Pre-selection approaches or some models with Mendelian sampling terms

I. Strandén[1], and E.A. Mäntysaari[1]

[1] *Natural Research Institute Finland (Luke) Tietotie 4, 31600 Jokioinen, Finland*

Abstract

Two equivalent models based on orthogonal random effects were presented. The first model was based on the LDL transformation of the relationship matrix of animals with observations, the second model was based on the LDL transformation of the full relationship matrix of all animals in the pedigree. The latter model yields directly the estimates of Mendelian Sampling terms, which can then be back transformed to breeding values. Models were tested using a small three country MACE protein data. Being analogous to the SNP-BLUP model, the transformed models were fitted using a regression design matrix approach with off-the-shelf breeding value estimation program MiX99. Both the orthogonal models and the original MACE model gave the same estimates for breeding values. From the new approaches, the full pedigree transformation was computationally more efficient although it required many more iterations to converge than the normal MACE model. The reason for better efficiency was postulated to the sparsity (low number of non-zeros) in the transformed design matrix. An approach to account for the reduction in MS-term variance due to genomic preselection was suggested.

Key words: MACE evaluations, Genomic preselection, Mendelian sampling term model

Introduction

The input phenotypes for the MACE are derived from national EBVs. In countries with genomic selection, these are biased because data from genomic preselection (GPS) of AI bulls is not included in evaluations. As a result, the EBVs deviate from the expected the more generations genomic selection has been applied. The GPS affects directly the Mendelian sampling (MS) terms: stronger is the selection, larger is E[MS], and smaller is Var[MS]. Because in most countries, the MACE results are used in genomic evaluations as reference phenotypes, the direct inclusion of GPS information into MACE input EBVs is not possible. Sullivan et al. (2022) proposed a new MACE approach which attempts to model the GPS effects. Their approach models the yearly averages of MS terms, and, after the analysis, the yearly averages are returned to EBVs. Sullivan's "Future MACE" is based on the ordinary EBV model. It should be possible to re-parametrize the model to operate directly on Mendelian sampling terms. Such an equivalent model was proposed by Quaas (1984) and acknowledged by Smith and Graser (1986). If the additional model terms in the Future MACE model would be included in the reparametrized model, the results would be equivalent. Furthermore, the MS reparameterization would make it easier to take into account the reduced variance in MS due to GPS.

This paper presents two models with transformed design matrices that will result unrelated random effects. In one of the models, the orthogonal unknown random effects are conventional Mendelian sampling terms. The models were tested with a small MACE data. Moreover, we present a possible approach for accounting for the reduction in Mendelian sampling variance due to genomic selection.

Materials and Methods

Models with unrelated random effects

Let the standard MACE BLUP model be

$$\boldsymbol{y} = \boldsymbol{X\mu} + \boldsymbol{Z_o u_o} + \boldsymbol{e}$$

where **y** has the country wise deregressed genetic predictions, $\boldsymbol{X\mu}$ describes the country/trait effects, and $\boldsymbol{u}_o$ and $\boldsymbol{e}$ the bull breeding values and residuals. The design matrix $\mathbf{Z}_o$ associates the breeding values to the observations. Assume $\boldsymbol{u_o} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{G_0} \otimes \boldsymbol{A_o})$ and $\boldsymbol{e} \sim \boldsymbol{N}(\mathbf{0}, \otimes \boldsymbol{R_i})$, where $\boldsymbol{G_0}$ and $\mathbf{A}_o$ are the genetic (co)variance matrix of traits and the relationship matrix of bulls; and $\mathbf{R}_i$ are the residual variances of observations in each country.

