Genet. Res., Camb. (2001), **77**, *pp.* 83–94. *With* 1 *figure*. Printed in the United Kingdom © 2001 Cambridge University Press

Inferring the trajectory of genetic variance in the course of artificial selection

D. SORENSEN*, R. FERNANDO[1] AND D. GIANOLA[2]
[1] *Department of Animal Science, Iowa State University, Ames, IA 50011-3150, USA*
[2] *Department of Animal Sciences, University of Wisconsin-Madison, Madison, WI 53706-1284, USA*

(*Received 29 October 1999 and in revised form 17 March and 31 July 2000*)

Summary

A method is proposed to infer genetic parameters within a cohort, using data from all individuals in an experiment. An application is the study of changes in additive genetic variance over generations, employing data from all generations. Inferences about the genetic variance in a given generation are based on its marginal posterior distribution, estimated via Markov chain Monte Carlo methods. As defined, the additive genetic variance within the group is directly related to the amount of selection response to be expected if parents are chosen within the group. Results from a simulated selection experiment are used to illustrate properties of the method. Four sets of data are analysed: directional selection with and without environmental trend, and random selection, with and without environmental trend. In all cases, posterior credibility intervals of size 95% assign relatively high density to values of the additive genetic variance and heritability in the neighbourhood of the true values. Properties and generalizations of the method are discussed.

1. Introduction

A seemingly unresolved problem in experimental quantitative genetics is as follows. Consider a selection experiment, or a livestock breeding programme, spanning several generations. How does one infer the genetic variance, the heritability, or any other genetic parameter of interest at any given generation (other than the initial generation) making use of all the data available?

A feature of most data sets collected from breeding programmes, or from designed selection experiments, is that not all individuals from a given generation are allowed to reproduce, thus leading to 'missing data' in some statistical sense. This problem has been studied from several perspectives, because such missing data are prevalent in genetic analysis. Pioneering work in genetics on inferences about parameters in populations undergoing selection has been that of Henderson *et al.* (1959) and of Curnow (1961). These authors did not make formal use of missing data theory, as developed by Rubin (1976) and Little (1976). Since then, there has been a growing literature on the analysis of missing or incomplete data, because this is a pervasive problem in survey sampling and in clinical trials. Reviews in the statistical literature up to the mid-1980s are in Little & Rubin (1987) and in Rubin (1987). Some of these ideas were introduced in animal breeding by Im *et al.* (1989). However, the topic of inferences under artificial selection continues to be of major interest in animal breeding and genetics, where considerable literature on the subject exists, from several different perspectives. A review can be found in Gianola *et al.* (1989). Gianola & Fernando (1986) proposed a Bayesian solution for a certain class of selection problems, where 'missingness' is ignorable, and Fernando & Gianola (1990) discussed it in detail. Sorensen *et al.* (1994) described how a Bayesian implementation via Markov chain Monte Carlo methods could be used to estimate the posterior distributions of base population parameters and of measures of the effectiveness of selection under ignorable 'missingness'. An application to a selection experiment for increased litter size in pigs is in Wang *et al.* (1994).

* Corresponding author. Present address: Section on Biometrical Genetics, Department of Animal Breeding and Genetics, Danish Institute of Agricultural Sciences, PB50, DK-8830 Tjele, Denmark. Tel: +45 89 99 12 15. Fax: +45 89 99 13 00. e-mail: snfds@genetics.agrsci.dk

In this work, a method is presented to infer the dynamics of genetic variance in the course of artificial selection. The approach is based on estimating the posterior distribution of the variance of additive genetic values at any point in time using the entire data. The paper is organized as follows. First, since problems related to inferences about parameters using selected data are central to the topic, conditions for ignorability of selection are briefly reviewed. Second, the logical basis of the procedure is developed, followed by a description of the model and of the inferential method. Third, results from a simulation study are presented, where it is shown that the method yields posterior distributions that cover well the true values of the desired variances. Generalizations and limitations of the approach are discussed in a concluding section.

2. Ignorable selection

Inferences drawn using the method proposed in this paper are valid only if it can be established that selection is ignorable. Here, a result in Gianola & Fernando (1986) is used, but a different (and more general) derivation is presented. Suppose that selection is based on a random vector $\mathbf{z}$, and that a discrete random variable $s(\mathbf{z})$, which we term the *selection function*, takes one of S disjoint values, leading to mutually exclusive and potentially observed data $\mathbf{y}_i$ ($i = 1, 2, \ldots, S$). Each of these data vectors $\mathbf{y}_i$ comprises the records that would be observed if a specific set of animals is selected, e.g. to produce additional records or to become parents of the following generation. For example, suppose that one of two cows is to be chosen to produce a second lactation record, based on first lactation yield. Let y_{jk} denote record k of cow j ($j = 1, 2$). Here $\mathbf{z} = [y_{11}, y_{21}]'$. If $y_{11} > y_{21}$, the additional record observed is y_{12} and $s(\mathbf{z}) = 1$, leading to $\mathbf{y}_1 = [y_{11}, y_{21}, y_{12}]'$. Conversely, if $y_{11} \leqslant y_{21}$, the additional observation is y_{22}, with $s(\mathbf{z}) = 2$ and $\mathbf{y}_2 = [y_{11}, y_{21}, y_{22}]'$. Hence, the sample space of the discrete random variable $s(\mathbf{z})$ represents all possible patterns ('breeding designs'), and the selection process amounts to choosing one among all possible patterns. More generally, $s(\mathbf{z})$ may pertain to 'breeding designs' in a multi-generation selection experiment. The random variable $s(\mathbf{z})$ has an associated probability distribution indexed by parameters $\boldsymbol{\varphi}$. These parameters are related to the selection process but are not necessarily of scientific interest.

The observed data vector is $\mathbf{y}_i$, and these records are used to infer a vector of parameters $\boldsymbol{\lambda}_i$; the parameter vector is data-specific, because, for example, unknown breeding values of individuals associated with $\mathbf{y}_1$, say, will be different from those associated with $\mathbf{y}_2$. In the absence of selection, inferences are based on the posterior distribution of $\boldsymbol{\lambda}_i$, with density $p(\boldsymbol{\lambda}_i|\mathbf{y}_i)$. Under selection, the posterior density of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\varphi}$ is proportional to:

$$p_{sel}(\boldsymbol{\lambda}_i, \boldsymbol{\varphi} \,|\, \mathbf{y}_i) \propto p(\boldsymbol{\lambda}_i, \boldsymbol{\varphi}) \sum_{j=1}^{S} p(\mathbf{y}_j, s(\mathbf{z}) = j \,|\, \boldsymbol{\lambda}_j, \boldsymbol{\varphi})\, \delta(i-j). \quad (1)$$

In (1), $p(\boldsymbol{\lambda}_i, \boldsymbol{\varphi})$ is the joint prior density of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\varphi}$, and $\delta(0) = 1$ (that is, when $s(\mathbf{z}) = i$ and breeding design i is chosen) and 0 for any other δ. This yields:

$$\begin{aligned} p_{sel}(\boldsymbol{\lambda}_i, \boldsymbol{\varphi} \,|\, \mathbf{y}_i) &\propto p(\boldsymbol{\lambda}_i, \boldsymbol{\varphi})\, p(\mathbf{y}_i, s(\mathbf{z}) = i \,|\, \boldsymbol{\lambda}_i, \boldsymbol{\varphi}) \\ &= p(\boldsymbol{\lambda}_i, \boldsymbol{\varphi})\, p(\mathbf{y}_i \,|\, \boldsymbol{\lambda}_i) \Pr[s(\mathbf{z}) = i \,|\, \mathbf{y}_i, \boldsymbol{\lambda}_i, \boldsymbol{\varphi}]. \end{aligned}$$

The posterior density of $\boldsymbol{\lambda}_i$ is obtained integrating $\boldsymbol{\varphi}$ out:

$$p_{sel}(\boldsymbol{\lambda} \,|\, \mathbf{y}_i) \propto p(\mathbf{y}_i \,|\, \boldsymbol{\lambda}_i) \int p(\boldsymbol{\lambda}_i, \boldsymbol{\varphi}) \Pr[s(\mathbf{z}) = i \,|\, \mathbf{y}_i, \boldsymbol{\lambda}_i \boldsymbol{\varphi}]\, d\boldsymbol{\varphi}. \quad (2)$$

The following two conditions are sufficient for ignorability of selection for Bayesian inference:

1. The distributions of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\varphi}$ are *a priori* independent.
2. Conditionally on the observed data $\mathbf{y}_i$ the probability of choosing design i does not depend on the parameters to be inferred, $\boldsymbol{\lambda}_i$, so that $\Pr[s(\mathbf{z}) = i \,|\, \mathbf{y}_i, \boldsymbol{\lambda}_i \boldsymbol{\varphi}] = \Pr[s(\mathbf{z}) = i \,|\, \mathbf{y}_i, \boldsymbol{\varphi}]$.

