Using Markov chain Monte Carlo in practice: A case study of deleterious gene structure in plants

Jaekyun Lee, Michael A. Newton, Erik V. Nordheim, Hyun Kang and Martin Lascoux

Technical Report # 944
February 1995
Abstract

The magnitude of the effect of deleterious genes on a population is characterized classically by the number of lethal equivalents. This number is a combination of several parameters that have different biological effects on genetic mortality. In conservation and breeding programs, it is important to be able to distinguish among different combinations of these parameters that lead to the same number of lethal equivalents, such as a large number of mildly deleterious genes or a few fully lethal genes. The capacity to distinguish such parameter combinations requires two generations of mating at least. Due to the complexity of the likelihood function and the existence of many unobservable data in this two-stage design, estimation is analytically intractable. We apply Bayesian Markov chain Monte Carlo to infer the model parameters. Our results show that beyond estimating the overall effects of deleterious genes, we can obtain information about an important genetic parameter expressing the intensity of mortality of the genes. Our Markov chain Monte Carlo algorithm includes a vector proposal to overcome inefficiency of the Gibbs sampler. Information about environmental effects is obtained from an outcrossing experiment. We investigate the performance of our approach for simulated data generated from the two-stage design.  

\[1\]

1Key words: Deleterious genes; Lethal equivalents; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Hierarchical modeling; Highest posterior density region.
1 Introduction

Biological populations contain deleterious genes, a point stressed forcefully by Fisher in his book “The Theory of Inbreeding” (Fisher, 1949). With animal and plant breeding in mind, Fisher also suggested the purging of these genes through generations of inbreeding:

“when an inbred line is formed from elite stock, it can contain no genes, however inferior it may appear, which were not present in its admired progenitors, nor can it hand any others on to its descendants.”

Deleterious genes may differ in their intensity of mortality, often defined by the selection coefficient. This coefficient, denoted $s$, is the probability of lethality of homozygotes having deleterious genes at a specific locus in a population. In a population where deleterious genes exist at $M$ loci with frequency $q_i$ and selection coefficient $s_i$ at the $i$-th locus, the number of lethal equivalents at the gametic (chromosome) level is expressed as,

$$\varepsilon = \sum_{i=1}^{M} s_i q_i$$

(Morton, Crow, and Muller, 1956). So, in the diploid individual, the number of lethal equivalents is $E = 2\varepsilon$. This can be interpreted as the average number of deleterious genes per individual weighted by their selection coefficients. In natural species, $E$ is known to be bounded -- mostly less than ten (Namkoong and Bishir, 1987).

Although it is difficult to identify all deleterious genes and to derive their frequencies and other characteristics, it was realized that their overall effect could be readily estimated. The first method to estimate the number of lethal equivalents in human populations was proposed by Morton, Crow, and Muller (1956). These authors, under certain assumptions, derived a point estimate of the number of lethal equivalents using infantile death data from consanguineous marriages. Alternatively, $E$ can be estimated using a combinatorial arguments (Bishir and Pepper, 1977; Namkoong and Bishir, 1987). However, in neither case was it possible to derive an interval estimate for the number of lethal equivalents. Recently, Lee, Nordheim, and Kang (1994), using both combinatorial arguments and sampling process from the base population, developed a method allowing asymptotic interval estimation.

For some applications, such as conservation of endangered species or breeding programs in small closely kept (sub)populations, it is desirable to go beyond an overall estimate of $E$ to identify the components $M, q,$ and $s$ separately. Different combinations of $M, q,$ and $s$ that result in the same number of lethal equivalents can lead to very distinct patterns of mortality when several generations are considered (Hedrick, 1994). Although a large number of mildly deleterious genes and a few lethal ones may have a similar impact in the first generation, the pattern of mortality in subsequent generations can be quite different. It takes fewer generations to “purge” lethal genes than mildly deleterious ones. Consequently, it makes sense to consider multi-generation data if the goal is to provide separate estimates of $M, q,$ and $s$. Simple experimental designs using one-stage of survival data as in, e.g., Sorensen (1969), Namkoong and Bishir (1987), and Lee, Nordheim, and Kang (1994), cannot provide
information on parameters of interest because the parameters are confounded when only one generation of data is available.

In this article, we consider the analysis of two-generation experiments designed to obtain separate information about $M$, $q$, and $s$. We focus on a particular experiment using the plant species *Brassica napus*, a rapidly cycling plant species. The likelihood function corresponding to this two-stage design is constructed by several steps of conditioning. A large number of unobserved (missing) data describe individual plant genotypes. Due to these missing data and the complexity of the likelihood, we pursue Bayesian analysis using Markov chain Monte Carlo (MCMC) (Gelfand and Smith, 1990; Smith and Roberts, 1993; Besag and Green, 1993).

The remainder of this article is organized as follows. In Section 2 we describe data from a particular two-stage experimental design on *Brassica napus*. A hierarchical model for the data is presented in Section 3. We explain our MCMC algorithm in Sections 4.1 and 4.2. In Section 4.3 we suggest a few methods that circumvent non-identifiability of environmental effects of mortality. Computations for the *Brassica napus* data are shown in Section 5. To validate our analysis, we perform a simulation study in Section 6. We present our conclusions and make some additional remarks in Section 7.

