A Test of Ignorability of Selection under the Infinitesimal Model: A Preliminary Report

Hossein Jorjani[1] & Daniel Gianola[2]

[1]Interbull Centre, Department of Animal Breeding & Genetics, Swedish University of Agricultural Sciences, Box 7023, S - 750 07, Uppsala, Sweden, [2]Department of Animal Sciences, University of Wisconsin, Madison 53706-1284, Wisconsin, USA

Abstract

Assessment of the constancy of variance of Mendelian sampling (MS) effects across generations is regarded as a useful tool for validation of genetic evaluation models that assume linearity of the offspring-parent regression or, more strongly, multivariate normality. Therefore, finding a well-posed method for inferring possible changes (or lack thereof) in the variance of MS effects over time is important. Here, we examine a procedure proposed by Sorensen *et al.* (2001) for inferring the trajectory of genetic variance under selection, which relies on absence of missing data and ignorability of selection. The method was used to assess changes in Mendelian sampling variance under some animal breeding situations with progressing amounts of missing data. Results from simulations indicate that, after discounting for inbreeding, the segregation variance was inferred well in most cases. However, constancy of variance of MS effects could not be corroborated under a sire model or when heritability was low. It seems that the variance of MS effects does not provide a sensitive enough end-point for discriminating between right and wrong evaluation models in many cases.

Introduction

The usefulness of the international genetic evaluations conducted at the Interbull Centre depends on the quality of input data provided by national genetic evaluation centers in the Interbull member countries. Another important factor is whether or not the assumptions made (implicitly or explicitly) in genetic evaluation models hold. Many methods have been suggested for validating genetic evaluation models (see Jorjani, 2003).

In particular, much emphasis has been placed on examining trends in the means and variances of predicted additive genetic values and of Mendelian sampling, MS, terms (e.g. Boichard *et al.*, 1995; Thompson, 2002; Klei *et al.*, 2002). This focus is related to the dependence of predicted breeding values on the input genetic parameters, especially in multi-trait models. The underlying rationale is: if genetic parameters are modeled and inferred appropriately, then there should be little reason for geneticists to worry about other parts of the evaluation models. However, this is debatable. Poor functional forms, lack of adequate transformation, non-linearity of effects, presence of limited-dependent variables (due to categorization and censoring), and effects of outliers may be as important as the input genetic parameters. On the other hand, the effect of genetic parameter values on the evaluations is very apparent, since slight changes in the values may produce differential shrinkage of deviations from contemporary groups. It must be kept in mind, however, that model selection and validation is a very complex issue (Burnham & Anderson, 2002; Sorensen & Gianola, 2002), and an examination of the behavior of putative MS terms may fail to reveal model inadequacies.

Here, we take the narrow point of view of a standard animal breeding model in which additive genetic effects are drawn from a multivariate normal process. The idea behind using MS for assessing model validity is as follows: under this infinitesimal model, and in the absence of inbreeding, the variance of MS (segregation variance) is assumed to be constant across generations. Let

$$\gamma_i = a_i - \left(\sum a_{i,j} a_j\right) \qquad [1]$$

where γ_i and a_j are the MS and breeding values of animal i (i=1, 2, …, N), respectively; a_j is the breeding value of ancestor j (usually sire and dam/maternal-grandsire), and $a_{i,j}$ is the additive genetic relationship between animal i and animal j. Under the "animal" model, an individual it (i=$1t$. $2t$,…, Nt) born in generation t, with known sire (s) and dam (d), has as Mendelian sampling term:

$$\gamma_{it} = a_{it} - \tfrac{1}{2} a_{s_{it}} - \tfrac{1}{2} a_{s_{it}} \qquad \textbf{[2]}$$

This is assumed to be distributed as:

$$\gamma_{it} \sim N\ (0,\ \sigma^2_a/2\ (1 - (F_s + F_d)/2)) \qquad \textbf{[3]}$$

where F_s and F_d are the sire's and dam's inbreeding coefficients, respectively, and σ^2_a is the base population additive genetic variance. As inbreeding accrues, the MS variance decreases, so the model may hold and, yet, the MS variance change due to inbreeding only. The effect of inbreeding can be discounted by noting that the rescaled MS term:

$$\gamma^*_{it} = \gamma_{it} \Big/ \sqrt{1 - (F_s + F_d)/2} \sim (0,\ \sigma^2_a/2) \qquad \textbf{[4]}$$

has a constant variance, across individuals and generations. If, given data, the variance of the rescaled γ^*_{it} effects is found to be constant over time, this should be construed as lack of evidence for refuting the infinitesimal model. The question is how the variance of γ^*_{it} can be inferred at any point of the selection process.

