interval. In the second approach, recombination events in two adjacent intervals are modeled, allowing for interference. This model is then used to develop a differential equation, the solution for which yields the map function. Both of these approaches are described in detail below.

Suppose that P_t is the probability of t crossovers in a chromosomal interval of length x Morgans. Recall that a recombination is observed when an odd number of crossovers occurs in this interval. Thus, probability r_x of a recombination in an interval of length x is

$$\begin{aligned} r_x &= P_1 + P_3 + P_5 + \cdots \\ &= \tfrac{1}{2}(1 - \sum_t P_t(-1)^t) \\ &= \tfrac{1}{2}(1 - P(-1)) \end{aligned} \tag{2}$$

where $P(S) = \sum_t P_t S^t$ is the probability generating function of the distribution of crossovers.

Haldane [7] used the Poisson distribution for P_t. This implies that crossovers in one interval are independent of those in another and that the probability of crossovers in a very short interval is proportional to the length of the interval. According to the Poisson distribution, the probability of t crossovers in an interval of length of x (in Morgan units) is

$$P_t = \frac{(\lambda x)^t e^{-\lambda x}}{t!} \tag{3}$$

The parameter λ in the Poisson distribution is the expected number of outcomes in a unit interval. Because map distance between two loci is defined as the expected number of crossovers between them, $\lambda = 1$, and

$$P_t = \frac{x^t e^{-x}}{t!} \tag{4}$$

The probability generating function for (4) is

$$\begin{aligned}P(S) &= \sum_t \frac{x^t e^{-x} S^t}{t!} \\ &= \sum_t \frac{(xS)^t e^{-xS}}{t!} \frac{e^{-x}}{e^{-xS}} \\ &= e^{x(S-1)}\end{aligned} \tag{5}$$

Using (5) in (2) gives Haldane's map function:

$$r_x = \tfrac{1}{2}(1 - e^{-2x}) \tag{6}$$

The inverse of (6) is

$$x = \begin{cases} -\frac{1}{2}\ln(1-2r_x) & \text{if } 0 \leq r_x < \frac{1}{2} \\ \infty & \text{if } r_x = \frac{1}{2} \end{cases}$$

Karlin [10] used the binomial distribution with parameters N and p for P_t. Thus, t is the number of successes in N Bernoulli trials each having probability p of success. From the definition of map distance, it follows that the map distance $x = \mathrm{E}(t) = Np$, and $p = x/N$. Now, the probability of t crossovers in an interval of length of x is

$$P_t = \binom{N}{t}(x/N)^t(1-x/N)^{N-t} \tag{7}$$

The probability generating function for (7) is

$$\begin{aligned}P(S) &= \sum_t \binom{N}{t}(x/N)^t(1-x/N)^{N-t}S^t \\ &= \sum_t \binom{N}{t}(xS/N)^t(1-x/N)^{N-t} \\ &= [xS/N + (1-x/N)]^N\end{aligned} \tag{8}$$

because $\sum_t \binom{N}{t} a^t b^{N-t} = (a+b)^N$. Using (8) in (2) gives the binomial map function:

$$r_x = \begin{cases} \frac{1}{2}[1-(1-2x/N)^N] & \text{if } x < N/2 \\ \frac{1}{2} & \text{if } x \geq N/2 \end{cases} \tag{9}$$

The inverse of (9) is

$$x = \tfrac{1}{2}N[1-(1-2r_x)^{1/N}] \tag{10}$$

In the second approach for deriving map functions, recombination is modeled in two adjacent intervals. Suppose three loci A, B, and C are ordered as ABC with a map distance of x between A and B, and a distance of h between B and C. Let $M(x)$ be the map function that we wish to derive that transforms map distances to recombination rates. It is assumed that when x is sufficiently small, $r_x = M(x) = x$.

Also, let $g_{\boldsymbol{\epsilon_i}}$ denote the probability of the recombination event indexed by $\boldsymbol{\epsilon_i}$; for example, g_{10} is the probability of a recombination in the first interval and no recombination in the second interval.

Using this notation, the probability r_{AC} of a recombination between A and C can be written as

$$r_{AC} = g_{10} + g_{01}$$

If there is no interference,

$$\begin{aligned} r_{AC} &= g_{10} + g_{01} \\ &= r_{AB}(1 - r_{BC}) + (1 - r_{AB})r_{BC} \\ &= r_{AB} + r_{BC} - 2r_{AB}r_{BC} \end{aligned} \tag{11}$$

Recall that $r_{AB}r_{BC} = g_{11}$ is the probability of a double recombination when interference is absent. When interference is present, the probability of a double recombination is given by (1). Thus, when interference is present, the probability of a recombination between A and C can be written as

$$r_{AC} = r_{AB} + r_{BC} - 2cr_{AB}r_{BC} \tag{12}$$

where c is the coefficient of coincidence. Now, (12) is rewritten using the map function $M(.)$ in place of the recombination rates:

$$M(x + h) = M(x) + M(h) - 2cM(x)M(h) \tag{13}$$

The above equation can be rearranged as

$$\frac{M(x + h) - M(x)}{h} = \frac{M(h) - 2cM(x)M(h)}{h} \tag{14}$$

As $h \to 0$, $\frac{M(h)}{h} \to 1$. Thus, taking the limit as $h \to 0$, on both sides of (14) gives

$$\frac{dr_x}{dx} = 1 - 2cr_x \tag{15}$$

Letting $c = 1$ and solving (15) gives Haldane's map function (6). When $c = 1$, recombination in the two intervals are independent; this assumption is implicit in the Poisson distribution.

Letting $c = 2r_x$ gives the Kosambi map function

$$r_x = \tfrac{1}{2}\frac{e^{4x} - 1}{e^{4x} + 1} \tag{16}$$

with inverse

$$x = \frac{1}{4} \ln \frac{1 + 2r_x}{1 - 2r_x} \tag{17}$$

Several other map functions derived from (15) by assuming different assumptions about c are given in AHGL (pp 14–17). Map functions derived from (15) with $c \neq 1$ are not suitable for linkage analysis with more than three loci (AHGL pp 124–127).

1.5 Computation of Recombination Probabilities

As mentioned earlier, probabilities of recombination events, denoted $g_{\boldsymbol{\epsilon_i}}$, play a key role in linkage analysis (e.g., AHGL pp 117-120). Here we describe the relationship between these recombination probabilities and map distances between loci. Using this relationship, all recombination probabilities can be computed given the map distances between loci.

To establish this relationship for k loci, we let $W_{\boldsymbol{\delta_i}}$ denote a region of the chromosome composed of inter-locus segments. Element j of the $k-1 \times 1$ vector $\boldsymbol{\delta_i}$ is 1 if the segment between loci j and $j+1$ is included in $W_{\boldsymbol{\delta_i}}$ and is 0 otherwise. The length of the region $W_{\boldsymbol{\delta_i}}$ is

$$x(\boldsymbol{\delta_i}) = \sum \boldsymbol{\delta_{i}}_j x_j$$

where x_j is the map distance between loci j and $j+1$.

The probability of an odd number of crossovers occurring in $W_{\boldsymbol{\delta_i}}$ is denoted $R(\boldsymbol{\delta_i})$ and is called the recombination value for $\boldsymbol{\delta_i}$. Given a map function $r_x = M(x)$, the recombination value for $\boldsymbol{\delta_i}$ can be computed as

$$R(\boldsymbol{\delta_i}) = M[x(\boldsymbol{\delta_i})] \tag{18}$$

The recombination value for $\boldsymbol{\delta_i}$ can also be computed as the sum of those $g_{\boldsymbol{\epsilon_i}}$ for which there is an odd number of recombinations in $W_{\boldsymbol{\delta_i}}$ (This rule works because the sum of an odd number of odd numbers is odd; the sum of an even number of odd numbers is even; and the sum of even numbers is always even.). For example, if $k = 4$ and $\boldsymbol{\delta_i} = [1, 0, 1]'$, $R(\boldsymbol{\delta_i}) = g_{001} + g_{011} + g_{100} + g_{110}$.

The number s_{ij} of recombinations in the region $W_{\boldsymbol{\delta_i}}$ given recombination event $\boldsymbol{\epsilon_j}$ is

$$s_{ij} = \boldsymbol{\delta_i}'\boldsymbol{\epsilon_j}$$

So, recombination value for $\boldsymbol{\delta_i}$ can be written as

$$\begin{aligned} R(\boldsymbol{\delta_i}) &= \sum_{j \text{ for } s_{ij}\text{odd}} g_{\boldsymbol{\epsilon_j}} \\ &= \tfrac{1}{2}[1 - \sum_{j=1}^{2^{k-1}} (-1)^{s_{ij}} g_{\boldsymbol{\epsilon_j}}] \end{aligned} \tag{19}$$

In matrix notation, the above relationship between the $R(\boldsymbol{\delta_i})$'s and the $g_{\boldsymbol{\epsilon_j}}$'s can be written as

$$\boldsymbol{r} = \tfrac{1}{2}[\mathbf{1} - \boldsymbol{A}\boldsymbol{g}] \tag{20}$$

where $\boldsymbol{r}$ is a $2^{k-1} \times 1$ vector of recombination values, $\mathbf{1}$ is a $2^{k-1} \times 1$ vector of 1's, the matrix $\boldsymbol{A} = \{(-1)^{s_{ij}}\}$, and $\boldsymbol{g}$ is a $2^{k-1} \times 1$ vector of recombination probabilities. Rearranging (20) gives

$$\boldsymbol{A}\boldsymbol{g} = \mathbf{1} - 2\boldsymbol{r}$$

Table 1: Recombination rate r_j and map length x_j for inter-locus segment $j = 1, 2, 3$. Map lengths are given for Haldane's map function and for the binomial map function with $N = 2$.

Segment	r_j	x_j	
j		Haldane	Binomial
1	0.1	0.1116	0.1056
2	0.05	0.0527	0.0513
3	0.2	0.2554	0.2254

The following properties can be shown to be true for the matrix $\boldsymbol{A}$: $\boldsymbol{A} = \boldsymbol{A}'$, $\boldsymbol{a}_i'\boldsymbol{a}_i = 2^{n-1}$ for $i = 1, \ldots, 2^{k-1}$, and $\boldsymbol{a}_i'\boldsymbol{a}_j = 0$ for $i \neq j$. So, $\boldsymbol{AA} = \boldsymbol{I}2^{k-1}$ and $\boldsymbol{A}^{-1} = \boldsymbol{A}\frac{1}{2^{k-1}}$. Now, $\boldsymbol{g}$ can be written in terms of $\boldsymbol{r}$ as

$$\begin{aligned} \boldsymbol{g} &= \boldsymbol{A}^{-1}(\boldsymbol{1} - 2\boldsymbol{r}) \\ &= \frac{\boldsymbol{A}(\boldsymbol{1} - 2\boldsymbol{r})}{2^{k-1}} \end{aligned} \tag{21}$$

In scalar notation, the recombination probability for $\boldsymbol{\epsilon_i}$ can be written as

$$g_{\boldsymbol{\epsilon_i}} = \sum_{j}^{2^{k-1}} (-1)^{s_{ij}} \frac{1 - 2R(\boldsymbol{\delta_j})}{2^{k-1}} \tag{22}$$

Equations (18) and (22) establish the relationship between map distances and recombination probabilities.

Consider the numerical example in AHGL (pp 125–126). Here, $k = 4$ and the recombination rates in the three intervals and the corresponding map distances for the Haldane and the binomial ($N = 2$) map functions are given in table (1).

The set of vectors $\boldsymbol{\epsilon_i}$ and $\boldsymbol{\delta}_j$ are given by the rows of the matrix $\boldsymbol{U}$:

$$\boldsymbol{U} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Thus, the matrix $\boldsymbol{S}$ of s_{ij}'s is

$$\boldsymbol{S} = \boldsymbol{UU'} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 2 & 1 & 2 \\ 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 \\ 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \end{bmatrix}$$

and the matrix $\boldsymbol{A}$ is

$$\boldsymbol{A} = \{(-1)^{s_{ij}}\} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \tag{23}$$

The map lengths for $W_{\boldsymbol{\delta_i}}$ and corresponding recombination values computed from Haldane and binomial ($N = 2$) map functions are in table (2). To obtain the recombination probabilities using the Haldane and binomial map functions, $\boldsymbol{r}$ vectors from the fifth and seventh columns of table (2) were used with the $\boldsymbol{A}$ matrix from (23) in equation (21). These probabilities are given in table (3).

The recombination probabilities under the Haldane map function can be computed much more simply due to the lack of interference. For example, $g_{110} = r_1 r_2 (1 - r_3) = (0.1)(0.05)(1 - 0.2) = 0.004$. However, this approach cannot be used when interference is present. Note that the Kosambi map function cannot be used for mapping more than 3 loci. When the Kosambi map function was used for this example, g_{111} was negative (AHGL p 126).

1.6 Linkage Disequilibrium in an Infinite Population

Gametic disequilibrium, which is more commonly referred to as linkage disequilirium (LD), is the statistical dependence between alleles in a haplotype. Under gametic equilibrium, alleles in a haplotype are independent. This is also called linkage equilibrium. Note that it is possible for two loci that are linked to be in gametic equilibrium; also, loci that are unlinked can be in gametic disequilibrium.

Suppose that starting from generations 0, all individuals are produced by random mating. Then, the probability of haplotype (A_i, B_j) in generation 1 can be written as

$$\Pr_1(A_i, B_j) = (1 - r) \Pr_0(A_i, B_j) + r \Pr(A_i) \Pr(B_j)$$

Table 2: Map length $x(\boldsymbol{\delta_i})$ and recombination value $R(\boldsymbol{\delta_i})$ for each $\boldsymbol{\delta_i}$, computed from Haldane and binomial ($N = 2$) map functions.

$\boldsymbol{\delta_i}$			Haldane		Binomial	
			$x(\boldsymbol{\delta_i})$	$R(\boldsymbol{\delta_i})$	$x(\boldsymbol{\delta_i})$	$R(\boldsymbol{\delta_i})$
0	0	0	0	0	0	0
0	0	1	0.2554	0.2	0.2254	0.2
0	1	0	0.0527	0.05	0.0513	0.05
0	1	1	0.3081	0.23	0.2767	0.2384
1	0	0	0.1116	0.1	0.1056	0.1
1	0	1	0.3670	0.26	0.3310	0.2762
1	1	0	0.1642	0.14	0.1569	0.1446
1	1	1	0.4197	0.284	0.3823	0.3092

Table 3: Probabilities of recombination events ($g_{\boldsymbol{\epsilon_i}}$) computed from Haldane and binomial ($N = 2$) map functions.

$\boldsymbol{\epsilon_i}$			$g_{\boldsymbol{\epsilon_i}}$	
			Haldane	Binomial
0	0	0	0.684	0.6704
0	0	1	0.171	0.1823
0	1	0	0.036	0.0415
0	1	1	0.009	0.0058
1	0	0	0.076	0.0854
1	0	1	0.019	0.0119
1	1	0	0.004	0.0027
1	1	1	0.001	0

where r is the recombination rate between loci A and B, and $\Pr_0(A_i, B_j)$ is the probability of (A_i, B_j) in generation 0. The disequilibrium in generation 1 is

$$\begin{aligned}\Delta_1 &= \Pr_1(A_i, B_j) - \Pr(A_i)\Pr(B_j)\\ &= (1-r)\Pr_0(A_i, B_j) + r\Pr(A_i)\Pr(B_j) - \Pr(A_i)\Pr(B_j)\\ &= (1-r)\Pr_0(A_i, B_j) - (1-r)\Pr(A_i)\Pr(B_j)\\ &= (1-r)\Delta_0\end{aligned}$$

where Δ_0 is the disequilibrium in generation 0. Similarly, the probability of haplotype (A_i, B_j) in generation 2 is

$$\Pr_2(A_i, B_j) = (1-r)\Pr_1(A_i, B_j) + r\Pr(A_i)\Pr(B_j)$$

and the disequilibrium in generation 2 is

$$\begin{aligned}\Delta_2 &= \Pr_2(A_i, B_j) - \Pr(A_i)\Pr(B_j)\\ &= (1-r)\Pr_1(A_i, B_j) + r\Pr(A_i)\Pr(B_j) - \Pr(A_i)\Pr(B_j)\\ &= (1-r)\Pr_1(A_i, B_j) - (1-r)\Pr(A_i)\Pr(B_j)\\ &= (1-r)\Delta_1\\ &= (1-r)^2\Delta_0\end{aligned}$$

It follows that in generation n, the disequilibrium is

$$\Delta_n = (1-r)^n\Delta_0$$

Thus, with each generation of random mating the haplotype distribution moves closer to equilibrium (statistical independence). For loci that are unlinked, $(1-r) = 1/2$ and equilibrium is reached quickly; for example, $(1/2)^{10} = 1/1024$. On the other hand, loci that are tightly linked will take much longer to reach equilibrium; for example, $(1-r)^{10} > 1/3$ for $r = 0.1$. In the limit, however, equilibrium is reached in an infinite population.

