Approximation of Standard Errors of Estimates as a By-Product for MC EM REML Analysis

K. Matilainen, I. Strandén and E.A. Mäntysaari.
MTT Agrifood Research Finland

ABSTRACT: We studied the possibility to utilize the Monte Carlo algorithm in estimation of standard errors for MC EM REML variance component estimates. Approach is based on the principle that the expected information matrix at the maximum likelihood estimate is equal to the variance of score function. While score functions include EM updates, the information matrix can be approximated as the variance of scaled EM updates over MC samples. Beef cattle data with birth weight and yearling weight observations was used to demonstrate the idea and to test the effectiveness of the method. The approximated standard errors agreed well with the asymptotic standard errors. Fairly large number of samples was needed to approximate variance of the (co)variance component estimates. The redeeming feature is that the approximated standard errors can be obtained as a by-product of variance component estimation.

Keywords: Monte Carlo; standard error; variance component.

INTRODUCTION
Currently average information (AI) REML (Gilmour et al. (1995)) has become a standard for likelihood based estimation of variance component estimates (VCE). As it uses approximated second derivative of likelihood with respect to variance components, it is much faster than Expectation Maximization (EM) based algorithms (Johnson and Thompson (1995)). Another benefit is that the AI matrix can be used to obtain the standard errors of VCE. The downside in AI REML, as well as in the first derivative based EM algorithm, is that the analytical estimation of first derivative requires inversion of the system of equations that has the size of corresponding mixed model equations.

Variance component estimation using an EM REML algorithm with estimation of prediction error variances by a Monte Carlo (MC) algorithm has shown to be an efficient method when coefficient matrix of a mixed effects model is huge (Matilainen et al. (2012)). Neither analytical EM REML nor MC EM REML gives standard errors of estimates directly. In this study we described how the MC algorithm in MC EM REML gives a tool for approximation of standard errors for estimated variance components in linear mixed effects model. Furthermore, we compared the approximated standard errors of MC EM REML estimates with analytically calculated standard errors using field beef cattle data.

MATERIALS AND METHODS
Data and model. Beef cattle data from Faba (Hollola, Finland) were used. Data had 25,220 birth weight observations and 7,715 yearling weight observations. Pedigree included 36,682 animals. Bivariate model was

\[ y = Xb + Z_a a_i + Z_m m_i + Z_p p_i + \varepsilon \]

where vector \( b \) had three fixed effects (sex, season and interaction of herd and year at birth or one year age), vectors \( a_i \), \( m_i \), \( p \) and \( \varepsilon \) had effects of random animal genetic, maternal genetic, maternal non-genetic environmental and residual, respectively. The incidence matrices \( X, Z_a, Z_m \) and \( Z_p \) relate the model effects to appropriate observations in the vector of observations \( y \). Random effects were assumed to be independently normally distributed with covariance matrices \( G_0 \otimes A, P_0 \otimes I \) and \( R_0 \otimes I \) for maternal and direct genetic, environmental and residual effects, respectively. Here,

\[ G_0 = \begin{pmatrix}
\sigma_{a}^2 & \sigma_{a,b} & \sigma_{a,m} & \sigma_{a,y} \\
\sigma_{a,b} & \sigma_{b}^2 & \sigma_{b,m} & \sigma_{b,y} \\
\sigma_{a,m} & \sigma_{b,m} & \sigma_{m}^2 & \sigma_{m,y} \\
\sigma_{a,y} & \sigma_{b,y} & \sigma_{m,y} & \sigma_{y}^2
\end{pmatrix} \]

\[ P_0 = \begin{pmatrix}
\sigma_{p}^2 & \sigma_{p,b} & \sigma_{p,m} & \sigma_{p,y} \\
\sigma_{p,b} & \sigma_{b}^2 & \sigma_{b,m} & \sigma_{b,y} \\
\sigma_{p,m} & \sigma_{b,m} & \sigma_{m}^2 & \sigma_{m,y} \\
\sigma_{p,y} & \sigma_{b,y} & \sigma_{m,y} & \sigma_{y}^2
\end{pmatrix} \]