Parameter estimation methodology employed currently in connection with national genetic evaluations is ambiguous with respect to which population the estimates pertain to. There is agreement in that "unbiased" estimates of the genetic parameters for an unselected, non-inbred, base population can be derived from likelihood functions through inclusion in the analysis of the numerator relationship matrix ($\boldsymbol{A}$) and of all data used for selection decisions. However, it is less clear whether or not the variance of predicted breeding values for animals born in a certain year can be regarded as an estimate of the genetic variance for that year, or as another estimate of the base population value based on a sample of animals born in the year in question, or as none of the above.

A simple method for inferring segregation variance at any time follows directly from Sorensen *et al.* (2001). These authors define the parameter "additive genetic variance at time t" as the dispersion of breeding values about the appropriate generation mean, conditionally on the genotypic values of the individuals born in the cohort in question. A similar definition applies to the MS variance at time t, after rescaling for the effect of inbreeding. Since the breeding values or the MS terms are unobserved, the latter (as well as the MS variance) must be inferred somehow. A simple frequentist procedure might consist of a method of moments fit, where the BLUP of the MS (after rescaling) is used in the formulae for the variance at time t. This does not take uncertainty into account, plus population parameters are estimated with error. A Bayesian treatment solves these problems, but at the expense of introducing prior distributions.

Sorensen *et al.* (2001) make use of missing data theory developed during the 1970's and later (Little, 1976; Rubin, 1976), and present a Bayesian procedure (see also Gianola & Fernando, 1986). A key assumption of Sorensen *et al.* (2001) is that whatever selection has taken place must be ignorable, in a well defined and precise sense. The paper should be consulted for details of the technical argument. In the proposed method, Sorensen *et al.* (2001) infer the base population parameters simultaneously with the variance of breeding values of any cohort of animals born during the selection process, with all parameters inferred from marginal posterior distributions (so uncertainty is fully accounted for). Their procedure makes use of all the available data, but selection must be "ignorable", as pointed out earlier. For example, if the conditional distribution of the missing data given the observed data is degenerate (*i.e.*, all data employed for selection decisions are included in the analysis), selection is ignorable. On the other hand, if selection is for 2 correlated traits, but only records for one of the traits are used in the analysis, the selection process cannot be ignored; this has also been shown in the literature using less formal procedures.

The purpose of the present study is to assess empirically the stringency of the ignorability conditions under some common

animal breeding situations. For this purpose, genetic parameters (including the variance of Mendelian sampling terms) were inferred in a series of *in silico* populations undergoing ignorable or non-ignorable selection.

Material and Methods

The infinitesimal model was used to simulate 40 ("small" population) or 100 ("large") animals per sex in an unselected, non-inbred, base population. For each animal, two phenotypes were simulated from the bivariate normal distribution:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{a1}+\sigma^2_{e1} & \rho\sigma_{a1}\sigma_{a2} \\ \rho\sigma_{a1}\sigma_{a2} & \sigma^2_{a2}+\sigma^2_{e2} \end{pmatrix}\right] \quad \textbf{[5]}$$

where the σ^2_{ij} (i=a, e; j=1, 2) are additive genetic or environmental variances, and ρ is the genetic correlation between traits; the environmental correlation was zero.

Values for σ^2_{ij} were chosen in such a way to give coefficients of heritability h^2=0.05, 0.25 or 0.50; simulated values of the genetic correlation were ρ=0.2, 0.5 or 0.8, and σ^2_{pj} =100 was the phenotypic variance for trait j (j=1, 2). Each generation, a fraction of animals (5%, 25% or 50%) was selected as parents, either at random (designated as I_5), or based on the indexes (I):

$I_4 = 4/4\ y_1 + 0/4\ y_2$
$I_3 = 3/4\ y_1 + 1/4\ y_2$
$I_2 = 2/4\ y_1 + 2/4\ y_2$
$I_1 = 1/4\ y_1 + 3/4\ y_2$
$I_0 = 0/4\ y_1 + 4/4\ y_2$

In I_4, selection is based entirely on trait 1, and in I_0 selection is on trait 2; the other three indexes represent 2-trait selection scenarios.