1.7 Linkage Disequilibrium in a Finite Population

In a closed finite population, in the absence of mutation, after a sufficient number of generations of random mating all alleles will become identical by descent; thus all alleles also will become identical in state. In other words, all loci will become “fixed”. In such a population genetic variability is absent and LD is not defined.

Loci that are moving toward fixation will pass through a phase where alleles that are identical by state (IBS) will also be identical by descent (IBD). In other words, all alleles that are IBS will trace back to a common ancestral mutant allele. Thus, at a biallelic locus A with alleles A_1 and A_2, all the A_1 alleles will have a common ancestor and all the A_2 alleles will have a different common ancestor. In such loci, Sved [17] has shown algebraically that the expected value of LD between loci A and B as measured by the squared correlation (ρ^2) between allele states is related to the probability of joint identity-by-descent at loci A and B for two randomly sampled gametes. At first, this relationship may not be obvious. To get an intuitive feel for this relationship, consider a population where alleles A_1 and A_2 are segregating at the A locus and alleles B_1 and B_2 at the B locus. Suppose all haplotypes with A_1 also have B_1 and those with A_2 have B_2. Then, ρ would be 1. If the allelic associations were reversed, i.e., A_1 goes with B_2 and A_2 goes with B_1, then ρ would be -1. In both cases ρ^2 is 1. In such a population for two randomly sampled gametes, the conditional probability would be one that alleles at the B locus are identical-by-state (IBS) given they are IBS at the A locus. On the other hand, if there are a few haplotypes where the association between alleles is different from the other haplotypes, then ρ^2 would be less than 1. Further, in this population for two randomly sampled gametes, the conditional probability would be less than 1 that alleles at the B locus are identical-by-state (IBS) given they are IBS at the A locus. Hopefully, this discussion helps to see that LD as measured by ρ^2 should be related to the probability of joint IBS. Recall, however, that we assumed that all A_1 alleles have descended from a common ancestor and similarly all A_2 alleles have descended from a different common ancestor. Thus, given alleles at the A locus are IBS, they have descended from a common ancestor (IBD), and provided that no recombination between the two loci has happened in the two paths descending from the common ancestor to the two randomly sampled haplotypes, the B alleles will also be IBD. Sved [17] denoted this conditional probability by Q, and reasoned that in the pool of gametes where alleles at the B locus are IBD given they are IBD at the A locus, LD as measured by the squared correlation (ρ^2) between allele states will be 1.0 [18]. On the other hand, in the pool of gametes where recombination has taken place between loci A and B, given random mating, ρ^2 is expected to be null [18].

Let $C = 1$ denote the condition that for a randomly sampled pair of gametes the alleles at locus B are IBD given that alleles are IBD at locus A, and $C = 0$ denote that this condition is not met. Then, the expected value of ρ^2 can be written as

$$\begin{aligned} E(\rho^2) &= \underset{C}{E}[E(\rho^2|C)] \\ &= E(\rho^2|C=1)\Pr(C=1) + E(\rho^2|C=0)\Pr(C=0) \\ &= 1Q + 0(1-Q) \\ &= Q. \end{aligned}$$

Let Q_t denote the probability of $C = 1$ in generation t. This probability can be recursively written as

$$Q_t = [\frac{1}{2N} + (1-\frac{1}{2N})Q_{t-1}](1-r)^2, \qquad (24)$$

where N is the effective population size, $\frac{1}{2N}$ is the probability that two randomly sampled gametes in generation t are both inherited from the same gamete in the previous generation, (1-$\frac{1}{2N}$) is the probability that they are inherited from different gametes in the previous generation, and $(1-r)^2$ and the probability that loci A and B do not recombine in these gametes in the last generation. At equilibrium, $Q_t = Q_{t-1} = Q_E$. So, setting Q_t and Q_{t-1} in (24) to Q_E gives

$$\begin{aligned} Q_E &= \frac{(1-r)^2}{2N-(2N-1)(1-r)^2} \\ &\approx \frac{1}{4Nr+1} \end{aligned} \tag{25}$$

In the derivation of (25) we only considered pairs of loci where alleles are segregating at each locus. Further, it was assumed that all haplotypes in the current generation descend from two ancestral haplotypes. So, if we code the four possible haplotypes at a pair of biallelic loci as: 00, 01, 10, an 11, the ancestral pair of haplotypes must be either (00,11) or (01,10). Any other pair would lead to one locus being fixed. For example, the pair (00,01) has locus one fixed for the allele coded as 0. So, if $4Nr$ is close to zero, most haplotypes in the current generations will be non-recombinants of the ancestral type and ρ^2 is close to 1. The consequences of a few recombinants are examined below using the following R function.

```
RSq = function(nij) {
    N = sum(nij)
    Exy = nij[4]/N
    Ex = (nij[3] + nij[4])/N
    Ey = (nij[2] + nij[4])/N
    Vx = Ex * (1 - Ex)
    Vy = Ey * (1 - Ey)
    Cxy = Exy - Ex * Ey
    res = Cxy^2/(Vx * Vy)
    return(res)
}
```

Here we only have non-recombinants:

```
nij = c(
80, # ancestral haplotype 00
0,  # recombinant         01
0,  # recombinant         10
20  # ancestral haplotype 11
)
RSq(nij)

## [1] 1
```

Now we introduce two recombinants:

```
nij = c(
80, # ancestral haplotype 00
1,  # recombinant         01
1,  # recombinant         10
18  # ancestral haplotype 11
)
RSq(nij)

## [1] 0.8743
```

Here is another example, where one of the ancestral haplotypes has a very low frequency:

```
nij = c(
98, # ancestral haplotype 00
1,  # recombinant         01
0,  # recombinant         10
1   # ancestral haplotype 11
)
RSq(nij)

## [1] 0.4949
```

In a finite population, however, most loci are fixed. Then, LD is not defined for these loci. When mutation introduces variability into such a locus, LD is defined but will be low. This is demonstrated in the following example:

```
# in this example the first locus is fixed until the mutant appears
nij = c(
80, # ancestral haplotype 00
19, # ancestral haplotype 01
1,  # mutant    haplotype 10
0   #
)
RSq(nij)

## [1] 0.002369
```

At mutation-drift equilibrium, most loci will be of this type, where mutation has recently introduced variability. Thus, $E(\rho^2)$ can be much lower in a population that has reached mutation-drift equilibrium than indicated by (25). To examine this further, the exact distribution of is ρ^2 is recursively computed next.

1.8 Distribution of ρ^2 in the Presence of Mutation

Computing the distribution of ρ^2 involves computing the joint distribution for allele frequencies at two loci. Thus, we will review first how to compute the distribution for allele frequency at a single locus.

1.8.1 Computing the distribution of allele frequency at a single locus

Consider a population of $2N$ gametes. Let Y be the number of A_1 alleles at locus A. The value of Y can take one of $2N+1$ values ranging from 0 to $2N$. Suppose the distribution of allele frequency in generation t is given by the vector $\boldsymbol{p}_t$ with $2N+1$ probabilities corresponding to each of the $2N+1$ possible values of Y. To model random mating, assume $2N$ gametes are sampled with replacement from the gametes of generation t. Then, ignoring mutation, migration and selection, the distribution of allele frequency in generation $t+1$ can be calculated as

$$\boldsymbol{p}_{t+1} = \boldsymbol{B}\boldsymbol{p}_t, \tag{26}$$

where $\boldsymbol{B}$ is a $(2N+1) \times (2N+1)$ matrix with element i, j containing the probability that a random variable from a Binomial$(2N, \frac{j}{2N})$ distribution would be equal to i for $i, j = 0, 1, 2, \ldots, 2N$. If this is not obvious, section 3.10.2 of the notes given here may be useful.

To model mutation, assume that an A_1 allele mutates to an A_2 with probability u and an A_2 mutates to an A_1 with probability v. Now, to accommodate mutation in computing the distribution of allele frequency, $\boldsymbol{B}$ is modified such that column j contains probabilities from the binomial distribution

$$\text{Binomial}[2N, \frac{j}{2N}(1-u) + (1-\frac{j}{2N})v].$$

Selection can be similarly accommodated by modifying the binomial probabilities for each j. See example here

1.8.2 Computing the joint distribution of allele frequencies at a two linked loci

Consider a locus A with alleles A_1 and A_2 and a linked locus B with alleles B_1 and B_2. A population of $2N$ gametes is now characterized by a vector $\boldsymbol{Y}$ with four elements containing the numbers of gametes with haplotypes: A_1B_1, A_1B_2, A_2B_1, and A_2B_2. Note that these four numbers must sum to $2N$. Let $\boldsymbol{X}$ be a $k \times 4$ matrix with each row representing a possible value of $\boldsymbol{Y}$, where the number k of rows in $\boldsymbol{X}$ is equal to

$$k = \frac{(2N+3)!}{3!(2N)!}.$$

As before let $\boldsymbol{p}_t$ denote the distribution of haplotype frequencies in generation t. Then, ignoring recombination, mutation, migration and selection, the distribution of haplotype frequencies in the next generation are given by

$$\boldsymbol{p}_{t+1} = \boldsymbol{M}\boldsymbol{p}_t, \tag{27}$$

where $\boldsymbol{M}$ is a $k \times k$ matrix with element i, j containing the probability that a random variable from a Multinomial($2N, \frac{\boldsymbol{x}'_j}{2N}$) distribution would be equal to $\boldsymbol{x}'_i$, for $i, j = 1, 2, \ldots, k$.

To model recombination, consider a population with frequency for haplotype A_iB_j given by f_{ij}. In gametes produced by this population, the probability of a non-recombinant A_1B_1 is $(1-r)\frac{f_{11}}{2N}$. A recombinant A_1B_1 gamete can be produced in one of four ways. They and their associated probabilities are:

1. alleles A_1 and B_1 originate from two different A_1B_1 haplotypes with associated probability $r\frac{f_{11}}{2N} \times \frac{(f_{11}-1)}{2N-1}$;

2. allele A_1 originates from an A_1B_1 haplotype and B_1 originates from an A_2B_1 with associated probability $r\frac{f_{11}}{2N} \times \frac{f_{21}}{2N-1}$;

3. allele A_1 originates from an A_1B_2 haplotype and B_1 originates from an A_1B_1 with associated probability $r\frac{f_{12}}{2N} \times \frac{f_{11}}{2N-1}$; and

4. allele A_1 originates from an A_1B_2 haplotype and B_1 originates from an A_2B_1 with associated probability $r\frac{f_{12}}{2N} \times \frac{f_{21}}{2N-1}$.

Combining these probabilities gives

$$\begin{aligned}\Pr(A_1B_1) =&(1-r)\frac{f_{11}}{2N}+ \\ &r\frac{f_{11}}{2N}[\frac{(f_{11}-1)}{2N-1}+\frac{f_{21}}{2N-1}]+ \\ &r\frac{f_{12}}{2N}[\frac{f_{11}}{2N-1}+\frac{f_{21}}{2N-1}].\end{aligned} \tag{28}$$

Similarly, probabilities of the remaining three types of gametes are:

$$\begin{aligned}\Pr(A_1B_2) =&(1-r)\frac{f_{12}}{2N}+ \\ &r\frac{f_{11}}{2N}[\frac{f_{12}}{2N-1}+\frac{f_{22}}{2N-1}]+ \\ &r\frac{f_{12}}{2N}[\frac{(f_{12}-1)}{2N-1}+\frac{f_{22}}{2N-1}],\end{aligned} \tag{29}$$

$$\begin{aligned}\Pr(A_2B_1) =&(1-r)\frac{f_{21}}{2N}+ \\ &r\frac{f_{21}}{2N}[\frac{f_{11}}{2N-1}+\frac{(f_{21}-1)}{2N-1}]+ \\ &r\frac{f_{22}}{2N}[\frac{f_{11}}{2N-1}+\frac{f_{21}}{2N-1}],\end{aligned} \tag{30}$$

and

$$\begin{aligned}\Pr(A_2B_2) =&(1-r)\frac{f_{22}}{2N}+\\ &r\frac{f_{21}}{2N}[\frac{f_{12}}{2N-1}+\frac{f_{22}}{2N-1}]+\\ &r\frac{f_{22}}{2N}[\frac{f_{12}}{2N-1}+\frac{(f_{22}-1)}{2N-1}].\end{aligned} \tag{31}$$

Let $\boldsymbol{\theta}_j$ be a vector with the four probabilities from equations (28) through (31) computed for haplotype frequencies from $\boldsymbol{x}'_j$. Now, haplotype probabilities following mutation can be modeled as

$$\boldsymbol{\beta}_j = \boldsymbol{T}\boldsymbol{\theta}_j, \tag{32}$$

where

$$\boldsymbol{T} = \begin{bmatrix} (1-u)^2 & (1-u)v & v(1-u) & v^2 \\ (1-u)u & (1-u)(1-v) & vu & v(1-v) \\ u(1-u) & uv & (1-v)(1-u) & (1-v)v \\ u^2 & u(1-v) & (1-v)u & (1-v)^2 \end{bmatrix}$$

To accommodate recombination and mutation in computing the distribution of haplotype frequencies, $\boldsymbol{M}$ is modified such that column j contains probabilities from the Multinomial$(2N, \boldsymbol{\beta}'_j)$ distribution.

Starting with an allele frequency of 0.5 at each locus and gametic equilibrium between the two loci, the expected value of ρ^2 was computed for 2000 generations given a mutation rate of $u = v = 1e^{-9}$, a recombination rate of $r = 0.002$ between the loci and an effective population size of $N_e = 5, 10, 25$, or 50. The results are shown in the figures 1 through 4. In addition to ρ^2, the figures also plot the frequencies of three groups of populations. In populations that belong to group 1, $\rho^2 < 1$. In populations that belong to group 2, two of the haplotypes are lost such that $\rho^2 = 1$ (for example, when haplotypes A_1B_2 and A_2B_1 are lost and only haplotypes A_1B_1 and A_2B_2 are segregating, $\rho = 1$). In populations of group 3, one of the loci is fixed, and therefore ρ is not defined.

In all these cases, group 1 starts out having frequency close to 1.0. Due to drift, however, the frequency in group 1 drops rapidly and frequencies in groups 2 and 3 rise. The expected value of ρ^2 depends to a large extent on the relative magnitudes of the frequencies of groups 1 and 2. Also, drift seems to reduce the frequency of group 1 faster than that of group 2. Therefore, ρ^2 rises rapidly and then stays high for a period. Once frequencies in groups 1 and 2 drop sufficiently low, changes in group frequencies due to mutation become significant. Mutation in group 3, which has the highest frequency, adds to group 1 faster than to group 2. Further, recombination in group 2 also contributes to group 1. The balance between these forces and drift back into group 3 from groups 1 and 2 determines the equilibrium value for ρ^2, which is reached by generation 2000 in all four cases.