The variance of breeding values associated with observations is

$$\boldsymbol{var}(\boldsymbol{Z_o u_o}) = \boldsymbol{Z_o}(\boldsymbol{G_0} \otimes \boldsymbol{A_o})\boldsymbol{Z_o}'$$

After decomposing the $\boldsymbol{A_o} = \boldsymbol{L_o D_o L_o}'$, we can derive an equivalent model where $\boldsymbol{Z_o u_o} = \boldsymbol{Z_o L_o m_o} = \widetilde{\boldsymbol{Z_o}} \boldsymbol{m_o}$, with $\boldsymbol{var}(\boldsymbol{m_o}) = \boldsymbol{G_0} \otimes \boldsymbol{D_o}$, and the matrix $\mathbf{D}_o$ diagonal. This gives us the first orthogonal random effect model:

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\mu} + \widetilde{\boldsymbol{Z}}_o \boldsymbol{m_o} + \boldsymbol{e} \qquad \textbf{(MS I)}$$

where now the $\boldsymbol{m}_0$ vector represents the deviations of individual bulls from their (earlier) relatives in $\boldsymbol{u}_o$.

In the above, the vector $\mathbf{u}_o$ included only the breeding values of bulls that had observations in at least one country. More common is an equivalent model where the **u** includes also the ancestors of the bulls with observations. Then $\boldsymbol{Z_o u_o} = \boldsymbol{Zu}$ and the variance remains the same

$$\boldsymbol{var}(\boldsymbol{Zu}) = \boldsymbol{Z}(\boldsymbol{G_0} \otimes \boldsymbol{A})\boldsymbol{Z}',$$

but here the matrix **A** includes all the ancestors of the bulls in $\mathbf{u}_o$. Analogously to MS I, the matrix **A** can be decomposed as **LDL**', and we can write second orthogonal term model:

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\mu} + \widetilde{\boldsymbol{Z}}\boldsymbol{m} + \boldsymbol{e} \qquad \textbf{(MS II)}$$

where now $\widetilde{\boldsymbol{Z}} = \boldsymbol{ZL}$. If the animals in **u** are ordered by age from oldest to young, the LDL decomposition can be written using the rules given by Quaas et al. (1984). Forming the **L** matrix can be done just by reading the pedigree file in age order. Then the **m** vector represents the MS-terms, and the **D** matrix their variance. With t traits and N_{obs} animals with data, the $\mathbf{Z}_o$ matrix has t x N_{obs} columns. The number of columns in the **Z** matrix is t x N_{ped}. Thereby also the number of equations in MME to be solved in MS II is larger than in MS I.

After solving the models, the original breeding value solutions are attained by $\widehat{\boldsymbol{u}} = \boldsymbol{L}\widehat{\boldsymbol{m}}$, i.e., standard BLUP and MS I and MS II models will give the same estimates for the breeding values.

Test data

The models were tested using a small data obtained by extracting 3 moderate size countries from the MACE research data used by Tyrisevä et al. (2018). Protein was used as a test trait. The total number of bulls with observations was 31,578 and the number of individuals in the full animal model pedigree was 66,775.

Matrices $\mathbf{L}_o$ and $\mathbf{D}_o$ were calculated from the $\mathbf{A}_0$ matrix using a short F90 program. The matrices **L** and **D** were obtained directly from the program Relax2 (Strandén & Vuori, 2006) after a small modification.

The original MACE model and the MS-term models were solved using MiX99 program (Pitkänen et al. 2022). The **L** matrices were given as regression design matrices using the same instructions that are used for the SNP-BLUP models (REGMATRIX, see MiX99 instructions in Appendix). Convergence of iteration was assumed when

$$\| \boldsymbol{rhs_{mme}} - \widehat{\boldsymbol{rhs}}_{mme} \| \Big/ \| \boldsymbol{rhs_{mme}} \| \leq \mathbf{10^{-6}},$$

where $\| \, . \, \|$ stands for 2-norm of a vector.

Results & Discussion

$\tilde{Z}$ matrix summaries

As the size of the $\mathbf{L}_0$ matrix was *31,578 x 31,578,* it consisted of 997M elements, but from which 492M (49%) were non-zero. This indicates that almost all animals with observations were somewhat related to each other. The **L** matrix had the same number of rows, but 66,776 columns. However, it was very sparse so that only 135,776 elements were non-zero (0.01%). This sparsity was exploited by instructing MiX99 to keep only the non-zeros in the memory.