If these two conditions are satisfied, (2) can be written as:

$$\begin{aligned} p_{sel}(\boldsymbol{\lambda}_i, \,|\, \mathbf{y}_i) &\propto p(\mathbf{y}_i \,|\, \boldsymbol{\lambda}_i)\, p(\boldsymbol{\lambda}_i) \int p(\boldsymbol{\varphi}) \Pr[s(\mathbf{z}) = i \,|\, \mathbf{y}_i, \boldsymbol{\varphi}]\, d\boldsymbol{\varphi} \\ &\propto p(\boldsymbol{\lambda}_i \,|\, \mathbf{y}_i). \end{aligned}$$

In this case the Bayesian analysis proceeds in the usual manner, as though selection had not taken place. The second condition is satisfied:

- when selection is at random, in which case the probability distribution of the selection function $s(\mathbf{z})$ does not depend on $\boldsymbol{\lambda}_i$ (or on $\mathbf{y}_i$);
- under data-based selection (all the data or the relevant subset of it used to make selection decisions is included in the analysis). In this case, $\Pr[s(\mathbf{z}) = i \,|\, \boldsymbol{\lambda}_i, \boldsymbol{\varphi}, \mathbf{y}_i] = 1$, a degenerate distribution (a constant);
- there are cases when selection is based on a variable $\mathbf{w}$, not included in $\mathbf{y}_i$. Here, selection will be ignorable only if $\Pr[s(\mathbf{w}) = i \,|\, \boldsymbol{\lambda}_i, \boldsymbol{\varphi}, \mathbf{y}_i] = \Pr[s(\mathbf{w}) = i \,|\, \boldsymbol{\varphi}, \mathbf{y}_i]$; that is, when the distribution of $\mathbf{w}$, given the observed data $\mathbf{y}_i$, is independent of $\boldsymbol{\lambda}_i$.

3. Model and method of inference

(i) *Bayesian structure*

The method used to infer the evolution of genetic parameters during the course of a genetic experiment is illustrated using a single-trait model, assuming that

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

selection is ignorable in the preceding sense. Extension to multiple traits is straightforward. It is assumed that, conditionally on a vector of location parameters $\boldsymbol{\theta}$, of order $(p+q)\times 1$, the sampling distribution of the $n\times 1$ data vector $\mathbf{y}$ is the Gaussian process:

$$\mathbf{y}\,|\,\boldsymbol{\theta}, \sigma_e^2 \sim N(\mathbf{W}\boldsymbol{\theta}, \mathbf{I}\sigma_e^2), \tag{3}$$

where $\mathbf{W}$ is a known incidence matrix, $\boldsymbol{\theta}$ is a location vector, $\mathbf{I}$ is an identity matrix and σ_e^2 is a residual, strictly positive, component of variance. Partition the location parameters as $\boldsymbol{\theta}' = (\mathbf{b}', \mathbf{a}')$, where $\mathbf{b}$ has order $p\times 1$ and contains parameters whose assigned prior distribution is a p-dimensional hyper-cube in the domain $\mathbf{b}_{\min}$, $\mathbf{b}_{\max}$, with boundaries chosen appropriately. The vector $\mathbf{a}$, of order $q\times 1$, contains additive genetic values whose prior distribution is assumed to be multivariate normal, that is:

$$\mathbf{a}\,|\,\mathbf{A}, \sigma_a^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2). \tag{4}$$

In (4), $\mathbf{A}$ is a known additive genetic relationship matrix among the q additive genetic values, and σ_a^2 is the unknown additive genetic variance. In this context σ_a^2 is interpreted as the additive genetic variance in the base population from which individuals are randomly sampled at the onset of the selection process.

The unknown variance components are assigned independent (*a priori*) scaled inverted chi-square prior distributions:

$$\sigma_i^2\,|\,\nu_i, S_i \sim S_i\,\chi_{\nu_i}^{-2}, \quad i = e, a, \tag{5}$$

where ν_i and S_i are known hyperparameters specifying the form of the distribution. An improper uniform distribution can be retrieved from (5) by setting $\nu_i = -2$ and $S_i = 0$.

Drawing Bayesian inferences about parameters of the probability model characterized by (3), (4) and (5) is well established in quantitative genetics, and a Markov chain Monte Carlo (MCMC) implementation can be found, for example, in Sorensen *et al.* (1994).

(ii) *Genetic considerations*

As discussed earlier, it is assumed that the data $\mathbf{y}$ accrue sequentially in time. At time 1, say, an arbitrary group of individuals is randomly drawn from some population. These individuals constitute a representative sample of all possible individuals born in generation 1. Within the individuals sampled, a fraction is chosen, either randomly or selectively (on the basis of their phenotypic values or of any function thereof), and allowed to reproduce, and a second generation of individuals is produced. Phenotypic records of non-parents are kept in the data. This process is repeated in each generation of offspring. As phenotypic records of individuals that do not contribute offspring are always kept in the data set, this satisfies the requirements for ignorability of selection, as discussed in the preceding section. Therefore, Bayesian inference about any function of the parameters can be drawn ignoring the selection process entirely.

The genetic model adopted is one of infinitesimally small, additive effects. In short, breeding values are postulated to stem from the sum of effects of alleles at an infinite number of loci, with each of the allelic contributions being infinitesimally small. The term $\mathbf{A}\sigma_a^2$ in (4) is the covariance matrix of the distribution of the vector of additive genetic values, for a fixed pedigree structure. Let $p(\mathbf{a})$ represent the probability density function of $\mathbf{a}$. Then, by definition:

$$\begin{aligned}\mathrm{Var}(\mathbf{a}\,|\,\mathbf{A}) &= \int \mathbf{a}\mathbf{a}' p(\mathbf{a})\,d\mathbf{a} - \left(\int \mathbf{a}\,p(\mathbf{a})\,d\mathbf{a}\right)\left(\int \mathbf{a}' p(\mathbf{a})\,d\mathbf{a}\right)\\ &= \mathbf{A}\sigma_a^2.\end{aligned} \tag{6}$$

The term σ_a^2 is a parameter of the marginal distribution of the scalar variable a_i, the i^{th} element of $\mathbf{a}$, which is the additive genetic value of an individual drawn at random from the population in question, such that:

$$\begin{aligned}\mathrm{Var}(a_i\,|\,A_{ii}) &= \int a_i^2 p(a_i)\,da_i - \left[\int a_i\,p(a_i)\,da_i\right]^2\\ &= (1+F_i)\,\sigma_a^2,\end{aligned} \tag{7}$$

where F_i is the inbreeding coefficient of an individual in the i^{th} position in $\mathbf{a}$, and A_{ii} is the i^{th} diagonal element of $\mathbf{A}$.

Consider a group or cohort consisting of a finite number of individuals. For example, the group could be composed of individuals born in a given generation, say t. Here, an experiment proceeding over time is envisaged, and suppose one is interested in characterizing the additive genetic variance of a breeding value randomly sampled at generation t. This variance will be denoted by $\sigma_a^{2(t)}$ and the group size by n_t. The additive genetic value of an individual sampled from generation t, a_t, is a random variable taking n_t possible values, each with probability $1/n_t$. By definition, the variance of a_t is:

$$\begin{aligned}\sigma_a^{2(t)} &= E(a_t^2) - [E(a_t)]^2\\ &= \frac{1}{n_t}\sum_{i=1}^{n_t} a_{i(t)}^2 - (\bar{a}_{(t)})^2,\end{aligned} \tag{8}$$

where $\bar{a}_{(t)} = E(a_t) = \frac{1}{n_t}\Sigma_{i=1}^{n_t} a_{i(t)}$ and $a_{i(t)}$ is the i^{th} additive genetic value in group t.