## 2 The Data

To study the efficacy of two-generation mating schemes for understanding lethal equivalents, we performed a simple experiment on the rapid-cycling plant species *Brassica napus* L. The base-population seeds were obtained from Crucifer Genetics Cooperative, Department of Plant Pathology, University of Wisconsin-Madison (Williams, 1985). We adopted selfing as a mating scheme - - a method possible with some organisms that produce both male and female zygotes (Kang, Lee, and Nordheim, 1994). The experiment was performed in Dr. Kang’s laboratory at the University of Wisconsin-Madison. To begin, 30 individuals (seeds) were randomly sampled from the base population. These seeds were planted and grown for selfing (resulting in 30 family lines). From each of these 30 selfed plants, a random sample of 48 seeds was collected and planted. We recorded two binary response variables that may be influenced by deleterious genes: germination and flowering (Table 1).

<table>
<thead>
<tr>
<th></th>
<th>Germination</th>
<th>Flowering</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From each family line that produced one or more mature plants in the first generation, one plant was selected randomly and selfed to produce seeds for the second generation of the family line. A random sample of 48 seeds was again chosen from each line in the second generation. (By accident, only 42 seeds were obtained from line 30 in the second generation.) Germination and flowering rates were recorded as above. If a family line went extinct in the first generation, we could not proceed to the second generation for that line and recorded all second generation counts as zero. (This happened for two lines.)

To control for environmental effects, 120 seeds from the base population were also planted in each generation. These base-population seeds were from outcrossing, a mating scheme in which the male and female zygotes come from different lines in relationship. In a greenhouse with full-time lighting and regular watering, it took 40 to 45 days to complete each of the
two generations. See Lascoux and Lee (in preparation) for a more detailed description of the experiment.

[Table 1 goes about here.]

3 A Hierarchical Model

Some assumptions are necessary to construct a statistical model. We assume that all different loci having deleterious genes act independently and that there is a constant deleterious gene frequency \( q \) per gamete per locus for all \( M \) loci carrying deleterious genes. We also assume a constant selection coefficient \( s \) for these \( M \) deleterious genes. The starting point for the model is a set of \( N \) parents (seeds) randomly drawn from the base population. The relevant genotype information from the first generation is carried by a set of trinomial random vectors

\[
\{v_{1i}, w_{1i}, M - (v_{1i} + w_{1i})\} \sim \text{Trinomial}\{M, \xi_{11}, \xi_{12}, 1 - (\xi_{11} + \xi_{12})\},
\]

for \( i = 1, 2, ..., N \) indexing the family line. The three counts, respectively, record the number of loci (in parent \( i \)) that are heterozygous for deleterious genes \( (v_{1i}) \), homozygous for deleterious \( (w_{1i}) \) genes, and homozygous for non-deleterious genes. Our assumptions imply that

\[
\xi_{11} = \frac{2q(1-q)}{1-sq^2} \quad \text{and} \quad \xi_{12} = \frac{q^2(1-s)}{1-sq^2}
\]

(cf. Falconer, 1981, pp. 26-27). Notice that these counts are unobservable in our mating experiments.

Selfing the \( i \)-th parent produces \( n_{1i} \) seeds, some subset of which will germinate. In the \( B. napus \) data, \( N = 30 \) and \( n_{1i} = 48 \) for all parents. The chance of genetic mortality depends on \( v_{1i}, w_{1i} \), and can be computed from a complementary argument as

\[
Q_{1i} = \Pr[ \text{seed from parent } i \text{ dies due to genetic causes} ]
= 1 - (1 - s/4)^{v_{1i}} (1 - s)^{w_{1i}}
\]

(cf. Lee, Nordheim, and Kang, 1994). However, environmental effects of mortality may also play a role here. Assuming that mortality due to environmental effects to each seed is a constant, say \( p_e \), and that genetic and environmental effects of mortality are independent, the probability of mortality \( Q_{1i} \) becomes

\[
Q_{1i} = 1 - (1 - p_e)(1 - s/4)^{v_{1i}} (1 - s)^{w_{1i}}.
\]

Thus, conditional on the genotypes \( G_1 = \{v_{1i}, w_{1i}\} \), the observed numbers \( D_1 = \{d_{1i}\} \) of nongerminating (or nonflowering) seeds are described by

\[
d_{1i} \sim \text{Binomial}(n_{1i}, Q_{1i}).
\]
The random variables for the second generation are similar, except that the process for the second generation is conditional on the results of the first generation. The numbers of heterozygous \( (v_{2i}) \) and homozygous \( (w_{2i}) \) loci at the second generation transmitted from the heterozygous loci \( (v_{1i}) \) at the first generation follow another trinomial distribution:

\[
\{v_{2i}, w_{2i}, v_{1i} - (v_{2i} + w_{2i})\} \sim \text{Trinomial}\{v_{1i}, \xi_{21}, \xi_{22}, 1 - (\xi_{21} + \xi_{22})\},
\]

where \( \xi_{21} = \frac{1/2}{1-s/4} \) and \( \xi_{22} = \frac{(1-s)/4}{1-s/4} \) are the relative frequencies after selection corresponding to the heterozygous and homozygous genotype individuals (refer to (2) and (3) when \( q = 1/2 \)). All the homozygous loci at the first generation will be transmitted as homozygous, so the total number of homozygous loci at the second generation is \( w_{2i}' = w_{1i} + w_{2i} \). Also, conditional on the genotypes \( G_s = \{v_{2i}, w_{2i}'\} \) of the selfed offspring at the first generation, the numbers of seeds \( D_s = \{d_{2i}\} \) that are nongerminating (or nonflowering) is each Binomial\( (n_{2i}, Q_{2i}) \) where

\[
Q_{2i} = 1 - (1 - p_e)(1 - s/4)^{v_{2i}}(1 - s)^{w_{2i}}.
\]