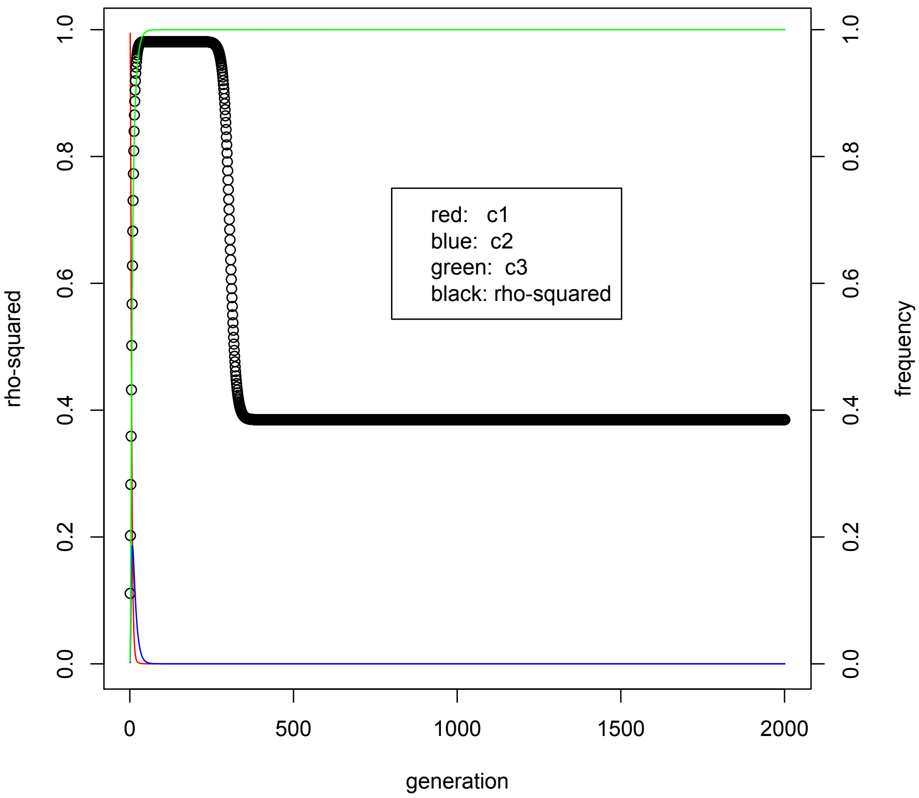

Figure 1: Expected value of ρ^2 by generation in population with effective population size $N_e = 5$, recombination rate $r = 0.002$, mutation rate $u = v = 1e^{-9}$. Frequencies in groups 1-3 are c1-c3.

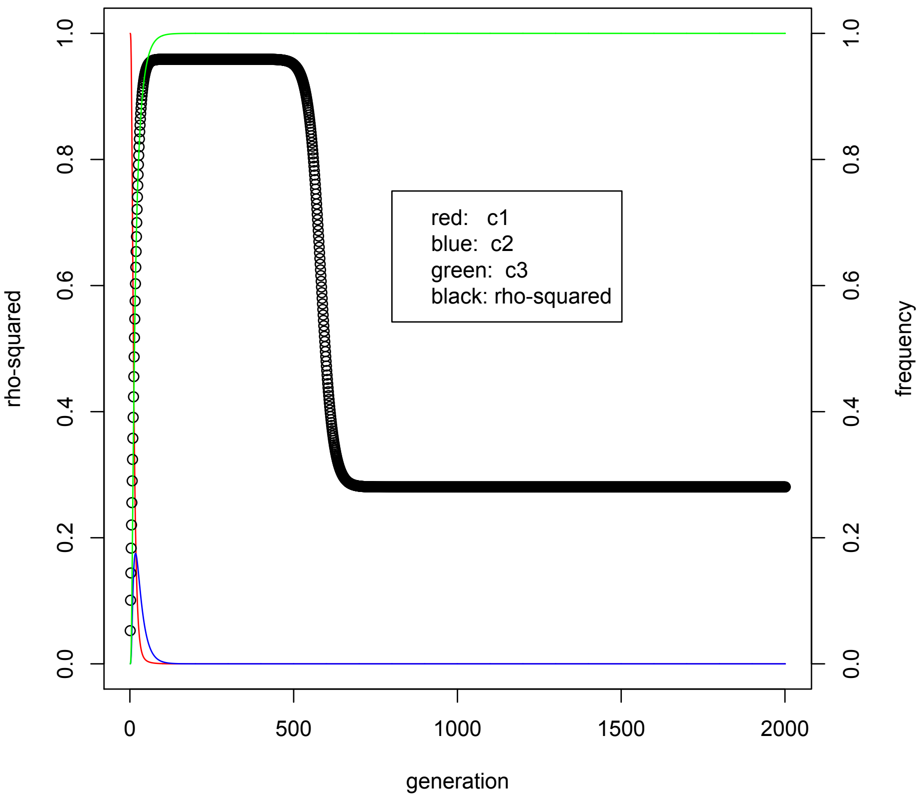

Figure 2: Expected value of ρ^2 by generation in population with effective population size $N_e = 10$, recombination rate $r = 0.002$, mutation rate $u = v = 1e^{-9}$. Frequencies in groups 1-3 are c1-c3.

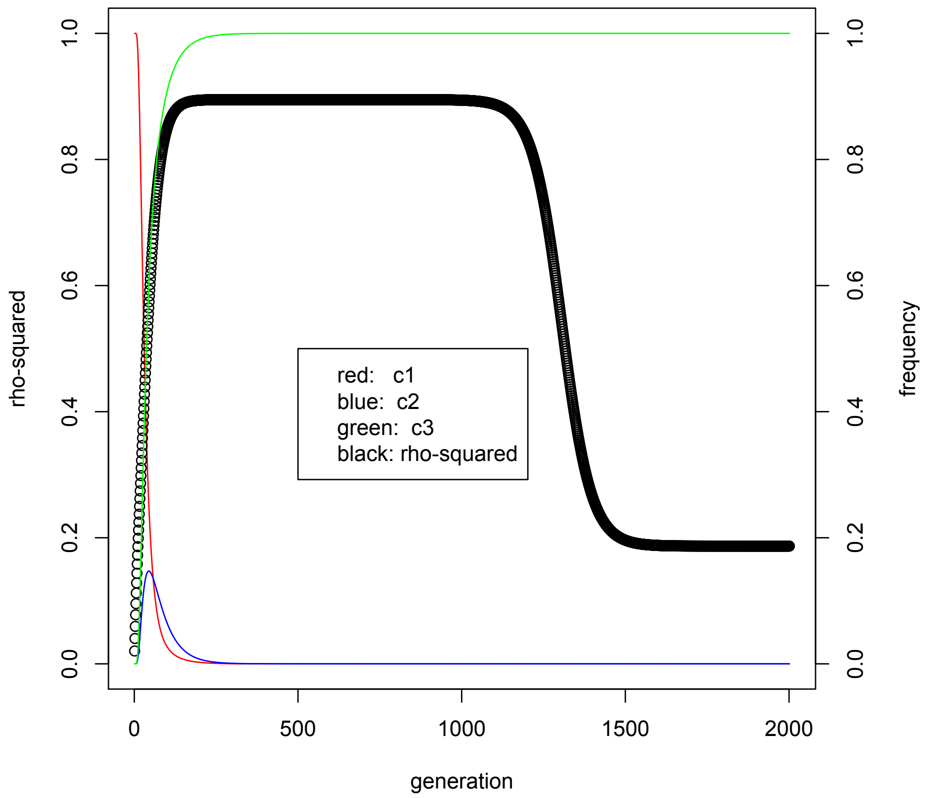

Figure 3: Expected value of ρ^2 by generation in population with effective population size $N_e = 25$, recombination rate $r = 0.002$, mutation rate $u = v = 1e^{-9}$. Frequencies in groups 1-3 are c1-c3.

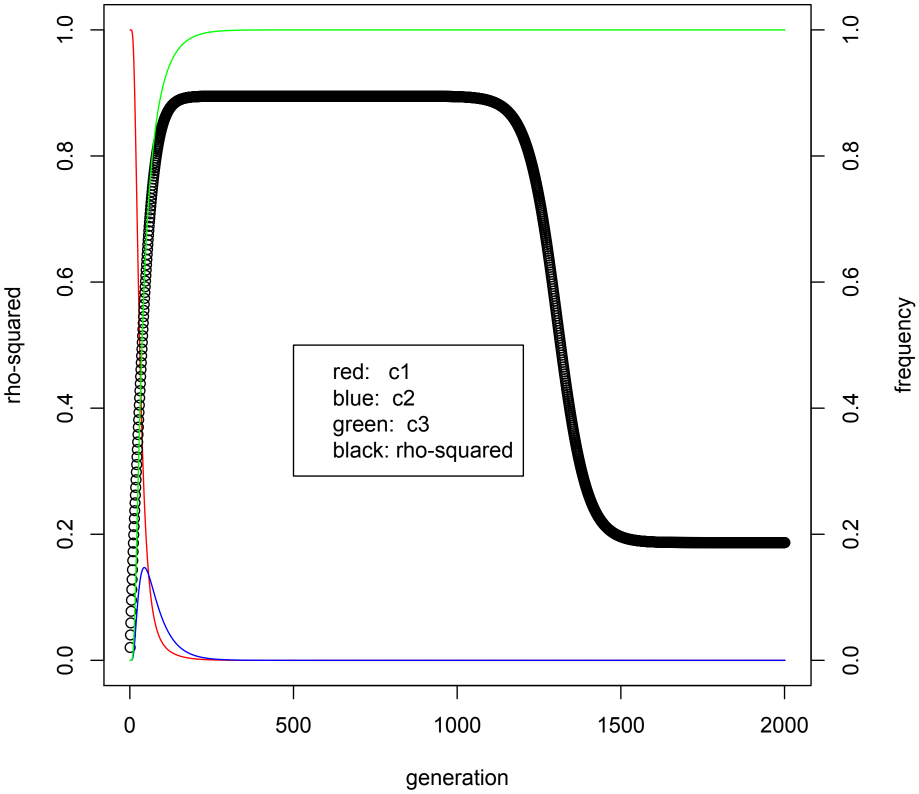

Figure 4: Expected value of ρ^2 by generation in population with effective population size $N_e = 50$, recombination rate $r = 0.001$, mutation rate $u = v = 1e^{-9}$. Frequencies in groups 1-3 are c1-c3.

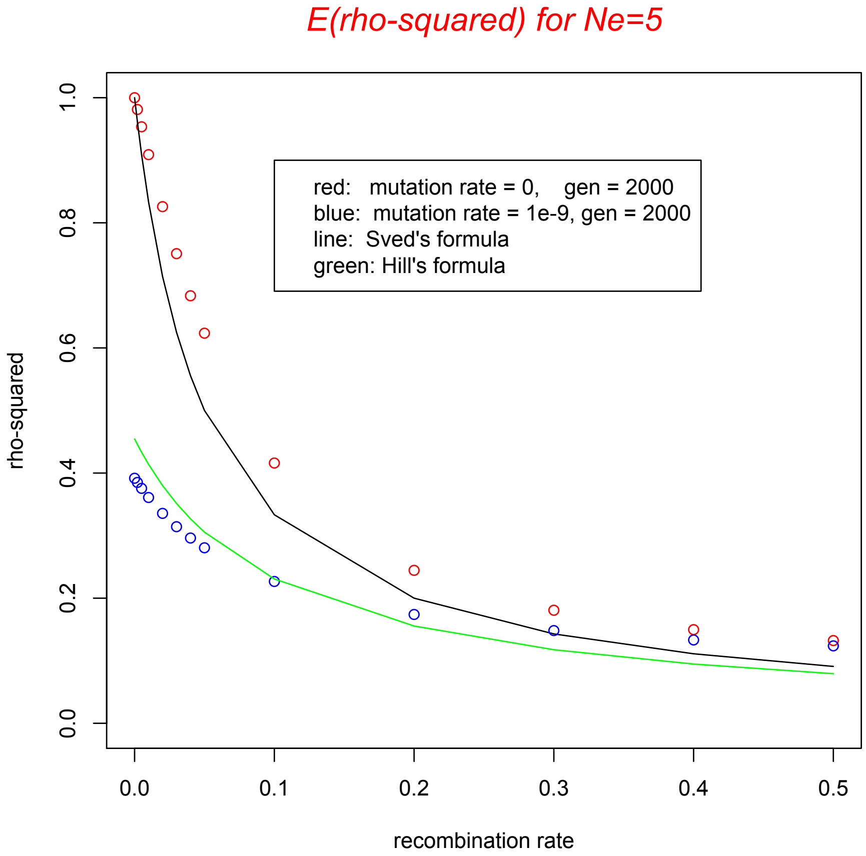

Figure 5: Equilibrium value of ρ^2 by recombination rate in population with effective population size $N_e = 5$ and mutation rate $u = v = 1e^{-9}$ or 0.

The relationship between the equilibrium value of ρ^2, the recombination rate, mutation rate, and effective population size can be seen from figures 5 through 7, where for comparison deterministic formulas by Sved [**?**] and Hill [**?**] are also plotted. When mutation rate is zero, there is good agreement with Sved's formula! When mutation is present, agreement is better with Hill's formula.

2 Bayes Theorem

2.1 Motivation

In whole-genome analyses, the number k of marker covariates typically exceeds the number of n of observations. In this situation, least squares methods cannot

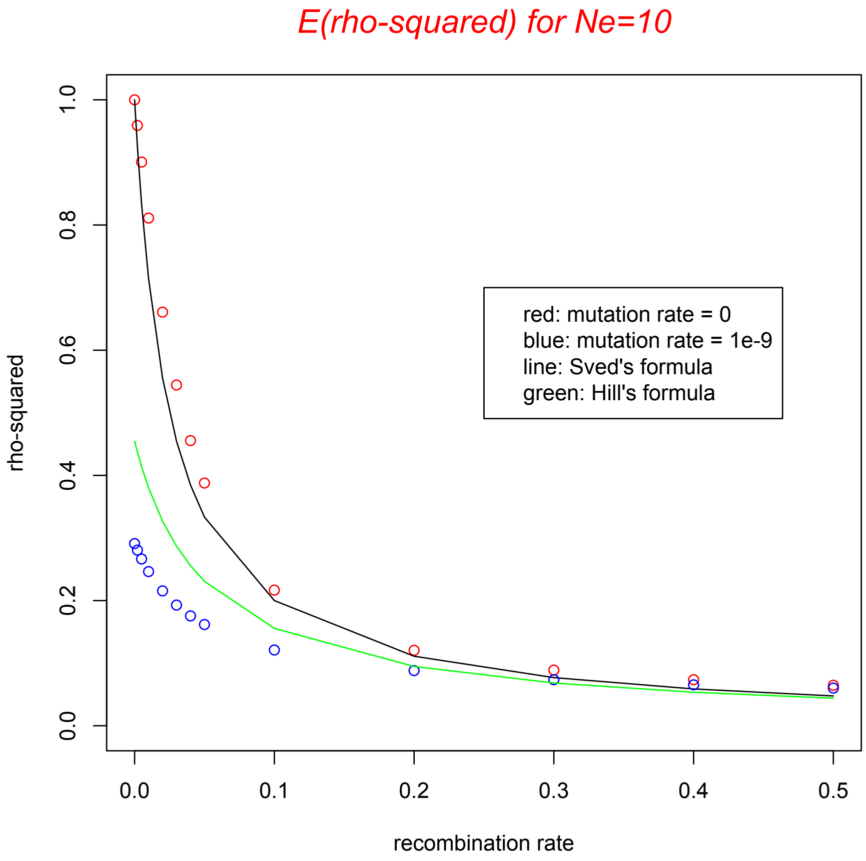

Figure 6: Equilibrium value of ρ^2 by recombination rate in population with effective population size $N_e = 10$ and mutation rate $u = v = 1e^{-9}$ or 0.

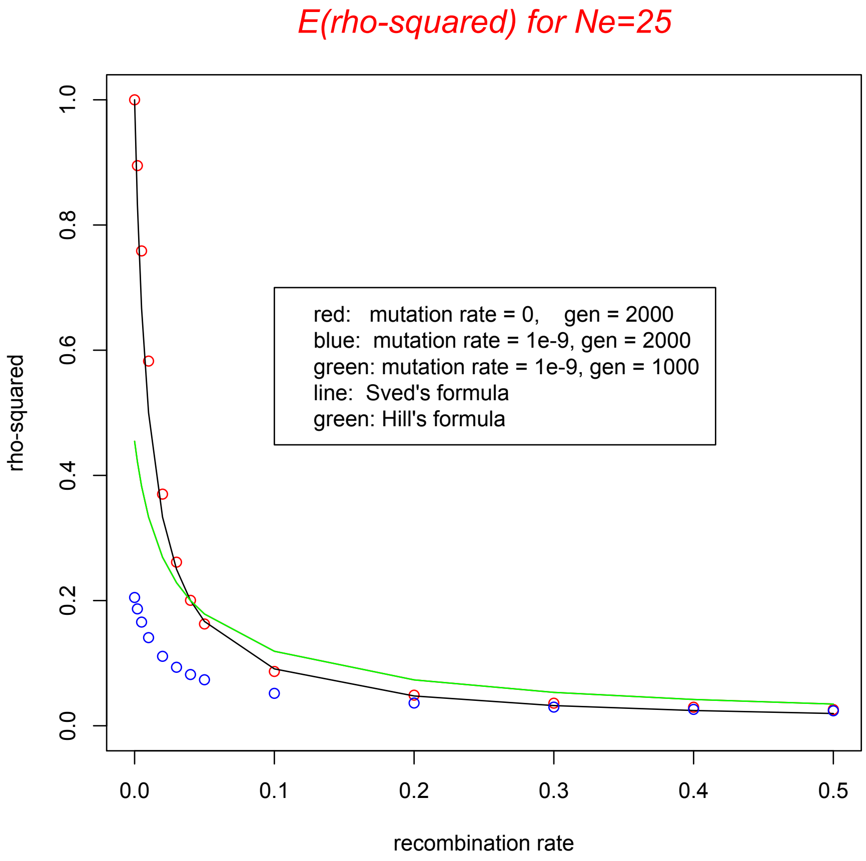

Figure 7: Equilibrium value of ρ^2 by recombination rate in population with effective population size $N_e = 25$ and mutation rate $u = v = 1e^{-9}$ or 0.

be used to simultaneously estimate the effects of all the k marker covariates. One of the most widely used methods to overcome this problem is Bayesian inference, where prior information about marker effects is combined with the data to make inferences about the marker effects. In Bayesian inference, inferences are based on conditional probabilities, and the Bayes theorem is a statement on conditional probability.