The two traits were either expressed in both sexes or only in females (in which case males were selected at random). Selected animals were mated randomly. For offspring, the breeding values were simulated using a standard rearrangement of Eq. **[2]**. The same procedure was repeated for 10 additional generations; each selection protocol was replicated 16 times.

Genetic analysis

We followed Sorensen *et al.* (2001), as outlined below. Given a vector of location parameters $\boldsymbol{\theta}$ (*i.e,* systematic effects and breeding values), the sampling distribution of the single-trait data vector **y** was a Gaussian process described by

$$\mathrm{y}|\theta,\sigma^2_{\mathrm{e}} \sim \mathrm{N}\,(\mathrm{W}\theta\ ,\ \mathrm{I}\sigma^2_{\mathrm{e}}) \quad \textbf{[6]}$$

where **W** is a known incidence matrix, **I** is an identity matrix of order n and σ^2_e is the residual variance. The $\boldsymbol{\theta}$ vector can be partitioned into two sub-vectors: **b** (the systematic, non-genetic, effects) and **a** (additive genetic values). The latter were assigned the multivariate normal prior

$$\mathrm{a}|\mathrm{A},\sigma^2_{\mathrm{a}} \sim \mathrm{MVN}\,(0,\ \mathrm{A}\sigma^2_{\mathrm{a}}) \quad \textbf{[7]}$$

where **A** is the additive genetic relationship matrix between all individuals, of order q. The vector **b** was assigned an improper uniform prior. The unknown variance components σ^2_e and σ^2_a were assigned, *a priori,* the independent scaled inverted chi- square distributions:

$$\sigma^2_{\mathrm{e}}\ |\ \nu_{\mathrm{e}},\mathrm{S}_{\mathrm{e}} \sim \nu_{\mathrm{e}}\ \mathrm{S}_{\mathrm{e}}\ \chi^{-2}_{\nu_{\mathrm{e}}} \quad \textbf{[8]}$$

and

$$\sigma^2_{\mathrm{a}}\ |\ \nu_{\mathrm{a}},\mathrm{S}_{\mathrm{a}} \sim \nu_{\mathrm{a}}\ \mathrm{S}_{\mathrm{a}}\ \chi^{-2}_{\nu_{\mathrm{a}}} \quad \textbf{[9]}$$

Here, ν_e, S_e, ν_a and S_a are known hyper-parameters specifying the form of the corresponding distributions. Hyper-parameter values were $\nu_e = \nu_a =$ -2 and $S_e = S_a = 0$.

Consider now a cohort of individuals born in generation t. The additive genetic variance at generation t is denoted as $\sigma_a^{2(t)}$, and the cohort size is denoted as 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:

$$\sigma_a^{2(t)} = E(a_t^2) - [E(a_t)]^2 = 1/n_t \sum_{i=1}^{n_t} a^2_{i(t)} - (\bar{a}_{(t)})^2 \quad \textbf{[10]}$$

where

$$\bar{a}_{(t)} = E(a_t) = 1/n_t \sum_{i=1}^{n_t} a_{i(t)} \qquad \textbf{[11]}$$

and $a_{i(t)}$ is the i^{th} additive genetic value in generation t. The variance of the segregation residuals at generation t is defined in a similar manner, but working with **[2]** or **[4]** instead of with the breeding values. The variance of the breeding values and of the segregation residuals at any time during the selection process was inferred from their corresponding marginal posterior distributions, using a Gibbs sampling procedure described below.