We note again that some of the lines from the first generation may be lost if there is no viable offspring of that line. All distributions for the second generation must be viewed as conditional on the results of the first generation; the second-generation trinomial distribution depends explicitly on \( s \) but depends on \( M \) and \( q \) only through \( v_{1i} \) and \( w_{1i} \) (Figure 1).

[Figure 1 goes about here.]

With \( k = 1, 2 \) and \( i = 1, ..., N \), define probabilities

\[
T_{1i} = \binom{M}{v_{1i}}(M - v_{1i})^{v_{1i} - 1}(1 - \xi_{11} - \xi_{12})^{w_{1i}}(1 - \xi_{11} - \xi_{12})^{w_{1i}}(1 - \xi_{11} - \xi_{12})^{w_{1i}},
\]

\[
T_{2i} = \binom{v_{1i}}{v_{2i}}(v_{1i} - v_{2i})^{v_{2i} - 1}(1 - \xi_{21} - \xi_{22})^{w_{2i}}(1 - \xi_{21} - \xi_{22})^{w_{2i}}(1 - \xi_{21} - \xi_{22})^{w_{2i}},
\]

\[
B_{ki} = \binom{n_{ki}}{d_{ki}}Q_{ki}^{d_{ki}}(1 - Q_{ki})^{n_{ki}}.
\]

The joint probability of genotypes and phenotypes becomes

\[
\Pr(G_1, G_2, D_1, D_2 | M, q, s, p_e) = \prod_{i=1}^{N} T_{1i}B_{1i}(T_{2i}B_{2i})^{U_i},
\]

where \( U_i = 1 \) if the line survives at the first generation and \( U_i = 0 \) otherwise. Note that genotype information is unobserved. The likelihood function for parameters \( M, q, s, \) and \( p_e \) is formed by summing (9) over all \( G_1, G_2 \) consistent with observed \( D_1, D_2 \). Because such a direct sum is rather difficult, we invoke a simulation method to enable computation. Thompson (1991) suggests the use of such a formulation of the likelihood function with both missing genotypes and observed phenotypic data in pedigree analysis.
4 Markov Chain Monte Carlo

4.1 Strategy

The Markov chain Monte Carlo (MCMC) method (Gelfand and Smith, 1990; Smith and Roberts, 1993; Besag and Green, 1993) can be used to overcome the fact that the likelihood is analytically intractable. Our strategy is to apply Bayesian analysis by formulating a prior distribution $\pi_0(M,q,s,p_e)$ over the parameter space and then using MCMC to simulate the joint posterior distribution of parameters and missing data. This posterior distribution has density function

$$\pi(G_1,G_2,M,q,s,p_e) \propto \Pr(G_1,G_2,D_1,D_2|M,q,s,p_e) \pi_0(M,q,s,p_e)$$  \hspace{1cm} (10)$$

and can be readily evaluated up to a constant. (We will use this notation $\pi$ for some of its marginal and conditional distributions without comment.) This posterior is defined over a $4 + 2 \sum_{i=1}^{N} n_{1i} + 2 \sum_{i=1}^{N} n_{2i}$ dimensional space of parameters and missing data. Parameter estimates and Bayesian confidence intervals will be formed from marginal distributions corresponding to (10).

We consider independent uniform priors on the three parameters: uniform (0,1) priors for the deleterious gene frequency and selection coefficient, and a uniform discrete prior on $M$ between $M_L$ and $M_U$, where $M_L$ and $M_U$ are the lower and upper bounds for the total number of deleterious loci of a natural species. We use $(M_L,M_U) = (100, 10,000)$ for our study. (We check sensitivity to the choice of $(M_L,M_U)$ in Sections 5 and 6.) The prior information of $p_e$ is derived from the data outcrossing experiment; we discuss this in Section 4.3.

The MCMC algorithm has four component chains, each modifying a different aspect of the state $x = (M,q,s,p_e,G_1,G_2)$. The MCMC method produces a sequence of states $x^1,x^2,...,x^n$, representing a realization of a Markov chain that has invariant distribution $\pi$ of Equation (10). Each step in this chain is produced by attempting four different changes to $x$. That is, the MCMC algorithm is formed from four component chains, each updating a different component of the state $x$. All four have the Metropolis-Hastings (MH) form (Hastings, 1970); we did not use the Gibbs sampler because conditional distributions are awkward. From the current state $x$, a proposal $x^*$ is sampled, its conditional density being $q(x,x^*)$. The MH ratio

$$r = \frac{\pi(x^*) q(x^*,x)}{\pi(x) q(x,x^*)}$$  \hspace{1cm} (11)$$

determines the chances that $x^*$ becomes the next step in the chain. With probability $\min(r,1)$, the chain moves to $x^*$ or it stays at the present value $x$.