It is important to emphasize the conceptual difference between (8) and σ_a^2 in (6) or (7). In the case of (8), the variance pertains to the distribution of a_t, conditionally on the particular realization of the n_t additive genetic values of individuals in the cohort. The stochastic element here is associated with the random choice of a particular additive genetic value. This variance is directly related to the amount of genetic variability available for selection when choosing among these n_t additive genetic values. As shown in Appendix A, the variance defined by (8) corresponds

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

to twice the covariance between a randomly chosen parent from generation t and its offspring. On the other hand, in (7), the variance is computed for a particular element i of the randomly sampled $\mathbf{a}$, for a given pedigree. Contrary to (8), i in (7) is fixed and the vector of additive genetic values varies from sample to sample. For example, a Monte Carlo estimate of (7) would involve sampling repeatedly the q additive genetic values, conditionally on the pedigree relationship, and then computing the sample variance of a_i from the realized values at position i.

(iii) *Implementation*

Since the additive genetic values $a_{i(t)}$ are not observed, in order to learn about $\sigma_a^{2(t)}$ from data $\mathbf{y}$, use is made of the Bayesian paradigm to construct the marginal posterior distribution of $[\sigma_a^{2(t)}\,|\,\mathbf{y}]$, that is, the distribution of $\sigma_a^{2(t)}$ conditionally on the entire data $\mathbf{y}$. Because selection is ignorable, the posterior distribution of interest is as in the absence of selection. This distribution is not in a recognizable form, but it can be estimated via a Bayesian MCMC approach. Here, Gibbs sampling is used because all conditional posterior distributions are recognizable. The scheme operates as follows: (8) is computed in each iteration, substituting $a_{i(t)}, i = 1, 2, \ldots, n_t$ by the Monte Carlo draws obtained from the fully conditional posterior distribution of each $a_{i(t)}$. The Monte Carlo estimate of $[\sigma_a^{2(t)}\,|\,\mathbf{y}]$ can be computed by iterating on the following Gibbs sampling loop:

- Sample $\boldsymbol{\theta}' = (\mathbf{b}', \mathbf{a}')$ from $N(\hat{\boldsymbol{\theta}}, \mathbf{C}^{-1}\sigma_e^2)$, where

  $$\mathbf{C} = [\mathbf{W}'\mathbf{W} + \Sigma]; \Sigma = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k \end{bmatrix}; k = \sigma_e^2/\sigma_a^2$$

  and $[\mathbf{W}'\mathbf{W} + \Sigma]\hat{\boldsymbol{\theta}} = \mathbf{W}'\mathbf{y}$.
- Sample σ_i^2 from $\tilde{S}_i \chi_{\tilde{\nu}_i}^{-2}, i = e, a$

where

$$\tilde{\nu}_e = n + \nu_e;\ \tilde{\nu}_a = q + \nu_a; \tilde{S}_e = (\mathbf{y} - \mathbf{W}\boldsymbol{\theta})'(\mathbf{y} - \mathbf{W}\boldsymbol{\theta})/\tilde{\nu}_e;$$
$$\tilde{S}_a = \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}/\tilde{\nu}_a.$$

- Compute $\sigma_a^{2(t)} = \frac{1}{n_t}\sum_{i=1}^{n_t} a_{i(t)}^2 - (\bar{a}_{(t)})^2$, where a_i is an appropriate element of $\boldsymbol{\theta}$ sampled above.
- Update and return to first step. Repeat as needed to meet requirements of convergence diagnostics and to attain a high enough Monte Carlo precision.

This is illustrated below using simulated data mimicking a typical laboratory genetic experiment.

4. Simulation study

(i) *Design and assumptions*

The method proposed is illustrated using simulated experiments which generated four sets of data as explained below, without attending here to issues related to optimality of designs. Each experiment consisted of a single line monitored over 21 generations. In the first experiment, labelled as 'Random line', generation 1 was formed by sampling 200 males and 200 females at random from a conceptual base population. From these, 10 males and 10 females were randomly sampled and mated to produce 200 male and 200 female offspring. These progeny constituted generation 2. This random selection and mating procedure was repeated 20 times, leading to 8400 records spread over 21 non-overlapping generations.

The second experiment, labelled as 'Selected line' and which was independent of the first, differed only in that the 10 males and 10 females chosen as parents were those having the largest phenotypic values. Selected males and females were mated randomly to produce offspring. As with the Random line, this was repeated 20 times, leading, again, to 8400 records spread over 21 non-overlapping generations.

For the 400 records of generation 1, in each of the two lines, additive genetic values and environmental deviates were sampled independently from (also independent) normal distributions with mean zero and variances $\sigma_a^2 = 10$ and $\sigma_e^2 = 10$, respectively, so heritability (h^2) was $\frac{1}{2}$. In subsequent generations, additive genetic values were drawn from a normal distribution, with mean equal to the average additive genetic values of the parents of the individual in question, and with variance equal to $\frac{1}{2}\sigma_a^2(1-\bar{F})$, where σ_a^2 is the base population genetic variance, and $\bar{F}$ is average inbreeding coefficient of the parents of the individual. Under the infinitesimal model, the segregation variance is not affected by selection.

In both the Random line and the Selected line, a phenotypic record was generated by summing the additive genetic value and the residual deviate. There was no environmental trend (or other nuisance location parameters) built in the simulations which generated these two data sets. The mean phenotypic value in the Random line was approximately zero in all generations. In order to study the performance of the proposed method in the presence of environmental trend, two additional data sets were generated as follows. Phenotypic records of the Random and Selected lines without environmental trend were added an increasing amount of 2·5 units, starting in generation 2. Thus in the Random line with environmental trend, the phenotypic mean at generation 21 was approximately equal to 50 units. Since this is of about the same magnitude as the total selection response obtained in the Selected line, the phenotypic mean of the latter with environmental trend was approximately 100 units. In summary, four sets of data were created: two Random lines with and without environmental trend, respectively, and two Selected

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

Table 1. *Random line. Average inbreeding* ($\bar{F}_t$), *additive genetic variance* ($\sigma_a^{2(t)}$) *and heritability* ($h^{2(t)}$) *by generation, and MCMC estimates of within-generation posterior means of additive genetic variances* ($\hat{E}(\sigma_a^{2(t)}|\mathbf{y})$) *and heritability* ($\hat{E}(h^{2(t)}|\mathbf{y})$. *Posterior standard deviations of additive genetic variance and heritability are in parentheses*

		Generation (t)			
		5	10	15	20
$\bar{F}_t$		0·058	0·152	0·228	0·317
$\sigma_a^{2(t)}$		8·69	8·47	5·08	5·78
$\hat{E}(\sigma_a^{2(t)}\|\mathbf{y})$	$-ET$	8·25 (0·710)	8·63 (0·751)	4·93 (0·549)	5·96 (0·615)
	$+ET$	8·43 (0·733)	8·55 (0·773)	5·39 (0·582)	5·62 (0·640)
$h^{2(t)}$		0·46	0·46	0·34	0·37
$\hat{E}(h^{2(t)}\|\mathbf{y})$	$-ET$	0·45 (0·029)	0·46 (0·029)	0·32 (0·029)	0·36 (0·028)
	$+ET$	0·46 (0·031)	0·45 (0·031)	0·35 (0·032)	0·35 (0·031)

$-ET$ ($+ET$): model without (with) environmental trend.

lines with and without environmental trend, respectively. The structure of the two sets of data within Random and Selected lines is identical.

(ii) *Model for analysis*

The data were analysed according to the model specified in (3), (4) and (5). In the absence of environmental trend, **b** in (3) is a scalar; otherwise it contains 20 elements representing the environmental effects peculiar to each generation. The additive genetic variances at intermediate generations were computed as described in Section 3(iii).

