# Strategies to Solve MACE Mixed Model Equations

***Freddy Fikse***
*Interbull Centre, Box 7023, 750 07 Uppsala, Sweden*

## Introduction

Routine genetic evaluations are typically subject to time constraints. For example, the Interbull evaluations have a turn around time of 10 days, from the deadline of data submission to Interbull to pre-release of international evaluation results to subscribing countries. During this time, 162 different sets of MACE mixed model equations (MME) need to be solved.

Klei (1995) presented a method to solve the MACE MME which requires an equation for a bull within a country only when he contributes additional information from that country. This is the case when he either has an own observation in a country or a descendant with information in a country. Since 80-90% of the bulls have information in one country only, this approach reduced the size of MME drastically.

Recently an increase in predictive ability of MACE evaluations was reported when sire-dam relationships were used in the additive genetic relationship matrix (Van der Linde and De Jong, 2005), rather than sire-maternal grandsire relationships as in the initial specification of Mace (Schaeffer, 1994). The data set used by Van der Linde and De Jong (2005) comprised seven countries, just over 60 thousand records for 56 thousand bulls with daughter information. The size of the pedigree file nearly doubled from 59 thousand records to 109 thousand records by adding bull dam pedigree information, and the time for solving the MME increased by a factor 9. When considering sire-dam relationships in the MACE evaluation for protein for Holstein, with 25 participating populations, it took approximately 23 hours to solve the MME, which would break the time limits for a routine evaluation.

The aim of this study was to investigate three strategies that reduce the resource (time and memory) to solve for MACE breeding values. The three strategies are: 1) setting up equations (in the MME) for parents only; 2) eliminating equations for bull dams with just on male offspring and 3) application of an alternative ordering algorithm.

## Methods

### *MACE*

In scalar form, the model for MACE can be written as:

$$y_{ij} = c_i + u_{ij} + e_{ij},$$

where: $y_{ij}$ is the observation (de-regressed national genetic evaluation) of bull $j$ in country $i$, $c_i$ is the mean for country $i$, $u_{ij}$ is the genetic merit of bull $j$ in country $i$, and $e_{ij}$ the residual pertaining to $y_{ij}$.

In matrix notation, the model can be written as:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_n \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_n \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix} \quad [2]$$
$$= \mathbf{Xc} + \mathbf{Z}_*\mathbf{u}_* + \mathbf{e}$$

where: $\mathbf{y}_i$, $\mathbf{u}_i$ and $\mathbf{e}_i$ are the vector of observations, genetic effects and residuals, respectively, in country $i$, $\mathbf{X}_i$ is an incidence matrix connecting observations to country mean effect, and $\mathbf{Z}_i$ is an incidence matrix connecting observations to genetic effects. The variance-covariance matrices for the random effects are: $\mathrm{var}(\mathbf{u}_*) = \mathbf{G} = \mathbf{A} \otimes \mathbf{G}_o$, where $\mathbf{A}$ is the relationship among animals in $\mathbf{u}_*$ and $\mathbf{G}_o$ the genetic variance-covariance matrix among traits, and $\text{var}\left(\mathbf{e}\right) = \mathbf{R} = \sum^{+} \mathbf{W}_i \sigma^2_{e_i}$, where $\mathbf{W}_i$ is a diagonal matrix with $EDC_{ij}^{-1}$ as elements, $EDC_{ij}$ being the effective daughter contribution of bull *j* in country *i*.

Considering that $\mathbf{u}_* = \mathbf{Qg}_* + \mathbf{s}_*$, solutions for country means and genetic effects can be obtained by solving the following MME:

$$\begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{0} & \mathbf{X'R^{-1}Z_*} \\ \mathbf{0} & \mathbf{Q'G^{-1}Q} & \mathbf{\text{-}Q'G^{-1}} \\ \mathbf{Z_*'R^{-1}X} & \mathbf{G^{-1}Q} & \mathbf{Z_*'R^{-1}Z_*} + \mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \mathbf{g}_* \\ \mathbf{u}_* \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{0} \\ \mathbf{Z_*'R^{-1}y} \end{pmatrix} \qquad [3]$$

### *Klei method*

By differentiating between local bulls and “international” bulls (i.e. bulls with progeny in >1 country), Klei (1995) rearranged [2] for a two-country case:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{Z}_{12} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_{21} & \mathbf{Z}_{23} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{11} \\ \mathbf{u}_{12} \\ \mathbf{u}_{21} \\ \mathbf{u}_{23} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \qquad [4]$$

$$= \mathbf{Xc} + \mathbf{Zu} + \mathbf{e}$$

Here, the first subscript to **Z** and **u** identifies the country for which breeding values are predicted, and the second subscript is an identifier for type of bull: 1 for international bulls, and *i+1* for local bulls in country *i*.

