The EM-REML Iteration Equations for Estimation of Variance Components in a Sire and Dam Model

Seiko Saito\textsuperscript{1} and Hiroaki Iwaisaki\textsuperscript{2}
\textsuperscript{1}Graduate School of Science and Technology, 
\textsuperscript{2}Department of Animal Science, Faculty of Agriculture, 
Niigata University, Niigata 950-21, Japan

SUMMARY

Assuming a specific type of data in the field of animal breeding, the iteration equations based on the expectation-maximization algorithm are derived for the restricted maximum likelihood estimation of variance components in a Sire and Dam Model. The application of the iteration equations to the data leads to the same estimates of additive genetic and environmental variances as those under the Individual Animal Model. With the procedure using the iteration equations, compared to the case for the Individual Animal Model, the size of the coefficient matrix to be inverted of the mixed model equations relatively becomes small, and the speed of convergence of the estimates becomes rather fast. Consequently, the total computational burden to obtain the proper estimates of additive genetic and environmental variances is expected to be considerably reduced in the proposed procedure. A numerical illustration, comparing the proposed procedure with the Individual Animal Model procedure, is given using simulated carcass data on beef cattle.

1. Introduction

In the field of animal breeding, the estimates of variances of the random effects such as breeding values and temporary environmental effects, i.e. the estimated components of variances like additive genetic and environmental ones, are often utilized: 1) to estimate genetic parameters such as heritabilities of, and genetic correlations among, economically important quantitative traits, 2) to conduct the (empirical) best linear unbiased prediction (BLUP; Henderson, 1975) of breeding values by the mixed model methodology, 3) to construct appropriate breeding plans, and so on.

For the method of variance component estimation, the restricted maximum likelihood (REML) procedure (Patterson and Thompson, 1971) is recently gaining favor among the animal breeders. In particular, REML with the Animal Model (AM) is becoming the method of choice. The REML approach under mixed linear models such as AM takes account of the loss in degrees of freedom that results from estimating fixed effects, and the REML estimates are relatively less biased by selection or assortative mating in the data (e.g., Gianola and Fernando, 1986).

Especially with AM, however, REML estimation using the expectation-maximization algorithm (Dempster et al., 1977), the so-called EM-REML, is frequently computationally very demanding for large datasets since it requires the calculation of a generalized inverse of the large coefficient matrix of mixed model equations (MME). Also, REML with the derivative-free algorithm (DF-REML), considered first by Graser et al. (1987), would not be applicable to very large datasets in this stage, although it involves Gaussian elimination rather than direct inversion of the coefficient matrix. Furthermore, Misztal and Perez-Enciso (1993) have reported for multiple-trait models that in certain cases the converged values in the DF-REML estimation were affected by the initial values. Thus, various prolific researches are required to overcome these difficulties in the REML estimation.

Key words: Carcass trait; EM algorithm; Iteration equation; REML; Sire-Dam Model; Variance component.
with AMs applied to large datasets.

In this paper, we propose a Sire and Dam Model in which the Mendelian sampling effect is appropriately accounted for and derive the EM-REML iteration equations for estimation of variance components in the proposed model. The model proposed is equivalent to the Individual Animal Model (IAM) for a specific type of data such as those of carcass traits in beef cattle, and the application of the equations to the data, with less computational burden, leads to the estimates of additive genetic and environmental variances equal to those under IAM.

2. EM-REML Estimation with IAM

First, we briefly review the EM-REML procedure under a sort of IAM. The IAM assumed here, in matrix notation, is

\[ y = X\beta + Z_a a + e, \]  

(1)

where \( y \) is the \( N \times 1 \) vector of observations, \( \beta \) is the \( r \times 1 \) vector of fixed effects, \( a \) is the \( p \times 1 \) vector of animal effects, that is, breeding values (random), \( e \) is the \( N \times 1 \) vector of residual effects (random), and \( X \) and \( Z_a \) are the \( N \times r \) and \( N \times p \) known incidence matrices, respectively. We assume the two conditions — 1) no maternal effects on \( y \) and 2) one observation per animal. Note that with observations maternally influenced or with more than two observations per animal we need to consider additional effects associated with maternal or permanent environmental ones in the model (1).