Our algorithm has four blocks of updating as mentioned above. The first updates missing genotypes $G_1,G_2$, the second updates selection coefficient $s$, the third jointly updates $(M,q)$, and the fourth updates environmental effects $p_e$. (The block to update $p_e$ can be omitted depending upon our assumptions about the environmental effects; see Section 4.3.) One cycle of the algorithm is created by applying the four updates in turn. To create posterior
summarizes, the chain is run for some large number of cycles - - 7,000 in our example (based on the covariance structure of the sample output; see Section 5).

4.2 Missing Data and Genetic Parameter Updates

In applying the MH algorithm to update each of the missing data values, \(v_{1i}, w_{1i}, v_{2i}, \) and \(w_{2i},\) we adopt as proposal distributions the trinomial distributions corresponding to probabilities \(T_{1i}, T_{2i}\) from Section 3. For instance, the proposal distribution for \(v_{1i}\) is Binomial\(\{v_{1i}, \xi_{1i}/(1-\xi_{1i})\}\) derived from the trinomial distribution corresponding to \(T_{1i}\) in (8). The MH ratio is

\[
r = \frac{\pi(v_{1i}^*) T_{1i}(v_{1i}|w_{1i})}{\pi(v_{1i}) T_{1i}(v_{1i}^*|w_{1i})},
\]

where \(T_{1i}(v_{1i}|w_{1i})\) is the probability function of \(v_{1i}\) for the binomial distribution, which is the conditional probability of \(v_{1i}\) given \(w_{1i}\). We move to a new value \(v_{1i}^*\) with probability \(\min(1,r)\); otherwise we stay at the present value \(v_{1i}\). The other missing genotypes are updated similarly. The acceptance rates of these MH updates are about 37 to 43%.

For updating \(s\), we use a random walk proposal distribution (Tierney, 1991). A candidate \(s^*\) is sampled uniformly from a small interval that has length \(2\delta_1\) centered at the current value \(s\). The resulting proposal density is symmetric as we allow proposals beyond the unit interval. Thus, the MH ratio is reduced to the posterior ratio \(\pi(s|\text{rest})/\pi(s|\text{rest})\) of the candidate \(s^*\) and the current state value \(s\). The width of the neighborhood is determined by trial and error, balancing the acceptance rate of proposals with the size of jumps to obtain good mixing (low autocorrelations) of the Markov chain (Roberts, Gelman, and Gilks, 1994). We adopt the interval length, \(\delta_1 = .04\), that produces an acceptance rate of 28%, in the example.

In a preliminary study, the \(M\) and \(q\) appeared to be highly correlated in their posterior; they are distributed in an attenuated region along a reciprocal line as is to be expected by inspection of the likelihood function. Such a high correlation among states is known to cause problems for any single-site updating scheme (e.g., Hills and Smith, 1992; Newton, Guttorm and Abkowitz, 1992). When we used two single-site proposal distributions, our MCMC implementation mixed extremely slowly (data not shown). To overcome this problem, we use a vector proposal distribution based on the knowledge that lines of constant \(Mq\) will have approximately constant posterior density:

1. Sample \(M^*\) from a discrete uniform distribution in the interval \((M_L, M_U)\).
2. For some constant \(\delta_2\), sample \(q^*\) uniformly from the interval \((M_L - \delta_2, M_L + \delta_2)\) where \(M, q\) are current values. When \(Mq - \delta_2 < 0\), \(q^*\) is sampled from the interval \((0, M_L + 2\delta_2)\).

The MH ratio is calculated as before. Noting the asymmetry of the proposal, we obtain

\[
r = \frac{\pi(M^*, q^*|\text{rest}) M}{\pi(M, q|\text{rest}) M^*} = \frac{\pi(G_1, G_2, M^*, q^*, s, p_c) M}{\pi(G_1, G_2, M, q, s, p_c) M^*}.
\]

The interval length \(\delta_2 = .6\) produces an acceptance rate of 31% in our example.
4.3 Environmental Effects

Even with the two-generation design, we do not obtain separate information about environmental effects because they are completely confounded with genetic effects of mortality. From the mortality probabilities in equations (4) and (7), \( p_e \) appears to contribute the same constant factor to both generations, so that this parameter results in the same proportion of reduction on the survival probabilities over the two generations. Thus, the mortality caused by environmental effects cannot be observed separately from that caused by genetic effects in this design. Using an asymptotic argument for mean convergence, Lee, Nordheim, and Kang (1994) show that genetic mortality is negligible in outcrossing data, and thus environmental effects of mortality can be estimated by using these data. Results from our outcrossing experiment are summarized in Table 1 (first row). In this section, using data from this outcrossing experiment, we obtain posterior information of environmental effects, which we use to form a prior on \( p_e \) in the main selfing design. This posterior distribution of \( p_e \) turns out to have approximately a Beta distribution.

