BAYESIAN INFERENCE ON GENETIC MERIT WITH UNCERTAIN PATERNITY

F.F. Cardoso¹, R. J. Tempelman¹ and E.J. Pollak²

¹Dept. of Animal Science, Michigan State University, East Lansing, MI, USA
²Dept. of Animal Science, Cornell University, Ithaca, NY, USA

INTRODUCTION

The use of multiple-sire mating is common in beef cattle, particularly in large populations raised on extensive pasture conditions. Hence, many pedigrees in these herds are uncertain. Best linear unbiased predictions (BLUP), based on the use of an average numerator relationship matrix (ANRM) incorporating Henderson's (1988) or related methods have been shown in recent simulation studies to be superior to genetic grouping methods for genetic evaluations on animals with uncertain paternity (Perez-Enciso and Fernando, 1992; Sullivan, 1995); however, these methods do not use phenotypic records for inferring sire assignments. The empirical Bayes procedure of Foulley et al. (1987) uses phenotypes to determine the posterior probabilities of sire assignments but its implementation is limited to sire models. The objectives of our study were: 1) to propose a Bayesian animal model for inference on genetic merit of animals with uncertain paternity and their posterior probabilities of sire assignments using Markov Chain Monte Carlo (MCMC) methods, and 2) to compare the proposed model with a model based on use of the ANRM.

METHODS AND MODELS

Construction of the Bayesian hierarchical model. In populations undergoing multiple-sire matings, some animals will have uncertain paternity with the sire of animal $i$ being one of a number ($n_i$) of candidate males (i.e. from the same mating pasture). To facilitate our MCMC implementation, we adopt the reduced animal model (RAM) of Quaas and Pollak (1980) as the first stage of our hierarchical model such that our procedure for inferring upon sire assignments for animals with uncertain paternity will depend on whether they are parent or non-parents (i.e. parents of other animals in the dataset or not). In a RAM, the conditional density of an observation ($y_i$) on animal $i$, given a certain sire assignment, can be written:

$$y_i | \beta, a_p, m_p, z'_u, \sigma^2_p, \sigma^2_e, \sigma^2_d ~ N \left( x'_i \beta + z'_i a_p + z'_i m_p, \sigma^2_p + d_i \sigma^2_e \right); \ i = 1, \ldots, n$$

Here, $\beta$, $a_p$, $m_p$ are, respectively, vectors of unknown “fixed” effects, additive and maternal genetic effects of $q_p$ parent animals, $\sigma^2_e$ is the residual variance, $x'_i$ and $z'_i$ are known row incidence vectors and the paternal element of $z'_u$ is either known or needs to be inferred for $q_u$ non-parent animals in the case of uncertain paternity. The term $d_i$ represents the fraction of the additive genetic variance ($\sigma^2_a$) due to Mendelian sampling for non-parent animal $i$ (being 0 otherwise) and is constant regardless of sire assignment in non-inbred populations. The prior probability for a candidate sire $j$ to be assigned to non-parent animal $i$ can be written as $\text{Prob}(z'_i = z'_{ij}) = \pi_{ij}$, where $z'_{ij}$ is the row incidence vector associating known RAM
maternal assignment but inferred paternal assignment to sire \( j, i = q_p + 1, \ldots, q_p q, j=1,2,\ldots, n \) with \( q = q_p + q_n \), typically greater than \( n \). Note that \( n_i = \pi_{ji} = 1 \) when paternity is certain for animal \( i \).

A subjective (typically flat) prior distribution may be assigned to \( \beta \) whereas a structural prior is jointly assigned to \( a_p \) and \( m_p \) with (co)variances determined by the Kronecker (\( \otimes \)) product of a genetic covariance matrix (\( G \)) with the \( q_p \times q_p \) numerator relationship matrix (\( A_{ppk} \)) among parent animals:

\[
P_p \left( a_p \begin{array} \mathbf{G} \mathbf{s} \end{array} \mathbf{m}_p \right) = \left( 2\pi \right)^{-q_p} \left| G \otimes A_{ppk} \right|^{-1/2} \exp \left\{ -\frac{1}{2} \left( a_p \mathbf{m}_p \right) \left( \mathbf{G}^{-1} \otimes A_{ppk}^{-1} \right) \left[ a_p \mathbf{m}_p \right] \right\} . \tag{2} \]