MCMC Gibbs sampling scheme

The Gibbs sampling scheme operated as follows:

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

where

$\mathbf{C} = [\mathbf{W}'\mathbf{W} + \Sigma]$;

$\Sigma = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k \end{bmatrix}$;

$k = \sigma^2_e / \sigma^2_a$; and

$[\mathbf{W}'\ \mathbf{W} + \boldsymbol{\Sigma}]\hat{\boldsymbol{\theta}} = \mathbf{W}'\ \mathbf{y}$;

2- Sample σ^2_i from $\tilde{S}_i \chi^{-2}_{\tilde{\nu}_i}$, $i = e, a$,

where

$\tilde{\nu}_e = n + \nu_e$;

$\tilde{\nu}_a = q + \nu_a$;

$\tilde{S}_e = (y - W\theta)'(y - W\theta)/\tilde{\nu}_e$;

$\tilde{S}_a = a' A^{-1} a / \tilde{\nu}_a$;

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

where

a_i is an element of $\boldsymbol{\theta}$;

4- Update and return to the first step.

Extension of the method to the situation where Mendelian sampling effects are drawn is straightforward, and changes are restricted to Step 3 (above), so now:

a) Calculate γ_{it} (Eq. **[2]**) from the sampled values of **a**;
b) Calculate $\sigma^{2(t)}_{\gamma}$ (similar to Eq. **[10]**);
c) Calculate γ^*_{it} as in Eq. **[4]**, and
d) Calculate $\sigma^{2(t)}_{\gamma^*}$ (similar to Eq. **[10]**).

In two-trait models, the additive genetic and residual variances are replaced by 2 x 2 genetic and residual covariance matrices, respectively. The mixed model equations are modified for a 2-trait situation. The sampling procedure is as before, except that the prior and fully conditional scaled inverted chi-squared distributions are replaced by appropriate inverse Wishart processes. A new parameter appearing in the 2-trait situation is the covariance between MS effects for the two traits; we did not monitor this during the simulations.

For implementing the above algorithm, the Gibbs program package (Misztal, 2002) was modified as needed. The Gibbs sampler was run for 75000 iterations, of which the first 25000 were discarded as burn-in, and 1 out of each 25 consecutive samples was kept for examination of posterior distributions, i.e., the nominal sample size for inferences was 2000.

Phenotypic records were analyzed with univariate or bivariate animal (or sire+maternal grand-sire) models including generation effects as "fixed". The fixed effects are environmental, since genetic trends should be captured in the posterior distributions of the breeding values, at least when selection is ignorable. For this preliminary report, 160 and 32 models in small and large populations, respectively, were analyzed (Table 1).

Results and Discussion

Only main tendencies are reported here. Generally speaking, posterior distributions captured well the simulated values and their fluctuations, though with some bias for individual models / replicates. Figure 1 shows an example.

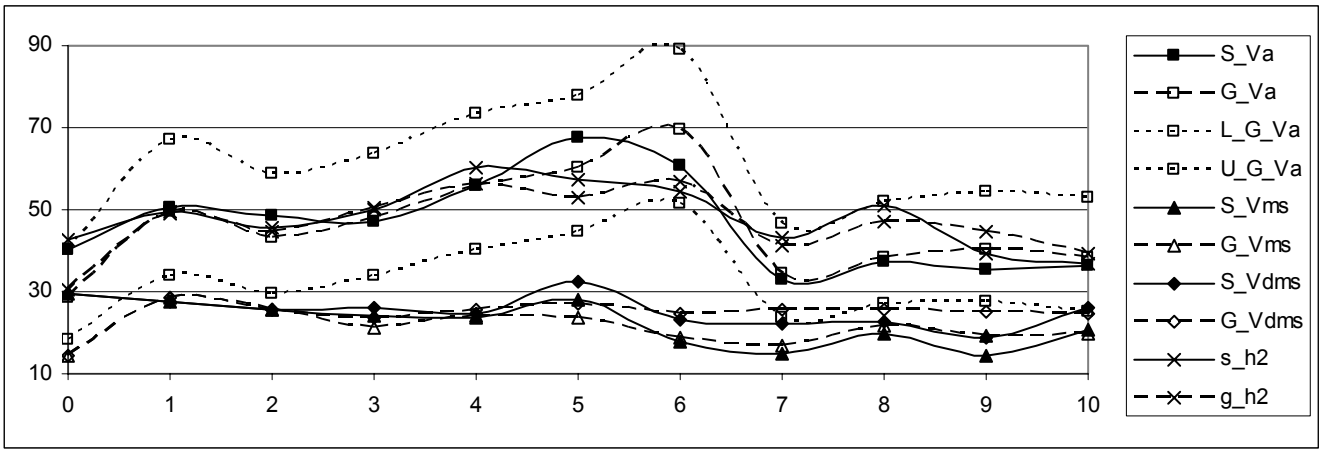

Figure 1. Simulated values (S) and posterior means (G) of genetic variance (V_a), Mendelian sampling variance (V_{ms}), Lower and Upper (L, U) 95% probability intervals for genetic variance, variance of discounted Mendelian sampling terms (V_{dms}), and heritability in a population of 40 animals/sex/generation. There were two traits (h^2=0.50 and r_g=0.8) expressed in both sexes, 25% of animals were selected at random (selection method I_5), and genetic evaluation was with a two-trait Animal Model.

