# 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 $\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{Z}_d\mathbf{a}_d+\mathbf{Z}_m\mathbf{a}_m+\mathbf{Z}_p\mathbf{p}+\mathbf{e}$, where vector $\mathbf{b}$ had three fixed effects (sex, season and interaction of herd and year at birth or one year age), vectors $\mathbf{a}_d$, $\mathbf{a}_m$, $\mathbf{p}$ and $\mathbf{e}$ had effects of random animal genetic, maternal genetic, maternal non-genetic environmental and residual, respectively. The incidence matrices $\mathbf{X}$, $\mathbf{Z}_d$, $\mathbf{Z}_m$ and $\mathbf{Z}_p$ relate the model effects to appropriate observations in the vector of observations $\mathbf{y}$. Random effects were assumed to be independently normally distributed with covariance matrices $\mathbf{G}_0\otimes\mathbf{A}$, $\mathbf{P}_0\otimes\mathbf{I}$ and $\mathbf{R}_0\otimes\mathbf{I}$ for maternal and direct genetic, environmental and residual effects, respectively. Here,

$$\mathbf{G}_0=\begin{bmatrix}\sigma^2_{m_b} & \sigma_{m_b,a_b} & \sigma_{m_b,m_y} & \sigma_{m_b,a_y}\\ \sigma_{a_b,m_b} & \sigma^2_{a_b} & \sigma_{a_b,m_y} & \sigma_{a_b,a_y}\\ \sigma_{m_y,m_b} & \sigma_{m_y,a_b} & \sigma^2_{m_y} & \sigma_{m_y,a_y}\\ \sigma_{a_y,m_b} & \sigma_{a_y,a_b} & \sigma_{a_y,m_y} & \sigma^2_{a_y}\end{bmatrix}$$

$$\mathbf{P_0}=\begin{bmatrix}\sigma^2_{p_b} & \sigma_{p_b,p_y}\\ \sigma_{p_y,p_b} & \sigma^2_{p_y}\end{bmatrix}$$

and

$$\mathbf{R_0}=\begin{bmatrix}\sigma^2_{e_b} & \sigma_{e_b,e_y}\\ \sigma_{e_y,e_b} & \sigma^2_{e_y}\end{bmatrix}$$

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 $\mathbf{G}_0$ as an example. Following Jensen et al. (1997) first derivatives of the REML log-likelihood with respect to elements in variance-covariance matrix $\mathbf{G}_0$ are

$$\frac{\partial logL}{\partial \mathbf{G}_0}=-\frac{1}{2}\{q\mathbf{G}_0^{-1}-\mathbf{G}_0^{-1}(\mathbf{S}+\mathbf{D})\mathbf{G}_0^{-1}\},$$

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

$$\mathbf{S}^*=q\mathbf{G}_0-\frac{1}{s}\sum_{h=1}^{s}\mathbf{SS}^h.$$

with elements $\mathbf{SS}^h_{ij}=\hat{\mathbf{a}}^{h\prime}_i\mathbf{A}^{-1}\hat{\mathbf{a}}^h_j$ and $\hat{\mathbf{a}}^h$ as BLUP estimates for simulated data $\mathbf{y}^h$, $h$=1,…,$s$, sampled from the same distribution as the original data $\mathbf{y}$.

First derivatives with respect to environmental and residual covariance components can be computed similarly. All together they form a 16 by 1 score function $\mathbf{J}(\theta)$ where

**Table 1.** REML estimates of covariance components (VC), their asymptotic standard errors (SE) and relative difference of approximated standard errors by MC method with respect to SE (rdASE).

| Parameter[1] | VC | SE | rdASE, % |
|---|---|---|---|
| $\sigma^2_{m_b}$ | 7.166 | 0.8408 | -0.3 |
| $\sigma_{m_b,a_b}$ | -7.797 | 1.027 | -0.3 |
| $\sigma^2_{a_b}$ | 19.28 | 1.735 | 1.8 |
| $\sigma_{m_b,m_y}$ | 36.08 | 6.681 | 0.9 |
| $\sigma_{m_b,a_y}$ | -42.22 | 9.564 | 1.4 |
| $\sigma^2_{m_y}$ | 297.3 | 81.52 | 0.5 |
| $\sigma_{a_b,m_y}$ | -43.84 | 8.312 | -0.6 |
| $\sigma_{a_b,a_y}$ | 87.13 | 12.61 | 3.4 |
| $\sigma_{a_y,m_y}$ | -359.6 | 89.38 | 1.6 |
| $\sigma^2_{a_y}$ | 926.7 | 142.5 | 1.9 |
| $\sigma^2_{p_b}$ | 3.043 | 0.3770 | -4.8 |
| $\sigma_{p_b,p_y}$ | 16.04 | 3.396 | -6.9 |
| $\sigma^2_{p_y}$ | 293.4 | 51.10 | 3.4 |
| $\sigma^2_{e_b}$ | 16.39 | 0.9160 | 2.2 |
| $\sigma_{e_b,e_y}$ | 27.12 | 7.073 | -1.4 |
| $\sigma^2_{e_y}$ | 1771 | 86.40 | 0.8 |

[1]Subscripts m, a, b, y stands for maternal genetic effect, animal genetic effect, birth weight and yearling weight, respectively.