Here \( s = \{s_{ij}\} \) denotes the list of candidate sires for each parent animal \( i, i = 1,2,\ldots, q_p \) for \( j = 1,2,\ldots, n_i \) candidate sires; i.e. some elements of \( A_{ppk} \) are uncertain (as implied by the conditioning on \( s \)) and need to be inferred. Let \( \gamma_i = a_i - 0.5 \left( a_{ij} + a_{di} \right) \) and \( \delta_i = m_i - 0.5 \left( m_{ij} + m_{di} \right) \) represent the additive and maternal genetic Mendelian sampling effects for animal \( i \) where, for example, \( a_{ij} \) and \( m_{ij} \) are the additive and maternal genetic effects pertaining to a candidate sire \( s_i \). Using Henderson's (1976) decomposition of \( A_{ppk}^{-1} \), the kernel portion of \( [2] \) can be rewritten:

\[
\exp \left\{ -\frac{1}{2} \left( a_p \mathbf{m}_p \right) \left( \mathbf{G}^{-1} \otimes A_{ppk}^{-1} \right) \left[ a_p \mathbf{m}_p \right] \right\} = \exp \left\{ \frac{1}{2} \left( \sum_{i=1}^{b} a_i^2 + \sum_{i=b+1}^{q_p} d_i^{-1} \gamma_i^2 \right) g_i^{11} \right\} \\
+ \frac{1}{2} \left( \sum_{i=1}^{b} m_i^2 + \sum_{i=b+1}^{q_p} d_i^{-1} \delta_i^2 \right) g_i^{22} \\
+ 2 \left( \sum_{i=1}^{b} a_i m_i + \sum_{i=b+1}^{q_p} d_i^{-1} \gamma_i \delta_i \right) g_i^{12} \tag{3} \]

where \( g_i^{ij} \) is the \( i,j \)th element of \( G^{-1} \) and \( b \leq q_p \) is the number of base (i.e. unrelated) population animals with no identifiable pedigree. Similarly, the determinant \( \left| G \otimes A_{ppk} \right| \) in \( [2] \) can be readily shown to be proportional to \( |G| \) and the product of Mendelian sampling fractions \( \prod_{i=1}^{q_p} d_i \) for the \( q_p \) parent animals. This alternative expression of \( [2] \) facilitates an explicit specification for modeling uncertainty on \( s \).

The third stage of the model specifies the prior probability of each of \( n_i \) males being the correct sire of animal \( i \), as given previously for non-parents by \( \text{Prob}(z_{ij} = z_{ij}') = \pi_{ji}, j = 1,2,\ldots, n_i \) and now \( \text{Prob}(s_j = s_{ij}) = \pi_{ji}, j = 1,2,\ldots, n_i \) for parents. In turn, each \( \pi_j = \{\pi_{ji}\}, i = b+1,\ldots,q \) are assigned independent Dirichlet prior distributions with hyperparameters that may be based on subjective (e.g. physical appearance) or molecular information, if available. For convenience, an inverted gamma density is typically specified on \( \sigma'e \) and an inverted Wishart prior is specified on \( G \).

The joint posterior density of the location parameters \( \beta, a_p, \) and \( m_p \), dispersion parameters, \( \sigma'e \) and \( G \), sire assignments through \( s \) (for parents) and \( \{z_{ij}'\} \) (for nonparents), and their
assignment probabilities \( \{ \pi_{ij} \}_{i=1}^{n_i} \) is simply specified as the product of the various aforementioned stages of the hierarchical model. A MCMC strategy involves sampling from full conditional distributions (FCD) of these parameters or groupings thereof. It can be readily shown that the FCD of \( \beta, \alpha_p \) and \( m_p \) is multivariate normal. Furthermore, the FCD of \( s \) or \( \{ z_{ij}' \} \) pertaining to animal \( i \) is a generalized Bernoulli distribution while the FCD for \( \pi \) can be shown to be a Dirichlet distribution. In a RAM, the FCD for \( \sigma^2_e \) and \( G \) are not readily recognizable such that a Metropolis Hastings update may appear necessary (Bink et al., 1998; van Kaam et al., 2002). Alternatively, we adopt the method of composition approach of Chib and Carlin (1999) to RAM by augmenting the joint posterior density with Mendelian sampling terms \( \delta_n = \{ \delta_{ij} \}_{1}^{q} \) and \( \gamma_n = \{ \gamma_{ij} \}_{1}^{q} \) for additive and maternal effects, respectively.