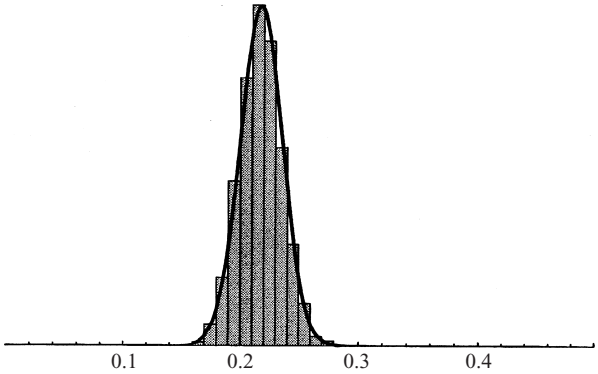

Fig. 1. Histogram of $[h^{2(20)}|\mathbf{y}]$, Selected line, with an overlaid normal distribution.

(iii) *Inferences and interpretation*

In the Random and Selected lines, inferences were drawn about the base population additive genetic variance (parameter σ_a^2 in (6) and (7)) and heritability, and about the additive genetic variance (the random variable $\sigma_a^{2(t)}$ defined in (8)) and heritability at generations 5, 10, 15 and 20. The base population additive genetic variance, σ_a^2, was inferred using established methods, i.e. Sorensen *et al.* (1994). The additive genetic variance ($\sigma_a^{2(t)}$) and heritability ($h^{2(t)} = \sigma_a^{2(t)}/(\sigma_a^{2(t)}+\sigma_e^2)$) at intermediate generations were inferred using the proposed method.

Results from these experiments are interpreted in the light of well-established quantitative genetic theory. With random mating, under the infinitesimal model, the additive genetic variance within a line declines as a consequence of inbreeding and of the correlation between additive genetic values that develops as family structure builds up. Additional positive or negative changes in additive genetic variance within the line may result from chance fluctuations in linkage disequilibrium (Avery & Hill, 1978). In conceptual replications of the experiment, the expected within-line additive genetic variance at generation t, $\sigma_a^{2(t)}$, is approximately equal to $\sigma_a^2(1-\bar{F}_t)$, where $\bar{F}_t$ is the average inbreeding coefficient in generation t (the exact expression is $\frac{\sigma_a^2}{n_t} tr[(\mathbf{I}-\frac{1}{n_t}\mathbf{J})\mathbf{A}_t]$, where $\mathbf{I}$ is an identity matrix, $\mathbf{J}$ is a square matrix of dimension n_t with all elements equal to 1, and $\mathbf{A}_t$ is the numerator relationship matrix between the n_t additive genetic values at generation t). In the Selected line, the same type of forces operate, but there is an additional reduction in additive genetic variance due to the generation of negative linkage (or joint) disequilibrium, the so-called Bulmer effect (Bulmer, 1971). Several simulation experiments in the literature have yielded results in agreement with these theoretical expectations, including Robertson (1977), Sorensen & Kennedy (1984) and Van der Werf & de Boer (1990).

(iv) *Results*

Results for the two Random lines are displayed in Table 1. The figures illustrate the decline in 'true'

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

Table 2. *Selected line. Average inbreeding* ($\bar{F}_t$), *additive genetic variance* ($\sigma_a^{2(t)}$) *and heritability* ($h^{2(t)}$) *by generation, and MCMC estimates of within-generation posterior means of additive genetic variances* ($\hat{E}(\sigma_a^{2(t)}|\mathbf{y})$) *and heritability* ($\hat{E}(h^{2(t)}|\mathbf{y})$. *Posterior standard deviations of additive genetic variance and heritability are in parentheses*

		Generation (t)			
		5	10	15	20
$\bar{F}_t$		0·214	0·313	0·478	0·565
$\sigma_a^{2(t)}$		6·37	5·63	3·77	3·18
$\hat{E}(\sigma_a^{2(t)}\|\mathbf{y})$	$-ET$	6·79 (0·543)	5·52 (0·464)	3·60 (0·316)	2·94 (0·285)
	$+ET$	6·02 (0·644)	5·99 (0·554)	3·87 (0·380)	3·27 (0·325)
$h^{2(t)}$		0·39	0·36	0·27	0·24
$\hat{E}(h^{2(t)}\|\mathbf{y})$	$-ET$	0·40 (0·021)	0·36 (0·020)	0·26 (0·018)	0·22 (0·018)
	$+ET$	0·38 (0·029)	0·38 (0·028)	0·29 (0·024)	0·25 (0·023)

$-ET$ ($+ET$): model without (with) environmental trend.

additive genetic variance and heritability (from basal levels of $\sigma_a^2 = 10$ and $h^2 = \frac{1}{2}$, respectively) under the infinitesimal model. As pointed out earlier, this is a result of the build up of covariances among additive genetic values, and of inbreeding. Also, results reflect random departures, peculiar to this line, from expectations over conceptual repeated sampling (Avery & Hill, 1978). The posterior distributions of $\sigma_a^{2(t)}$ and of $h^{2(t)}$ were well approximated by a normal distribution (not presented) and therefore only posterior standard deviations are shown. The Monte Carlo estimates of the mean, standard deviation, and of the 95% posterior intervals of the posterior distributions of $\sigma_a^{2(t)}$ and of $h^{2(t)}$ (mean $\pm$ 1·96 posterior standard deviations) indicate very good coverage of the true additive genetic variance and heritability at each of the generations monitored. This holds also for the line in which there is environmental trend. Here, as expected, the posterior standard deviation is a little larger than in the line without environmental trend. In short, the trajectories of additive genetic variance and of heritability under random selection with or without environmental trend are inferred correctly using the method proposed.

Results for the two Selected lines are shown in Table 2. As expected, inbreeding developed at a faster rate, and the additive genetic variance and heritability fell more rapidly in the Selected than in the Random line. Part of the extra decline in additive genetic variance and heritability is due to the Bulmer effect. As for the Random line, the trajectories of genetic variance and heritability in the course of selection (with or without environmental trend) were captured well by the proposed Bayesian method. The difference in posterior uncertainty with and without environmental trend is considerably larger under directional selection than under random selection.

As was the case with the Random lines, the posterior probability intervals were symmetric, which was confirmed graphically. To illustrate, a histogram of the posterior distribution of heritability at generation 20 from the Selected line without environmental trend is shown in Fig. 1, overlaid against a normal distribution fit based on the posterior mean and variance. Departures from normality seemed negligible despite the fact that only 10 pairs of parents contributed offspring to the next generation. A very similar pattern but with larger variation was obtained from the Selected line with environmental trend (not shown).

The estimate of the posterior mean of the base population additive genetic variance in the Random line without environmental trend was 9·58, and the posterior standard deviation was 0·99. With environmental trend, the posterior mean and standard deviation were 10·72 and 1·06, respectively. Corresponding values for the Selected line without environmental trend were 9·80 and 0·45 and with environmental trend 9·67 and 0·81. The posterior intervals assign relatively high density to values of the additive genetic variance in the neighbourhood of the true value of 10. As was the case with the additive variance at intermediate generations, the fall in precision of inferences about the base population additive genetic variance in the presence of environmental trend is relatively larger in the Selected lines than in the Random lines.

A comparison of these measures of spread and those in Tables 1 and 2, indicates that posterior inferences about additive genetic variance and heritability in the base population (or in any of the generations monitored) were sharper in the Selected than in the Random line. This is because the variance of the posterior distribution of additive genetic

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

variance decreases as the correlation among the additive genetic values increases. With the type of selection practised, this correlation is lower under random mating, and thus the posterior variance is higher in the Random line. This is discussed in Appendix B. On the other hand, the posterior variance of genetic means increases as additive genetic values become more correlated (Sorensen *et al*., 1994).

5. Discussion

A method for inferring the variance within an arbitrarily defined group of individuals that makes use of all data (that is, observations from other groups are also used in the analysis) was presented and evaluated. In a genetic context, the group is usually composed of individuals belonging to a certain generation or time period. Hence, the method can be used to study how genetic variance evolves during the course of a selection experiment. As defined, the additive genetic variance within the group is directly related to the amount of selection response to be expected if parents are chosen within that group. A genetic model of infinitesimally additive effects was posited, and the validity of the results presented here rests on this assumption. In order to study the evolution of genetic variance under other forms of gene action, or if a finite number of loci is assumed, the same conceptual framework applies. However, the form of the joint posterior distribution would differ as well as, possibly, the implementation. Clearly, if an additive model is to be used, this must be done after it has been found to be more plausible than competing models relying on different genetic assumptions. This process of model comparison would benefit from the presence of a control line. Techniques for model selection are not dealt with in this paper.

