EFFECTS OF NON TRANSLATION INVARIANT SELECTION ON ESTIMATES OF VARIANCE COMPONENTS

F. S. Schenkel\textsuperscript{1,2} and L.R. Schaeffer\textsuperscript{1}

\textsuperscript{1}CGIL, University of Guelph, Canada, N1G2W 1
\textsuperscript{2}Federal University of Rio Grande Sul (UFRGS), Brazil. Sponsored by Brazilian Agency for Graduate Studies (CAPES)

SUMMARY
The effect of non translation invariant phenotypic selection on estimates of variance components using Bayesian and REML estimation was assessed by simulation. Combinations of two selection schemes, two models, and two levels of pedigree information were generated one hundred times. Populations consisted of 450 animals spread over 6 generations. The simulated trait had heritability of 0.50. The methods were empirically not biased by selection when all information was available. With 15% randomly missing \( \pi_i \), the methods were biased by selection when contemporary group (CG) effects were present. For both methods the mean square error of the estimates increased with selection when CG were present.

Keywords: Variance components, Selection, Bayes, REML, Gibbs sampling

INTRODUCTION
Genetic parameters for production traits are commonly estimated from data on selected animals of commercial livestock operations. Accurate estimates of variance components \((\sigma^2)\) are important because the prediction error variances for predicted genetic values increase as \(\sigma^2\) deviates from true values (Henderson 1975). Van Tassel et al. (1995) compared effects of selection on \(\sigma^2\) using REML and Bayesian estimates assuming a model without fixed effects except the overall mean. The selection criterion (phenotypic observations) was translation invariant to the overall mean. Theoretical arguments (Gianola and Fernando 1986; Gianola et al. 1989; Fernando and Gianola 1990) show that if all data that led to the current population are available, then inferences can be made using the likelihood as if selection has not occurred and without requiring translation invariance of the selection criterion. However, under selection, the sampling properties of likelihood methods such as REML are not known (Fernando and Gianola 1990). Im et al. (1989) showed that when data are missing at random and the marginal prior densities of the parameters being evaluated and of the parameters that describe the missing data process are independent, then inferences could be made using the likelihood as though selection has not occurred. The results of Im et al. (1989) hold for Bayesian inference provided the parameters are distinct, i.e., their prior distributions are independent.

The objective of this investigation was to assess the effects of selection on REML and Bayesian estimates of variance components when the selection criterion is not translation invariant. The impact of missing pedigree information was also of interest.

MATERIAL AND METHODS
Stochastic simulation. Data were generated using a stochastic procedure similar to van de Werf and de Boer (1990), Pieramati and van Vleck (1993) and van Tassell et al. (1995). A large number of
unlinked loci was assumed. The base population consisted of 10 males and 40 females sampled randomly from a conceptually infinite population of unrelated and unselected individuals. Base animals (generation 0) were mated at random (4 females per male) to produce 40 males and 40 females of generation 1. The 10 phenotypically best males were selected in each of the 4 subsequent generations. Each of the females was mated at random to one of the selected males to produce two offspring, one of each gender.

The model for simulating a record was

$$y_i = \beta_i + a_j + e_{ij},$$

where $y_i$ is the phenotypic observation of the animal $j$, $\beta_i$ is the fixed effect of the contemporary group (CG) $i$, $a_j$ is the additive genetic value of the animal $j$ and $e_{ij}$ is the random residual effect. Values for $e_{ij}$ were drawn from a $N(0, \sigma^2_e = 10.0)$. For base animals, $a_j$ were drawn from a $N(0, \sigma^2_a = 10.0)$. Genetic values for animals of later generations were simulated as $a_j = \frac{1}{\sqrt{2}}a_{s_j} + \frac{1}{\sqrt{2}}a_{d_j} + m_j$, where $a_{s_j}$ and $a_{d_j}$ are genetic values of sire and dam of the individual $j$. The coefficient of inbreeding of the parents was taken into account when the Mendelian sampling effect ($m_j$) was simulated.

Several different groups of data were generated. The random model (RM) did not include CG effects. In this case $\beta_i$ was equal to 0 for every $i$. The mixed model (MM) included CG effects that were simulated in the first replicate and kept constant over replicates. There were 4 CG of each gender per generation. Their effects ($\beta_i$) were drawn from a uniform distribution with $\sigma^2_\beta = 10.0$, and with parameters $\alpha = -\beta$ and $\beta = 5.5$. Animals were assigned randomly to CG within sex and generation in each replicate. Two selection methods were used. The selected population (SP) was created by selection of 25% of the best males across CG based on phenotypic values. The random population (RP) was simulated by choosing a random fraction of 25% of the males. In both cases the mating of sires to dams was random. For both SP and RP, either full pedigree information (FP) was used or random missing pedigree information (MP) were generated, erasing at random the sire and dam of 15% of the animals. The number of replicates was 100 per combination of models, selection methods, and pedigree information.