2.2 Conditional Probability of X Given Y

Suppose X and $Y$are two random variables with joint probability distribution $\Pr(X, Y)$. Then, the conditional probability of X given Y is given by Bayes theorem as

$$\Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)} \tag{33}$$

where $\Pr(Y)$ is the probability distribution of Y. Similarly, the the conditional probability of Y given X is

$$\Pr(Y|X) = \frac{\Pr(X, Y)}{\Pr(X)},$$

which upon rearranging gives

$$\Pr(X, Y) = \Pr(Y|X)\Pr(X). \tag{34}$$

Then, substituting (34) in (33) gives

$$\begin{aligned}\Pr(X|Y) &= \frac{\Pr(X, Y)}{\Pr(Y)} \\ &= \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)},\end{aligned}$$

which is the form of the formula that is used for inference of X given Y.

2.3 Bayes Theorem by Example

Here we consider a simple example to justify the formula (33). The following table gives the joint distribution of smoking and lung cancer in a hypothetical population of 1,000,000 individuals.

	Smoking		
Cancer	Yes	No	
Yes	42,500	7,500	50,000
No	207,500	742,500	950,000
	250,000	750,000	

Given these numbers, consider how you would compute the relative frequency of lung cancer among smokers. There are a total of 250,000 smokers in this population, and among these 250,000 individuals, 42,500 have lung cancer. So, relative frequency of lung cancer among smokers is $\frac{42{,}500}{250{,}000}$. As we reason below, this relative frequency is also the conditional probability of lung cancer given the individual is a smoker.

1. The frequentist definition of probability of an event is the limiting value of its relative frequency in a "large" number of trials.

2. Suppose we sample with replacement individuals from the 250,000 smokers and compute the relative frequency of the incidence of lung cancer.

3. It can be shown that as the sample size goes to infinity, this relative frequency will approach $\frac{42{,}500}{250{,}000} = 0.17$.

4. This ratio can also be written as

$$\frac{42,500/1,000,000}{250,000/1,000,000} = 0.17.$$

5. The ratio in the numerator is the joint probability of smoking and lung cancer, and the ratio in the denominator is the marginal probability of smoking.

3 Bayesian Inference

3.1 Meaning of Probability in Bayesian Inference

In the frequentist approach, probability is a limiting frequency. Thus, probabilities are always associated with random events. In Bayesian inference, on the other hand, probability is used to quantify your belief that an unobservable variable has a particular value. For example a Bayesian can ask questions such as:

- What is the probability that heritability for milk yield is larger than 0.5?

- What is the probability that variability in milk yield is due to more than 100 loci?

These Bayesian probabilities are not necessarily associated with a random experiment that assigns values to the variables in question.

3.2 Essential Elements of Bayesian Inference

- Bayesian inference starts by specifying what you believe about the parameters or unknowns through prior probabilities. In whole-genome analyses, we will use a prior probability density to quantify our belief that the effect of most marker covariates is zero or close to zero and only a few covariates have effects that deviate from zero.

- These parameters are related to the data through the model or "likelihood", which are conditional probabilities for the data given the parameters. In whole-genome analyses, this is usually a multiple regression model with normally distributed residuals.
- The prior and the likelihood are combined using Bayes theorem to obtain posterior probabilities, which are conditional probabilities for the parameters given the data.
- Inferences about the parameters are based on the posterior.

3.3 Use of Bayes Theorem

- Let $f(\boldsymbol{\theta})$ denote the prior probability density for $\boldsymbol{\theta}$.
- Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ denote the likelihood
- Then, the posterior probability of $\boldsymbol{\theta}$ is:

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \\ &\propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \end{aligned}$$

4 Markov Chain Monte-Carlo Methods

- Often no closed form for $f(\boldsymbol{\theta}|\boldsymbol{y})$
- Further, even if computing $f(\boldsymbol{\theta}|\boldsymbol{y})$ is feasible, obtaining $f(\theta_i|\boldsymbol{y})$ would require integrating over many dimensions
- Thus, in many situations, inferences are made using the empirical posterior constructed by drawing samples from $f(\boldsymbol{\theta}|\boldsymbol{y})$
- Gibbs sampler is widely used for drawing samples from posteriors

4.1 Gibbs Sampler

- Want to draw samples from $f(x_1, x_2, \ldots, x_n)$
- Even though it may be possible to compute $f(x_1, x_2, \ldots, x_n)$, it is difficult to draw samples directly from $f(x_1, x_2, \ldots, x_n)$
- Gibbs:
  - Get valid a starting point $\mathbf{x}^0$

- Draw sample $\mathbf{x}^t$ as:

$$\begin{array}{lll} x_1^t & \text{from} & f(x_1|x_2^{t-1}, x_3^{t-1}, \ldots, x_n^{t-1}) \\ x_2^t & \text{from} & f(x_2|x_1^{t}, x_3^{t-1}, \ldots, x_n^{t-1}) \\ x_3^t & \text{from} & f(x_3|x_1^{t}, x_2^{t}, \ldots, x_n^{t-1}) \\ \vdots & & \vdots \\ x_n^t & \text{from} & f(x_n|x_1^{t}, x_2^{t}, \ldots, x_{n-1}^{t}) \end{array}$$

- The sequence $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^n$ is a Markov chain with stationary distribution $f(x_1, x_2, \ldots, x_n)$

4.2 Making Inferences from Markov Chain

Can show that samples obtained from a Markov chain can be used to draw inferences from $f(x_1, x_2, \ldots, x_n)$ provided the chain is:

- Irreducible: can move from any state i to any other state j
- Positive recurrent: return time to any state has finite expectation
- *Markov Chains*, J. R. Norris (1997)

5 Bayesian Inference by Application to Simple Linear Regression

Simple linear regression is used to illustrate Bayesian inference, using the Gibbs sampler. The Gibbs sampler is used to draw samples from the posterior distribution of the intercept, the slope and the residual variance.

5.1 The Model

Consider the linear model:

$$y_i = \beta_0 + x_i\beta_1 + e_i. \tag{35}$$

where for observation i, y_i is the value of the dependent variable, β_0 is the intercept, x_i is the value of the independent variable and e_i is a residual. Flat priors are used for the intercept and slope, and the residuals are assumed to be identically and independently distributed normal random variables with mean zero and variance σ_e^2. A scaled inverted chi-square prior is used for σ_e^2.

5.2 Simulation of Data

```
n = 20  # number of observations
k = 1  # number of covariates
x = matrix(sample(c(0, 1, 2), n * k, replace = T), nrow = n, ncol = k)
X = cbind(1, x)
head(X)

##      [,1] [,2]
## [1,]    1    1
## [2,]    1    1
## [3,]    1    0
## [4,]    1    0
## [5,]    1    1
## [6,]    1    1

betaTrue = c(1, 2)
y = X %*% betaTrue + rnorm(n, 0, 1)
head(y)

##          [,1]
## [1,]  3.25781
## [2,]  2.98891
## [3,]  0.64307
## [4,] -0.01805
## [5,]  3.01461
## [6,]  3.55116
```

5.3 Least Squares Estimation

In matrix notation, the model (35) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Then, the least-squares estimator of β is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

and the variance of this estimator is

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2.$$

5.3.1 Calculations in R:

```
XPX = t(X) %*% X
rhs = t(X) %*% y
(XPXi = solve(XPX))

##           [,1]     [,2]
## [1,]  0.09848 -0.06061
## [2,] -0.06061  0.07576

(betaHat = XPXi %*% rhs)

##       [,1]
## [1,] 0.417
## [2,] 2.366

eHat = y - X %*% betaHat
(resVar = t(eHat) %*% eHat/(n - 2))

##       [,1]
## [1,] 1.122
```

5.4 Bayesian Inference

Consider making inferences about $\boldsymbol{\beta}$ from $f(\boldsymbol{\beta}|\mathbf{y},\sigma_e^2)$. By using the Bayes theorem, this conditional density is written as

$$
\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{y},\sigma_e^2) &= \frac{f(\mathbf{y}|\boldsymbol{\beta},\sigma_e^2)f(\boldsymbol{\beta})f(\sigma_e^2)}{f(\mathbf{y},\sigma_e^2)} \\
&\propto f(\mathbf{y}|\boldsymbol{\beta},\sigma_e^2)f(\boldsymbol{\beta})f(\sigma_e^2) \\
&\propto f(\mathbf{y}|\boldsymbol{\beta},\sigma_e^2) \\
&= (2\pi\sigma_e^2)^{-n/2}\exp\left\{-\frac{1}{2}\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma_e^2}\right\}, \qquad (36)
\end{aligned}
$$

which looks like the n-dimensional normal density of $\mathbf{y}$ with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{I}\sigma_e^2$. But, $f(\boldsymbol{\beta}|\mathbf{y},\sigma_e^2)$ should be a two-dimensional density. So, the quadratic $Q = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ in the exponent of (36) is rearranged as

$$
\begin{aligned}
Q &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} \\
&= \mathbf{y}'\mathbf{y} + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) - \hat{\boldsymbol{\beta}}^{'}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}},
\end{aligned}
$$

where $\hat{\boldsymbol{\beta}}$ is the solution to $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$, which is the least-squares estimator of $\boldsymbol{\beta}$. In this expression, only the second term depends on $\boldsymbol{\beta}$. Thus, $f(\boldsymbol{\beta}|\mathbf{y},\sigma_e^2)$

can be written as

$$f(\boldsymbol{\beta}|\mathbf{y},\sigma_e^2) \propto \exp\left\{-\frac{1}{2}\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{\sigma_e^2}\right\},$$

which can be recognized as proportional to the density for a two-dimensional normal distribution with mean $\hat{\boldsymbol{\beta}}$ and variance $(\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2$. Thus, in this simple setting, the posterior mean of $\boldsymbol{\beta}$ is given by the least-squares estimate, and drawing samples from the posterior are not needed. But, to illustrate the Gibbs sampler, we will apply it to this simple example.

5.4.1 Gibbs Sampler for $\boldsymbol{\beta}$

The simple regression model can be written as

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{x}\beta_1 + \mathbf{e}.$$

In the Gibbs sampler, β_0 is sampled from its full-conditional posterior: $f(\beta_0|\mathbf{y},\beta_1,\sigma_e^2)$. This conditional distribution is computed for the current values of β_1 and σ_e^2. So, we can write the model as

$$\mathbf{w}_0 = \mathbf{1}\beta_0 + \mathbf{e},$$

where $\mathbf{w}_0 = \mathbf{y} - \mathbf{x}\beta_1$. Then, the least-squares estimator of β_0 is

$$\hat{\beta}_0 = \frac{\mathbf{1}'\mathbf{w}_0}{\mathbf{1}'\mathbf{1}},$$

and the variance of this estimator is

$$Var(\hat{\beta}_0) = \frac{\sigma_e^2}{\mathbf{1}'\mathbf{1}}.$$

By applying the strategy used to derive $f(\boldsymbol{\beta}|\mathbf{y},\sigma_e^2)$ above, the full-conditional posterior for β_0 can be shown to be a normal distribution with mean $\hat{\beta}_0$ and variance $\frac{\sigma_e^2}{\mathbf{1}'\mathbf{1}}$. Similarly, the full-conditional posterior for β_1 is a normal distribution with mean

$$\hat{\beta}_1 = \frac{\mathbf{x}'\mathbf{w}_1}{\mathbf{x}'\mathbf{x}}$$

and variance $\frac{\sigma_e^2}{\mathbf{x}'\mathbf{x}}$, where $\mathbf{w}_1 = \mathbf{y} - 1\beta_0$. In the calculations below, we will use the true value of σ_e^2.

5.4.2 Calculations in R:

```
beta = c(0, 0)  # starting values for beta
# loop for Gibbs sampler
niter = 10000  # number of samples
```

```
meanBeta = c(0, 0)
for (iter in 1:niter) {
    # sampling intercept
    w = y - X[, 2] * beta[2]
    x = X[, 1]
    xpxi = 1/(t(x) %*% x)
    betaHat = t(x) %*% w * xpxi
    beta[1] = rnorm(1, betaHat, sqrt(xpxi))  # using residual var = 1
    # sampling slope
    w = y - X[, 1] * beta[1]
    x = X[, 2]
    xpxi = 1/(t(x) %*% x)
    betaHat = t(x) %*% w * xpxi
    beta[2] = rnorm(1, betaHat, sqrt(xpxi))  # using residual var = 1
    meanBeta = meanBeta + beta
    if ((iter%%1000) == 0) {
        cat(sprintf("Intercept = %6.3f \n", meanBeta[1]/iter))
        cat(sprintf("Slope     = %6.3f \n", meanBeta[2]/iter))
    }
}

## Intercept =  0.429
## Slope     =  2.352
## Intercept =  0.408
## Slope     =  2.371
## Intercept =  0.413
## Slope     =  2.371
## Intercept =  0.413
## Slope     =  2.371
## Intercept =  0.409
## Slope     =  2.374
## Intercept =  0.409
## Slope     =  2.375
## Intercept =  0.412
## Slope     =  2.372
## Intercept =  0.413
## Slope     =  2.370
## Intercept =  0.414
## Slope     =  2.370
## Intercept =  0.414
## Slope     =  2.369
```

5.4.3 Full-conditional Posterior for σ_e^2

Recall that we assumed a scaled inverted chi-square prior for σ_e^2. The density function for this is:

$$f(\sigma_e^2) = \frac{(S_e^2\nu_e/2)^{\nu_e/2}}{\Gamma(\nu_e/2)}(\sigma_e^2)^{-(2+\nu_e)/2}\exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\}, \tag{37}$$

where S_e^2 and ν_e are the scale and the degrees of freedom parameters for this distribution. Applying Bayes theorem to combine this prior with the "likelihood" (given in (36)), the full-conditional posterior for the residual variance can be written as

$$\begin{aligned}
f(\sigma_e^2|\mathbf{y},\boldsymbol{\beta}) &= \frac{f(\mathbf{y}|\boldsymbol{\beta},\sigma_e^2)f(\boldsymbol{\beta})f(\sigma_e^2)}{f(\mathbf{y},\boldsymbol{\beta})} \\
&\propto f(\mathbf{y}|\boldsymbol{\beta},\sigma_e^2)f(\boldsymbol{\beta})f(\sigma_e^2) \\
&\propto (\sigma_e^2)^{-n/2}\exp\left\{-\frac{1}{2}\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma_e^2}\right\} \\
&\times (\sigma_e^2)^{-(2+\nu_e)/2}\exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\} \\
&= (\sigma_e^2)^{-(n+2+\nu_e)/2}\exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\nu_e S_e^2}{2\sigma_e^2}\right\}. \quad (38)
\end{aligned}$$

Comparing (38) with (37), can see that it is proportional to a scaled inverse chi-squared density with $\tilde{\nu}_e = n+\nu_e$ degrees of freedom and $\tilde{S}_e^2 = \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\nu_e S_e^2}{\tilde{\nu}_e}$ scale parameter. A sample from this density can be obtained as $\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\nu_e S_e^2}{\chi^2_{\tilde{\nu}_e}}$, where $\chi^2_{\tilde{\nu}_e}$ is a chi-squared random variable with $\tilde{\nu}_e$ degrees of freedom.

5.4.4 Exercise

In the R script given here, the simulated value of the residual variance was used in the sampling of $\boldsymbol{\beta}$. Extend this script to also sample σ_e^2 from its full-conditional posterior given above. In R, rchisq(1,ν) gives a chi-squared random variable with ν degrees of freedom.