The method can be extended in a straightforward manner to infer genetic covariances or correlations at given time periods, provided that selection is ignorable. This would require computing, for example:

$$\mathrm{Cov}(a_i, a_j)^{(t)} = \frac{1}{n_t}\sum_k a_{k,i(t)}\, a_{k,j(t)} - (\bar{a}_{i(t)}, \bar{a}_{j(t)})$$

rather than (8), for additive genetic values of traits i and j in group t.

Inferences about heritability pertaining to intermediate generations of a selection experiment can be drawn using offspring–mid-parent regressions. This is a form of inference based on the conditional distribution of offspring means, given the means of the parental phenotypic values. That is, to infer heritability at generation t, say, data from generations t and $t+1$ only are used, in contrast to the proposed method, which uses all data available. The posterior uncertainty about heritability derived from offspring–parent regressions is inversely proportional to the variance among parents. If these are selected, resulting in reduced parental variation, inferences using this simple method will tend to be very imprecise. This was examined in the Random and Selected lines (without environmental trend) at generations 5, 10, 15 and 20. If non-informative priors are used, the posterior distribution of heritability is a truncated-t in the interval $(0, 1)$ (Box and Tiao, 1973); the degrees of freedom are the number of pairs of parents, minus 2. We drew 30000 samples from this distribution using the method of composition (Tanner, 1996). Results indicated clearly that the method can be extremely imprecise. For example, in the Random line at generation 20, the 2·5 and 97·5 percentiles of the posterior distribution of heritability were 0·18 and 0·63, respectively. In the Selected line, the respective values were 0·03 and 0·98. For this setting, the regression procedure is worthless, and much sharper inferences can be obtained with the proposed Bayesian method (Tables 1, 2).

A more elaborate approach was proposed by Sorensen & Kennedy (1984) using a likelihood-type estimator. In their study, individuals of the generation whose variance was of interest (the base individuals) were treated as unrelated, and data from earlier generations were omitted from the analysis. Simulation experiments suggested that their method retrieved estimates of the additive genetic variance that were in good agreement with the true variance, on average. In a later study, Van der Werf & de Boer (1990) found the agreement to be less satisfactory, however. The *ad hoc* method lacks the formal theoretical foundation of the present approach, and does not make use of all the data available.

A model to study the evolution of genetic parameters with selection was presented by Beniwal *et al*. (1992) and by Heath *et al*. (1995). In this model the conditional additive genetic variance of the offspring, given the parents, changes with time. Estimation of these additive genetic variances at each generation can disclose whether predictions based on the infinitesimal model hold; the latter assumes that this Mendelian sampling variance remains constant throughout the selection process, after accounting for inbreeding. The model proposed by these authors is different from the one presented here, both statistically and genetically. It is different statistically because the variances at each generation are parameters and, as such, are part of a likelihood equation. In the method presented here, the genetic variance within a cohort is a random variable, and it is therefore not part of a likelihood equation. It is different genetically, because in the model of Beniwal *et al*. (1992), the variances inferred are Mendelian sampling variances, whereas those inferred with the proposed method represent the variances available for selection within the group in question. Of course in principle the method proposed here could also be used

with the model proposed by Beniwal *et al*. (1992). It would yield a Monte Carlo estimate of the posterior distribution of the variance available for selection in the group, under the model posed.

The performance of the new method was illustrated doing the analysis within the randomly selected line or within the directionally selected line. While a single selected line does not necessarily constitute a good design to infer genetic parameters, especially in the presence of environmental trend, the selected line may be all that there is available to draw inferences. It was shown that even in this unfavourable situation, the proposed method can give a very adequate picture of the variance at a particular generation. Often a selection experiment comprises data from both a selected and a control line started from a common base population. The new method can be used to do a joint analysis using data from both lines. To illustrate, the additive genetic variance at generations 5, 10, 15 and 20 was inferred in the Selected line, making use of data from both the Selected and the Random line, resulting in 16800 records. In the absence of environmental trend, the mean of the posterior distribution of the additive genetic variance (posterior standard deviation) in the Selected line at generations 5, 10, 15 and 20, was 6·75 (0·53), 5·47 (0·45), 3·64 (0·32) and 2·91 (0·27), respectively. A comparison with the results using data from the Selected line only in Table 2 (*–ET*), shows that the gain in efficiency from the extra data is relatively modest. In the presence of environmental trend, the corresponding figures are 6·72 (0·56), 5·43 (0·47), 3·57 (0·33) and 2·98 (0·29); the gain in efficiency is considerably more pronounced (compare with results in Table 2, (*+ET*)).

The method proposed was formulated within the Bayesian framework, using a MCMC sampling scheme; this gives a simple solution. A relevant question is whether a satisfactory non-Bayesian alternative exists. From a sampling theory point of view, any function of the additive genetic variance at time t, as defined here, is an unobservable random variable. Hence it is natural to infer this variable from its conditional distribution, given the data. Specifically, we are interested in the conditional distribution of the random variable:

$$\sigma_a^{2(t)} = \frac{1}{n_t}\sum_{i=1}^{n_t} a_{i(t)}^2 - (\bar{a}_{(t)})^2,$$

given the observations. This conditional distribution is also a posterior distribution (given the data, the fixed effects and the dispersion components), so it is unaffected by ignorable selection, as discussed earlier. As in the Bayesian setting, the density of the target distribution cannot be derived in closed form, and depends on unknown parameters, that is, the fixed effects and the variance components. However, if one replaces such parameters by, for example, maximum likelihood estimates, it is possible to estimate the conditional distribution of $\sigma_a^{2(t)}$ using MCMC methods. The difference compared with the Bayesian approach is that the frequentist parameters would not be involved (updated) in the sampling procedure. Conceptually, one draws m samples from the conditional distribution:

$$\mathbf{a}\,|\,\mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y} \sim N\left[\left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1} \times \mathbf{Z}'(\mathbf{y} - \mathbf{X}\mathbf{b}), \left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2\right],$$

either directly, if feasible, or using MCMC otherwise. From the sampled $\mathbf{a}$, and for each sample k ($k = 1, 2, \ldots, m$), one stores the subvector of breeding values at generation t, $\mathbf{a}_{(t)}$, and forms:

$$\{\sigma_a^{2(t)}\}^{[k]} = \frac{1}{n_t}\left[\mathbf{a}'_{(t)}\,\mathbf{a}_{(t)} - \frac{(\mathbf{1}'\mathbf{a}_{(t)})^2}{n_t}\right]^{[k]}; k = 1, 2, \ldots, m.$$

Thus, $\{\sigma_a^{2(t)}\}^{[k]}$ is a draw from the conditional distribution $\sigma_a^{2(t)}\,|\,\mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y}$. From the collection of draws, the desired conditional distribution can be estimated, given values of $\mathbf{b}, \sigma_a^2, \sigma_e^2$. If a point predictor of the random variable $\sigma_a^{2(t)}$ is sought, one can follow Henderson (1973, 1975), and use the best predictor in the mean squared error sense, that is, the conditional mean. Alternative point predictors are the conditional median (minimizing the expected value of the absolute error of prediction) or the conditional mode. The conditional mean can be expressed analytically as:

$$\begin{aligned} E[\sigma_a^{2(t)}\,|\,\mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y}] &= \frac{1}{n_t}E\left[\left(\mathbf{a}_{(t)} - \mathbf{1}\frac{(\mathbf{1}'\mathbf{a}_{(t)})}{n_t}\right)' \times \left(\mathbf{a}_{(t)} - \mathbf{1}\frac{(\mathbf{1}'\mathbf{a}_{(t)})}{n_t}\right)\right] \\ &= \frac{1}{n_t}E\left[\mathbf{a}'_{(t)}\left(\mathbf{I} - \frac{\mathbf{J}}{n_t}\right)\mathbf{a}_{(t)}\right] \\ &= \frac{1}{n_t}\left\{\hat{\mathbf{a}}'_{(t)}\left(\mathbf{I} - \frac{\mathbf{J}}{n_t}\right)\hat{\mathbf{a}}_{(t)} + tr\left[\left(\mathbf{I} - \frac{\mathbf{J}}{n_t}\right) \times \mathrm{Var}(\mathbf{a}_{(t)}\,|\,\mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y})\right]\right\}, \end{aligned}$$

where $\mathbf{1}$ is an n_t times 1 vector of ones, $\mathbf{J}$ is an n_t times n_t matrix of ones,

$\hat{\mathbf{a}}_{(t)} = E(\mathbf{a}_{(t)}\,|\,\mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y})$

is an appropriate subvector of

$$\left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

and $\mathrm{Var}(\mathbf{a}_{(t)}\,|\,\mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y})$ is an appropriate submatrix of $(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2})^{-1}\sigma_e^2$. Analytical calculation of the best predictor requires computing the inverse of the part of the mixed model equations pertaining to the

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

additive genetic values only. This can be avoided using a MCMC approximation to $\mathrm{Var}(\mathbf{a}_{(t)} \mid \mathbf{b}, \sigma_a^2, \sigma_e^2, \mathbf{y})$. Clearly, knowledge of $\mathbf{b}, \sigma_a^2, \sigma_e^2$ (base population parameters) is needed for constructing the conditional distribution, and for deriving a suitable point predictor. If maximum likelihood estimates are used, this frequentist approach does not take into account the uncertainty in the estimation of these parameters.