**Bayesian analyses.** The mixed linear model was $y = X\beta + Za + e$, where $y$ is the vector of observations, $\beta$ is the vector of fixed effects, treated in the Bayesian estimation as a vector of random effects with a prior distribution representing prior knowledge about their values, $a$ is the vector of random genetic additive effects, $e$ is the vector of random residuals, and $X$ and $Z$ are appropriate incidence matrices. For RM, $\beta$ was the overall mean. For MM, $\beta$ contained CG effects. The distributional assumptions were $a \sim N(0, A\sigma^2_a)$, and $e \sim N(0, I\sigma^2_e)$, where $A$ is the numerator of Wright’s relationship matrix. The vectors $a$ and $e$ were assumed independent. The inverse of $A$ was obtained taking into account inbreeding. Besides the distributional assumptions made previously, prior densities (PD) were assigned for all variance components and the location parameter $\beta$. Two different priors were assumed for $\beta$: a flat improper prior $p(\beta) = \text{constant}$, indicating no prior knowledge about its effect and a proper prior, $\beta \sim N(0, I\sigma^2_\beta)$. For variance components, independent scaled inverse chi-square distributions (ix) were assumed:

$$p(\sigma^2_i | \nu, \nu_0) = (\nu_0/\nu)^{\nu/2-1} \exp\left(-\nu \sigma^2_i / \nu_0 \sigma^2_i \right)$$

for $i = \beta, a, e$, where $\nu_i$ is a degree of belief parameter and $\nu_0^2$ can be thought as a prior value for the variance. Notice that when a flat improper prior was assumed for $\beta$, in [1] did not include $\beta$. A prior value for $\nu_i$ equal to 5 was used for all
variances in order to have the variance of the prior $\mathbf{x}$ as large as possible (200.0) but finite. Given the value for $\mathbf{v}$, the prior values for $s_i^2$ equal to 6.0 were specified such that expected values of the prior $\mathbf{x}$ were equal to the simulated values (10.0). The implementation of Gibbs sampling (GS) scheme followed van Tassel et al. (1995). The Gibbs sampling loop was repeated 10,000 times. A burn-in period of 1,000 rounds was used. All samples after the burn-in period were used to estimate the mean ($\text{BME}$) of the posterior marginal distribution of $\sigma_i^2$. For estimating the mode ($\text{BMD}$) samples from every 20th round after burn-in period were used.

**REML analyses.** Multiple trait derivative-free REML programs (Boldman et al. 1995) were used to estimate $\sigma_i^2$ using the same models as in Bayesian analyses. The starting values of variances were the true simulation values.

**RESULTS AND DISCUSSION**

The average value of $\sigma_i^2$ for the methods $\text{BME}$, $\text{BMD}$, and REML with their empirical standard error (se) and mean square error (mse) are presented in Table 1, as well the true simulated $\sigma_i^2$ in the base population and its standard deviation (sd). The true average $\sigma_i^2$ was calculated using quadratic form (quad). The MSE was defined as $\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\text{estimate}_i - \text{quad})^2$, where $\text{quad} = \frac{1}{r} A^{-1} A$, where $\mathbf{a}$ is the vector of simulated breeding values, $r$ is the number of animals, and $n$ is the number of replicates. As expected by theory (Gianola et al. 1989; Fernando and Gianola 1990) and by simulation results (Jansen and Mao 1991; Van Tassell et al. 1995) when all information was available both Bayesian and REML estimates were empirically not biased by selection (lines 1 vs 3 and 5 vs 7). However under selection with 15% randomly missing pedigree information (pi), 62, was severely biased downwards in the MM data sets and slightly biased upwards in the RM data sets (lines 6 vs 8 ($p<0.01$) and 2 vs 4 ($p<0.01$), respectively). Results are not in agreement with the theoretical arguments presented by Im et al. (1989). The downward bias in the MM was severe enough to produce all methods smaller than the true variance in the last generation (8.58). Part of $\sigma_i^2$ was retained by the fixed effects solutions, since in the RM data set the bias was smaller and upwards. In absence of selection, randomly missing pi did not cause bias (lines 1 vs 2 and 5 vs 6). The use of a proper informative prior for $\mathbf{b}$ or considering $\mathbf{b}$ as a random variable in REML reduced the bias due to selection with missing pi in the MM data set (line 9). Regarding the residual variances (results not shown), the bias due to selection when there was randomly missing pi was in the opposite direction (upwards) and occurred only in the MM data set.