The genetic effects for country 1 for the local bulls in country 2 ($\mathbf{u}_{13}$), and the genetic effects for country 2 for local bulls in country 1 ($\mathbf{u}_{22}$) are not needed to model the observations. Consequently, no equations for those effects need to be included in the MME. Since a relatively small proportion of bulls are international bulls, this parameterization of the model leads to a drastically reduced dimension of the MME compared to [3].

The only caveat of this approach is that $\mathbf{G}^{-1}$ is slightly more complicated to build (i.e. no longer $\mathbf{A}^{-1} \otimes \mathbf{G}_o^{-1}$): $\mathbf{G}^{-1} = \sum_i \mathbf{t}_i \mathbf{t}_i' d_i^{-1} \otimes \mathbf{G}_k$. Here, $\mathbf{A}^{-1} = \sum_i \mathbf{t}_i \mathbf{t}_i' d_i^{-1}$, where the summation is over all animals in the pedigree, $\mathbf{t}_\text{i}$ is the $i^{\text{th}}$ row of matrix $\mathbf{T}^{-1}$ which has ones on the diagonals and negative values to the left of the diagonal in the columns corresponding to the parents of animal *i* (Quaas, 1988), and $\mathbf{G}_k = \mathbf{H}_k \left(\mathbf{H}_k' \mathbf{G}_o \mathbf{H}_k\right)^{-1} \mathbf{H}_k'$, where $\mathbf{H}_\text{k}$ is a picker matrix obtained by deleting a column from $\mathbf{I}_2$ corresponding to the country in which animal *i* does not have information (Klei, 1995). This structure of $\mathbf{G}^{-1}$ is such that it is easy to build by processing the list of animals (with their parents) and an indicator in which countries the animal has information.

### *Reduced model*

For animal model evaluations Quaas and Pollak (1980) showed that the size of the MME can be reduced by setting up the equations for parents only. The technique is based on the general result that any random effect can be put into the residual term as long as they are not correlated to the random effects remaining in the model. Quaas and Pollak (1980) state that the MME formed from the reduced equivalent model will be exactly the same as those obtained by absorbing equations from the full set of equations.

For a single observation, the statistical-genetic model can be written as:

$$y_{ij} = c_i + a_{ij} + e_{ij}$$

Breeding value $a_{ij}$ can be expressed as:

$$a_{ij} = \tfrac{1}{2}\left(a_{is} + a_{id}\right) + m_{ij} \qquad [5]$$

where $m_{ij}$ is the Mendelian sampling deviation for animal *j* in country *i*. Combining the previous two equations leads to:

$$y_{ij} = c_i + ½\left(a_{is} + a_{id}\right) + m_{ij} + e_{ij}$$

For non-parents, $m_{ij}$ and $e_{ij}$ can be combined to form a single residual term:

$$e^*_{ij} = m_{ij} + e_{ij}$$

with variance:

$$\begin{aligned}\text{var}\left(e^*_{ij}\right) &= \text{var}\left(m_{ij}\right) + \text{var}\left(e_{ij}\right) \\ &= d_j^{-1}\sigma^2_{a_i} + w_{ij}^{-1}\sigma^2_{e_i}\end{aligned}$$

In matrix notation, the observations for country 1 for local bulls without male offspring are modeled as:

$$\mathbf{y}_{12_n} = \mathbf{X}_{12_n} c_1 + \mathbf{Z}_{12_n}\mathbf{u}_{12_p} + \mathbf{e}_{12_n}, \qquad [6]$$

and for local bulls with male offspring:

$$\mathbf{y}_{12_p} = \mathbf{X}_{12_p} c_1 + \mathbf{Z}_{12_p}\mathbf{u}_{12_p} + \mathbf{e}_{12_p} \qquad [7]$$

Combining [6] and [7], and adding the observations for country 1 for the international bulls leads to the following model for all observations in country 1:

$$\mathbf{y}_1 = \begin{pmatrix} \mathbf{y}_{11} \\ \mathbf{y}_{12_p} \\ \mathbf{y}_{12_n} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{12_p} \\ \mathbf{X}_{12_n} \end{pmatrix} c_1 + \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{12_p} \\ \mathbf{0} & \mathbf{Z}_{12_n} \end{pmatrix} \begin{pmatrix} \mathbf{u}_{11} \\ \mathbf{u}_{12_p} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_{11} \\ \mathbf{e}_{12_p} \\ \mathbf{e}_{12_n} \end{pmatrix} \qquad [8]$$

The variance-covariance matrix of residuals in [8] becomes:

$$\text{var}\left(\mathbf{e}_1\right) = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{12_p} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}^{-1}_{12_n}\sigma^2_{a_1} + \mathbf{W}_{12_n}\sigma^2_{e_1} \end{pmatrix}$$

***Absorption of dam***

Dams do not have own observations in MACE, but enter the picture through the additive numerator relationship matrix which is based on sire-dam relationships.