To obtain a posterior distribution from the outcrossing data, we construct the likelihood function as follows. We define two distributions: 1) the sampling distribution of an outcrossing individual with respect to the genotypes of parents, and 2) the survival distribution of the individual. The former is a multinomial distribution of \( \{x_i, y_i, z_i, M = (x_i + y_i + z_i)\} \) corresponding to the numbers of loci of both parents that are heterozygous \( (x_i) \), one heterozygous and the other homozygous \( (y_i) \) of deleterious genes, and both homozygous \( (z_i) \), respectively. This multinomial distribution has parameters \( M, 2^2, 2^1, 2^1, 2^2 \), and \( 1 - (2^2 + 2^1, 2^1 + 2^2) \) using definitions in Section 3. The survival distribution is a Bernoulli trial of \( d_{oi} \) for the mortality occurrence of each seed. This Bernoulli trial has parameter \( Q_{oi} = 1 - (1 - p_e)(1 - s/4)^{x_i}(1 - s/2)^{y_i}(1 - s)^{z_i} \). Denoting \( M_{oi}, B_{oi} \) for the probability expressions of these two distributions, the joint probability of genotypes and phenotypes in this outcrossing design having \( N_o \) seeds is

\[
\Pr[G_o, D_o | M, q, s, p_e] = \prod_{i=1}^{N_o} M_{oi} B_{oi} \quad (14)
\]

where \( G_o = \{x_i, y_i, z_i\} \) and \( D_o = \{d_{oi}\} \) with \( i = 1, \ldots, N_o \).

Using the same independent uniform priors on the three parameters \( M, q, \) and \( s \) as in Section 4.2, we consider another independent uniform prior on \( p_e \), so that the posterior is proportional to \( (14) \). We need to derive the (marginal) posterior of \( p_e \) from this joint posterior. A direct integration or approximation to the marginal posterior is not tractable as with the case of the selfing design. However, because the deleterious gene frequency is conjectured to be very low in nature, less than \( 10^{-d} \) due to mutation-selection balance (Namkoong and Bishir, 1987), the genetic effects of mortality are negligible in outcrossing. Consequently, the posterior likelihood function of \( p_e \) from \( (14) \) is approximated by Binomial\( (N_o, p_e) \) (as the probability function of \( d_o \)), where \( d_o = \sum_{i=1}^{N_o} d_{oi} \) (Appendix 1). Therefore, the outcrossing posterior of \( p_e \) is approximately Beta with parameters \( d_o + 1 \) and \( N_o - d_o + 1 \). For germination and flowering data respectively (Table 1), these two posteriors are estimated as Beta\( (22, 220) \) and Beta\( (28, 214) \) distributions.
We treat each of these Beta posteriors in turn as the prior of \( p_e \) in the main selfing study. This parameter in the MCMC run is updated by applying an MH algorithm with the random walk proposal distribution on \( p_e \), similar to updating \( s \) in Section 4.2. The MH ratio in this case is simply the posterior ratio of a new candidate value \( p_e^* \) and the current value \( p_e \). The acceptance rate of this MH step is about 38%, in our example.

5 Results

We consider three possible assumptions about environmental effects in analyzing \( B. \ napus \) data. The first assumption is to ignore environmental effects altogether by supposing \( p_e = 0 \). The second is to assume \( p_e = \hat{p_e} \), the (approximate) MLE from outcrossing data. The third is to use the Beta priors of \( p_e \) obtained from the outcrossing experiment as described in the previous section.

We performed six MCMC runs on the \( B. \ napus \) data, one for each combination of viability measure (germination, flowering) and assumption about environmental effects. These are summarized in Table 2. For each run, we (somewhat conservatively) discarded the 2,000 initial iterations because time series plots of the samples warn of slow mixing. We found high autocorrelations in the time series plots of the state parameters from the MCMC run and thus applied subsampling at every 10-th cycle. This leads to 500 states \( x = (G_1, G_2, M, q, s, p_e) \) from the 7,000 cycles. Because autocorrelation inflates the variance of the estimates from the MCMC run, we estimated excess variance using a Tukey-Hanning window estimator (Geyer, 1992). The variance inflation indicates that the 500 states are equivalent to at least 125 independent samples from the posterior. Each run took about 20 minutes on a Dec Alpha system running a Fortran program.

As noted, \( p_e = 0 \) and \( p_e = \hat{p}_e \) use the MCMC implementation without considering environmental effects of mortality. For \( p_e = \hat{p}_e \), we substitute \( p_e \) with the fixed estimates of environmental effects \( p_e \), which were estimated as .0875 (\( = 21/240 \)) and .1125 (\( = 27/240 \)) from the germination and flowering data of outcrossing in Table 1, respectively.

[Table 2 goes here.]

In Table 2, we can see how the MCMC estimates of the number of lethal equivalents are affected by our assumptions about environmental effects. Assuming \( p_e = 0 \) gives different results from the others. If we use \( p_e = \hat{p}_e \), the results are very close to those from \( p_e \sim \text{Beta} \), suggesting that the Beta variance has relatively little impact. The estimates of the number of lethal equivalents (selection coefficient) when \( p_e = 0 \) are bigger (smaller) than those of both when \( p_e = \hat{p}_e \) and when \( p_e \sim \text{Beta} \). However, the estimates in the latter two cases are similar. We cannot obtain precise estimates of the parameters \( M \) and \( q \) because the posterior variances are very large. As pointed out earlier, these parameters are highly correlated along a reciprocal line (Figure 2). However, we do obtain good estimates of the product of these two parameters (Table 2).