Except when there was selection and missing pi in the MM data set, BME estimates were closer to REML estimates than BMD estimates. Van Tassell et al. (1995) found similar results. This happened because the parameters of the prior $\mathbf{i}X$ were defined based on expected value of the distributions to be equal to the simulated values. The higher empirical $\text{MSE}$ of $\sigma_i^2$ found in the RP than in SP for the MM data sets agreed with results of Van Tassell et al. (1995). However in the MM data set, the empirical $\text{MSE}$ of $\sigma_i^2$ was higher in SP than in RP. In general the presence of fixed effects increased the $\text{MSE}$. Randomly missing pi had a large impact on the $\text{MSE}$ of $\sigma_i^2$ for all methods, being the increase in $\text{MSE}$ larger in SP. The use of proper informative prior for $\mathbf{b}$ decreased greatly the $\text{MSE}$ for both BME and BMD. Considering $\mathbf{b}$ as random reduced the $\text{MSE}$ of REML estimates. In general $\text{MSE}$ of the BME was considerably smaller than those of BMD and REML estimates. Though not shown $\sigma_e^2$ followed a similar pattern.
Table 1. Estimated additive genetic variances ($\delta^2_a$), their empirical standard error (SE) and mean square error (MSE) given by Bayesian estimates via GS using mean (BME) or mode (BMD) and given by REML for combination of MO, SM, PI and PD

<table>
<thead>
<tr>
<th>Line</th>
<th>MO</th>
<th>SM</th>
<th>PI</th>
<th>PD</th>
<th>$\delta^2_a$</th>
<th>SE</th>
<th>MSE</th>
<th>$\delta^2_a$</th>
<th>SE</th>
<th>MSE</th>
<th>$\delta^2_a$</th>
<th>SE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9.98</td>
<td>.20</td>
<td>3.63</td>
<td>9.56#</td>
<td>.22</td>
<td>4.37</td>
<td>10.04</td>
<td>.22</td>
<td>4.35</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9.67</td>
<td>.21</td>
<td>4.18</td>
<td>9.32#</td>
<td>.24</td>
<td>5.83</td>
<td>9.72</td>
<td>.23</td>
<td>4.97</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10.25</td>
<td>.19</td>
<td>2.95</td>
<td>10.11</td>
<td>.20</td>
<td>3.35</td>
<td>10.24</td>
<td>.20</td>
<td>3.26</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1.100</td>
<td>.24</td>
<td>5.77</td>
<td>10.75#</td>
<td>.26</td>
<td>6.42</td>
<td>11.08</td>
<td>.25</td>
<td>6.47</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9.68</td>
<td>.22</td>
<td>4.37</td>
<td>9.33#</td>
<td>.24</td>
<td>5.94</td>
<td>10.08</td>
<td>.25</td>
<td>5.58</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>10.01</td>
<td>.23</td>
<td>4.53</td>
<td>9.28</td>
<td>.25</td>
<td>6.51</td>
<td>9.72</td>
<td>.25</td>
<td>5.97</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>10.00</td>
<td>.26</td>
<td>6.49</td>
<td>9.46#</td>
<td>.29</td>
<td>8.53</td>
<td>9.90</td>
<td>.30</td>
<td>8.68</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>7.49*</td>
<td>.21</td>
<td>11.12</td>
<td>6.98*</td>
<td>.21</td>
<td>14.00</td>
<td>7.03*</td>
<td>.26</td>
<td>16.06</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>8.46</td>
<td>.23</td>
<td>7.58</td>
<td>7.88*</td>
<td>.24</td>
<td>10.28</td>
<td>7.83#</td>
<td>.26</td>
<td>11.70</td>
</tr>
</tbody>
</table>

MO = model: 0 = RM, 1 = MM; SM = selection method: 0 = RP, 1 = SP (25% best males); PI = pedigree information: 0 = FP, 1 = MP (15% randomly); PD = prior density for $\beta$: 0 = flat improper prior, 1 = proper prior (normal density).

True average $\sigma^2_a$ in the base generation was $10.08 \pm 0.07$ for all data sets.

#Empirically significantly or highly significantly different from the base generation true value.

CONCLUSIONS

For the conditions of this simulation study, for all methods, non translation invariant phenotypic selection increased the MSE of $\delta^2_a$ and $\delta^2_e$, and biased downwards $\delta^2_a$ when there was 15% randomly missing pi. For $\delta^2_e$, the bias was smaller and upwards.

REFERENCES