The variance-covariance matrix for Mendelian sampling effects is (block) diagonal (i.e., Mendelian sampling effects of any two different animals are uncorrelated), and this effect can in a sense be viewed as a residual effect.

Equation [5] is recursive, which can be exploited by decomposing it as:

$$\begin{aligned} a_{ij} &= ½\left(a_{is} + ½\left(a_{i,mgs} + a_{i,mgd} + m_{id}\right)\right) + m_{ij} \\ &= ½a_{is} + ¼a_{i,mgs} + ¼a_{i,mgd} + ½m_{id} + m_{ij} \end{aligned} \qquad [9]$$

If the bull *j* is the only offspring of this dam, then $½m_{id}$ and $m_{ij}$ can be combined to form a single Mendelian sampling term:

$$m^*_{ij} = m_{ij} + ½m_{id}$$

with variance:

$$\begin{aligned}\text{var}\left(m^*_{ij}\right) &= \text{var}\left(m_{ij}\right) + \text{var}\left(½m_{id}\right) \\ &= \left(d_j^{-1} + ½d_d^{-1}\right)\cdot\sigma^2_a\end{aligned}$$

The same principle as for the reduced model is applied here: a random effect not correlated to any other random effects in the model can be put in the residual.

The rules for building $\mathbf{A}^{-1}$ based on sire-dam relationships stem from [5] (Henderson, 1976). Similarly, rules for building $\mathbf{A}^{-1}$ when absorbing dam equations are based on [9]. In fact, these rules are essentially the same as for building $\mathbf{A}^{-1}$ based on sire-mgs relationships.

### *Ordering strategies*

Two different approaches for symbolic factorization of the MME were compared: the multiple minimum degree (MMD) algorithm (Liu, 1985), which is the default in Fspak90 (Misztal and Perez-Enciso, 1988), and a multilevel partitioning algorithm implemented in the Metis software package (Karypis and Kumar, 1998). Several of the tuning parameters of the Metis software were varied to investigate their effect on the (cpu-) time and memory needed to solve the MME.

## Material

Two different data sets were used to illustrate the effect of the three strategies. First, the data set used for the AM analysis by Van der Linde and De Jong (2005) was available (referred to as Small). Second, the national evaluation results used in the Interbull test evaluation of March 2007 for Holstein, protein was used (referred to as Large). Countries had been requested to provide bull dam pedigree in addition to bull pedigree that is routine collected. Summary statistics for both data sets are in Table 1.

**Table 1** Summary of data sets

| | Small | Large |
|---|---|---|
| No of countries | 7 | 25 |
| No of genetic groups | 228 | 835 |
| No of animals in pedigree | 108648 | 197716 |
| Males | 61987 | 105396 |
| Females | 46870 | 92946 |
| with 1 offspring | 32072 | 66977 |
| with >1 offspring | 15008 | 26596 |
| No of observations | 60859 | 110034 |
| No of bulls with obs. | 56775 | 96502 |
| with obs. in 1 country | 54024 | 89609 |
| with obs. in >1 country | 2751 | 6893 |
| without offspring | 55455 | 93901 |
| with offspring | 1320 | 2601 |

## Results

### *Small data set*

The rank of the MME was reduced by 36% and 22% when putting Mendelian sampling terms into the residuals for non-parent bulls reduced and absorbing equations for bull dams with just one progeny (Table 3). When both approaches were combined, the reduction was 57%.

The reduction in the number of non-zero elements in the MME ranged between 9 and 28%, which was much smaller than the reduction of rank (Table 3). This indicates that both techniques lead to an increased density of the MME.

The number of non-zero elements in the factor matrix was between 3 and 6% lower than in the reference situation. Unexpectedly, the combination of putting Mendelian sampling terms in the residual and using Metis to order the equations resulted into more non-zero elements in the factor compared with the reference situation.

When using the MMD algorithm for symbolic factorization, the alternative where Mendelian sampling terms were put into the residuals was fastest to solve the MME (Table 3). When bull dam equations were also absorbed, slightly more time was needed to obtain solutions, probably because the number of non-zero elements in factor was higher.

The choice of ordering algorithm had the largest influence on the time needed to solve the MME (Table 3). The combination of the absorption techniques and Metis ordering roughly halved the time needed obtain solutions.