[Figure 2 goes about here.]
The selection coefficient is very closely estimated in the two approaches $p_c=\hat{p}_c$ and $p_c\sim\text{Beta}$. Both produce a very high estimate of selection coefficient. Thus, our $B.\ \text{napus}$ result shows that this species carries a few lethal ($s \approx 1$) genes rather than many mildly deleterious ($s \ll 1$) ones. Figure 3 shows the (marginal) posterior densities of $s$ for both germination and flowering data using $p_c\sim\text{Beta}$ approach. From these posterior densities the confidence bounds of $s$ can be obtained by the HPD (highest posterior density) region, which are $(.956, 1)$ and $(.964, 1)$ for germination and flowering data, respectively.

[Figure 3 goes about here.]

We check sensitivity to the prior assumptions using two different choices of lower and upper bounds of parameter $M$, $(M_L, M_U) = (100, 5,000)$ and $(1,000, 50,000)$. The estimates of $s$, $Mq$, and $E$ are not sensitive to choices of these bounds. The estimate of $s$ is highly concentrated close to 1 and does not vary much by the different choices of the bounds, which leads to stable estimates of $Mq$ (and so of $E$). For instance, $\hat{s} = .99, \hat{Mq} = 2.30$, and $\hat{E} = 4.55$ for $(100, 5,000)$ and $\hat{s} = .99, \hat{Mq} = 2.32$, and $\hat{E} = 4.57$ for $(1,000, 50,000)$. However, separate estimates of $M$ and $q$ strongly depend on these choices; for the two choices, $M$ and $q$ are estimated as $3,259.5, 8.75e-4$ and $33,456.6, 8.7e-5$, respectively.

6 A Simulation Study

To validate our MCMC approach, we apply it to five sets of simulated data generated from the same model -- design as that for the real experiment. We chose parameter values $M = 3000, q = .002$, and $s = .45$ (so, $E = 2Mqs = 5.4$ and $Mq = 6$); these are values of interest for some applications to forest genetics; the value for $E$ is similar than the estimated value for $B.\ \text{napus}$ data, although $s$ is different. We ignore the environmental effects of mortality in this simulation. We apply MCMC to each data set using the same burn-in time, subsampling rule, and total run length as before to obtain 500 draws of parameter states from each run. On average, the posterior means of the parameters $E$, $s$, and $Mq$ are close to their true parameter values (Table 3).

[Table 3 goes about here.]

For all five data sets, interval estimates of $E$, $s$, and $Mq$ contain the corresponding true parameter values. For instance, in Figure 4 we plot the (posterior) marginal density of $s$ for the first simulated data set; the 95 % HPD region is computed as $(.445, .480)$ in this case.

[Figure 4 goes about here.]

The joint posterior of $M$ and $q$ appears to be concentrated along a reciprocal line as in Figure 2. A significant bias is observed in the estimates of $M$ and $q$ separately; the sample mean of $M$ ($q$) is overestimated (underestimated). However, these estimates are not precise; their estimated values by the MCMC samples are quite different from the true parameter values.
for the simulated data and their sample variances are relatively large. Furthermore, as with the real data case, these estimates of $M$ and $q$ strongly depend on the prior distribution of $M$.

Using the same alternative combinations of $(M_L, M_U)$ as in the real data, we checked sensitivity to the prior assumptions. For the first simulated data set, using $(100, 5,000)$ and $(1,000, 50,000)$, $M$ and $q$ are estimated as 3,380.8, 1.9e-3 and 33,241.5, 2.7e-5, respectively; these estimates seem to be kept roughly proportional to the upper bound $M_U$ of $M$. The estimates of $E$, $s$, and $Mq$ are 5.75, .42, 5.28 (std. error .31, .064, .58) and 5.69, .54, 6.92 (std. error .35, .048, .99) for the two combinations of bounds, respectively. Thus, the estimates of these three parameters are also affected by the choice of the bounds, but the impacts are relatively small compared to those of $M$ and $q$. Note that for these simulated data, the estimate of $s$ slightly increases as does the upper bound $M_U$; this estimate was not sensitive to the choice of the bounds for the real data.

7 Discussion

7.1 Statistical Issues

The two-stage design for our selfing experiment allows us to obtain separate information about the number of lethal equivalents ($E$) and the selection coefficient ($s$). This design does not give separate information about parameters $M$ and $q$, although we do obtain a precise estimate of their products of these two parameters. The MCMC approach enables us to infer parameter values in the presence of a complicated likelihood function that has a large number of missing data.

In our MCMC implementation, the two-dimensional proposal distribution of $M$ and $q$ dramatically improves the efficiency of the algorithm over a single component updating scheme such as the Gibbs sampler (data not shown). Often reparametrization is suggested to circumvent such difficulty (Hills and Smith, 1992). This may be necessary in a one-dimensional updating scheme. However, our vector proposal updating for $(M, q)$ does not attempt to reparametrize. Instead, we devise a proposal distribution that can move along the original space of these non-linearly correlated parameters, avoiding the problem of a slow mixing Gibbs sampler.

We also note that the interval widths of the random walk proposals for $s$ and $(M, q)$ were chosen carefully such that the choices balance rate of mixing with acceptance rate for efficiency; the interval widths were chosen to give acceptance rates of 28 and 31% (Roberts, Gelman, and Gilks, 1994). The acceptance rates of updating missing genotypes were between 37 to 43% and that of $p_e$ was 38%.