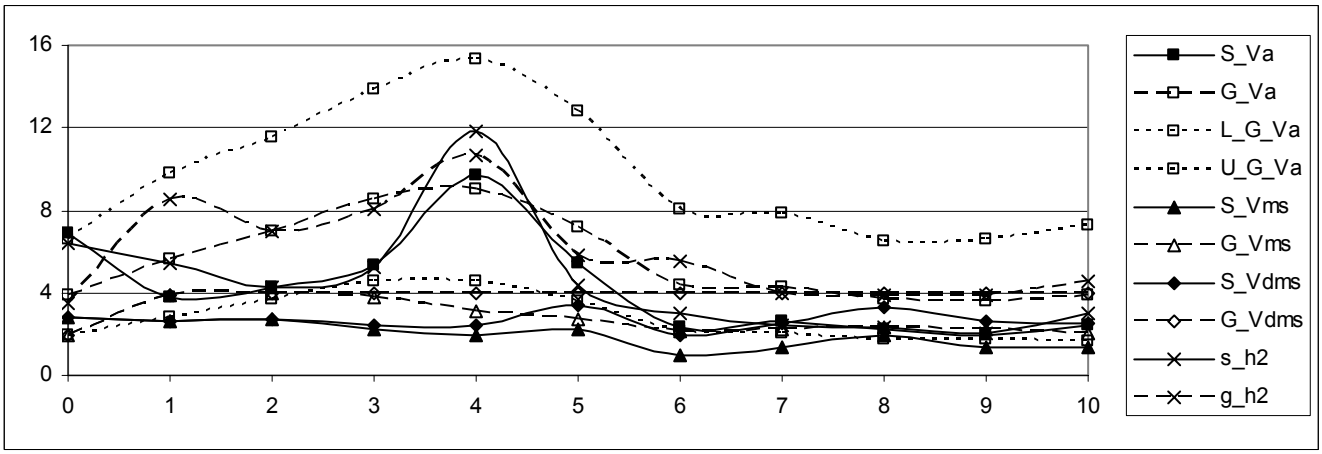

Figure 2a. Animal model analysis of a population of 40 animals / sex / generation. There were two traits (h^2=0.05 and r_g=0.2) expressed in both sexes, 25% of animals were selected for an index of the two traits (selection method I_3), and genetic evaluation included both traits. Descriptions of curves are as in Figure 1.