5.5 Model with Normal Prior for Slope

Here we consider a model with a flat prior for β_0 and a normal prior for the slope:

$$\beta_1 \sim N(0,\sigma_\beta^2),$$

where σ^2_β is assumed to be known. Then, the full-conditional posterior for $\theta' = [\boldsymbol{\beta}, \sigma^2_e]$ is

$$\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\
&\propto \left(\sigma^2_e\right)^{-n/2} \exp\left\{-\frac{(\mathbf{y}-\mathbf{1}\beta_0-\mathbf{x}\beta_1)'(\mathbf{y}-\mathbf{1}\beta_0-\mathbf{x}\beta_1)}{2\sigma^2_e}\right\} \\
&\times \left(\sigma^2_\beta\right)^{-1/2} \exp\left\{-\frac{\beta_1^2}{2\sigma^2_\beta}\right\} \\
&\times (\sigma^2_e)^{-(2+\nu_e)/2} \exp\left\{-\frac{\nu_e S^2_e}{2\sigma^2_e}\right\}.
\end{aligned}$$

5.5.1 Full-conditional for β_1:

The full-conditional for β_1 is obtained by dropping all terms and factors that do not involve β_1:

$$\begin{aligned}
f(\beta_1|\text{ELSE}) &\propto \exp\left\{-\frac{(\mathbf{y}-\mathbf{1}\beta_0-\mathbf{x}\beta_1)'(\mathbf{y}-\mathbf{1}\beta_0-\mathbf{x}\beta_1)}{2\sigma^2_e}\right\} \\
&\times \exp\left\{-\frac{\beta_1^2}{2\sigma^2_\beta}\right\} \\
&\propto \exp\left\{-\frac{\mathbf{w}'\mathbf{w}-2\mathbf{w}'\mathbf{x}\beta_1+\beta_1^2(\mathbf{x}'\mathbf{x}+\sigma^2_e/\sigma^2_\beta)}{2\sigma^2_e}\right\} \\
&\propto \exp\left\{-\frac{\mathbf{w}'\mathbf{w}-(\beta_1-\hat{\beta}_1)^2(\mathbf{x}'\mathbf{x}+\sigma^2_e/\sigma^2_\beta)-\hat{\beta}_1^2(\mathbf{x}'\mathbf{x}+\sigma^2_e/\sigma^2_\beta)}{2\sigma^2_e}\right\} \\
&\propto \exp\left\{-\frac{(\beta_1-\hat{\beta}_1)^2}{\frac{2\sigma^2_e}{(\mathbf{x}'\mathbf{x}+\sigma^2_e/\sigma^2_\beta)}}\right\},
\end{aligned}$$

where

$$\hat{\beta}_1 = \frac{\mathbf{x}'\mathbf{w}}{(\mathbf{x}'\mathbf{x}+\sigma^2_e/\sigma^2_\beta)},$$

and $\mathbf{w} = \mathbf{y}-\mathbf{1}\beta_0$. So, the full-conditional posterior for β_1 is a normal distribution with mean $\hat{\beta}_1$ and variance $\frac{\sigma^2_e}{(\mathbf{x}'\mathbf{x}+\sigma^2_e/\sigma^2_\beta)}$.

5.5.2 Exercise

1. Use R to simulate a vector of 1000 values for β_1 from a normal distribution with mean zero and variance 3. Use the rnorm command for this. Plot a histogram of these values. Use the command hist for drawing a histogram.

2. Using a value of 1 for β_0 and one of the sampled values of β_1, generate a vector of observations, $\mathbf{y}$, that follows a simple linear regression model. Use $\sigma_e^2 = 5$ to simulate $\mathbf{y}$.

3. Use the Gibbs sampler to draw 10,000 samples for β_1 from its posterior distribution.

   (a) Compute the mean and variance of the sampled values.

   (b) Draw a histogram of the sampled values. Compare with prior.

6 Extension to Multiple Linear Regression

Consider the multiple regression model

$$y_i = \beta_0 + \sum_j x_j \beta_j + e_i, \tag{39}$$

which extends model (35) to include multiple covariates x_j. In matrix notation, this model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \ldots, \beta_k]$ and the matrix $\mathbf{X}$ contains the corresponding covariates.

6.1 Model with Normal Prior for Regression Coefficients

Here we consider a model with a flat prior for β_0 and iid normal priors for the slopes:

$$\beta_j \sim N(0, \sigma_\beta^2) \,\text{for}\, j = 1, 2, \ldots, k,$$

where σ_β^2 is assumed to be known. The residuals are assumed iid normal with null mean and variance σ_e^2, which itself is assigned a scaled inverted chi-square prior. Then, the joint posterior for $\boldsymbol{\theta}$ is

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \\ &\propto \left(\sigma_e^2\right)^{-n/2} \exp\left\{ -\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_e^2} \right\} \\ &\times \left(\sigma_\beta^2\right)^{-k/2} \exp\left\{ -\frac{\sum_{j=1}^k \beta_j^2}{2\sigma_\beta^2} \right\} \\ &\times \left(\sigma_e^2\right)^{-(2+\nu_e)/2} \exp\left\{ -\frac{\nu_e S_e^2}{2\sigma_e^2} \right\}. \end{aligned}$$

The posterior distribution for $\boldsymbol{\beta}$ can be written as

$$\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{y},\sigma^2_\beta,\sigma^2_e) &= \frac{f(\mathbf{y}|\boldsymbol{\beta},\sigma^2_\beta,\sigma^2_e)f(\boldsymbol{\beta}|\sigma^2_\beta)f(\sigma^2_e)}{f(\mathbf{y},\sigma^2_\beta,\sigma^2_e)} \\
&\propto f(\mathbf{y}|\boldsymbol{\beta},\sigma^2_\beta,\sigma^2_e)f(\boldsymbol{\beta}|\sigma^2_\beta)f(\sigma^2_e) \\
&\propto f(\mathbf{y}|\boldsymbol{\beta},\sigma^2_\beta,\sigma^2_e)f(\boldsymbol{\beta}|\sigma^2_\beta) \\
&\propto \left(\sigma^2_e\right)^{-n/2}\exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2_e}\right\} \\
&\times \left(\sigma^2_\beta\right)^{-k/2}\exp\left\{-\frac{\sum_{j=1}^k\beta_j^2}{2\sigma^2_\beta}\right\} \\
&\propto \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\sum_{j=1}^k\beta_j^2\frac{\sigma^2_e}{\sigma^2_\beta}}{2\sigma^2_e}\right\} \\
&\propto \exp\left\{-\frac{\mathbf{y}'\mathbf{y}-2\mathbf{y}'\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X}+\mathbf{D}\frac{\sigma^2_e}{\sigma^2_\beta})\boldsymbol{\beta}}{2\sigma^2_e}\right\} \\
&\propto \exp\left\{-\frac{\mathbf{y}'\mathbf{y}-(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X}+\mathbf{D}\frac{\sigma^2_e}{\sigma^2_\beta})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})-\hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X}+\mathbf{D}\frac{\sigma^2_e}{\sigma^2_\beta})\hat{\boldsymbol{\beta}}}{2\sigma^2_e}\right\} \\
&\propto \exp\left\{-\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X}+\mathbf{D}\frac{\sigma^2_e}{\sigma^2_\beta})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{2\sigma^2_e}\right\},
\end{aligned}$$

for

$$(\mathbf{X}'\mathbf{X}+\mathbf{D}\frac{\sigma^2_e}{\sigma^2_\beta})\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{y}, \tag{40}$$

where $\mathbf{D}$ is a diagonal matrix with zero on the first diagonal and ones on the remaining diagonals. Thus, the full-conditional posterior for $\boldsymbol{\beta}$ is a normal distribution with mean given by (40) and variance $(\mathbf{X}'\mathbf{X}+\mathbf{D}\frac{\sigma^2_e}{\sigma^2_\beta})^{-1}\sigma^2_e$.

6.1.1 Full-conditionals:

The full conditionals for β_0 and σ^2_e are identical to those in simple linear regression.

Full-conditional for β_j

The full-conditional for β_j is obtained by dropping from the joint posterior all terms and factors that do not involve β_j:

$$\begin{aligned}
f(\beta_1|\text{ELSE}) &\propto \exp\left\{-\frac{(\mathbf{w}_j - \mathbf{x}_j\beta_j)'(\mathbf{w}_j - \mathbf{x}_j\beta_j)}{2\sigma_e^2}\right\} \\
&\times \exp\left\{-\frac{\beta_j^2}{2\sigma_\beta^2}\right\} \\
&\propto \exp\left\{-\frac{\mathbf{w}_j'\mathbf{w}_j - 2\mathbf{w}_j'\mathbf{x}_j\beta_j + \beta_j^2(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\beta^2)}{2\sigma_e^2}\right\} \\
&\propto \exp\left\{-\frac{\mathbf{w}_j'\mathbf{w}_j - (\beta_j - \hat{\beta}_j)^2(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\beta^2) - \hat{\beta}_j^2(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\beta^2)}{2\sigma_e^2}\right\} \\
&\propto \exp\left\{-\frac{(\beta_j - \hat{\beta}_j)^2}{\frac{2\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\beta^2)}}\right\},
\end{aligned}$$

where

$$\hat{\beta}_j = \frac{\mathbf{x}_j'\mathbf{w}_j}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\beta^2)},$$

and $\mathbf{w}_j = \mathbf{y} - \sum_{l \neq j} \mathbf{x}_l\beta_l$. So, the full-conditional posterior for β_j is a normal distribution with mean $\hat{\beta}_j$ and variance $\frac{\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\beta^2)}$.

6.2 Exercise

1. Generate data from model (39) with $k = 10$ covariates.
2. Setup and solve the mixed model equations given by (40).
3. Sample the elements of $\boldsymbol{\beta}$ using Gibbs.
4. Compute the posterior mean of $\boldsymbol{\beta}$ from the samples and compare with mixed model solutions.
5. Compute the posterior covariance matrix from the sampled values. Compare results with inverse of the mixed-model coefficient matrix.

6.3 Model with unknown σ_β^2

In the previous section, we assumed that σ_β^2 in the prior of the slopes was known. Here, we will consider this variance to be unknown with a scaled inverted chi-square prior with scale parameter S_β^2 and degrees of freedom ν_β. The joint posterior for this model is

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\
&\propto \left(\sigma_e^2\right)^{-n/2} \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma_e^2}\right\} \\
&\times \left(\sigma_\beta^2\right)^{-k/2} \exp\left\{-\frac{\sum_{j=1}^k \beta_j^2}{2\sigma_\beta^2}\right\} \\
&\times \left(\sigma_\beta^2\right)^{-(2+\nu_\beta)/2} \exp\left\{-\frac{\nu_\beta S_\beta^2}{2\sigma_\beta^2}\right\} \\
&\times \left(\sigma_e^2\right)^{-(2+\nu_e)/2} \exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\}.
\end{aligned}
$$

Then, the full-conditional posterior for σ_β^2 is

$$
\begin{aligned}
f(\sigma_\beta^2|\mathbf{y},\boldsymbol{\beta},\sigma_e^2) &\propto \left(\sigma_\beta^2\right)^{-k/2} \exp\left\{-\frac{\sum_{j=1}^k \beta_j^2}{2\sigma_\beta^2}\right\} \\
&\times \left(\sigma_\beta^2\right)^{-(2+\nu_\beta)/2} \exp\left\{-\frac{\nu_\beta S_\beta^2}{2\sigma_\beta^2}\right\} \\
&\propto \left(\sigma_\beta^2\right)^{-(2+k+\nu_\beta)/2} \exp\left\{-\frac{\sum_{j=1}^k \beta_j^2 + \nu_\beta S_\beta^2}{2\sigma_\beta^2}\right\},
\end{aligned}
$$

which can be recognized as a scaled inverted chi-square distribution with $\tilde{\nu}_\beta = k + \nu_\beta$ degrees of freedom and scale parameter $\tilde{S}_\beta^2 = (\sum_{j=1}^k \beta_j^2 + \nu_\beta S_\beta^2)/\tilde{\nu}_\beta$. A sample from this posterior can be obtained as $\frac{\sum_{j=1}^k \beta_j^2 + \nu_\beta S_\beta^2}{\chi_{\tilde{\nu}_\beta}^{-2}}$.

6.3.1 Exercise

Extend the sampler used in the previous section to treat σ_β^2 as an unknown. Plot the posterior distribution for σ_β^2.

6.4 Model with unknown covariate-specific variances

Here we consider a model where the prior for the slope corresponding to covariate j is normal with mean 0 and variance σ_j^2, where σ_j^2 has scaled inverted chi-square prior with scale parameter S_β^2 and degrees of freedom ν_β. The joint posterior for this model is

$$\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\
&\propto \left(\sigma_e^2\right)^{-n/2}\exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma_e^2}\right\} \\
&\times \prod_{j=1}^{k}\left(\sigma_j^2\right)^{-1/2}\exp\left\{-\frac{\beta_j^2}{2\sigma_j^2}\right\} \\
&\times \prod_{j=1}^{k}\left(\sigma_j^2\right)^{-(2+\nu_\beta)/2}\exp\left\{-\frac{\nu_\beta S_\beta^2}{2\sigma_j^2}\right\} \\
&\times (\sigma_e^2)^{-(2+\nu_e)/2}\exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\}.
\end{aligned}$$

It can be shown that:

1. The full-conditional posterior for β_j is normal with mean

$$\hat{\beta}_j = \frac{\mathbf{x}_j'\mathbf{w}_j}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_j^2)},$$

and variance $\frac{\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j+\sigma_e^2/\sigma_j^2)}$.

2. The full-conditional posterior for σ_j^2 is a scaled inverted chi-square distribution with $\tilde{\nu}_\beta = 1 + \nu_\beta$ degrees of freedom and scale parameter $\tilde{S}_\beta^2 = (\beta_j^2 + \nu_\beta S_\beta^2)/\tilde{\nu}_\beta$. A sample from this posterior can be obtained as $\frac{\beta_j^2+\nu_\beta S_\beta^2}{\chi_{\tilde{\nu}_\beta}^{-2}}$.

3. Marginally, the prior for β_j is a scaled t distribution with ν_β degrees of freedom, mean 0 and scale parameter S_β^2.

6.4.1 Exercise

1. Derive the full-conditional posterior for β_j.

2. Derive the full-conditional posterior for σ_j^2.

3. Use a Gibbs sampler to compute the posterior mean of $\boldsymbol{\beta}$.

6.5 Model with Mixture Prior for Regression Coefficients

As before, a flat prior is used for the intercept, μ. The prior for slope j is a mixture:

$$\beta_j = \begin{cases} 0 & \text{probability}\,\pi \\ \sim N(0,\sigma_\beta^2) & \text{probability}\,(1-\pi) \end{cases},$$

where σ^2_β has a scaled inverted chi-square prior with scale parameter S^2_β and degrees of freedom ν_β. In order to use the Gibbs sampler, it is convenient to write β_j as

$$\beta_j = \delta_j \gamma_j,$$

where δ_j is a Bernoulli variable with probability $1 - \pi$ of being 1:

$$\delta_j = \begin{cases} 0 & \text{probability}\,\pi \\ 1 & \text{probability}\,(1-\pi) \end{cases},$$

and γ_j is normally distributed with mean zero and variance σ^2_β. Then, the model for the phenotypic values can be written as

$$y_i = \mu + \sum_{j=1} X_{ij}\gamma_j\delta_j + e_i.$$

6.5.1 Full-conditionals:

The joint posterior for all the parameters is proportional to

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\ &\propto \left(\sigma^2_e\right)^{-n/2} \exp\left\{-\frac{\mathbf{y} - \mathbf{1}\mu - \sum \mathbf{X}_j\gamma_j\delta_j}{2\sigma^2_e}\right\} \\ &\times \prod_{j=1}^{k} \left(\sigma^2_\beta\right)^{-1/2} \exp\left\{-\frac{\gamma^2_j}{2\sigma^2_\beta}\right\} \\ &\times \prod_{j=1}^{k} \pi^{(1-\delta_j)}(1-\pi)^{\delta_j} \\ &\times (\sigma^2_\beta)^{-(\nu_\beta+2)/2} \exp\left\{-\frac{\nu_\beta S^2_\beta}{2\sigma^2_\beta}\right\} \\ &\times (\sigma^2_e)^{-(2+\nu_e)/2} \exp\left\{-\frac{\nu_e S^2_e}{2\sigma^2_e}\right\}, \end{aligned}$$

where $\boldsymbol{\theta}$ denotes all the unknowns.