Usual assumptions on expectation and dispersion parameters are

\[ E \begin{pmatrix} y \\ a \\ e \end{pmatrix} = \begin{pmatrix} X\beta \\ 0 \\ 0 \end{pmatrix} \]  

(2)

and

\[ \text{Var} \begin{pmatrix} a \\ e \end{pmatrix} = \begin{pmatrix} A_a\sigma_a^2 & 0 \\ 0 & R_a \end{pmatrix}; R_a = I\sigma_e^2, \]  

(3)

respectively, where \( \sigma_a^2 \) and \( \sigma_e^2 \) are variance components for animal and residual effects, respectively, and \( A_a \) is the additive relationship matrix for animals.

With these assumptions, MME to be solved in each EM-REML iteration are expressed as

\[ \begin{pmatrix} X'\widetilde{R}_a^{-1}X \\ Z_a'\widetilde{R}_a^{-1}X \\ Z_a'\widetilde{R}_a^{-1}Z_a + A_a^{-1}\sigma_a^{-2} \end{pmatrix} \begin{pmatrix} \beta \\ \tilde{a} \\ \tilde{e} \end{pmatrix} = \begin{pmatrix} X'\widetilde{R}_a^{-1}y \\ Z_a'\widetilde{R}_a^{-1}y \end{pmatrix}, \]  

(4)

where \( \tilde{\sigma}_a^{-2} \) and \( \tilde{R}_a^{-1} \) are the prior information for \( \sigma_a^{-2} \) and \( R_a^{-1} \), respectively. Then the EM-REML iteration equations are given by

\[ \begin{pmatrix} \tilde{\sigma}_a^2 \\ \tilde{\sigma}_e^2 \end{pmatrix} = \begin{pmatrix} p & 0 \\ 0 & N \end{pmatrix}^{-1} \begin{pmatrix} \tilde{a}'A_a^{-1}\tilde{a} + \text{tr}(A_a^{-1}C_{22}) \\ \tilde{e}'\tilde{e} + \text{tr}(WCW') \end{pmatrix}, \]  

(5)

where \( C \) stands for a generalized inverse of the coefficient matrix of eq.(4), \( C_{22} \) is the sub-matrix corresponding to \( a \) of the partitioned \( C \) matrix, and \( W = (X \ Z_a) \).

The IAM usually includes random effects of animals with records and animals which are their parents and ancestors, incorporating all known relationship information in the analysis, and EM-REML estimation with IAM requires the direct inversion of the coefficient matrix of the MME of size equal to the sum of the number of fixed effects \( r \) and that of animal effects \( p \). With the model (1), we have \( \sigma_A^2 = \sigma_a^2 \) and \( \sigma_E^2 = \sigma_e^2 \), where \( \sigma_A^2 \) and \( \sigma_E^2 \) are the additive genetic and environmental variances, respectively. The additive genetic variance is the variance of the breeding values of animals.
3. EM-REML Iteration Equations in a Sire and Dam Model

In this section, we propose a Sire and Dam Model in which the Mendelian sampling effect is appropriately accounted for and derive a EM-REML estimation formula for the model whose application yields the same estimates of additive genetic and environmental variances as those under IAM.

In order to estimate $\sigma^2_G$ and $\sigma^2_E$ properly, we first need to assume the following conditions on the trait analyzed and the type of data actually collected: 1) no maternal effects affect the trait; 2) the number of observations per animal is only one; 3) none of the animals with an observation ever becomes a sire or a dam; 4) animals producing progeny have no own observations. Furthermore, we assume 5) both sires and dams of the animals with an observation are completely identified. The first two conditions are the same as those assumed with the IAM model (1), and the condition 5) is fundamental to include sire and dam effects in a model. The conditions, 3) and 4), are essential to obtaining the same estimates of additive genetic and environmental variances as those under IAM.