An alternative frequentist approach would be to infer the variance at time t, from its conditional distribution given a vector $\mathbf{v}$ of $n-\mathrm{rank}(\mathbf{X})$ linearly independent 'error contrasts', in the REML sense. Conceptually, one would draw samples of $\mathbf{a}$ from the conditional distribution:

$$\mathbf{a} \mid \sigma_a^2, \sigma_e^2, \mathbf{v} \sim N\left[\left(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1} \times \mathbf{Z}'\mathbf{M}\mathbf{y}, \left(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2\right]$$

and form the corresponding draws from the conditional distribution of $\sigma_a^{2(t)}$, as before. Above, $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the usual projection matrix. This sampling procedure permits estimating the entire conditional distribution. Analytically, the conditional mean is:

$$E[\sigma_a^{2(t)} \mid \sigma_a^2, \sigma_e^2, \mathbf{v}] = \frac{1}{n_t}\left\{\hat{\mathbf{a}}'_{(t)}\left(\mathbf{I} - \frac{\mathbf{J}}{n_t}\right)\hat{\mathbf{a}}_{(t)} + tr\left[\left(\mathbf{I} - \frac{\mathbf{J}}{n_t}\right)\mathrm{Var}(\mathbf{a}_{(t)} \mid \sigma_a^2, \sigma_e^2, \mathbf{v})\right]\right\},$$

where now, $\hat{\mathbf{a}}_{(t)} = E(\mathbf{a}_{(t)} \mid \sigma_a^2, \sigma_e^2, \mathbf{v})$ is an appropriate subvector of

$$\left(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\mathbf{Z}'\mathbf{M}\mathbf{y}$$

and $\mathrm{Var}(\mathbf{a}_{(t)} \mid \sigma_a^2, \sigma_e^2, \mathbf{v})$ is an appropriate submatrix of

$$\left(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_a^2}\right)^{-1}\sigma_e^2.$$

This *ad hoc* method eliminates the requirement of having to specify values of the fixed effects (elimination of nuisance parameters), but still depends on knowledge of the variance components of the base population, which must be estimated by some method (e.g. REML). If such estimates are used in lieu of the true variance parameters, the method does not take into account their error of estimation, as in the case of maximum likelihood. It should be noted that this second frequentist method also has a Bayesian interpretation, because the conditional distribution $\mathbf{a} \mid \sigma_a^2, \sigma_e^2, \mathbf{v}$ is identical to the posterior distribution $\mathbf{a} \mid \sigma_a^2, \sigma_e^2, \mathbf{y}$ obtained by integrating the joint posterior distribution $\mathbf{a}, \mathbf{b} \mid \sigma_a^2, \sigma_e^2, \mathbf{y}$ with respect to $\mathbf{b}$ in a Bayesian probability model, where $\mathbf{b}$ is assigned a flat (improper) distribution (Gianola & Fernando, 1986).

It is not obvious which of the two frequentist methods, i.e. infer $\sigma_a^{2(t)}$ from an estimated conditional distribution using maximum likelihood versus restricted maximum likelihood estimates of the base population parameters, has better repeated sampling properties. It is even less obvious whether a better method (with respect to some loss function) exists, in which case neither of the two frequentist point predictors would be admissible (Gianola, 1990). Contrary to the self-contained fully Bayesian approach, no account is taken of the uncertainty associated with estimates of the base population parameters (heritability, residual variance and fixed effects in an univariate model). Depending on the dimension of the problem at hand, the frequentist approaches may be computationally less time-consuming than the fully Bayesian solution, in which case they may be used for approximate inferences. However, their inability to account for uncertainty of estimates can be a potentially serious problem in a highly parameterized, multi-trait model.

It can not be overemphasized that the method proposed here works provided selection is ignorable, as discussed earlier. When this is not the case, but something is known about the form in which selection alters the data-generating mechanism, attempts should be made to incorporate such knowledge in the modelling process. Otherwise, inferences are liable to be distorted.

Appendix A. The covariance between a randomly chosen parent at generation t and its offspring is equal to $\frac{1}{2}\sigma_a^{2(t)}$

It is shown that the covariance between a randomly chosen parent from generation t and its offspring is $\frac{1}{2}\sigma_a^{2(t)}$, where $\sigma_a^{2(t)}$ is the variance of the additive genetic values of the individuals in generation t.

It is assumed that the trait is additive and that parents are randomly sampled from generation t. Without loss of generality, it is supposed that the trait is controlled by only two loci. The paternal and maternal alleles at locus $l (l = 1, 2)$ are denoted as p_i^l and m_i^l. For simplicity non-genetic effects on phenotypic values will be ignored.

The phenotypic value of individual i in generation t can be written as:

$$y_i = \mu + \alpha_{p_i^1} + \alpha_{m_i^1} + \beta_{p_i^2} + \beta_{m_i^2} + e_i, \qquad \text{(A1)}$$

where μ is the phenotypic mean in generation 0, $\alpha_{p_i^1}$ and $\alpha_{m_i^1}$ are the additive effects at the first locus, and $\beta_{p_i^2}$ and $\beta_{m_i^2}$ are the additive effects at the second locus, defined using the genetic frequencies in generation 0, and e_i is an independently distributed residual. In generation 0, the additive effects have null expectations. With selection, the expected values of additive effects will not be null but, in all generations, e_i will

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

have null expectations and are assumed to be uncorrelated with all other effects. Covariances between additive effects within and across loci are not null because parents are sampled from a finite population and, also, because of selection. Let a_i be the sum of the additive effects of individual i. Then the variance of a_i is:

$$\begin{aligned}\mathrm{Var}(a_i) = {} & \mathrm{Var}(\alpha_{p_i^1}) + \mathrm{Cov}(\alpha_{p_i^1}, \alpha_{m_i^1}) \\ & + \mathrm{Cov}(\alpha_{p_i^1}, \beta_{p_i^2}) + \mathrm{Cov}(\alpha_{p_i^1}, \beta_{m_i^2}) \\ & + \mathrm{Cov}(\alpha_{m_i^1}, \alpha_{p_i^1}) + \mathrm{Var}(\alpha_{m_i^1}) \\ & + \mathrm{Cov}(\alpha_{m_i^1}, \beta_{p_i^2}) + \mathrm{Cov}(\alpha_{m_i^1}, \beta_{m_i^2}) \\ & + \mathrm{Cov}(\beta_{p_i^2}, \alpha_{p_i^1}) + \mathrm{Cov}(\beta_{p_i^2}, \alpha_{m_i^1}) \\ & + \mathrm{Var}(\beta_{p_i^2}) + \mathrm{Cov}(\beta_{p_i^2}, \beta_{m_i^2}) \\ & + \mathrm{Cov}(\beta_{m_i^2}, \alpha_{p_i^1}) + \mathrm{Cov}(\beta_{m_i^2}, \alpha_{m_i^1}) \\ & + \mathrm{Cov}(\beta_{m_i^2}, \beta_{p_i^2}) + \mathrm{Var}(\beta_{m_i^2}). \end{aligned} \tag{A2}$$