Considering our priors as noninformative for the three parameters $M$, $q$, and $s$, we have checked sensitivity to the priors only for the upper and lower bounds $M_U, M_L$ of $M$ using two different choices of $(M_L, M_U)$ - - (100, 5,000) and (1,000, 50,000). The MCMC estimates of $E$ were not sensitive to the choices of these bounds. The estimates of $s$ and $Mq$ seem to be slightly affected by these choices when $s$ is small (for the simulated data in Section
6). However, for both real and simulated data, the separate estimates of \( M \) and \( q \) strictly depend on these choices; these estimates seem to vary proportional to the upper bound \( M_U \) of \( M \).

Using only one generation of data, the adjustment to the number of lethal equivalents by environmental effects can be estimated by methods given in Lee, Nordheim, and Kang (1994). In this method, the numerical impact of the environmental effects can be approximated by a logarithmic function on the survival from the outcrossing data; these values are .36 and .47 corresponding to the germination and flowering cases, respectively. The asymptotic estimates of the number of lethal equivalents after adjusting for environmental effects are obtained as 3.95 and 4.07, with sample variances .176 and .185, for the two data sets. The MCMC estimates of the numbers of lethal equivalents with \( p_e \sim \text{Beta} \) are slightly larger than those obtained by the asymptotic method, but differences are not significant with respect to their sample variations (see Table 2). However, because the asymptotic estimates are derived from data based only on the first generation, these two methods are difficult to compare directly.

### 7.2 Genetic Interpretation

For \( B. napus \) data, our results imply that there are a few fully deleterious (lethal) genes rather than a large number of mildly deleterious genes in the base population. The reduction of mortality between the first and second generations is caused by the impact of the selection coefficient \(-\) the intensity of mortality of deleterious genes (Hedrick, 1994). In general, the higher the selection coefficient, the bigger the reduction of mortality at the second generation. Given the number of lethal equivalents, a high intensity of mortality due to deleterious genes results in those genes being placed among fewer loci than for a low intensity case. Thus, in a population that has a large number of mildly deleterious genes, each progeny may preserve more deleterious loci, which may be more difficult to purge from a breeding stock than in a population that has a few lethals. Therefore, in different breeding programs, the selection coefficient can suggest using different breeding strategies.

We did not obtain very much separate information on parameters \( M \) and \( q \) in our example. However, because \( M \) and \( q \) have their own biological importance, geneticists are often interested in the characteristics of these parameters themselves. Thus, a further study can be undertaken to seek out possible likelihood structures that allow separate identifications.

To make separate inferential statements of the parameters about the number of lethal equivalents \( E \) and the selection coefficient \( s \), we needed to consider more than a one-stage mating scheme. We considered a two-generation mating design, but our model can be generalized to multi-generation designs. For simplicity we restricted attention to selfing (as the type of mating), although the model could be extended to other types of mating, e.g., full-sibs or half-sibs (cf. Lee, Nordheim, and Kang, 1994). In a further study we intend to relax some restrictions, e.g., constant frequency and selection coefficient of deleterious genes. We did not consider another important genetic parameter, the dominance coefficient \( h \) of deleterious genes. Essentially, we have assumed \( h = 0 \); that is, an individual heterozygous for a deleterious gene is always viable. This also needs to be relaxed in a further study.
because deleterious genes often are not recessive and may play an important role even in heterozygous states. The difficulty may be overcome by using a combination of different mating systems and/or different types of data, e.g., molecular marker data.
Appendix: Approximation of Outcrossing Likelihood

In Equation (14) $M_{oi}$ has its support on $\{(x_i, y_i, z_i)|0 \leq x_i + y_i + z_i \leq M\}$. However, the probability that at least one of $x_i, y_i, z_i$, is greater than zero is negligible.

$$\Pr(\{(x_i, y_i, z_i)|x_i + y_i + z_i > 0\}) = 1 - \Pr(\{(x_i, y_i, z_i)|x_i + y_i + z_i = 0\})$$
= $$1 - (1 - \xi_{11} - \xi_{11}\xi_{12} - \xi_{12}^2)^M$$
= $$1 - \sum_{k=0}^{M}(\frac{M!}{k!})(-1)^k(\xi_{11}^2 + \xi_{11}\xi_{12} + \xi_{12}^2)^k$$
= $$M(\xi_{11}^2 + \xi_{11}\xi_{12} + \xi_{12}^2) + O(M^2\xi_{11}^4)$$
= $$M\xi_{11}^2 + O(M\xi_{11}^2) \text{ (since } \xi_{12} << \xi_{11})$$

In natural plant species, the average number of deleterious genes carried by an individual ($M\xi_{11}$) is small. Assuming that the frequency of heterozygous individuals of a deleterious gene ($\xi_{11}$) is less than $10^{-4}$ at most, the probability $M\xi_{11}^2$ is very small - in general, less than $10^{-2}$. Therefore, the probability that at least one of $x_i, y_i, z_i$ is greater than zero is negligible, and we can approximate the multinomial probability $M_{oi}$ in (14) to the case only when $\{(x_i, y_i, z_i)|x_i + y_i + z_i = 0\}$.

References


Table 1. Survival Data for a Two Generation Selfing Experiment of *Brassica Napus* L. The experiment was conducted simultaneously with 120 outcrossing seeds and 30 selfed families. Each selfed family had a random sample of 48 seeds per generation; the second generation selfing was assigned to a randomly sampled offspring of the first-generation survivors of each family.