When the above conditions are fulfilled, the effect (the breeding value) of an individual with one record defined in the IAM can be replaced by the sum of the average of the breeding values of the parents (the sire and the dam) without records and a Mendelian sampling effect. In a mathematical expression, we have

$$ a_{(i)} = \frac{1}{2} (a_{s(i)} + a_{d(i)}) + m_{(i)} $$

$$ = u_{s(i)} + u_{d(i)} + m_{(i)}, \quad (6) $$

where $a_{(i)}$ is the breeding value of an animal, $i$, with one record, $a_{s(i)}$ and $a_{d(i)}$ are the breeding values of the sire and the dam, respectively, $u_{s(i)}$ and $u_{d(i)}$ are the effects associated with the sire and the dam, respectively, and $m_{(i)}$ is the Mendelian sampling effect. Therefore, the vector of observations $y$ in the model (1) can be rewritten by the so-called Sire and Dam Model, as follows

$$ y = X\beta + Z_s u_s + Z_d u_d + Z_0 m + e, \quad (7) $$

where $y$ is the $N \times 1$ vector of observations, $\beta$ is the $r \times 1$ vector of fixed effects, $u_s$ is the $q_s \times 1$ vector of sire effects (random), $u_d$ is the $q_d \times 1$ vector of dam effects (random), $m$ is the $N \times 1$ vector of individual Mendelian sampling effects (random), $e$ is the $N \times 1$ vector of residual effects (random) with $\text{Var}(e) = I \sigma^2_e$, and $X$, $Z_s$, $Z_d$ and $Z_0$ are the corresponding incidence matrices, and then $Z_0 = I_q$, where $I_q$ is the $N \times N$ identity matrix, since each animal is assumed to have only one observation.

Then, eq.(7) can be rewritten as

$$ y = X\beta + (Z_s Z_d) \begin{pmatrix} u_s \\ u_d \end{pmatrix} + e $$

$$ = X\beta + Zu + e, \quad (8) $$

where $u$ is the $q \times 1$ vector of effects of sires and dams (random) with $q = q_s + q_d$, $e$ is the $N \times 1$ vector of residuals (random), and $Z$ is the corresponding incidence matrix. With this model, because each effect of sires and dams represents the transmitting ability, that is, one half of the breeding value, assumptions for expectation and dispersion parameters of $u$ and $e$ are generally written as

$$ \mathbb{E} \begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (9) $$

and

$$ \text{Var} \begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} A \sigma^2 \ \\
0 \ \\
R \end{pmatrix}, \quad (10) $$

- 3 -
respectively, where \( \sigma_t^2 \) stands for the variance of transmitting abilities with \( \sigma_t^2 = (1/4)\sigma_A^2 \), and \( \mathbf{A} \) is the additive relationship matrix for sires and dams. For the matrix \( \mathbf{R} \), the usual assumption is

\[
\mathbf{R} = (2\sigma_t^2 + \sigma_e^2)\mathbf{I}
\]

with \( \sigma_e^2 = \sigma_R^2 \), which means that parent animals are assumed to be non-inbred.

Hence, for the case of parent animals inbred, we now need to consider the general form of the variance of Mendelian sampling effects and to replace eq.(11) by the appropriate \( \mathbf{R} \) based on the general form. The variance of \( \mathbf{m} \) is expressed by

\[
\text{Var}(\mathbf{m}) = 4\mathbf{D}\sigma_t^2,
\]

where \( \mathbf{D} \) is a diagonal matrix since Mendelian sampling effects occur independently each other, and the diagonal elements \( d_i, i = 1, 2, ..., N \), are given by

\[
d_i = 1 - .25(a_{ss} + a_{dd}) = .5 - .25(f_s + f_d),
\]

where \( f_s \) and \( f_d \) are inbreeding coefficients for the sire and the dam, respectively, and \( a_{ss} = 1 + f_s \) and \( a_{dd} = 1 + f_d \). Note that if all parents are non-inbred animals, then the diagonal elements of the \( \mathbf{D} \) matrix are all .5, and that \( \text{Var}(\mathbf{m}) \) is simply expressed as \( 2\mathbf{I}\sigma_t^2 \).

