DMU - A Package for Analyzing Multivariate Mixed Models in quantitative Genetics and Genomics

P. Madsen, J. Jensen, R Labouriau, O. F. Christensen and G. Sahana

Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University

ABSTRACT: The DMU-package for Analyzing Multivariate Mixed Models has been developed over a period of more than 25 years. This paper gives an overview of new features and the recent developments around the DMU-package, including: Genomic prediction (SNP-BLUP, G-BLUP and “One-Step”), Genome-wide association studies, Survival models and double hierarchical generalized linear mixed models.

Key words: genomic; survival-analysis; DHGLM.

INTRODUCTION

DMU is a package primarily used for applications in quantitative genetics and genomics. The package implements powerful tools to estimate variance components, fixed effects (BLUE), and to predict random effects (BLUP). Developments of DMU have been driven by needs of research projects in applied quantitative animal genetics over a period of more than 25 years. A general overview of modules in DMU and their functions are discussed in Madsen et. al (2010) and in Madsen & Jensen (2013). This paper describes some of the new features added in the recent years.

Genomic Prediction

SNP-BLUP

The basic statistical model is given by

$$y = \mu + \sum_{j=1}^{M} X_j a_j + e \,, \qquad [1]$$

where y is a vector of phenotypes with length N ; a_j is the effect of SNP $_j$, X_j is a vector of length N of genotypes of the individuals for SNP j, where $-1/\sqrt{H_j}$ denotes homozygous for the first allele; 0 denotes heterozygous; $1/\sqrt{H_j}$ denotes homozygous for the second allele, and H_j is the heterozygosity for the j'th SNP.

The modules DMUAI, DMU4 and DMU5 that are used for variance components estimation, hypothesis testing and for obtaining solutions to MME in large datasets can handle models with: Common variance components for all SNP effects and any number of SNP groups with common variance per SNP group. This latter functionality can be used to group SNPs in genomic feature models where SNPs are grouped according to gene functioning, metabolic pathways, or effect on specific traits.

G-BLUP

In G-BLUP the additive relationship matrix **A,** which express the expected relationship is replaced by the genomic relationship matrix **G**, which express the realized relationship among individuals. **G** itself is relatively straightforward to calculate based on VanRaden (2008), but its inverse cannot be set up directly, and must be computed using brute force methods. This can be computationally demanding and will limit the size of **G** matrices that can be handled to a few hundred-thousand individuals, both due to computer time requirement and numerical accuracy. Furthermore, **G** as well as $\mathbf{G}^{-1}$ may be dense and **G** does not necessarily have full rank and no exact inverse exists.

The statistical model is given by

$$y = X\beta + Zu + e \,, \qquad [2]$$

where y is a vector of phenotypes, β is the vector of fixed effect(s) and u is the vector of breeding values, **X** and **Z** are appropriate design matrices.

The MME for model [2] is

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \otimes G_o^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix},$$

where G is the genomic relationship matrix constructed from SNP's and G_0 is the genetic co-variance matrix.

The variance components of model [2] can be estimated using DMUAI, and prediction can be performed with DMU4 and DMU5. The inverse genomic relationship matrix ($\mathbf{G}^{-1}$) is used as a variance structure. The model can include a number of genomic relationship matrices with corresponding co-variance matrices. As for the SNP-BLUP model this can be used for genomic feature models where the genomic relationship matrices are based on groups of SNPs. To facilitate model comparison of nested models by Likelihood Ratio test, log(|**G**|) can be specified as the first record in the file containing $\mathbf{G}^{-1}$.

One-Step Method

The One-Step procedure combining genomic and additive relationship developed by Christensen & Lund (2010) has been implemented for variance component estimation by DMUAI and for prediction in DMU4 and DMU5. The One-Step approach can be used for all types of model that can be handled by DMU using the combined relationship matrix instead of the additive genetic relationship matrix based on pedigree information. The combined relationship matrix used in One-step (**H** matrix) can be setup with different weight on the additive relationship matrix for typed individuals $\mathbf{A}_{11}$ and with an adjusted genomic relationship matric $\mathbf{G}^{*}$ so that the Avg.diag($\mathbf{G}^{*}$) = Avg.diag($\mathbf{A}_{11}$) and Avg.offdiag(**G***) = Avg.offdiag($\mathbf{A}_{11}$).