6.5.2 Full-conditional for μ

The full-conditional for μ is a normal distribution with mean $\hat{\mu}$ and variance $\frac{\sigma^2_e}{n}$, where $\hat{\mu}$ is the least-squares estimate of μ in the model

$$\mathbf{y} - \sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{X}_j\gamma_j\delta_j = \mathbf{1}\mu + \mathbf{e},$$

and $\frac{\sigma^2_e}{n}$ is the variance of this estimator (n is the number of observations).

6.5.3 Full-conditional for γ_j

$$
\begin{aligned}
f(\gamma_j|\text{ELSE}) &\propto& \exp\left\{-\frac{(\mathbf{w}_j - \mathbf{X}_j\gamma_j\delta_j)'(\mathbf{w}_j - \mathbf{X}_j\gamma_j\delta_j)}{2\sigma_e^2}\right\} \\
&\times& \exp\left\{-\frac{\gamma_j^2}{2\sigma_\beta^2}\right\} \\
&\propto& \exp\left\{-\frac{[\mathbf{w}_j'\mathbf{w}_j - 2\mathbf{w}_j'\mathbf{X}_j\gamma_j\delta_j + \gamma_j^2(\mathbf{x}_j'\mathbf{x}_j\delta_j + \sigma_e^2/\sigma_\beta^2)]}{2\sigma_e^2}\right\} \\
&\propto& \exp\left\{-\frac{(\gamma_j - \hat{\gamma}_j)^2}{\frac{2\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j\delta_j + \sigma_e^2/\sigma_\beta^2)}}\right\},
\end{aligned}
$$

where

$$\mathbf{w}_j = \mathbf{y} - \mathbf{1}\mu - \sum_{l\neq j}\mathbf{X}_l\gamma_l\delta_l.$$

So, the full-conditional for γ_j is a normal distribution with mean

$$\hat{\gamma}_j = \frac{\mathbf{X}_j'\mathbf{w}_j\delta_j}{(\mathbf{x}_j'\mathbf{x}_j\delta_j + \sigma_e^2/\sigma_\beta^2)}$$

and variance $\frac{\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j\delta_j + \sigma_e^2/\sigma_\beta^2)}$.

6.5.4 Full-conditional for δ_j

$$\Pr(\delta_j = 1|\text{ELSE}) \propto \frac{h(\delta_j = 1)}{h(\delta_j = 1) + h(\delta_j = 0)},$$

where

$$h(\delta_j) = \pi^{(1-\delta_j)}(1-\pi)^{\delta_j}\exp\left\{-\frac{\mathbf{w}_j - \mathbf{X}_j\gamma_j\delta_j}{2\sigma_e^2}\right\}.$$

6.5.5 Full-conditional for σ_β^2

$$
\begin{aligned}
f(\sigma_j^2|\text{ELSE}) &\propto& \left(\sigma_\beta^2\right)^{-k/2}\exp\left\{-\frac{\sum_{j=1}^k\gamma_j^2}{2\sigma_\beta^2}\right\} \\
&\times& (\sigma_\beta^2)^{-(\nu_\beta+2)/2}\exp\left\{-\frac{\nu_\beta S_\beta^2}{2\sigma_\beta^2}\right\} \\
&\propto& (\sigma_\beta^2)^{-(k+\nu_\beta+2)/2}\exp\left\{-\frac{\sum_{j=1}^k\gamma_j^2 + \nu_\beta S_\beta^2}{2\sigma_j^2}\right\},
\end{aligned}
$$

and this is proportional to a scaled inverted chi-square distribution with $\tilde{\nu}_j = \nu_\beta + k$ and scale parameter $\tilde{S}_j^2 = (\sum_{j=1}^k \gamma_j^2 + \nu_\beta S_\beta^2)/\tilde{\nu}_j$.

6.5.6 Full-conditional for π

$$f(\pi|ELSE) \propto \pi^{(k-\sum_{j=1}^k \delta_j)}(1-\pi)^{\sum_{j=1}^k \delta_j},$$

which is proportional to a Beta distribution with parameters $a = k - \sum_{j=1}^k \delta_j + 1$ and $b = \sum \delta_j + 1$.

6.5.7 Full-conditional for σ_e^2

$$\begin{aligned} f(\sigma_e^2|\text{ELSE}) &\propto (\sigma_e^2)^{-n/2} \exp\left\{-\frac{\mathbf{y} - \mathbf{1}\mu - \sum \mathbf{X}_j\gamma_j\delta_j}{2\sigma_e^2}\right\} \\ &\times (\sigma_e^2)^{-(2+\nu_e)/2} \exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\} \\ &\propto (\sigma_e^2)^{-(n+2+\nu_e)/2} \exp\left\{-\frac{(\mathbf{y} - \mathbf{1}\mu - \sum \mathbf{X}_j\gamma_j\delta_j)'(\mathbf{y} - \mathbf{1}\mu - \sum \mathbf{X}_j\gamma_j\delta_j) + \nu_e S_e^2}{2\sigma_e^2}\right\}, \end{aligned}$$

which is proportional to a scaled inverted chi-square density with $\tilde{\nu}_e = n + \nu_e$ degrees of freedom and $\tilde{S}_e^2 = \frac{(\mathbf{y}-\mathbf{1}\mu-\sum \mathbf{X}_j\gamma_j\delta_j)'(\mathbf{y}-\mathbf{1}\mu-\sum \mathbf{X}_j\gamma_j\delta_j)+\nu_e S_e^2}{\tilde{\nu_e}}$ scale parameter.

7 Bayesian Regression Models for Whole-Genome Analyses

Meuwissen et al. [12] introduced three regression models for whole-genome prediction of breeding value of the form

$$y_i = \mu + \sum_{j=1}^k X_{ij}\alpha_j + e_i,$$

where y_i is the phenotypic value, μ is the intercept, X_{ij} is j^{th} marker covariate of animal i, α_j is the partial regression coefficient of X_{ij}, and e_i are identically and independently distributed residuals with mean zero and variance σ_e^2. In most current analyses, X_{ij} are SNP genotype covariates that can be coded as 0, 1 and 2, depending on the number of B alleles at SNP locus j.

In all three of their models, a flat prior was used for the intercept and a scaled inverted chi-square distribution for σ_e^2. The three models introduced by Meuwissen et al. [12] differ only in the prior used for α_j.

7.1 BLUP

In their first model, which they called BLUP, a normal distribution with mean zero and known variance, σ^2_α, is used as the prior for α_j.

7.1.1 The meaning of σ^2_α

Assume the QTL are in the marker panel. Then, the genotypic value g_i for a randomly sampled animal i can be written as

$$g_i = \mu + \mathbf{x}'_i\boldsymbol{\alpha},$$

where $\mathbf{x}'_i$ is the vector of SNP genotype covariates and $\boldsymbol{\alpha}$ is the vector of regression coefficients. Note that randomly sampled animals differ only in $\mathbf{x}'_i$ and have $\boldsymbol{\alpha}$ in common. Thus, genotypic variability is entirely due to variability in the genotypes of animals. So, σ^2_α is not the genetic variance at a locus [2, 4].

7.1.2 Relationship of σ^2_α to genetic variance

Assume loci with effect on trait are in linkage equilibrium. Then, the additive genetic variance is

$$V_A = \sum_j^k 2p_jq_j\alpha_j^2,$$

where $p_j = 1 - q_j$ is gene frequency at SNP locus j. Letting $U_j = 2p_jq_j$ and $V_j = \alpha_j^2$,

$$V_A = \sum_j^k U_jV_j.$$

For a randomly sampled locus, covariance between U_j and V_j is

$$C_{UV} = \frac{\sum_j U_jV_j}{k} - (\frac{\sum_j U_j}{k})(\frac{\sum_j V_j}{k})$$

Rearranging this expression for C_{UV} gives

$$\sum_j U_jV_j = kC_{UV} + (\sum_j U_j)(\frac{\sum_j V_j}{k})$$

So,

$$V_A = kC_{UV} + (\sum_j 2p_jq_j)(\frac{\sum_j \alpha_j^2}{k}).$$

Letting $\sigma^2_\alpha = \frac{\sum_j \alpha_j^2}{k}$ gives

$$V_A = kC_{UV} + (\sum_j 2p_jq_j)\sigma^2_\alpha$$

and

$$\sigma^2_\alpha = \frac{V_A - kC_{UV}}{\sum_j 2p_jq_j},$$

which gives

$$\sigma^2_\alpha = \frac{V_A}{\sum_j 2p_jq_j},$$

if gene frequency is independent of the effect of the gene.

7.1.3 Full-conditionals:

The joint posterior for all the parameters is proportional to

$$\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\
&\propto \left(\sigma^2_e\right)^{-n/2} \exp\left\{-\frac{\mathbf{y} - \mathbf{1}\mu - \sum \mathbf{X}_j\alpha_j}{2\sigma^2_e}\right\} \\
&\times \prod_{j=1}^{k} \left(\sigma^2_\alpha\right)^{-1/2} \exp\left\{-\frac{\alpha_j^2}{2\sigma^2_\alpha}\right\} \\
&\times \left(\sigma^2_\alpha\right)^{-(\nu_\alpha+2)/2} \exp\left\{-\frac{\nu_\alpha S^2_\alpha}{2\sigma^2_\alpha}\right\} \\
&\times \left(\sigma^2_e\right)^{-(2+\nu_e)/2} \exp\left\{-\frac{\nu_e S^2_e}{2\sigma^2_e}\right\},
\end{aligned}$$

where $\boldsymbol{\theta}$ denotes all the unknowns.

7.1.4 Full-conditional for μ

The full-conditional for μ is a normal distribution with mean $\hat{\mu}$ and variance $\frac{\sigma^2_e}{n}$, where $\hat{\mu}$ is the least-squares estimate of μ in the model

$$\mathbf{y} - \sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{X}_j\alpha_j = \mathbf{1}\mu + \mathbf{e},$$

and $\frac{\sigma^2_e}{n}$ is the variance of this estimator (n is the number of observations).

7.1.5 Full-conditional for α_j

$$\begin{aligned}
f(\beta_j|\text{ELSE}) &\propto \exp\left\{-\frac{(\mathbf{w}_j - \mathbf{X}_j\alpha_j)'(\mathbf{w}_j - \mathbf{X}_j\alpha_j)}{2\sigma_e^2}\right\} \\
&\times \exp\left\{-\frac{\alpha_j^2}{2\sigma_\alpha^2}\right\} \\
&\propto \exp\left\{-\frac{[\mathbf{w}_j'\mathbf{w}_j - 2\mathbf{w}_j'\mathbf{X}_j\alpha_j + \alpha_j^2(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\alpha^2)]}{2\sigma_e^2}\right\} \\
&\propto \exp\left\{-\frac{(\alpha_j - \hat{\alpha_j})^2}{\frac{2\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\alpha^2)}}\right\},
\end{aligned}$$

where

$$\mathbf{w}_j = \mathbf{y} - \mathbf{1}\mu - \sum_{l \neq j} \mathbf{X}_l\alpha_l.$$

So, the full-conditional for α_j is a normal distribution with mean

$$\hat{\alpha}_j = \frac{\mathbf{X}_j'\mathbf{w}_j}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\alpha^2)}$$

and variance $\frac{\sigma_e^2}{(\mathbf{x}_j'\mathbf{x}_j + \sigma_e^2/\sigma_\alpha^2)}$.

7.1.6 Full-conditional for σ_α^2

$$\begin{aligned}
f(\sigma_\alpha^2|\text{ELSE}) &\propto \times \prod_{j=1}^{k} \left(\sigma_\alpha^2\right)^{-1/2} \exp\left\{-\frac{\alpha_j^2}{2\sigma_\alpha^2}\right\} \\
&\times (\sigma_\alpha^2)^{-(\nu_\alpha+2)/2} \exp\left\{-\frac{\nu_\alpha S_\alpha^2}{2\sigma_\alpha^2}\right\} \\
&\propto (\sigma_\alpha^2)^{-(k+\nu_\alpha+2)/2} \exp\left\{-\frac{\sum_{j=1}^{k}\alpha_j^2 + \nu_\alpha S_{\beta\alpha}^2}{2\sigma_\alpha^2}\right\},
\end{aligned}$$

and this is proportional to a scaled inverted chi-square distribution with $\tilde{\nu}_j = \nu_\alpha + k$ and scale parameter $\tilde{S}_j^2 = (\sum_k \alpha_j^2 + \nu_\alpha S_\alpha^2)/\tilde{\nu}_j$.

7.1.7 Full-conditional for σ_e^2

$$\begin{aligned}
f(\sigma_e^2|\text{ELSE}) &\propto \left(\sigma_e^2\right)^{-n/2}\exp\left\{-\frac{\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\alpha_j}{2\sigma_e^2}\right\}\\
&\times \left(\sigma_e^2\right)^{-(2+\nu_e)/2}\exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\}\\
&\propto \left(\sigma_e^2\right)^{-(n+2+\nu_e)/2}\exp\left\{-\frac{(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\alpha_j)'(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\alpha_j)+\nu_e S_e^2}{2\sigma_e^2}\right\},
\end{aligned}$$

which is proportional to a scaled inverted chi-square density with $\tilde{\nu}_e = n + \nu_e$ degrees of freedom and $\tilde{S}_e^2 = \frac{(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\alpha_j)'(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\alpha_j)+\nu_e S_e^2}{\tilde{\nu_e}}$ scale parameter.

7.2 BayesB

7.2.1 Model

The usual model for BayesB is:

$$y_i = \mu + \sum_{j=1}^{k} X_{ij}\alpha_j + e_i,$$

where the prior μ is flat and the prior for α_j is a mixture distribution:

$$\alpha_j = \begin{cases} 0 & \text{probability}\,\pi \\ \sim N(0,\sigma_j^2) & \text{probability}\,(1-\pi) \end{cases},$$

σ_j^2 has a scaled inverted chi-square prior with scale parameter S_α^2 and ν_α degrees of freedom. The residual is normally distributed with mean zero and variance σ_e^2, which has a scaled inverted chi-square prior with scale parameter S_e^2 and ν_e degrees of freedom. Meuwissen et al. [12] gave a Metropolis-Hastings sampler to jointly sample σ_j^2 and α_j. Here, we will show how the Gibbs sampler can be used in BayesB.

In order to use the Gibbs sampler, the model is written as

$$y_i = \mu + \sum_{j=1}^{k} X_{ij}\beta_j\delta_j + e_i,$$

where $\beta_j \sim N(0,\sigma_j^2)$ and δ_j is Bernoulli$(1-\pi)$:

$$\delta_j \;=\; \begin{cases} 0 & \text{probability}\,\pi \\ 1 & \text{probability}\,(1-\pi) \end{cases}.$$

Other priors are the same as in the usual model. Note that in this model, $\alpha_j = \beta_j \delta_j$ has a mixture distribution as in the usual BayesB model.

7.2.2 Full-conditionals:

The joint posterior for all the parameters is proportional to

$$\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}) \\
&\propto \left(\sigma_e^2\right)^{-n/2} \exp\left\{-\frac{\mathbf{y} - \mathbf{1}\mu - \sum \mathbf{X}_j \beta_j \delta_j}{2\sigma_e^2}\right\} \\
&\times \prod_{j=1}^{k} \left(\sigma_j^2\right)^{-1/2} \exp\left\{-\frac{\beta_j^2}{2\sigma_j^2}\right\} \\
&\times \prod_{j=1}^{k} \pi^{(1-\delta_j)}(1-\pi)^{\delta_j} \\
&\times \prod_{j=1}^{k} (\sigma_j^2)^{-(\nu_\beta+2)/2} \exp\left\{-\frac{\nu_\beta S_\beta^2}{2\sigma_j^2}\right\} \\
&\times (\sigma_e^2)^{-(2+\nu_e)/2} \exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\},
\end{aligned}$$

where $\boldsymbol{\theta}$ denotes all the unknowns.

7.2.3 Full-conditional for μ

The full-conditional for μ is a normal distribution with mean $\hat{\mu}$ and variance $\frac{\sigma_e^2}{n}$, where $\hat{\mu}$ is the least-squares estimate of μ in the model

$$\mathbf{y} - \sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{X}_j \beta_j \delta_j = \mathbf{1}\mu + \mathbf{e},$$

and $\frac{\sigma_e^2}{n}$ is the variance of this estimator (n is the number of observations).