Thus, instead of eq.(11), we assume

\[
\mathbf{R} = 4\mathbf{D}\sigma_t^2 + \mathbf{I}\sigma_e^2
\]

in eq.(10). The corresponding MME under the current Sire and Dam Model with eq.(14) are given by

\[
\begin{pmatrix}
X\widetilde{\mathbf{R}}^{-1}X \\
Z\widetilde{\mathbf{R}}^{-1}X
\end{pmatrix}
\begin{pmatrix}
X\widetilde{\mathbf{R}}^{-1}Z \\
Z\widetilde{\mathbf{R}}^{-1}Z + \mathbf{A}^{-1}\sigma_t^2
\end{pmatrix}
\begin{pmatrix}
\hat{\beta} \\
\hat{\upsilon}
\end{pmatrix}
= 
\begin{pmatrix}
X\widetilde{\mathbf{R}}^{-1}\mathbf{y} \\
Z\widetilde{\mathbf{R}}^{-1}\mathbf{y}
\end{pmatrix},
\]

where \( \sigma_t^{-2} \) and \( \widetilde{\mathbf{R}}^{-1} \) are the prior information for \( \sigma_t^{-2} \) and \( \mathbf{R}^{-1} \), respectively.

With eq.(15), by taking the quadratic forms for the REML estimation and then by deriving their pseudoexpectations, assuming that the prior values for the dispersion parameters are equal to the true values, we have a form of the EM-REML iteration equations in the current Sire and Dam Model, as follows

\[
\begin{pmatrix}
\vec{\sigma}_t^2 \\
\vec{\sigma}_e^2
\end{pmatrix}
= 
\begin{pmatrix}
q\vec{\sigma}_t^{-4} + 4\text{tr}(\mathbf{Q}_d\mathbf{D}) \text{tr}(\mathbf{Q}_a) \\
4\text{tr}(\mathbf{Q}_e\mathbf{D}) \text{tr}(\mathbf{Q}_e)
\end{pmatrix}^{-1}
\begin{pmatrix}
\hat{\upsilon}'\mathbf{Q}_d\hat{\upsilon} + \text{tr}(\mathbf{Q}_d\mathbf{C}_{22}) + \vec{\varepsilon}'\mathbf{Q}_e\vec{\varepsilon} + \text{tr}(\mathbf{Q}_e\mathbf{WCW}^') \\
\vec{\varepsilon}'\mathbf{Q}_e\vec{\varepsilon} + \text{tr}(\mathbf{Q}_e\mathbf{WCW}^')
\end{pmatrix},
\]

with

\[
\mathbf{Q}_t = \mathbf{A}^{-1}\vec{\sigma}_t^{-4}, \mathbf{Q}_a = 4\widetilde{\mathbf{R}}^{-1}\mathbf{D}\widetilde{\mathbf{R}}^{-1}, \text{and} \mathbf{Q}_e = \widetilde{\mathbf{R}}^{-1}\widetilde{\mathbf{R}}^{-1},
\]

where \( \vec{\sigma}_t^{-4} \) and \( \widetilde{\mathbf{R}}^{-1} \) are the prior information for \( \sigma_t^{-4} \) and \( \mathbf{R}^{-1} \), respectively, \( \mathbf{C} \) stands for a generalized inverse of the coefficient matrix of eq.(15), \( \mathbf{C}_{22} \) is the sub-matrix corresponding to \( \upsilon \) of the partitioned \( \mathbf{C} \) matrix, and \( \mathbf{W} = (\mathbf{X} \quad \mathbf{Z}) \). The iterative use of eq.(16) with eq.(15) until the final convergence of the estimates leads to the REML estimates of \( \sigma_t^2 \) and \( \sigma_e^2 \), and we obtain the REML estimates of \( \sigma_A^2 \) and \( \sigma_D^2 \) as \( 4\vec{\sigma}_t^2 \) and \( \vec{\sigma}_e^2 \), respectively, which are equal to the final solutions for \( \vec{\sigma}_d^2 (= \vec{\sigma}_A^2) \) and \( \vec{\sigma}_b^2 (= \vec{\sigma}_D^2) \) by eq.(5).
Figure 1. Mating structure for animals. The letters S and D represent a sire and a dam without observations, respectively. The letter P stands for a progeny with one observation. The two groups of animals, P_{17} to P_{23} and P_{24} to P_{30}, were assumed to be slaughtered in Years I and II, respectively.