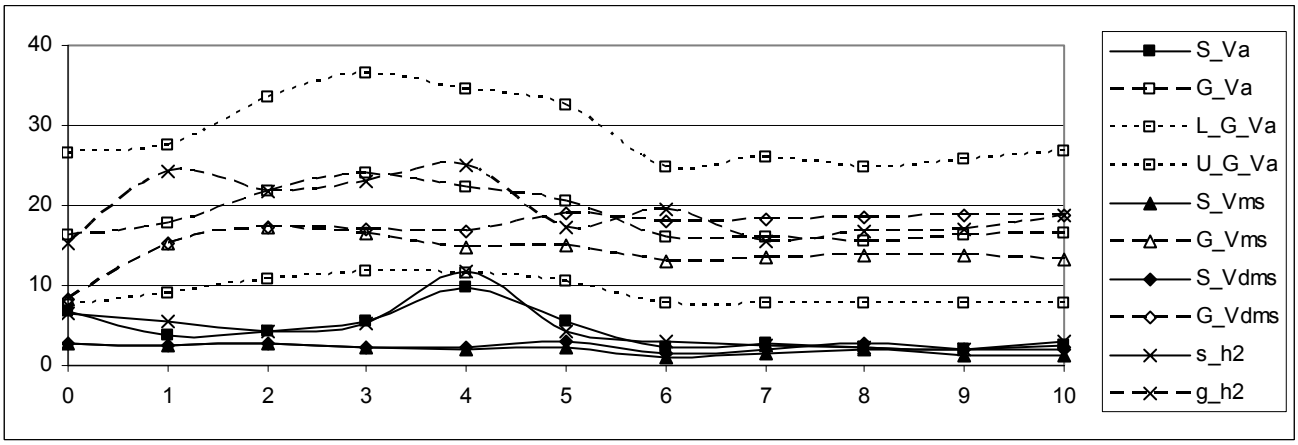

Figure 2b. Sire model evaluation of a population of 40 animals / sex / generation. There were two traits (h^2=0.05 and r_g=0.2) expressed in both sexes, 25% of animals were selected for an index of the two traits (selection method I_3), and genetic evaluation considered both traits. Descriptions of curves are as in Figure 1. Please notice differences of scale between Figures 2a and 2b.

Table 1. Percentage of models with a non-zero trend (over generations) in variance of rescaled Mendelian sampling terms ($\sigma^{2(t)}_{\gamma*}$) for small or large populations and several simulation settings.

	Small Population Size			Large Population size		
		% of models with non-zero trends			% of models with non-zero trends	
	Total # of models	Simulated values	Posterior values	Total # of models	Simulated values	Posterior values
h^2						
0.05	80	2.5	42.5			
0.25				32	12.5	15.6
0.50	80	15.0	18.8			
r_g						
0.2	80	7.5	33.8			
0.5				32	12.5	15.6
0.8	80	10.0	27.5			
Selection						
5	32	12.5	25.0	8	0.0	12.5
4	32	18.8	28.1	8	25.0	25.0
3	32	0.0	40.6			
2	32	6.3	34.4	8	25.0	25.0
1	32	6.3	25.0			
0				8	0.0	0.0
Trait						
1	80	7.5	30.0	16	12.5	0.0
2	80	10.0	31.3	16	12.5	31.3
Selection						
0.05						
0.25	160	8.8	30.6	32	12.5	15.6
0.50						
Evaluation						
1 Only	80		35.0	16		18.8
1 & 2	80		26.3	16		12.5
Evaluation						
AM	80		3.8	16		0.0
SM	80		57.5	16		31.3
Total	**160**	**8.8**	**30.6**	**32**	**12.5**	**15.6**

The example in Figure 1 represents a situation without "missing data" so selection is "ignorable". Posterior means of $\sigma^{2(t)}_{a}$ (G_Va) follow Monte Carlo fluctuations of the simulated values (S_Va), and there is little difference between the two trajectories. The same is true for posterior means of $\sigma^{2(t)}_{\gamma}$, $\sigma^{2(t)}_{\gamma*}$ and heritability. Posterior means of $\sigma^{2(t)}_{\gamma}$ (G_Vms) and variance of the simulated MS terms (S_Vms) have a decreasing trend, because of inbreeding. When MS terms are rescaled (see Eq. **[4]**), very little fluctuation is observed; posterior means of $\sigma^{2(t)}_{\gamma*}$ (G_Vdms) and the variance of simulated values (S_Vdms) show a flat trend, especially during later generations. Upper and lower bounds for the 95% posterior probability interval for additive genetic variance (U_G_Va, and L_G_Va, respectively) are also shown in the figure. The "true" trajectory is inside of the bounds, so the Bayesian analysis gives a correct inference in this case.

There are several factors in the simulation process (h^2, r_g, selection method, sex-limited expression of the trait and intensity of selection) and in the genetic evaluations

(number of traits and the evaluation model used, i.e., "sire" or "animal" model). These create a sequence of "missing data" situations and, thus, of degree of "ignorability". There is also the effect of population size: in small populations the volatility of the simulation and estimation processes is larger. As the amount of "missing data" increases, the bias in estimation of variances increases as well. Also, 95% probability intervals widen and, in some situations, the simulated values tend to fall outside of the interval. An example of this is shown in Figures 2a and 2b. In Figure 2a (animal model, 2-trait analysis) the "true" trajectory of the genetic variance is inside of the credibility interval. In Figure 2b (sire model, 2-trait analysis), the genetic variance and the variance of the MS effects are overstated throughout the selection process; this is a situation where all data are used in the analysis but the relationship information employed is incomplete.