7.2.4 Full-conditional for β_j

$$\begin{aligned}
f(\beta_j|\text{ELSE}) &\propto \exp\left\{-\frac{(\mathbf{w}_j-\mathbf{X}_j\beta_j\delta_j)'(\mathbf{w}_j-\mathbf{X}_j\beta_j\delta_j)}{2\sigma^2_e}\right\} \\
&\times \exp\left\{-\frac{\beta_j^2}{2\sigma^2_j}\right\} \\
&\propto \exp\left\{-\frac{[\mathbf{w}_j'\mathbf{w}_j-2\mathbf{w}_j'\mathbf{X}_j\beta_j\delta_j+\beta_j^2(\mathbf{x}_j'\mathbf{x}_j\delta_j+\sigma^2_e/\sigma^2_j)]}{2\sigma^2_e}\right\} \\
&\propto \exp\left\{-\frac{(\beta_j-\hat{\beta}_j)^2}{\frac{2\sigma^2_e}{(\mathbf{x}_j'\mathbf{x}_j\delta_j+\sigma^2_e/\sigma^2_j)}}\right\},
\end{aligned}$$

where

$$\mathbf{w}_j=\mathbf{y}-\mathbf{1}\mu-\sum_{l\neq j}\mathbf{X}_l\beta_l\delta_l.$$

So, the full-conditional for β_j is a normal distribution with mean

$$\hat{\beta}_j=\frac{\mathbf{X}_j'\mathbf{w}_j\delta_j}{(\mathbf{x}_j'\mathbf{x}_j\delta_j+\sigma^2_e/\sigma^2_j)}$$

and variance $\frac{\sigma^2_e}{(\mathbf{x}_j'\mathbf{x}_j\delta_j+\sigma^2_e/\sigma^2_j)}$.

7.2.5 Full-conditional for δ_j

$$\Pr(\delta_j=1|\text{ELSE})\propto\frac{h(\delta_j=1)}{h(\delta_j=1)+h(\delta_j=0)},$$

where

$$h(\delta_j)=\pi^{(1-\delta_j)}(1-\pi)^{\delta_j}\exp\left\{-\frac{\mathbf{w}_j-\mathbf{X}_j\beta_j\delta_j}{2\sigma^2_e}\right\}.$$

7.2.6 Full-conditional for σ^2_j

$$\begin{aligned}
f(\sigma^2_j|\text{ELSE}) &\propto \left(\sigma^2_j\right)^{-1/2}\exp\left\{-\frac{\beta_j^2}{2\sigma^2_j}\right\} \\
&\times (\sigma^2_j)^{-(\nu_\beta+2)/2}\exp\left\{-\frac{\nu_\beta S^2_\beta}{2\sigma^2_j}\right\} \\
&\propto (\sigma^2_j)^{-(1+\nu_\beta+2)/2}\exp\left\{-\frac{\beta_j^2+\nu_\beta S^2_\beta}{2\sigma^2_j}\right\},
\end{aligned}$$

and this is proportional to a scaled inverted chi-square distribution with $\tilde{\nu}_j=\nu_\beta+1$ and scale parameter $\tilde{S}^2_j=(\beta_j^2+\nu_\beta S^2_\beta)/\tilde{\nu}_j$.

7.2.7 Full-conditional for σ_e^2

$$
\begin{aligned}
f(\sigma_e^2|\text{ELSE}) &\propto \left(\sigma_e^2\right)^{-n/2}\exp\left\{-\frac{\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\beta_j\delta_j}{2\sigma_e^2}\right\} \\
&\times \left(\sigma_e^2\right)^{-(2+\nu_e)/2}\exp\left\{-\frac{\nu_e S_e^2}{2\sigma_e^2}\right\} \\
&\propto \left(\sigma_e^2\right)^{-(n+2+\nu_e)/2}\exp\left\{-\frac{(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\beta_j\delta_j)'(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\beta_j\delta_j)+\nu_e S_e^2}{2\sigma_e^2}\right\},
\end{aligned}
$$

which is proportional to a scaled inverted chi-square density with $\tilde{\nu}_e = n + \nu_e$ degrees of freedom and $\tilde{S}_e^2 = \frac{(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\beta_j\delta_j)'(\mathbf{y}-\mathbf{1}\mu-\sum\mathbf{X}_j\beta_j\delta_j)+\nu_e S_e^2}{\tilde{\nu_e}}$ scale parameter.

8 Single-Step GBLUP

Introduction

- Genotypes are available only on a few thousand (non-random) individuals
- Phenotype and pedigree information available on millions
- Often, phenotypes are not available on genotyped individuals (sires)
- Training (estimation of marker effects) based on deregressed EBV
- Marker-based EBV combined with pedigree-based EBV using selection index theory
- An alternative, single-step approach proposed by Legarra et al. [11, 13, 1]

8.1 Theory

8.1.1 Marker Effect Model

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{e}, \tag{41}
$$

- $\boldsymbol{\beta}$ are fixed effects
- $\mathbf{X}$ incidence matrix for fixed effects
- $\mathbf{M}$ marker covariates
- $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}\sigma_\alpha^2)$
- $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_{\mathbf{e}}^{\mathbf{2}})$

8.1.2 Breeding Value Models

- Two mixed linear models are linearly equivalent and will lead to the same inferences if the vector $\mathbf{y}$ of observations has the same first and second moments in both models [8].
- In this sense, a model that is equivalent to (41) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{e}, \tag{42}$$

- $\mathbf{g} = \mathbf{M}\boldsymbol{\alpha}$ has
  - null means and
  - covariance matrix

$$\begin{aligned} Var(\mathbf{g}|\mathbf{M}) &= Var(\mathbf{M}\boldsymbol{\alpha}) \\ &= \mathbf{M}Var(\boldsymbol{\alpha})\mathbf{M}'. \end{aligned}$$

- Then, in both models (41) and (42), the mean of $\mathbf{y}$ is $\mathbf{X}\boldsymbol{\beta}$ and
- the covariance matrix is

$$Var(\mathbf{y}|\mathbf{M}) = \mathbf{M}Var(\boldsymbol{\alpha})\mathbf{M}' + \mathbf{I}\sigma_e^2.$$

  Thus, these two models are linearly equivalent and will lead to the same inferences.

- When the number of markers is large relative to the size of $\mathbf{g}$, BLUP of $\mathbf{g}$ can be obtained efficiently [19, 16] by solving the MME that correspond to (42).
- Under some assumptions,

$$\sigma_\alpha^2 = \frac{\sigma_g^2}{\sum_j 2p_j(1-p_j)} \tag{43}$$

- So,

$$\begin{aligned} Var(\mathbf{g}|\mathbf{M}) &= \frac{\mathbf{M}\mathbf{M}'}{\sum_j 2p_j(1-p_j)}\sigma_g^2 \\ &= \mathbf{G}\sigma_g^2. \end{aligned} \tag{44}$$

8.2 BLUP combining genotype and pedigree relationships

- Suppose $\mathbf{g}$ is partitioned as

$$\mathbf{g} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{M}_2\boldsymbol{\alpha} \end{bmatrix},$$

- $\mathbf{g_1}$ are the genomic BVs of the animals with missing genotypes $\mathbf{M}_1$
- $\mathbf{g}_2$ are the BVs of those with observed genotypes $\mathbf{Z}_2$.
- Following Legarra et al. [11], the vector $\mathbf{g}_1$ can be written as

$$\begin{aligned} \mathbf{g}_1 &= \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2 + (\mathbf{g}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2) \\ &= \hat{\mathbf{g}}_1 + \boldsymbol{\epsilon}, \qquad (45) \end{aligned}$$

- $\mathbf{A}_{ij}$ are partitions of $\mathbf{A}$ that correspond to $\mathbf{g}_1$ and $\mathbf{g}_2$.
- The first term in (45) is the best linear predictor (BLP) of $\mathbf{g}_1$ given $\mathbf{g}_2$,
- the second is a residual.
- It is easy to see that $\boldsymbol{\epsilon}$ in (45) is uncorrelated to $\mathbf{g}_2$,
- therefore if $\mathbf{g}_1$ and $\mathbf{g}_2$ are multivariate normal, ϵ and $\mathbf{g}_2$ are independent.
- Consider first the conditional distribution of $\mathbf{g}_1$ given $\mathbf{P}$. Then, as expected, the
  - variance of $\mathbf{g}_1$ is

$$\begin{aligned} Var(\mathbf{g_1}|\mathbf{P}) &= [\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})]\sigma_g^2 \qquad (46) \\ &= \mathbf{A}_{11}\sigma_g^2, \qquad (47) \end{aligned}$$

    where the first term of (46) is the variance of the $\hat{\mathbf{g}}_1$ and the second term is the variance of $\boldsymbol{\epsilon}$.
  - Similarly, $Var(\mathbf{g}_2|\mathbf{P}) = \mathbf{A}_{22}\sigma_g^2$.
- Consider now the conditional distribution of $\mathbf{g}_1$ given $\mathbf{M}_2$.
- Note that, given the observed genotypes $\mathbf{M}_2$, the distribution of $\mathbf{g}_2$ changes to a multivariate normal with
  - mean $\mathbf{0}$ and
  - covariance matrix $\mathbf{M}_2\mathbf{M}_2'\sigma_\alpha^2$.
- As explained below, this change in the distribution of $\mathbf{g}_2$, produces an associated change in the distribution of $\mathbf{g}_1$ to a normal with
  - mean $\mathbf{0}$ and
  - covariance matrix:

$$Var(\mathbf{g_1}|\mathbf{M}_2) = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2\mathbf{M}_2'\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\sigma_\alpha^2 + (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\sigma_g^2, \qquad (48)$$

    - where now the vector $\hat{\text{g}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2$ has covariance matrix given by the first term of (48),

  - ∗ because $\boldsymbol{\epsilon}$ is independent of $\mathbf{g}_2$, the second term of (48) remains identical to that of (46).

  - Similarly, the covariance between $\mathbf{g}_1$ and $\mathbf{g}_2$ conditional on $\mathbf{M}_2$ is

$$Cov(\mathbf{g}_1, \mathbf{g}_2) = \mathbf{A_{12}}\mathbf{A}_{22}^{-1}\mathbf{M}_2\mathbf{M}_2'\sigma_\alpha^2.$$

- Further, assuming (43), the above results can be combined to show that conditional on $\mathbf{Z}_2$,

  - $\mathbf{g}$ has a multivariate normal distribution with null mean and covariance matrix:

$$Var(\mathbf{g}|\mathbf{Z}_2) = \mathbf{H} = \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} + (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}) & \mathbf{A_{12}}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathrm{A}_{21} & \mathbf{G} \end{bmatrix} \sigma_g^2, \tag{49}$$

    where $\mathbf{G} = \mathbf{M}_2\mathbf{M}_2'/[\sum 2p_i(1-p_i)]$.

  - The inverse of this matrix is needed to setup the MME, and this is computed as

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A_{22}^{-1}} \end{bmatrix}.$$

- Note that this requires computing both $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$, which are dense and not easy to compute.

- Due to the increased adoption of SNP genotyping in livestock, $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$ are becoming too large for SS-GBLUP to remain computationally feasible (e.g. ¿100,000 animals).

- A second problem in SS-GBLUP is related to the scaling that is done using the SNP frequencies.

- As mentioned earlier, when all data that were used for selection are available for computing the conditional mean, it can be computed as if selection had not taken place [5, 9, 14].

- If selection has taken place, this requires using SNP frequencies from the founders, as these frequencies are not changed by selection.

- In most situations, however, SNP genotypes are not available in the founders and frequencies observed in the genotyped animals are used, which can lead to biased evaluations, particularly in a multi-breed context.

- An approach very similar to that of using (45) to model missing genotypes was proposed by Fernando (see equation (43) in[6]) in the context of genomic prediction using kernel regression, where missing genotypes were replaced by their conditional expectation computed using all available information, in contrast to using BLP as in SS-GBLUP.

- Also, a residual that is similar to $\boldsymbol{\epsilon}$ was included in the model.
- When these residuals are modeled exactly, the inverse of their covariance matrix is not sparse and SS-GBLUP would not be computationally feasible.

9 Single-Step Bayesian Regression

9.1 Theory

The mixed linear model for the phenotypic values can be expressed in terms of a BVM (50) or an MEM (51) as

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{g}+\mathbf{e} && (50)\\ &= \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{M}\boldsymbol{\alpha}+\mathbf{e}, && (51)\end{aligned}$$

where we have introduced the incidence matrix $\mathbf{Z}$ to accommodate animals with repeated records or animals without records. As in SSBV-BLUP, suppose $\mathbf{M}_1$ is not observed. Then it is not possible to use (51) as the basis for the MEM. Note that $\mathbf{M}_1\boldsymbol{\alpha}$ is equal to $\mathbf{g}_1$. So, using (45) for $\mathbf{g}_1$ and writing $\mathbf{g}_2 = \mathbf{M}_2\boldsymbol{\alpha}$, the model for the phenotypic values becomes

$$\begin{aligned}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} &= \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\end{bmatrix}+\mathbf{e} && (52)\\ &= \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2\boldsymbol{\alpha}+\boldsymbol{\epsilon}\\ \mathbf{M}_2\boldsymbol{\alpha}\end{bmatrix}+\mathbf{e} && (53)\\ &= \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta}+\begin{bmatrix}\mathbf{Z}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{Z}_2\end{bmatrix}\begin{bmatrix}\hat{\mathbf{M}}_1\boldsymbol{\alpha}+\boldsymbol{\epsilon}\\ \mathbf{M}_2\boldsymbol{\alpha}\end{bmatrix}+\mathbf{e} && (54)\\ &= \mathbf{X}\boldsymbol{\beta}+\mathbf{W}\boldsymbol{\alpha}+\mathbf{U}\boldsymbol{\epsilon}+\mathbf{e}, && (55)\end{aligned}$$

where

$$\mathbf{U}=\begin{bmatrix}\mathbf{Z}_1\\ 0\end{bmatrix},\ \mathbf{X}=\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\text{ and }\mathbf{W}=\begin{bmatrix}\mathbf{W}_1\\ \mathbf{W}_2\end{bmatrix}=\begin{bmatrix}\mathbf{Z}_1\hat{\mathbf{M}}_1\\ \mathbf{Z}_2\mathbf{M}_2\end{bmatrix}.$$

The matrix $\hat{\mathbf{M}}_1 = \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{M}_2$ of imputed SNP covariates can be obtained efficiently, using partitioned inverse results, by solving the easily formed very sparse system:

$$\mathbf{A}^{11}\hat{\mathbf{M}}_1 = -\mathbf{A}^{12}\mathbf{M}_2, \qquad (56)$$

where $\mathbf{A}^{ij}$ are partitions of $\mathbf{A}^{-1}$ that correspond to partitioning $\mathbf{g}$ into $\mathbf{g}_1$ and $\mathbf{g}_2$.

The differences between this MEM (55) and the model that is currently used for Bayesian regression (BR) are: 1) that some of the covariates in (55) are imputed, and 2) a residual term $\boldsymbol{\epsilon}$ has been introduced to account for the deviations of the imputed genotype covariates from their unobserved, actual values. Regardless of the prior used for $\boldsymbol{\alpha}$, the distribution of the vector $\boldsymbol{\epsilon}$ of imputation residuals will be approximated by a multivariate normal vector with null mean and covariance matrix $(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})\sigma_g^2$ (see equation 48), where σ_g^2 is assigned a scaled inverse chi-square distribution with scale parameter S_g^2 and degrees of freedom ν_g. Imputing the covariates needs to be done only once, and it can be done efficiently in parallel. Imputation of unobserved SNP covariates will not add significantly to overall computing time.