The computation of the inverse combined relationship matrix $\mathbf{H}^{-1}$ is done using DMU1. As the genomic relationship matrix for genotyped animals is typically dense, the number of typed animals can be a limiting factor for size of problems that can be handled due to memory requirement. For situations with 20.000 and 40.000 genotyped animals, the additional memory needed by DMU1 for computing $\mathbf{H}^{-1}$ are ~3.5 and 13 GB, respectively. The fact that $\mathbf{H}^{-1}$ is less sparse than A^{-1} also increases the memory requirement for prediction (DMU4 and DMU5).

Parallel Computation

The LHS of the MME system in SNP- and G-BLUP are in general dense and memory requirements are considerably larger compared to traditional animal model BLUP. Another consequence of the dense LHS is that the traditional use of sparse matrix techniques becomes inefficient, therefore, options for using dense matrix operation have been implemented in DMUAI and DMU4. The implementation is based on LAPACK (Anderson et. al (1999)) subroutines parallelized for multi-core (SMP) computers based on shared memory architecture.

The use of dense matrix operations combined with parallel computation has shown to give a considerable reduction in wall clock computer time. In a commercial application (NAV genomic evaluation) with ~ 24.000 typed bulls, the execution time for a single trait analysis was reduced from ~60 hours to ~45 minutes).

Asymmetric MME

As shown by Henderson (1984), the MME can be rearranged into an un-symmetric system by multiplying the random (genomic) part with **G.**

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ GZ'R^{-1}X & GZ'R^{-1}Z + I\otimes G_0^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ GZ'R^{-1}y \end{bmatrix}$$

This formulation do not involve $\mathbf{G}^{-1}$, therefore **G** do not need to be positive definite. Due to the multiplication by **G**, the "Genomic" part of the un-symmetric MME will typically be the major part of the system.

Rearranging the asymmetric MME as

$$\begin{bmatrix} X'R^{-1}X & 0 \\ 0 & GZ'R^{-1}Z + I\otimes G_0^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}(y - Z\hat{u}) \\ GZ'R^{-1}(y - X\hat{\beta}) \end{bmatrix}$$

leads to a double iterative solving algorithm, where each global iteration consists solving for the non-genomic effects followed by solving for the genomic effects.

Tests of the asymmetric solvers have been performed on the G-BLUP model used by NAV for Nordic Red Cattle. The data consists of deregressed protein proofs for 3662 bulls and the genomic relationship matrix (G) included 5287 typed animals. Solving the un-symmetric MME required considerable more iterations than solving the symmetric MME (Table 1). The solutions from symmetric and asymmetric MME were identical and total computing time was reduced due to avoiding computing G-inverse.

Genome-Wide Association Studies

Linear mixed models (LMM) are the method of choice for genetic association studies in human and other organisms due to the advantage of control of false positive (FP) associations due to population structure, family relatedness and/or cryptic relatedness (Yang et. al (2014)). However, LMM for genome-wide association studies (GWAS) including large sample sizes is computationally exhaustive as millions of genetic markers are analyzed individually. DMUAI is efficient to run LMM for GWAS studies where the candidate SNP is fitted as fixed regression and the random polygenic effect through the relationship matrix. The relationships among study individuals can be based on either pedigree records or on genome-wide markers. A number of haplotype-based models can also be analyzed, for example, random haplotype model to avoid FPs due to confounding of haplotypes within families (Boleckova et. al (2012)) and genealogy-based haplotype grouping.

Table 1. Number of iteration needed to obtain converges for symmetric and un-symmetric MME.