**Table 3** Size of the mixed model equations and time for solving for combinations of three “reduction” approaches for the Small data set.

| Alternative | | | | Rank | NZE (million) | | Work space needed by Fspak (Mb) | Time (s) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Klei | MS | Dam | Order | | Upper triangular | Factor | | Total | Order & symb fac | Num fac |
| y | n | n | MMD | 147337 | 1.35 | 20.19 | 179 | 382.9 | 74.2 | 308.2 |
| y | y | n | MMD | 94069 | 1.12 | 19.46 | 169 | 338.9 | 56.3 | 282.1 |
| y | n | y | MMD | 115887 | 1.23 | 19.60 | 172 | 357.7 | 65.5 | 291.8 |
| y | y | y | MMD | 62619 | 0.98 | 19.66 | 168 | 343.7 | 44.5 | 298.8 |
| y | y | n | Metis[a] | 94069 | 1.12 | 21.23 | 189 | 251.5 | 1.9 | 249.2 |
| y | y | y | Metis[a] | 62619 | 0.98 | 19.00 | 165 | 199.9 | 1.5 | 198.1 |

[a] Metis ordering with default values for parameters

**Table 3** Size of the mixed model equations and time for solving for combinations of three “reduction” approaches for the Large data set.

| Alternative | | | | Rank | NZE (million) | | Work space needed by Fspak (Gb) | Time (s) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Klei | MS | Dam | Order | | Upper triangular | Factor | | Total | Order & symb fac | Num fac |
| y | n | n | MMD | 390723 | 10.37 | 627.00 | 4.83 | 30806.9 | 281.3 | 30519.6 |
| y | y | y | Metis[a] | 207514 | 9.29 | 517.56 | 4.06 | 19169.4 | 4.6 | 19159.8 |
| y | y | y | Metis[b] | 207514 | 9.29 | 408.72 | 3.23 | 12812.0 | 60.2 | 12747.8 |

[a] Metis ordering with default values for parameters; [b] Metis ordering, obtained using 20 graph separators in each dissection step, and considering vertices with degree 50 times higher than average as dense and placing them at the bottom of the graph.

### *Large data set*

The rank was nearly halved by putting Mendelian sampling terms into the residuals for non-parent bulls reduced and absorbing equations for bull dams with just one progeny (Table 3). Like for the small data set, the reduction in the non-zero elements in the upper triangular of the left hand side and the factor matrix was relatively smaller. The amount of workspace needed by Fspak90 reduced by 17 to 35%, but depending on the parameters for Metis.

The time needed for solving the MME reduced by 17% when putting Mendelian sampling into the residuals and absorbing bulldam equations, and using the default values for the parameters in Metis. Another considerable reduction was achieved after specifying different parameters. For the Large data set it is worthwhile to note that increased time for ordering the MME can result in much faster numerical factorization.

## Discussion

For the Large data set, with data from 25 populations, just over 60 thousand records and nearly 200 thousand animals in the pedigree, the time needed to solve the MME was just below 4 hrs for the fastest alternative. The program to solve the MME required approximately 5 Gb of RAM. With this performance it becomes feasible to consider MACE with sire-dam relationships for routine international genetic evaluations.

Absorption of effects implies that all predicted breeding values are not directly solved through the MME. However, “missing” breeding values for the absorbed equations can be predicted by back-solving, and is implemented in the software.

Application of LU decomposition for direct solving of the MME (by Fspak90) may need to be given up due to non-linear increases of resources required when amount of data and pedigree increases. Iterative methods, for example preconditioned conjugate gradient, come into the picture; the absorption techniques presented here are applicable in that case as well, especially when iteration is on the coefficient matrix, but the effect on the time needed to solve the MME may need to be re-evaluated.

## References

Klei, L. 1995. Evaluating Holstein sires for conformation traits using data from the United States and The Netherlands. *PhD thesis*. Cornell University, Ithaca, New York.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics 32*, 69-83.

Liu, J.W.H. 1985. Modification of the minimum degree algorithm by multiple elimination. *ACM Trans. Math. Soft. 11*, 141-153.

Karypis, G. & Kumar, V. 1998. A software package or partitioning unstructured graphs, partitioning meshes and computing fill-reducing orderings of sparse matrices. Version 4.0. http://www.cs.umn.edu/~metis.

Misztal, I. & Perez-Enciso, M. 1998. FSPAK90: A Fortran90 interface to parse-matrix package FSPAK with dynamic memory allocation and sparse matrix structure. *Proc 6th WCGALP 27*, 467-468.

Quaas, R.L. 1988. Additive genetic model with groups and relationships. *J. Dairy Sci. 71*, 1338-1345.

Quaas, R.L. & Pollak, E.J. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. *J. Anim. Sci. 51*, 1277-1287.

Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. J. Dairy Sci. 77, 2671-2678.

Van der Linde C. & De Jong, G. 2005. Mace with sire-mgs and animal pedigree. http://www.interbull.org