The MME that correspond to (55) for BayesC with $\pi = 0$ are

$$\left[\begin{array}{lll} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} & \mathbf{X}_1'\mathbf{Z}_1 \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{I}\frac{\boldsymbol{\sigma}^{\mathbf{2}}_{\mathbf{e}}}{\boldsymbol{\sigma}^{\mathbf{2}}_{\alpha}} & \mathbf{W}_1'\mathbf{Z}_1 \\ \mathbf{Z}_1'\mathbf{X}_1 & \mathbf{Z}_1'\mathbf{W}_1 & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{11}\frac{\sigma_e^2}{\sigma_g^2} \end{array}\right] \left[\begin{array}{c} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{array}\right] = \left[\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y}_1 \end{array}\right]. \qquad (57)$$

The submatrix of these MME that correspond to $\boldsymbol{\epsilon}$ are identical to those for $\mathbf{g}_1$ from a pedigree-based analysis and are very sparse. Thus as explained in the next section, conditional on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, $\boldsymbol{\epsilon}$ can sampled efficiently by using either a blocking-Gibbs sampler [3, 15] or a single-site, Gibbs sampler used in pedigree-based analyses [15]. Note that these MME, which do not have $\mathbf{G}$ or its inverse, may be used to overcome the computational problems facing SSBV-BLUP. The predicted BVs can be written as

$$\tilde{\mathbf{g}} = \left[\begin{array}{c} \hat{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{array}\right] \hat{\boldsymbol{\alpha}} + \mathbf{U}\hat{\boldsymbol{\epsilon}} = \left[\begin{array}{c} \hat{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{array}\right] \hat{\boldsymbol{\alpha}} + \left[\begin{array}{c} \mathbf{Z}_1 \\ 0 \end{array}\right] \hat{\boldsymbol{\epsilon}}. \qquad (58)$$

A similar system of MME without $\boldsymbol{\epsilon}$ was solved by iteration for a MEM [19] but using only genotyped animals. The MME given by (57) has the advantage that it will not grow in size as more animals are genotyped, in contrast to the MME corresponding to (52) that is given by Aguilar et al. [1], but assuming (**??**) they give identical EBV.

9.2 Numerical Example

```r
# results from this script are stored in 'smallExWkSp.RData'
setwd("/Users/rohan/AeroFS/latex/courses/WinterWorkshp2013/R")
set.seed(12345)
library("RSim")
data(sim)
genParms = list(nChr = 1, chrLength = 0.1, nLoci = 11, mutRate = 1e-10)
# simulating from pedigree
ind = c(1, 2, 3, 4, 5, 6)
sire = c(0, 0, 0, 1, 1, 1)
dam = c(0, 0, 0, 2, 2, 3)
size = length(ind)
ped = cbind(ind, sire, dam)
```

```
mySim = RSim(genParms)
mySim$pedSample(ped)

## [1] 6

snpMat = mySim$getGenotypes(1, size)
G = cor(t(snpMat))
Gi = solve(G)
Z = snpMat[1:3, ]
QPos = c(5)
QTL = snpMat[, QPos]
Z = Z[, -QPos]
G = Z %*% t(Z)/ncol(Z)
# round(G,3)
Gi = solve(G)
nMarkers = ncol(Z)
ped.df = data.frame(ped)
names(ped.df) = c("ind", "sire", "dam")
write.table(ped.df, file = "pedSmall.dat", row.names = F, col.names = F)
require(Matrix)

## Loading required package:  Matrix
## Loading required package:  methods
## Loading required package:  lattice

require(RMatvec)

## Loading required package:  RMatvec
## Loading required package:  Rcpp

rPed = new(RPed, "pedSmall.dat", TRUE)
subA = function(ids) {
    n = length(ids)
    A = matrix(nrow = n, ncol = n)
    for (i in 1:n) {
        for (j in i:n) {
            A[i, j] = rPed$getAij(ids[i], ids[j])
            A[j, i] = A[i, j]
        }
    }
    colnames(A) = ids
    rownames(A) = ids
    return(A)
}
ids = row.names(ped.df)
A = subA(ids)
# put parents at bottom
```

```
x = c(4, 5, 6, 1, 2, 3)
A = A[x, x]
Ai = solve(A)
A11i = Ai[1:3, 1:3]
A12 = A[1:3, 4:6]
A22 = A[4:6, 4:6]
A22i = solve(A22)
require(foreach)

## Loading required package:  foreach

require(doMC)

## Loading required package:  doMC
## Loading required package:  iterators
## Loading required package:  parallel

registerDoMC()
res = foreach(i = 1:nMarkers, .combine = cbind) %dopar% {
    g2 = Z[, i]
    g1 = A12 %*% A22i %*% g2
}
WDat = c(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1,
    0, 0, 0, 0, 0, 0, 1)
W = matrix(WDat, 5, 6, byrow = T)
Zh0 = rbind(res, Z)
Zh = W %*% Zh0
X = rep(1, 5)
r1 = ncol(Z)
r2 = 1
mmeRow1 = cbind(t(X) %*% X, t(X) %*% Zh, t(X[1:3]))
mmeRow2 = cbind(t(Zh) %*% X, t(Zh) %*% Zh + diag(nMarkers) * r1, t(Zh[1:3, ]))
mmeRow3 = cbind(X[1:3], Zh[1:3, ], diag(3) + A11i * r2)
mme = rbind(mmeRow1, mmeRow2, mmeRow3)
# generating phenotypes
y = QTL + rnorm(6)
y = y[x]  # reordering (offspring first, then sire and dam)
y = y[-4]  # removing sire phenotype
rhs = c(t(X) %*% y, t(Zh) %*% y, y[1:3])
sol = solve(mme, rhs)
ghat1 = sol[1] + Zh[1:3, ] %*% sol[2:(1 + nMarkers)] + sol[(2 + nMarkers):(2 +
    nMarkers + 2)]
ghat2 = sol[1] + Z %*% sol[2:(1 + nMarkers)]
markerNames = c()
for (i in 1:nMarkers) {
    name = paste("m", i, sep = "")
    markerNames = c(markerNames, name)
}
colNames = c("mu", markerNames, "e4", "e5", "e6")
colnames(mme) = colNames
rownames(mme) = colNames
# (round(mme,3))
ghat = rbind(ghat1, ghat2)
# SSBLUP
HTest = Zh0 %*% t(Zh0)/ncol(Z)
HTest[1:3, 1:3] = HTest[1:3, 1:3] + solve(A11i)
HRow1 = cbind(A12 %*% A22i %*% G %*% A22i %*% t(A12) + solve(A11i), A12 %*%
    G)
Hrow2 = cbind(G %*% A22i %*% t(A12), G)
H = rbind(HRow1, Hrow2)  # checked that the inverse of this matches Hi
Hi = Ai
Hi[4:6, 4:6] = Hi[4:6, 4:6] + Gi - A22i
ssbRow1 = cbind(t(X) %*% X, t(X) %*% W)
ssbRow2 = cbind(t(W) %*% X, t(W) %*% W + Hi)
mmeSSB = rbind(ssbRow1, ssbRow2)
rhsSSB = c(t(X) %*% y, t(W) %*% y)
solSSB = solve(mmeSSB, rhsSSB)
ghatSSB = solSSB[1] + solSSB[2:7]
ghats = cbind(ghat, ghatSSB)
colnames(ghats) = c("SSBReg", "SSBLUP")
(ghats)

##   SSBReg SSBLUP
## 4 0.9812 0.9812
## 5 0.9075 0.9075
## 6 0.7959 0.7959
## 1 0.8997 0.8997
## 2 0.7883 0.7883
## 3 0.8137 0.8137

colnames(Z) = markerNames
colNamesSSB = c("mu", "4", "5", "6", "1", "2", "3")
colnames(mmeSSB) = colNamesSSB
rownames(mmeSSB) = colNamesSSB
save.image(file = "/Users/rohan/AeroFS/latex/courses/WinterWorkshp2013/R/smallExWkSp.RData")
```

Consider the pedigree in Table 4.

Suppose genotypes are available only on the parents, individuals 1, 2, and 3. Genotypes ($\mathbf{M}_2$) at 10 markers are given in Table 5.

Following Legarra et al. [11], the relationship matrix is rearranged such that $\mathbf{A}_{11}$ are relationships among individuals 4, 5, and 6, which do not have genotypes, and $\mathbf{A}_{22}$ are relationships among the parents, 1, 2, and 3, which have genotypes given in Table 5. The inverse of the rearranged relationship

individual	sire	dam	phenotypes
1	0	0	-999.00
2	0	0	0.45
3	0	0	0.87
4	1	2	1.26
5	1	2	1.03
6	1	3	0.67

Table 4: Pedigree for numerical example. Genotypes are available only on parents: 1, 2 and 3. Phenotypes are available on all except the sire.

individual	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
1	1	2	1	1	0	0	1	2	1	0
2	2	1	1	1	2	0	1	1	1	1
3	0	1	0	0	2	1	2	1	1	1

Table 5: Marker genotypes at ten markers on parents: 1, 2, and 3.

matrix is given in Table 6.

The imputed genotypes $\hat{\mathbf{M}}_1$ could be obtained efficiently by solving the system (56) and are given in Table 7. %

Now, to setup the MME we will assume that $\sigma^2_\alpha = \frac{\sigma^2_g}{10}$, $\sigma^2_g = \sigma^2_e$, and that μ, an effect common to all the observations, is the only fixed effect. Then, the MME (57) and solutions corresponding to the marker effects model (55) is given in Table 8. For comparison, the MME and solutions for the single-step BV model are given in Table 9. The solutions for μ are identical from the two sets of MME, and BLUP of $\mathbf{g}$ obtained as (58), using the solutions to (57), are identical to the solutions to $\mathbf{g}$ given in Table 9.

	4	5	6	1	2	3
4	2.00	0.00	0.00	-1.00	-1.00	0.00
5	0.00	2.00	0.00	-1.00	-1.00	0.00
6	0.00	0.00	2.00	-1.00	0.00	-1.00
1	-1.00	-1.00	-1.00	2.50	1.00	0.50
2	-1.00	-1.00	0.00	1.00	2.00	0.00
3	0.00	0.00	-1.00	0.50	0.00	1.50

Table 6: Inverse of rearrannged relationship matrix. Row and column labels are the individual IDs. The matrix has been rearanged such that the three individuals without genotypes are in the first block, and the genotyped individuals are in the next block.

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
4	1.50	1.50	1.00	1.00	1.00	0.00	1.00	1.50	1.00	0.50
5	1.50	1.50	1.00	1.00	1.00	0.00	1.00	1.50	1.00	0.50
6	0.50	1.50	0.50	0.50	1.00	0.50	1.50	1.50	1.00	0.50

Table 7: Imputed genotypes at the ten markers for individuals 4, 5, and 6.

References

[1] I Aguilar, I Misztal, D L Johnson, A Legarra, S Tsuruta, and T J Lawlor. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of holstein final score. *J Dairy Sci*, 93(2):743–752, Feb 2010. 8, 9.1

[2] R. L. Fernando, D. Habier, C. Stricker, J. C. M. Dekkers, and L. R. Totir. Genomic selection. *Acta Agriculturae Scandinavica, Section A - Animal Science*, 57(4):192–195, 2007. 7.1.1

[3] L.A. Garcia-Cortes and D. Sorensen. On a multivariate implementation of the gibbs sampler. *Genet. Sel. Evol.*, 28:121–126, 1996. 9.1

[4] D Gianola, G de los Campos, W G Hill, E Manfredi, and R Fernando. Additive genetic variability and the bayesian alphabet. *Genetics*, 183(1):347–363, Sep 2009. 7.1.1

[5] D. Gianola and R. L. Fernando. Bayesian methods in animal breeding. *J. Anim. Sci.*, 63:217–244, 1986. 8.2

[6] D. Gianola, R. L. Fernando, and A. Stella. Genomic assisted prediction of genetic value with semi-parametric procedures. *Genetics*, pages 1761–1776, 2006. 8.2

[7] J. B. S. Haldane. The combination of linkage values, and the calculation of distances between the loci of linked factors. *J. of Genetics*, VIII:299–309, 1919. 1.4

[8] C. R. Henderson. *Applications of Linear Models in Animal Breeding.* Univ. Guelph, Guelph, Ontario, Canada, 1984. 8.1.2

[9] S. Im, R. L. Fernando, and D. Gianola. Likelihood inferences in animal breeding under selection: a missing-data theory view point. *Genet. Sel. Evol.*, 21:399–414, 1989. 8.2

[10] S. Karlin. Theoretical aspects of genetic map functions in recombination processes. In A. Chakravarti, editor, *Human Population Genetics: The Pitsburgh Symposium*, pages 209–228, New York, NY, 1984. Van Nostrand Reinhold. 1.4

[11] A Legarra, I Aguilar, and I Misztal. A relationship matrix including full pedigree and genomic information. *J Dairy Sci*, 92(9):4656–4663, Sep 2009. 8, 8.2, 9.2

[12] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001. 7, 7.2.1

[13] I Misztal, A Legarra, and I Aguilar. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J Dairy Sci*, 92(9):4648–4655, Sep 2009. 8

[14] D. Sorensen, R. L. Fernando, and D. Gianola. Inferring the trajectory of genetic variance in the course of artificial selection. *Genetical Research*, 77:83–94, 2001. 8.2

[15] D. A. Sorensen and D. Gianola. *Likelihood,Bayesian,and MCMC Methods in Quantitative Genetics.* Springer, 2002. 9.1

[16] I Strandén and D J Garrick. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *J Dairy Sci*, 92(6):2971–2975, Jun 2009. 8.1.2

[17] JA Sved. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical population biology*, 2(2):125–141, 1971. 1.7

[18] John A Sved. Linkage disequilibrium and its expectation in human populations. *Twin Research and Human Genetics*, 12(01):35–43, 2009. 1.7

[19] P M VanRaden. Efficient methods to compute genomic predictions. *J Dairy Sci*, 91(11):4414–4423, Nov 2008. 8.1.2, 9.1

	μ	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	ϵ_4	ϵ_5	ϵ_6
μ	5.00	5.50	6.50	3.50	3.50	7.00	1.50	6.50	6.50	5.00	3.50	1.00	1.00	1.00
m_1	5.50	18.75	7.25	5.25	5.25	7.50	0.25	5.75	7.25	5.50	3.75	1.50	1.50	0.50
m_2	6.50	7.25	18.75	4.75	4.75	8.50	1.75	8.25	8.75	6.50	4.25	1.50	1.50	1.50
m_3	3.50	5.25	4.75	13.25	3.25	4.50	0.25	3.75	4.75	3.50	2.25	1.00	1.00	0.50
m_4	3.50	5.25	4.75	3.25	13.25	4.50	0.25	3.75	4.75	3.50	2.25	1.00	1.00	0.50
m_5	7.00	7.50	8.50	4.50	4.50	21.00	2.50	9.50	8.50	7.00	5.50	1.00	1.00	1.00
m_6	1.50	0.25	1.75	0.25	0.25	2.50	11.25	2.75	1.75	1.50	1.25	0.00	0.00	0.50
m_7	6.50	5.75	8.25	3.75	3.75	9.50	2.75	19.25	8.25	6.50	4.75	1.00	1.00	1.50
m_8	6.50	7.25	8.75	4.75	4.75	8.50	1.75	8.25	18.75	6.50	4.25	1.50	1.50	1.50
m_9	5.00	5.50	6.50	3.50	3.50	7.00	1.50	6.50	6.50	15.00	3.50	1.00	1.00	1.00
m_{10}	3.50	3.75	4.25	2.25	2.25	5.50	1.25	4.75	4.25	3.50	12.75	0.50	0.50	0.50
ϵ_4	1.00	1.50	1.50	1.00	1.00	1.00	0.00	1.00	1.50	1.00	0.50	3.00	0.00	0.00
ϵ_5	1.00	1.50	1.50	1.00	1.00	1.00	0.00	1.00	1.50	1.00	0.50	0.00	3.00	0.00
ϵ_6	1.00	0.50	1.50	0.50	0.50	1.00	0.50	1.50	1.50	1.00	0.50	0.00	0.00	3.00
rhs	4.29	4.67	5.77	3.08	3.08	5.61	1.21	5.49	5.77	4.29	2.80	1.26	1.03	0.67
sol	0.86	-0.01	0.01	0.00	0.00	-0.03	-0.00	-0.00	0.01	0.00	-0.01	0.14	0.06	-0.06

Table 8: Mixed equations for marker effects model with observed and imputed genotype covariates. The last two rows give the right-hand-side and the solutions.

	μ	g_4	g_5	g_6	g_1	g_2	g_3
μ	5.00	1.00	1.00	1.00	0.00	1.00	1.00
g_4	1.00	3.00	0.00	0.00	-1.00	-1.00	0.00
g_5	1.00	0.00	3.00	0.00	-1.00	-1.00	0.00
g_6	1.00	0.00	0.00	3.00	-1.00	0.00	-1.00
g_1	0.00	-1.00	-1.00	-1.00	3.08	0.00	0.42
g_2	1.00	-1.00	-1.00	0.00	0.00	4.00	-1.00
g_3	1.00	0.00	0.00	-1.00	0.42	-1.00	3.08
rhs	4.29	1.26	1.03	0.67	0.00	0.45	0.87
sol	0.86	0.12	0.05	-0.06	0.04	-0.07	-0.04

Table 9: Mixed equations for single-step BV model. The last two rows give the right-hand-side and the solutions.