<table>
<thead>
<tr>
<th>outcrossing</th>
<th>generation 1</th>
<th>generation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>germ</td>
</tr>
<tr>
<td>selfing 1</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>48</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>48</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>38</td>
</tr>
<tr>
<td>7</td>
<td>48</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>48</td>
<td>25</td>
</tr>
<tr>
<td>9</td>
<td>48</td>
<td>16</td>
</tr>
<tr>
<td>10</td>
<td>48</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>48</td>
<td>30</td>
</tr>
<tr>
<td>14</td>
<td>48</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>48</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>48</td>
<td>29</td>
</tr>
<tr>
<td>17</td>
<td>48</td>
<td>42</td>
</tr>
<tr>
<td>18</td>
<td>48</td>
<td>31</td>
</tr>
<tr>
<td>19</td>
<td>48</td>
<td>22</td>
</tr>
<tr>
<td>20</td>
<td>48</td>
<td>41</td>
</tr>
<tr>
<td>21</td>
<td>48</td>
<td>36</td>
</tr>
<tr>
<td>22</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>48</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>48</td>
<td>4</td>
</tr>
<tr>
<td>27</td>
<td>48</td>
<td>34</td>
</tr>
<tr>
<td>28</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>48</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>48</td>
<td>26</td>
</tr>
</tbody>
</table>
Table 2. MCMC Estimates of Genetic Parameters and Environmental Effects for Germination and Flowering Data of *Brassica Napus* L. The MCMC runs were performed separately on three models considering three different assumptions of environmental effects $p_e$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$p_e = 0$</th>
<th>$p_e = \hat{p}_e$</th>
<th>$p_e \sim \text{Beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{\mu}_{</td>
<td>y}$</td>
<td>$\bar{\sigma}_{</td>
</tr>
<tr>
<td>Germination</td>
<td>$E$</td>
<td>4.74</td>
<td>.36</td>
</tr>
<tr>
<td></td>
<td>$s$</td>
<td>.70</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td>$Mq$</td>
<td>3.39</td>
<td>.37</td>
</tr>
<tr>
<td></td>
<td>$M$</td>
<td>6531.8</td>
<td>2409.4</td>
</tr>
<tr>
<td></td>
<td>$q$</td>
<td>6.64e-4</td>
<td>5.18e-4</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Flowering</td>
<td>$E$</td>
<td>4.92</td>
<td>.39</td>
</tr>
<tr>
<td></td>
<td>$s$</td>
<td>.75</td>
<td>.048</td>
</tr>
<tr>
<td></td>
<td>$Mq$</td>
<td>3.28</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td>$M$</td>
<td>6614.2</td>
<td>2362.3</td>
</tr>
<tr>
<td></td>
<td>$q$</td>
<td>6.36e-4</td>
<td>5.34e-4</td>
</tr>
<tr>
<td></td>
<td>$p_e$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3. MCMC Estimates of Genetic Parameters for Simulated Data. Each of the 5 simulated data sets was generated using parameter values \((M, q, s) = (3,000, .002, .45)\).

<table>
<thead>
<tr>
<th>set / par</th>
<th>(E)</th>
<th>(s)</th>
<th>(Mq)</th>
<th>(M)</th>
<th>(q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.55(.35)</td>
<td>.471(.059)</td>
<td>5.98(.75)</td>
<td>6592.1(2328.4)</td>
<td>1.12e-3(.76e-3)</td>
</tr>
<tr>
<td>2</td>
<td>5.02(.33)</td>
<td>.527(.085)</td>
<td>4.89(.89)</td>
<td>6607.7(2261.7)</td>
<td>8.51e-4(5.9e-4)</td>
</tr>
<tr>
<td>3</td>
<td>5.62(.33)</td>
<td>.415(.078)</td>
<td>7.02(1.43)</td>
<td>6755.4(2267.6)</td>
<td>1.30e-3(1.0e-3)</td>
</tr>
<tr>
<td>4</td>
<td>5.53(.32)</td>
<td>.439(.079)</td>
<td>6.52(1.31)</td>
<td>6711.3(2331.4)</td>
<td>1.22e-3(9.7e-4)</td>
</tr>
<tr>
<td>5</td>
<td>5.46(.32)</td>
<td>.390(.067)</td>
<td>7.23(1.50)</td>
<td>6713.6(2316.1)</td>
<td>1.35e-3(1.0e-3)</td>
</tr>
</tbody>
</table>

NOTE: Each value represents \(\mu_{\|y} (\hat{\sigma}_{\|y})\) - empirical mean (standard error) of MCMC samples.
Figure 1: Hierarchical Structure of Two-Generation Selfing Experiment.
Figure 2: Attenuated Joint Posterior of $(M, q)$ along a Reciprocal Line from MCMC Run for Germination Data of *Brassica Napus* L. Five hundred points of $(M, q)$ are plotted.
Figure 3: Marginal Posterior Densities of Selection Coefficient Estimated by Rao-Blackwellization: (a) Germination Data, (b) Flowering Data.
Figure 4: Marginal Posterior Density of Selection Coefficient Estimated by Rao-Blackwellization for the First Simulated Data Set. The data are generated by using parameter values \((M, q, s) = (3000, .002, .45)\).