and

\[ R_0 = \begin{pmatrix}
\sigma_{e}^2 & \sigma_{e,b} & \sigma_{e,m} & \sigma_{e,y} \\
\sigma_{e,b} & \sigma_{b}^2 & \sigma_{b,m} & \sigma_{b,y} \\
\sigma_{e,m} & \sigma_{b,m} & \sigma_{m}^2 & \sigma_{m,y} \\
\sigma_{e,y} & \sigma_{b,y} & \sigma_{m,y} & \sigma_{y}^2
\end{pmatrix} \]

where sub-subscripts \( b \) and \( y \) refer to birth and yearling weight, respectively. Thus, there were 16 unique elements of covariance components to be estimated.

Method. Consider approximation of standard errors for the REML estimates of genetic covariance matrix \( G_0 \) as an example. Following Jensen et al. (1997) first derivatives of the REML log-likelihood with respect to elements in variance-covariance matrix \( G_0 \) are

\[ \frac{\partial \log L}{\partial G_0} = -\frac{1}{2} \{ q G_0^{-1} - G_0^{-1} (S + D) G_0^{-1} \} \]

where \( q \) is the number of levels in random genetic effect, and \( S \) and \( D \) are 4 by 4 matrices with elements \( S_{ij} = \text{tr}(A^{-1} C^{a,bj}) \) and \( D_{ij} = \bar{a}^{i}_l A^{-1} \bar{a}_j \). Here, \( \bar{a}_i \) is a sub-vector of \( \bar{a} \) corresponding to the BLUP solution of \( i^\text{th} \) trait and effect combination in the model using current variance components and data, and \( C^{a,bj} \) is the part of the inverse of the coefficient matrix of the mixed model equations corresponding to \( \bar{a}_i \) and \( \bar{a}_j, i,j=1,...,4 \). In calculation of first derivatives, matrix \( D \) is easy to compute even for large problems but matrix \( S \) requires an inverse of possibly very large matrix. As shown by García-Cortés et al. (1995), \( S \) can be approximated by the MC algorithm:

\[ S' = q G_0 - \sum_{h=1}^{4} S_h G_h \]

with elements \( S_h^h = \bar{a}^h A^{-1} \bar{a}^h \) and \( \bar{a}^h \) as BLUP estimates for simulated data \( y^h, h=1,...,s \), sampled from the same distribution as the original data \( y \).

First derivatives with respect to environmental and residual covariance components can be computed similarly. All together they form a 16 by 1 score function \( I(\theta) \) where
REML step by MiX99 gave approximated standard errors (ASE). Table 1 shows the relative differences of approximated standard errors in percents: 100*(SE-ASE)/SE. Approximated standard errors by MC method with 1,000 MC samples correspond quite well with the asymptotic standard errors by DMU. Maximum, minimum and mean of relative differences were 3.4, -6.9 and 0.2 percent, respectively. These differences are not only due to the MC sampling algorithm, but DMU gives standard errors based on average information matrix whereas the MC algorithm in MiX99 gives standard errors based on expected information matrix.

Figure 1 shows the approximated standard error along the number of MC samples for animal genetic covariance between birth weight and yearling weight $\sigma_{ab,y}$. Straight line is the asymptotic standard error by DMU.

**RESULTS AND DISCUSSION**

Table 1 shows the REML estimates of 16 different variance components (VC) and the asymptotic standard errors of covariance parameters (SE) calculated by DMU. With the same starting values one additional MC EM
CONCLUSION

It is possible to calculate approximate standard errors of REML estimates using a MC method. The method is fairly easy to implement. Compared to calculations of MC EM REML estimates only, additional calculation is needed to obtain the variances over scaled gradients. The method is, however, computationally intensive. This study proposed that at least 200 MC samples were needed to approximate variance of gradients. Presumably the approximation is more precise when the number of MC samples is increased.

LITERATURE CITED


Madsen, P., and Jensen, J. (2012). DMU. Version 6, release 5.1