4. A Numerical Illustration

The conditions 1) to 5) assumed in Section 3 on the trait analyzed and the data collected would generally be fulfilled, for instance, in the case of carcass traits in beef cattle, since generally animals slaughtered leave carcass records and they never become parents. Accordingly, a data-sample on the degree of marbling (a typical example of carcass traits) of 30 beef cattle was simulated. The fourteen animals with only one observation each were assumed to be slaughtered in two years (the two levels of the fixed factor), and the remaining sixteen animals were assumed to have no observations. The mating structure for the 30 animals is shown in Figure 1. Normality was assumed on the distribution of the trait. In the generation of data, the true values assumed were 2.5, .4, .9333 and .3 for population mean, additive genetic variance, environmental variance and heritability (\( \sigma_A^2/(\sigma_A^2 + \sigma_E^2) \)), respectively. The proportion of the variation due to the fixed effects (\( \kappa_f^2 \)) to the total variation, \( \kappa_f^2/(\kappa_f^2 + \sigma_A^2 + \sigma_E^2) \), was assumed to be 30\%, and the two actual values of the fixed effects were set to be \( \pm \kappa_f \). The actual data generated are listed in Table 1.

With the IAM, that is, eq.(1), the incidence matrices \( X \) and \( Z_a \) are given as

\[
X' = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\]

and

\[
Z_a = \begin{pmatrix} 0 \end{pmatrix} \begin{pmatrix} 1 \end{pmatrix},
\]

respectively, where \( 0 \) is the 14 \times 16 matrix which has 0 value in all elements, and \( I \) is the 14 \times 14 identity matrix. The additive relationship matrix \( A_a \) in eq.(3) is for all animals shown in Figure 1, and is obtained according to Henderson (1976) and Quaas (1976). On the other hand, with the Sire and Dam Model given as eq.(7), the incidence
Table 1
The data generated (y)

<table>
<thead>
<tr>
<th>Animal No.</th>
<th>y(Fattening year I)</th>
<th>Animal No.</th>
<th>y(Fattening year II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>4.26</td>
<td>24</td>
<td>2.02</td>
</tr>
<tr>
<td>18</td>
<td>5.70</td>
<td>25</td>
<td>1.80</td>
</tr>
<tr>
<td>19</td>
<td>3.22</td>
<td>26</td>
<td>2.57</td>
</tr>
<tr>
<td>20</td>
<td>2.91</td>
<td>27</td>
<td>3.99</td>
</tr>
<tr>
<td>21</td>
<td>2.55</td>
<td>28</td>
<td>2.22</td>
</tr>
<tr>
<td>22</td>
<td>2.45</td>
<td>29</td>
<td>1.02</td>
</tr>
<tr>
<td>23</td>
<td>3.05</td>
<td>30</td>
<td>.32</td>
</tr>
</tbody>
</table>

Table 2
EM-REML estimates by the three models

<table>
<thead>
<tr>
<th>Model</th>
<th>IAM</th>
<th>Sire-Dam</th>
<th>Sire-Dam</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_a ) or ( R )</td>
<td>( \sigma_e^2 )</td>
<td>( 4D \sigma_t^2 + \sigma_e^2 )</td>
<td>( (2\sigma_t^2 + \sigma_e^2)I )</td>
</tr>
<tr>
<td>Size of MME *</td>
<td>32</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>No. of iterations</td>
<td>645</td>
<td>305</td>
<td>314</td>
</tr>
<tr>
<td>( \sigma_a^2 ) or ( \sigma_t^2 )</td>
<td>.64171</td>
<td>.16043</td>
<td>.15562</td>
</tr>
<tr>
<td>( \sigma_e^2 )</td>
<td>.73219</td>
<td>.73219</td>
<td>.74543</td>
</tr>
<tr>
<td>( \sigma_A^2 )</td>
<td>.64171</td>
<td>.64171</td>
<td>.62248</td>
</tr>
<tr>
<td>( \sigma_E^2 )</td>
<td>.73219</td>
<td>.73219</td>
<td>.74543</td>
</tr>
</tbody>
</table>

* All animals with no observation were included in the MME.