IOD solver (DMU5)	# of iterations
Symmetric MME	60
Un-symmetric MME	
Global	173
Non-genomic part	173
Genomic part	570

Survival Models

A flexible class of multivariate mixed survival models for continuous and discrete time with a complex covariance structure is implemented in DMU. The framework allows properly to handle right censoring, truncation, late entry and time-dependent explanatory variables. It is possible to combine models based on continuous time with models based on discrete time, and even generalized linear mixed models, all in a joint analysis. This allows to properly treat competing risks (Maia et. al 2014 and 2014b). The continuous time models implemented are approximations of the frailty model in which the baseline hazard function is piece-wise constant. The discrete time models used are multivariate variants of the discrete relative risk models. The survival models implemented include a dispersion parameter, which is essential for obtaining a decomposition of the variance of the trait of interest as a sum of parcels representing the additive genetic effects, environmental effects and unspecified sources of variability; as required in quantitative genetic applications.

Double Hierarchical Generalized Linear Mixed Model

The Generalized Linear Mixed Models (GLMM) facilities in DMU can be used for analyzing double hierarchical linear models (DHGLM) as proposed by Lee & Nelder (2006). Additionally, it is also possible to fit models that include genetic effects on both the mean and the residual variance (e.g. Rönnegård et. al (2010)). This can be utilized in analyzing effects of canalization. Initial results have shown good performance of these methods. Unbiased estimates of genetic influence on residual variance has been obtained in simulation studies.

R Wrapper for Special Applications

General R-interface

Specifying complex models in DMU can be a challenge for the less experienced user. Therefore a general R interface (Rdmu) has been developed. This allows specifying models in the usual R-format and the interface will automatically set up correct models, run the analysis and retrieve all results in R objects.

GWAS

A setup for GWAS named GENMIX (Genealogy Based Mixed Model) has been developed by Sahana et. al (2011). It is based on the general DMU R-interface (Rdmu) and the Blossoc software (Mailund et. al (2006)).

Survival Models

A R-library (survDMU) was developed for assisting in making analysis involving the survival models described above.

Hierarchical Models (Mean and Variance)

A R-setup for running DHGLM is under development.

AVAILABILITY

The DMU-packaged is distributed as executable files for Linux, Windows and Mac (http://dmu.agrsci.dk). There are no charges for using DMU for research purposes, but DMU should be referred in publications by citing the "DMU Users Guide". The terms of conditions for commercial use (e.g. routine genetic evaluation) can be obtained from Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University (Per.Madsen@agrsci.dk). The general R-interface can be obtained from OleF.Christensen@agrsci.dk and the survDMU R-library can be obtained from Rodrigo.Labouriau@agrsci.dk. For information on GENMIX contact Goutam.Sahana@agrsci.dk

LITERATURE CITED

Anderson, E., Bai, Z., Bischof, C. et. al, (1999). LAPACK Users' Guide, 3'rd edition, ISBN=0-89871-447-8.

Boleckova, J., Christensen, O.F., Sørensen, P. et al. (2012). Czech J. Anim. Sci., 57: 1–9

Christensen, O.F. and Lund, M. S. (2010). Genet. Sel. Evol. 42:2.

Henderson, C.R. (1984). Applications of Linear Models in Animal Breeding. University of Guelph Press, Guelph, Canada.

Lee Y., Nelder J.A. (2006). Appl Stat 2006, 55:139-185.

Madsen, P., and Jensen, J. (2013). DMU Ver. 6, rel. 5.2 http://dmu.agrsci.dk/DMU/Doc/Current/dmuv6_guide.5.2.pdf

Madsen, P., Su, G., Labouriau, R. and Christensen, O.F. (2010). Proc. 9'th WCGALP, Proceedings CD - paper 732

Maia, R.P., Madsen, P and Labouriau, R. (2014). Journal of Applied Statistics. http://dx.doi.org/10.1080/02664763.2013.868416.

Maia, R.P., Ask, B, Madsen, P, Pedersen, J and Labouriau, R. (2014b). J. Dairy Sci. 97: 1753-1761.

Mailund T. Besenbacher S., Schierup M.H. (2006) Bmc Bioinformatics 7: 454.

Rönnegård, L., Felleki, M., Fikse, F. et. al. (2010). Genet. Sel. Evol. 42:8

Sahana, G., Mailund, T., Lund, M.S. et al. (2011). PLoS ONE 6(11): e27061

VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

Yang, J., Zaitlen, N.A., Goddard, M.E. et al. (2014). Nature Genet. 46: 100-106