Iteration times and PCG Convergence

Both MS models showed poor convergence (Figure 1). When the MACE model required 178 PCG iterations, the MS I model required 2,164 iterations and the MS II model 1,098 iterations. This was expected since similar differences have been seen in the comparison of GBLUP and SNPBLUP. As has been found for SNP-BLUP the convergence can likely be improved by a better PCG preconditioner matrix or using so called second level preconditioner (Vandenplas et al. 2019).

Also, the full iteration time was much less (44 sec) for the MS II model than for MS I (53 min). This was because of fewer iterations, but also the iteration rounds were much faster because of only a fraction of non-zero elements in the **L** compared to $\mathbf{L}_o$. In a true implementation of MS II, the multiplication of direction vector **v** by ***ZL*** in the PCG algorithm could be done using the sparse $\mathbf{L}^{-1}$ matrix build in-the-fly while reading pedigree.

The current approach was only an equivalent model for the ordinary MACE. To take into account the effect of GPS on the expected value of MS terms, an additional fixed effect needs to be added to the model. This could be a bull birth year within the country of selection. To take into account the effect of GPS on var(MS), a two level iterative approach could be tested:

1) Solve the MS model, get solutions for the MS term **m**
2) Compute the SD (and average) of **m** within predefined groups
3) Adjust the variance terms in **D** for individuals with deviating **m** using the information in step 2)
4) Go to step 1) with the new **D** or stop after some rounds.

If the GPS would lead to less variation in MS-terms, the updated **D** would lead to higher regression of MS-terms towards parent averages and corresponding birth-year means.

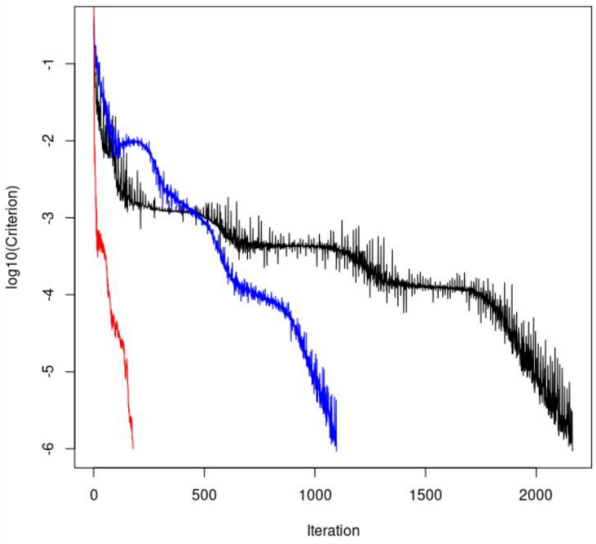

Figure 1. Convergence patterns of different solving approaches presented as a relative norm of the difference of true and predicted RHS of MME (log(Cr)). The red line is for the normal MACE, blue represents the orthogonal terms for bulls with observations, and the black line is for the MS-term model with the solutions for all the ancestors.

Conclusions

Two different models based on orthogonal random effects gave the same estimates for breeding values. The first model was based on the LDL transformation of the relationship matrix of animals with observations, the second model was based on the LDL transformation of the full relationship matrix with all animals in the pedigree. The full pedigree transformation was computationally more efficient when fitted

with off-the-shelf genetic/genomic evaluation software MiX99.

Acknowledgments

Peter Sullivan (Lactanet, Canada) for his comments and help in understanding the Future MACE model.

References

Quaas, R. L., Anderson, R. D., & Gilmour, A. R. 1984. *BLUP school handbook. Use of mixed models for prediction and for estimation of (co) variance components*. Animal Genetics and Breeding Unit, University of New England.

Pitkänen, T. J., Gao, H., Kudinov, A., Taskinen, M., Mäntysaari, E. A., Lidauer, M. H., & Strandén, I. 2022. From data to genomic breeding values with the MiX99 software suite. In *Proc 12th WCGALP,* pp. 1534-1537. Wageningen Academic Publishers.