Suppose i is the father of j. Then, because of random mating, the alleles of maternal origin received by j are uncorrelated with those of i. This will be true even in a finite population as long as the parents of j are mated at random. Further, by assumption, the residuals of i and j are uncorrelated. Thus, the covariance between phenotypic values i and j is:

$$\begin{aligned}\mathrm{Cov}(y_i, y_i) = {} & \mathrm{Cov}(\alpha_{p_j^1}, \alpha_{p_i^1}) + \mathrm{Cov}(\alpha_{p_j^1}, \alpha_{m_i^1}) \\ & + \mathrm{Cov}(\alpha_{p_j^1}, \beta_{p_i^2}) + \mathrm{Cov}(\alpha_{p_j^1}, \beta_{m_i^2}) \\ & + \mathrm{Cov}(\beta_{p_j^2}, \alpha_{p_i^1}) + \mathrm{Cov}(\beta_{p_j^2}, \alpha_{m_i^1}) \\ & + \mathrm{Cov}(\beta_{p_j^2}, \beta_{p_i^2}) + \mathrm{Cov}(\beta_{p_j^2}, \beta_{m_i^2}) \end{aligned} \tag{A3}.$$

Because i is the father of j, the set of paternal alleles (p_j^1, p_j^2) of j is a copy of one of the four sets (p_i^1, p_i^2), (p_i^1, m_i^2), (m_i^1, p_i^2) or (m_i^1, m_i^2) of i. Let the value of the random variable Z indicate the set of alleles j received from i. For example, when j receives (p_i^1, m_i^2) from i, Z will be equal to 2. Using Z, the above covariance can be written as:

$$\begin{aligned}\mathrm{Cov}(y_i, y_i) = {} & E[\mathrm{Cov}(y_i, y_j) \mid Z] \\ & + \mathrm{Cov}[E(y_i \mid Z), E(y_j \mid \mid Z)]. \end{aligned} \tag{A4}$$

The second term of (A4) is null because the expected value of y_i (or of y_j) does not depend on Z. Now, when $Z = 1$, $p_j^1 = p_i^1$ and $p_j^2 = p_i^2$. It follows that:

$$\begin{aligned}\mathrm{Cov}(y_i, y_j \mid Z = 1) = {} & \mathrm{Var}(\alpha_{p_i^1}) + \mathrm{Cov}(\alpha_{p_i^1}, \alpha_{m_i^1}) \\ & + \mathrm{Cov}(\alpha_{p_i^1}, \beta_{p_i^2}) + \mathrm{Cov}(\alpha_{p_i^1}, \beta_{m_i^2}) \\ & + \mathrm{Cov}(\beta_{p_i^2}, \alpha_{p_i^1}) + \mathrm{Cov}(\beta_{p_i^2}, \alpha_{m_i^1}) \\ & + \mathrm{Var}(\beta_{m_i^2}) + \mathrm{Cov}(\beta_{p_i^2}, \beta_{m_i^2}). \end{aligned} \tag{A5}$$

Note that the first line of (A5) is identical to the first line of (A2), and the second line of (A5) is identical to the third line of (A2). The first line of (A2) is the sum of the covariances of $\alpha_{p_i^1}$ with itself and with the effects of the remaining alleles. The second line of (A2) is the sum of the covariances of $\alpha_{m_i^1}$ with itself and with the effects of the remaining alleles. In a purebred population, the paternal and maternal alleles are identically distributed. Therefore, the sum of the covariances of line 1 of (A2) is equal to the sum of the covariances on line 2. Similarly, the sum of the covariances on line 3 of (A2) is equal to the sum of the covariances on line 4. Thus, (A5) $= \frac{1}{2}\mathrm{Var}(a_i)$. When $Z = 2$, $p_j^1 = p_i^1$ and $p_j^2 = m_i^2$. So, $\mathrm{Cov}(y_i, y_j \mid Z = 2)$ is equal to the sum of the covariances on lines 1 and 4, and this is equal to $\frac{1}{2}\mathrm{Var}(a_i)$. Similarly, $\mathrm{Cov}(y_i, y_j \mid Z = k) = \frac{1}{2}\mathrm{Var}(a_i)$ for $k = 3, 4$. Therefore the first term of (A4) is $\frac{1}{2}\mathrm{Var}(a_i)$, from which it follows that the covariance between a randomly chosen parent from generation t and its offspring is $\frac{1}{2}\sigma_a^{2(t)}$.

Appendix B. The posterior variance of $\sigma_a^{2(t)}$ as a function of the posterior correlation between breeding values

If a random vector $\mathbf{a}$ is distributed as $\mathbf{a} \sim N(\boldsymbol{\mu}, \mathbf{V})$, and $\mathbf{Q}$ is a known square matrix, the quadratic form $\mathbf{a}'\mathbf{Q}\mathbf{a}$ has variance (Searle, 1971):

$$\mathrm{Var}(\mathbf{a}'\mathbf{Q}\mathbf{a}) = 2tr(\mathbf{Q}\mathbf{V})^2 + 4\boldsymbol{\mu}'\mathbf{Q}\mathbf{V}\mathbf{Q}\boldsymbol{\mu}. \tag{B1}$$

Assume location and the dispersion parameters of the base population to be known. Then, the posterior distribution of the breeding values is multivariate normal, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$, say. The relationship between the variance of $\sigma_a^{2(t)}$ and the posterior correlation between breeding values is illustrated first in a simple situation and, subsequently, in a slightly more general setting.

Suppose that at generation t there are two individuals, with breeding values a_1 and a_2, respectively. Then:

$$\sigma_a^{2(t)} = \frac{1}{4}(a_1 - a_2)^2,$$

which is a quadratic form on the two breeding values. Using (B1), the posterior variance of $\sigma_a^{2(t)}$ is:

$$\begin{aligned}\mathrm{Var}[\sigma_a^{2(t)}] \propto \mathrm{Var}(a_1 - a_2)^2 = {} & 2\mathrm{Var}^2(a_1 - a_2) \\ & + 4(\mu_1 - \mu_2)^2\mathrm{Var}(a_1 - a_2). \end{aligned} \tag{B2}$$

If $\mu_1 = \mu_2$, the second term to the right of the equals sign vanishes. Now,

$$\begin{aligned}\mathrm{Var}(a_1 - a_2) = {} & \mathrm{Var}(a_1) + \mathrm{Var}(a_2) \\ & - 2\rho\sqrt{\mathrm{Var}(a_1)\,\mathrm{Var}(a_2)}, \end{aligned}$$

where ρ is the posterior correlation between breeding values. In the course of directional selection, the breeding values become less dispersed, their posterior correlation increases, and $\mathrm{Var}(a_1 - a_2)$ decreases; the increase in correlation is slower under random selection. Hence, (B2) decreases as ρ increases from 0 to 1. In the limit, when $\rho = 1$ (fixation), $\mathrm{Var}(a_1) = \mathrm{Var}(a_2)$, so $\mathrm{Var}(a_1 - a_2) = 0$, and, consequently, $\mathrm{Var}[\sigma_a^{2(t)}] = 0$.

