Wrestling with a WOMBAT: Selected new features for linear mixed model analyses in the genomic age

Karin Meyer

Animal Genetics and Breeding Unit,University of New England, Armidale, NSW 2351, Australia

Summary

The software package WOMBAT has undergone considerable changes since its release in 2006. We describe its adaptation to multi-threaded processing and outline selected capabilities that have been added recently. Firstly, expanded features for multivariate analyses considering more than a few traits are available. These offer the facility for restricted maximum likelihood estimation subject to a penalty on the likelihood to reduce sampling errors. Furthermore options for analyses of overlapping subsets of traits and subsequent 'pooling' of estimates based on likelihood principles are provided. Secondly, special modules have been added for the iterative solution of larger sets of mixed model equations, incorporating genomic information in a single-step analysis. These include options for analyses fitting the standard breeding value model or the so-called hybrid model. In addition, auxiliary calculations to obtain the genomic or joint relationship matrix (or inverse thereof) for animals with and without genotypes are available.

Keywords: Software, WOMBAT*, penalised REML, pooling covariances, single step genomic BLUP*

Introduction

WOMBAT is a freely available software package for linear mixed model analysis in quantitative genetics with focus on estimation of covariance components and genetic parameters via restricted maximum likelihood (REML), aimed primarily at problems common to animal breeding applications. It was first presented at the 8th WCGALP in Belo Horizonte (Meyer, 2006) and, to date, has been cited in well over 750 scientific publications. Since then, there have been continuous improvements and additions to it's capabilities, in response to scientific developments, changes in computing facilities and requests of colleagues. This paper highlights selected changes and features added in recent years.

Multi-threaded processing

Computer hardware has undergone dramatic changes in the last decade: whilst the increase in processing speed for individual CPUs has slowed, multiple cores and processors, co-processors (e.g. GPUs), extensive in-core memory and large amounts of disk space are readily available today. This has altered programming paradigms so that parallel execution to minimize elapsed ('wall') time is often more important than – previously prized – conservative memory requirements or fast, sequential execution. This has been accompanied by an increasing need for analyses involving mixed model equations (MME) comprising some large, dense blocks, as typically arising for genomic data. Fortunately, 'off-the-shelf' linear algebra routines are readily exploited for some of the resulting 'dense' calculations required and, moreover, are available in highly optimized forms for multi-threaded execution.

WOMBAT employs both BLAS (Blackford *et al.*, 2002) and LAPACK (Anderson *et al.*,

1999) routines for calculations involving dense matrix blocks. By default, the coefficient matrix in the MME (**C**) is considered sparse and stored as such. However, identifying dense sub-matrices arising in sparse matrix factorization or inversion for REML analyses allows these routines to be exploited in what is generally referred to as 'super-nodal' approach (see Masuda *et al.*, 2014, for an outline). In addition, WOMBAT has a 'dense' mode (invoked via run option --dense) in which **C** is stored in full. This eliminates the overhead from addressing operations in sparse storage and is thus advantageous when fitting covariance structures with many non-zero coefficients, such as a genomic relationship matrix.

Parallel execution for Linux executables is achieved by loading BLAS and LAPACK routines from the multi-threaded version of the Intel Math Kernel library (MKL) together with use of OpenMP instructions for selected operations. In addition, sparse BLAS routines (from MKL) are used for large, sparse problems. The number of threads utilized can be regulated through appropriate environment variables.

Penalized estimation of covariance components

Modern computing has made multivariate analyses to estimate covariance components involving quite a number of traits simultaneously technically feasible, and allows appropriately large data sets to be considered. However, the number of parameters to be estimated increases quadratically with the number of traits, so that estimates are afflicted by substantial sampling variation. REML estimation subject to a penalty on the likelihood function can 'improve' estimates by reducing sampling variances markedly at the expense of little additional bias. This yields estimates that are, on average, closer to their population values than unpenalized values (Meyer, 2011, 2016a; Meyer & Kirkpatrick, 2010). Suitable penalties can be derived as proportional to (minus) the logarithm of the probability density of assumed prior distributions for selected functions of the parameters. For instance, canonical eigenvalues may be shrunk towards their mean or genetic correlations be shrunk towards their phenotypic counterparts.

WOMBAT provides facilities for penalized REML estimation for multivariate analyses, offering a choice of penalties on (canonical) eigenvalues, correlations or covariance matrices and of stringency of penalization. Further options to assist with cross-validation to determine the optimal strength of penalization are available. Relevant options are acquired from the parameter file. However, whilst conceptually appealing, attempts to determine the optimal strength of penalties via cross-validation tend to complicate analyses and can be laborious. More recently, a 'mild' default penalty based on the assumption of a Beta distribution for canonical eigenvalues or partial correlations has been shown to be simple and effective, achieving a substantial proportion of benefits possible without the need for additional analyses (Meyer, 2016a). This is recommended for routine analyses involving five or more traits and can be selected by adding a simple, single line to the parameter file for WOMBAT.