Matrix \( X \) is the same with that in the IAM. The incidence matrix \( Z_a \) (\( Z_d \)) contains one 1 value in the column pertaining to the effect of its sire (its dam) and zero values in the other columns. Also, \( Z_a \) is the 14 × 14 identity matrix. The matrix \( A \) in eq.(10) is an appropriate sub-matrix of \( A_a \), and the diagonal matrix \( D \) in eq.(14) is given as

\[
D = \text{diag} (.5 .5 .5 .5 .4375 .5 .5 .5 .5 .5 .5 .4375 .5 .5 )
\]

For the data, EM-REML estimation using IAM and the current Sire and Dam Model with eq.(14) was conducted. The analysis by the usual Sire and Dam Model with eq.(11) was also carried out. The initial values for the variance components were set to true values. Convergence was assumed when change of the variance ratio between two successive iterations was less than 10^{-6}.

The EM-REML estimates of variance components in the IAM and the two Sire and Dam Models are given in Table 2. The sizes of MME were 32 and 18 for IAM and the Sire-Dam Models, respectively, when all animals with no observation were included in the equations. While the number of iterations required was 645 for REML estimation with IAM, that for the current Sire and Dam Model with eq.(14) was 305. The estimates of animal and residual components, \( \sigma_a^2 (= \sigma_A^2) \) and \( \sigma_e^2 (= \sigma_E^2) \), were .64171 and .73219,
respectively. The EM-REML equations in the current model with eq.(14) gave .16043 and .73219 for $\tilde{\sigma}_e^2$ and $\tilde{\sigma}_{e_i}^2$, respectively, and led to the same estimates of $\sigma_A^2$ and $\sigma_E^2$ as those with IAM. Note that the estimates of $\sigma_A^2$ and $\sigma_E^2$ by the usual Sire and Dam Model with $R = (2\sigma_I^2 + \sigma_e^2)I$ are different from those with IAM.

5. Discussion

In the field of animal breeding, the dataset-size used for estimating variance components with the REML method is becoming larger and larger. Hence, we definitely need to study less expensive strategies for handling large datasets in the REML estimation.

In this study, assuming a Sire and Dam Model, we describe one form of the iteration equations for EM-REML estimation of the variance components. For the procedure with the equations, all of the five conditions defined in Section 3 must be satisfied so that we are able to obtain the same estimates of additive genetic and environmental variances as those under IAM. A typical example in which all conditions are generally fulfilled is the dataset on carcass traits in beef cattle. As stated in Section 3, $4\sigma_I^2$ and $\sigma_e^2$ in the proposed Sire and Dam Model are theoretically identical to $\sigma_A^2 (= \sigma_a^2)$ and $\sigma_E^2$, respectively, and for a given sample, in spite of different initial values in the parameter spaces for $\sigma_I^2$ and $\sigma_e^2$ their converged values correspond to one fourth of the converged value for $\sigma_a^2$ and the value for $\sigma_e^2$ in the IAM, respectively.

One advantage of the procedure with the current Sire and Dam Model is that the number of the MME to be solved is obviously reduced compared with the case of IAM. For the case of carcass traits, the difference between the sizes of the MMEs for the current model and IAM would generally become large as the number of slaughter animals with one observation each in the data increases. Another advantage of the present procedure is that the speed of convergence to attain final estimates becomes faster than that under IAM. This fact is due to a relatively larger proportion of non-zero elements in the coefficient matrix of the MME, compared with the case of IAM. With the present procedure applied to the specific type of data, therefore, the computational burden to obtain the proper solutions for additive genetic and environmental variances is expected to be considerably reduced. Additional work with the Wagyu Cattle Registry Association of Japan is required to reveal the extent to which the application of the current EM-REML approach, utilizing a sparse matrix solver, to official large datasets for carcass traits reduces the computational burden in the conventional approach with IAM.

Finally, even for carcass traits, there might be some cases that the condition 4) in Section 3 is not satisfied. The REML procedure described by Henderson (1986), using a Reduced Animal Model (Quaas and Pollak, 1980), should be used for such a case.

REFERENCES


*Received June 1995; Accepted January 1996*