Since \( \delta_n \) and \( \gamma_n \) are completely aliased with the respective genetic effects \( \alpha_a \) and \( m_m \), given \( \alpha_p \) and \( m_p \), the FCD of \( \sigma^2_e \) and \( G \) based on this augmented joint posterior density can be readily shown to be standard inverted-gamma and inverted Wishart, respectively.

**Simulation study.** Ten populations, each consisting of five overlapping generations of 80 sires, 400 dams (480 parents) and 2000 non-parent animals with 30% of animals having uncertain paternity, were simulated to compare our model (FULL), with a ANRM model as described by Henderson (1988) but using MCMC. For all populations, \( \sigma^2_e = 60 \) and \( G = \begin{bmatrix} 30 & -5 \\ -5 & 20 \end{bmatrix} \).

Multiple sire group sizes ranged from \( n_i = 2 \) to 10. Model comparison was based on mean squared error of prediction (MSEP) and rank correlations between posterior mean and true genetic values and by the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2001).

**RESULTS AND DISCUSSION**

Since it was unclear to us whether \( s_{ij} \) (parents)/ \( z_{ij}' \) (non-parents) or \( \pi_{ij} \) should be used for inferring uncertainty with respect to assignment of sire \( j \) to animal \( i \), posterior means of both using FULL, averaged across all animals with uncertain paternity and categorized as parent or non-parent, are presented in Table 1. The average posterior probabilities attributed to the true sire were between 1 and 10% larger than the respective priors (\( 1/n_i \) for respective mating group) with posterior means based on \( s_{ij} \) and \( z_{ij}' \) having a slightly better general performance than those based on \( \pi_{ij} \). Posterior means of additive and maternal genetic effects were very similar, and no significant difference in MSEP and rank correlations involving these posterior means were detected between models. There was, however, a tendency of having smaller MSEP and higher rank correlation under FULL, especially for animals with uncertain paternity. The calculated DIC’s were also always in favor of the FULL model compared to the ANRM model, thereby indicating a better fit of FULL to the data.
Table 1. Average posterior means of probabilities ($\pi_{ij}$) and assignments ($s_{ij}/z_{1ij}$) of true sires by multiple-sire group size and parents versus non-parent animals using full model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Category</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_{ij}$</td>
<td>Parents</td>
<td>0.513</td>
<td>0.341</td>
<td>0.259</td>
<td>0.175</td>
<td>0.126</td>
<td>0.105</td>
</tr>
<tr>
<td>$\pi_{ij}$</td>
<td>Non-parents</td>
<td>0.509</td>
<td>0.339</td>
<td>0.259</td>
<td>0.172</td>
<td>0.130</td>
<td>0.103</td>
</tr>
<tr>
<td>$s_{ij}$</td>
<td>Parents</td>
<td>0.525</td>
<td>0.349</td>
<td>0.269</td>
<td>0.183</td>
<td>0.127</td>
<td>0.110</td>
</tr>
<tr>
<td>$z_{1ij}$</td>
<td>Non-parents</td>
<td>0.517</td>
<td>0.345</td>
<td>0.268</td>
<td>0.178</td>
<td>0.134</td>
<td>0.105</td>
</tr>
</tbody>
</table>

We have also applied these methods to the analysis of weaning weight records of 5,973 Hereford calves, raised in Southern Brazil, obtaining favorable results in comparison with the ANRM approach based on DIC, despite unusually low posterior mean heritability for weaning weight in the sample data. We also noted posterior standard deviations of genetic merits were noticeably smaller with FULL compared to ANRM.

CONCLUSION

Our FULL model represents an important alternative for genetic prediction on populations having multiple-sire mating. Genetic markers potentially provide a source of prior information for sire assignments that could be incorporated in our model, particularly when the markers are not distinctive enough to make paternity assignments to one of several related candidate sires but are distinctive enough to exclude others. The impact of the FULL model is expected to be particularly important for large herds. These herds provide sizable gene pools for selection and great potential for genetic progress, but very often the exclusive use of single mating is very costly, maybe impractical, due to size of the operation and labor required.

ACKNOWLEDGEMENT

F. F. Cardoso acknowledges financial support from CAPES – Brasilia/Brazil.

REFERENCES