In order to see if $\sigma^{2(t)}_{\gamma*}$ has a flat trend over generations, the ratio of the posterior mean of this variable to the base population total additive genetic variance was regressed on generation number. Table 1 shows percentages of models having regression coefficients significantly ($p < 0.05$) different from 0.

There were 1296 possible models from combination of all factors. However, only 160 and 32 models in the small and large population sizes, respectively, are included in this report (Table 1). On average, about 70% of the models in the small populations and 87% of the models in the large populations had a flat trend for the variance of rescaled MS. Note, however, that the ratio used as response variable is correlated over generations, so the null hypothesis "flat trend" is expected to be rejected too often by ordinary least-squares. Using a non-parametric test would have been more sensible here. Therefore, the following results and conclusions should be treated cautiously.

There was a clear association between evaluation model (animal model vs sire model) and non-zero trends in the variance of MS terms. For animal models, 3.8% and 0.0% of the models in the small and large populations, respectively, showed non-zero trends; for the sire models, the corresponding figures were 57.5% and 31.3%. Within models where there was a "significant" trend in the posterior mean of $\sigma^{2(t)}_{\gamma*}$ (G_Vdms), such trends were between 1% to 5% of the base population variance per generation. However, trends as large as 20% of the base population values were observed as well.

An association between heritability value and a trend of the variance of MS effects was found also. In small populations, the difference in percentage of models with non-zero trend between simulated and posterior values was only 3.8% (18.8-15.0) for h^2=0.50, while it amounted to 40% (42.5-2.5) for h^2=0.05. This may be related to the precision of the analysis, as one would expect genetic variances to be more difficult to infer when the signal/noise ratio is small, *i.e.*, low heritability.

On the basis of the results from sire models and low heritability values, it appears that one should be careful about assuming constancy of the variance of Mendelian sampling effects across all evaluation models and traits.

We found one counter-intuitive result. Differences between analyses with only one trait vs. two traits were unclear, even when selection was based entirely or partially on the second trait, with the records on the second trait ignored in the evaluation. This will be investigated in the future.

Irrespective of the amount of bias in any of the parameters monitored, the posterior means of $\sigma^{2(t)}_{\gamma*}$ showed much less fluctuation (especially in later generations), despite some non-zero trends. Perhaps monitoring the variance of MS effects is not a sensitive enough criterion for discriminating between "right" and "wrong" genetic evaluation models.

The simulation strategies and evaluation models adopted may be did not provide a stringent enough test of ignorability. For example, missing pedigree information may play a vital role, and this was not considered here. A lack of a test statistic is also a point of concern, because in the analysis of field data one is interested to know when a departure from expectation can be established. These and related questions will be addressed in future studies.

Acknowledgment

HJ acknowledges fruitful discussions and advice from Ignacy Misztal, Yu-Mei (Ruby) Chang and Freddy Fikse on modifications needed for adapting Misztal's Gibbsf90 program package to this study.

References

Boichard, D., Bonaiti, B., Barbat, A. & Mattalia, S. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. *J. Dairy Sci. 78,* 431-437.

Burnham, K.P. & Anderson, D.R. 2002. *Model selection and multimodel inference: A practical information-theoretic approach.* Springer-Verlag, New York.

Gianola, D. & Fernando, R.L. 1986. Bayesian methods in animal breeding theory. *J. Anim. Sci. 63,* 217-244.

Jorjani, H. 2003. An overview of validation issues in national genetic evaluation systems (N-GES). *Interbull Bulletin 30*, 49-55.

Klei, L., Mark, T., Fikse, F. & Lawlor, T. 2002. A method for verifying genetic evaluation results. *Interbull Bulletin 29,* 178-182.

Misztal, I. 2002. http://nce.ads.uga.edu/~ignacy/

Sorensen, D., Fernando, R. & Gianola, D. 2001. Inferring the trajectory of genetic variance in the course of artificial selection. *Gen. Res. Camb. 77,* 83-94.

Sorensen, D. & Gianola, D. 2002. *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics.* 740 pp. Springer, New York.

Thompson, R. 2001. Statistical validation of genetic models. *Livest. Prod. Sci. 72,* 129-134.