Smith, S. P., & Graser, H. U. 1986. Estimating variance components in a class of mixed models by restricted maximum likelihood. *Journal of Dairy Science*, *69*(4), 1156-1165.

Strandén, I., & Vuori, K. 2006. RelaX2: pedigree analysis programme. Proc of the 8th WCGALP, Belo Horizonte, Brazil, 13-18 August, 2006 (pp. 27-30).

Sullivan, P. G., Mäntysaari, E., de Jong, G., & Savoia, S. 2022. Using genetic regressions to account for genomic preselection effects in MACE. Interbull Bulletin, (57), 117-124.

Tyrisevä, A. M., Mäntysaari, E. A., Jakobsen, J., Aamand, G. P., Dürr, J., Fikse, W. F., & Lidauer, M. H. 2018. Detection of evaluation bias caused by genomic preselection. J Dairy Sci, 101(4), 3155-3163.

Vandenplas, J., Calus, M. P., Eding, H., & Vuik, C. (2019). A second-level diagonal preconditioner for single-step SNPBLUP. Genetics Selection Evolution, 51, 1-16.

.

Appendix I.

MiX99 instruction codes.

Multi-trait MACE model for 3 countries

```
DATAFILE  ../MACE_smaller_123_MT.dat
INTEGER   BULL CTRY1 CTRY2 CTRY3
REAL      dyd1 W1 dyd2 W2 dyd3 W3
MISSING   -8192.0

PARFILE   MACE_smaller.var   # Variance component file

PEDFILE   ../amped_selected.ped # Pedigree file
PEDIGREE  BULL am  # Genetics associated with pedigree

TMPDIR ./tmp

MODEL
   dyd1 = CTRY1  -     -    BULL ! weight= W1
   dyd2 =   -  CTRY2   -    BULL ! weight= W2
   dyd3 =   -    -   CTRY3 BULL ! weight= W3
```

MS I model

```
DATAFILE  ../MACE_smaller_123_MT.dat
INTEGER   BULL CTRY1 CTRY2 CTRY3
REAL      dyd1 W1 dyd2 W2 dyd3 W3
MISSING   -8192.0

PARFILE   ../MACE_regr_MT.var   # Variance component file

REGMATRIX   heterogeneous reg FIRST=1 LAST=31578
REGFILE     ../amatrix_MT.L.txt
REGPARFILE  amatrixD_MT_3tr.txt

TMPDIR ./tmp

MODEL
   dyd1 = CTRY1  -     -    ! weight=W1
   dyd2 =   -  CTRY2   -    ! weight=W2
   dyd3 =   -    -   CTRY3 ! weight=W3
```

MS II model

```
DATAFILE  ../MACE_smaller_123_MT.dat
INTEGER   BULL CTRY1 CTRY2 CTRY3
REAL      dyd1 W1 dyd2 W2 dyd3 W3
MISSING   -8192.0

PARFILE   ../MACE_regr_MT.var   # Variance component file

REGMATRIX   heterogeneous reg first=2 last=66776
REGFILE     ../lmatrix_MT.reg
REGPARFILE  ../lmatrix_MT.par

TMPDIR ./tmp

MODEL
   dyd1 = CTRY1  -     -    ! weight=W1
   dyd2 =   -  CTRY2   -    ! weight=W2
   dyd3 =   -    -   CTRY3 ! weight=W3
```

The MACE model using trait groups

```
DATAFILE  MACE_smaller.dat
INTEGER   BULL COUNTRY
REAL      dyd_PROT WEIGHT
MISSING   -8192.0

TRAITGROUP COUNTRY

PEDFILE   amped_selected.ped # Pedigree file
PEDIGREE  BULL am  # Genetics associated with pedigree

PARFILE   MACE_smaller.var   # Variance component file

TMPDIR ./tmp

MODEL
   dyd_PROT(1) = COUNTRY BULL ! weight= WEIGHT
   dyd_PROT(2) = COUNTRY BULL ! weight= WEIGHT
   dyd_PROT(3) = COUNTRY BULL ! weight= WEIGHT
```