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

A slightly more general situation follows. Given the location and dispersion parameters (of the base population), let the breeding values at generation t have the posterior distribution $\mathbf{a}_t \sim N(\boldsymbol{\mu}_t, \mathbf{V}_t)$. Then:

$$\sigma_a^{2(t)} = \frac{1}{n_t}\sum_{i=1}^{n_t} a_{i(t)}^2 - (\bar{a}_{(t)})^2 = \frac{1}{n_t}(\mathbf{a}_t - \mathbf{1}\bar{a}_t)'(\mathbf{a}_t - \mathbf{1}\bar{a}_t)$$
$$= \frac{1}{n_t}\mathbf{a}_t'\mathbf{M}\mathbf{a}_t, \qquad \text{(B3)}$$

where $\mathbf{1}$ is a vector of ones, and $\mathbf{M} = \mathbf{I} - \frac{1}{n_t}\mathbf{J}$ is a projection matrix (idempotent), with $\mathbf{J}$ being an $n_t \times n_t$ matrix of ones. If $\boldsymbol{\mu}_t = \mathbf{1}\mu_t$, that is, if all breeding values of individuals in generation t have the same posterior expectation, then, *a posteriori*, from (B1) one can write:

$$\mathrm{Var}(\sigma_a^{2(t)}) = 2\left(\frac{1}{n_t}\right)^2 tr(\mathbf{M}\mathbf{V}_t)^2$$
$$= 2\left(\frac{s^2}{n_t}\right)^2 tr(\mathbf{M}\mathbf{R}\mathbf{M}\mathbf{R}), \qquad \text{(B4)}$$

where it is assumed that breeding values at generation t have the same posterior variance (s^2) and posterior correlation matrix $\mathbf{R}$, with constant coefficient of correlation ρ in all its off-diagonal elements. As the posterior correlation between breeding values goes to zero (representing the situation before selection, if a sample of unrelated animals is drawn), $\mathbf{R} \to \mathbf{I}$ (identity matrix), so $tr\,(\mathbf{MRMR}) \to tr(\mathbf{M}) = n_t - 1$. Likewise, $s^2 \to \sigma_a^2$, the base population (or prior) variance. Hence:

$$\lim_{\rho \to \mathbf{0}} \mathrm{Var}(\sigma_a^{2(t)}) = \frac{2\sigma_a^4(n_t - 1)}{n_t^2}.$$

As data accrue and selection proceeds, $\mathbf{R} \to \mathbf{J}$ (matrix of ones) because breeding values become intercorrelated, so $tr\,(\mathbf{MRMR}) \to tr\,(\mathbf{MJMJ}) = \mathbf{0}$. Further, in a Gaussian model, $s^2 \leqslant \sigma_a^2$, necessarily, as Bayesian learning always reduces variance. Hence:

$$\lim_{\rho \to \mathbf{1}} \mathrm{Var}(\sigma_a^{2(t)}) = 0.$$

However, the trajectory of $\mathrm{Var}(\sigma_a^{2(t)})$ as ρ increases is not trivial. For this specific set of assumptions (common posterior variance, equi-correlation) consider (B4). Note that:

$$tr(\mathbf{MRMR}) = tr(\mathbf{R}^2) - \frac{2}{n_t}tr(\mathbf{JR}^2) + \frac{1}{n_t^2}tr(\mathbf{JRJR}). \qquad \text{(B5)}$$

After algebra, (B5) can be written as:

$$tr(\mathbf{MRMR}) = n_t[1 + (n_t - 1)\rho^2] - [1 + (n_t - 1)\rho]^2.$$

Using this in (B4):

$$\mathrm{Var}(\sigma_a^{2(t)}) = 2\left(\frac{s^2}{n_t}\right)^2 (n_t - 1)(1 - \rho^2). \qquad \text{(B6)}$$

The relationship between $\mathrm{Var}(\sigma_a^{2(t)})$ and ρ is not explicit because the posterior variance of breeding values (s^2) depends on ρ as well. Typically, in a Gaussian linear model such as the one employed here, s^2 is an increasing function of ρ, but the relationship depends on the data structure and the degree of relatedness between relatives (through the matrix $\mathbf{A}$). For example, suppose that the available information consists of phenotypic records on two related individuals, having an additive relationship equal to a_{12}. Using standard theory, and assuming a known mean (0) and that the residual variance is equal to 1, the posterior covariance matrix between breeding values a_1 and a_2 is equal to:

$$[\mathbf{I} + \mathbf{A}^{-1}]^{-1} = \frac{1 - a_{12}^2}{(1 - a_{12}^2 + \alpha)^2 - a_{12}^2\alpha} \times \begin{bmatrix} 1 - a_{12}^2 + \alpha & a_{12}\alpha \\ a_{12}\alpha & 1 - a_{12}^2 + \alpha \end{bmatrix},$$

where $\alpha = \frac{1 - h^2}{h^2}$. The posterior correlation is then:

$$\rho = \frac{a_{12}\,\alpha}{1 - a_{12}^2 + \alpha}$$

going to 1 as $a_{12} \to 1$. The posterior variance of any of the two breeding values is expressible as:

$$s^2 = \frac{\alpha^{-1} - \alpha^{-1}a_{12}^2}{\alpha^{-1} - a_{12}\rho},$$

this being an increasing function of ρ. However, the dependence can be mild, as $a_{12}\rho$ may be negligible relative to α^{-1}, especially for traits of moderate to high heritability and with a sparse relationship structure. This indicates that if the posterior variance of the breeding values is mildly dependent on ρ, the posterior variance of $\mathrm{Var}(\sigma_a^{2(t)})$ will decrease with directional selection, as it is a decreasing function of ρ. This result should not be construed as general because, in less stylized settings, the necessary expressions cannot be obtained in closed form.

Our interest in this problem was rekindled by discussion with Bruce Walsh. Bernt Guldbrandtsen wrote the code for the generation of the simulated data. The support of the Wisconsin Agriculture Experiment Station and grant NRICGP/USDA 99-35205-8162 to D. G. is acknowledged.

References

Avery, P. & Hill, W. G. (1978). Variability in genetic parameters among small populations. *Genetical Research* **29**, 193–213.

Beniwal, B. K., Hastings, I. M., Thompson, R. & Hill, W. G. (1992). Estimation of changes in genetic parameters in selected lines of mice using REML with an animal model. 1. Lean mass. *Heredity* **69**, 352–360.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

https://doi.org/10.1017/S0016672300004845 Published online by Cambridge University Press

Bulmer, M. G. (1971). The effect of selection on genetic variability. *American Naturalist* **105**, 201–211.

Curnow, R. N. (1961). The estimation of repeatability and heritability from records subject to culling. *Biometrics* **17**, 553–566.

Fernando, R. L. & Gianola, D. (1990). Statistical inferences in populations undergoing selection or non-random mating. In *Advances in Statistical Methods for Genetic Improvement of Livestock* (ed. D. Gianola & K. Hammond), pp. 437–453. Berlin: Springer.

Gianola, D. (1990). Can BLUP and REML be improved upon? Proceedings of the 4th World Congress of Genetics Applied to Livestock Production **XIII**, 445–449.

Gianola, D. & Fernando, R. L. (1986). Bayesian methods in animal breeding theory. *Journal of Animal Science* **63**, 217–244.

Gianola, D., Fernando, R., Im, S. & Foulley, J. L. (1989). Likelihood estimation of quantitative genetic parameters when selection occurs: models and problems. *Genome* **31**, 768–777.

Heath, S. C., Bulfield, G., Thompson, R. & Keightley, P. D. (1995). Rates of change of genetic parameters of body weight in selected mouse lines. *Genetical Research* **66**, 19–25.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Proceedings of the Animal Breeding and Genetics Symposium in Honor of J. L. Lush*, pp. 10–41. Champaign, IL: American Society of Animal Science and American Dairy Science Association.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **55**, 423–447.

Henderson, C. R., Kempthorne, O., Searle, S. R. & von Krosigk, C. M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192–218.

Im, S., Fernando, R. & Gianola, D. (1989). Likelihood inferences under selection: a missing-data theory view. *Genetics Selection Evolution* **21**: 399–414.

Little, R. J. A. (1976). Inferences about means from incomplete multivariate data. *Biometrika* **63**, 593–604.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Robertson, A. (1977). Artificial selection with a large number of linked loci. In: *Proceedings of the International Conference on Quantitative Genetics* (ed. E. J. Pollak, O. Kempthorne & T. B. Bailey), pp. 307–322. Ames, Iowa: Iowa State University Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sorensen, D. & Kennedy, B. W. (1984). Estimation of genetic variances from selected and unselected populations. *Journal of Animal Science* **59**, 1213–1223.

Sorensen, D., Wang, C. S., Jensen, J. & Gianola, D. (1994). Bayesian analysis of genetic change due to selection using Gibbs sampling. *Genetics Selection Evolution* **26**, 333–360.

Tanner, M. A. (1996). *Tools for Statistical Inference*. Berlin: Springer.

Van der Werf, J. H. J. & De Boer, I. J. M. (1990). Estimation of additive genetic variance when base populations are selected. *Journal of Animal Science* **68**, 3124–3132.

Wang, C. S., Gianola, D., Sorensen, D. A., Jensen, J., Christensen, A. & Rutledge, J. J. (1994). Response to selection for litter size in Danish Landrace pigs: a Bayesian analysis. *Theoretical and Applied Genetics* **88**, 220–230.