θ has all the 16 variance components. $\mathbf{J}(\theta)$ can be approximated by a score function $\mathbf{J}^h(\theta)$ using MC sample $\mathbf{y}^h$, $h$=1,…,$s$. Following Matilainen et al. (2013) the information matrix can be approximated by the variance of the score functions over MC samples

$$\mathrm{Cov}([\mathbf{J}^1(\theta) \quad \cdots \quad \mathbf{J}^s(\theta)]'),$$

where the function Cov(**J**) returns a 16 by 16 matrix with a diagonal element as variance within a column in **J**, and an off-diagonal element as covariance between a pair of columns in **J**.

**Analyses.** Our goal was to compare the standard error estimation only. Therefore we wanted both the analytical and the MC estimation methods to base to the same VCE. We started analysis with DMU software package (Madsen and Jensen (2012)) to obtain the REML VCE. To assure that standard errors were based on final VCE, DMU was restarted after the convergence. This gave the asymptotic standard errors based on average information matrix.

Similarly, the same AI REML solutions of VCE by DMU were used as starting values in MC EM REML and one REML iterate was done to obtain approximated standard errors. The procedure for calculation of variance of gradients over MC samples was included in the implementation of MC EM REML in MiX99 (Lidauer et al. (2011), Matilainen et al. (2012)). Number of MC samples was set to 1,000 because of the small size of the data.

## 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 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.

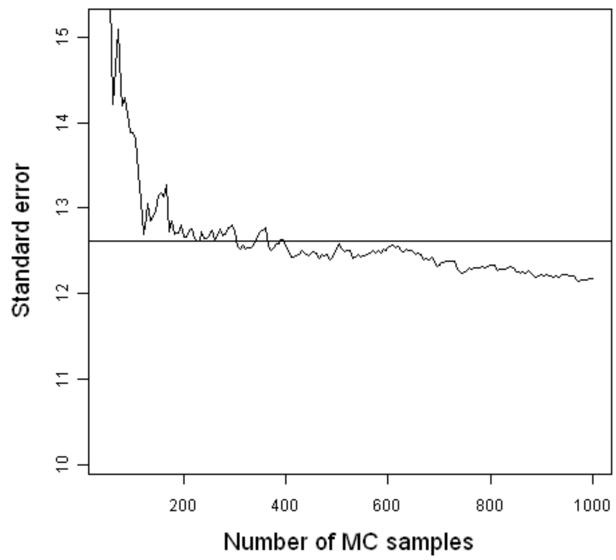


**Figure 1.** Approximated standard errors along the number of MC samples for animal genetic covariance between birth weight and yearling weight $\sigma_{a_b,a_y}$. Straight line is the asymptotic standard error by DMU.

Figure 1 shows the approximated standard error along the number of MC samples for animal genetic covariance between birth weight and yearling weight, which had the largest relative difference among genetic parameters. Although the trend of this parameter seems to continue downwards, additional MC samples showed that level did not change anymore and 1,000 MC samples were enough. Actually, already 400 MC samples would have given similar results as presented in Table 1 and reasonable approximations would have obtained by 200 MC samples.

In practice, use of analytical REML and standard error calculation has limits imposed by memory need of coefficient matrix of the mixed model equations. Such limits do not exist for the MC approach. MC approach can be used for the variance component estimation and standard error calculation for even large problems. While the computing time for each REML iterate in AI REML is proportional to third power of the number of equations in mixed model equations, the MC EM REML scales close to linear with respect to size of the data and number of equations.

## 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

García-Cortés, L. A., Moreno, C., Varona, L. et al. (1995). *J. Anim. Breed. Genet.* 112:176-182.

Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995). *Biometrics* 51: 1440–1450.

Jensen, J., Mäntysaari, E. A., Madsen, P. et al. (1997). *J. Indian Soc. Agric. Stat.* 49:215-236.

Johnson, D. L. and Thompson, R (1995). *J. Dairy Sci.* 78:449-456.

Lidauer, M. H., Matilainen, K., Mäntysaari, E. A. et al (2011). Technical reference guide for MiX99, Release VI/2011.

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

Matilainen, K., Mäntysaari, E. A., Lidauer, M. H. et al. (2012). *J. Anim. Breed. Genet.* 129:457-468.

Matilainen, K., Mäntysaari, E. A., Lidauer, M. H. et al. (2013) *PLoS ONE* 8(12): e80821. doi:10.1371/journal.pone.0080821