Pooling estimates from analyses of parts

In spite of technical advances for multivariate analyses, there are scenarios where the number of traits of interest is too large for a joint analysis or computational demands are excessive. A standard approach in this case is to carry out analyses of small, overlapping subsets of traits. In the absence of potential selection bias, this may be as simple as performing all pairs of bivariate analyses, $q(q-1)/2$ for q traits. A post-estimation step is then needed to 'pool' individual results into valid, 'complete' covariance matrices. Often, this is done using simple strategies, such as averaging of covariance components, truncating or regressing the eigenvalues of the combined

matrix towards their mean to ensure matrices are positive definite. Alternative, likelihood based approaches have been described (Mäntysaari, 1999; Thompson *et al.*, 2005) but are rarely applied. These may yield combined matrices that are on average closer to the corresponding population values, especially when matrices for all sources of variation are considered simultaneously (Meyer, 2013). Thompson *et al.* (2005) outlined how to pool partial results using standard software for variance component estimation by transforming results from individual part-analyses into pseudo-observations.

Options are available in WOMBAT to make analyses by parts straightforward. To begin with, the run option --subset facilitates generating all the parameter files needed for subset analyses from a single file, set up for all traits. Similarly, only the overall data and pedigree files are required, and subset analyses write out additional files, suitable as input for pooling subsequent. Methods implemented for pooling are 'iterative summation of part matrices' (Mäntysaari, 1999) (run option --itsum) and newer, likelihood-based procedures, invoked by --pool. For the latter, additional options allow selection of mildly penalized estimation or assumptions for a pseudo pedigree structure, as proposed by Meyer (2013) to 'improve' estimates, similar to multivariate analyses of all traits.

Single-step genomic evaluation

Whilst WOMBAT is primarily a program for REML analyses, a number of modules have been added for iterative solutions of linear mixed models. These modules are meant as research tools (not to compete in efficiency or capacity with commercial or custom software), using mostly in-core storage of the MME. However, depending on computer resources available, they can readily accommodate moderately large problems with tens of millions equations in the model.

Breeding value model

Iterative solutions for 'standard' breeding value models have been available for a number of years and are specified using run option --solvit. This holds the complete MME in core, using compressed sparse row (CSR) format for the (typically) sparse coefficient matrix. It provides the choice of Gauss-Seidel and pre-conditioned conjugate gradient (PCG) algorithms, with diagonal, block-diagonal or successive over-relaxation (SSOR; Meyer, 2016b) pre-conditioners for the latter. Further, it allows fitting of random effects with user-defined general covariance matrix and provides the option to parameterize to genetic principal components.

Recently, there has been increasing interest in genetic evaluation using the so-called single step genomic BLUP (SS-GBLUP), with involves an inverse relationship matrix ($\mathbf{H}^{-1}$) which, with appropriate ordering, is comprised of a dense block for genotyped animals and sparse blocks otherwise; see, for instance, Legarra *et al.* (2014). This structure is exploited in a modified module, invoked with --s1step. As for --solvit, it allows for input of the matrix $\mathbf{H}^{-1}$ element-wise (non-zero elements of the half-stored matrix only), but the parts of $\mathbf{C}$ for genotyped animals are stored in full, for all traits. While this can result in substantial memory requirements, it facilitates effective use of BLAS routines as well as the SSOR preconditioner.

An alternative implementation is selected using run option --s2step. This differs by using 'iteration on data' to calculate the bulk of the product of $\mathbf{C}$ and a vector $\mathbf{r}$, as required in each PCG iterate. Only a diagonal pre-conditioning scheme is implemented. Moreover, for $\mathbf{H}^{-1} = \mathbf{A}^{-1} + \mathbf{\Delta}$, the inverse of the numerator relationship matrix (NRM), $\mathbf{A}^{-1}$, is set up directly from pedigree information and only the 'add-on' part for genotyped individuals, $\mathbf{\Delta}$, is to be read from file, with input in binary format. The 'add-on' is again stored in full but with storage requirements

proportional to the number of genotyped animals only.

Existence of multiple modules for essentially the same task is a reflection of how additional capabilities were developed over time and some consolidation may take place in future.

'Hybrid model'

Recently, the so-called hybrid model (HM) has been suggested as an equivalent model for SS-GBLUP (Fernando *et al.*, 2014). In brief, this fits breeding values for non-genotyped animals and marker effects for genotyped individuals. The HM does not require calculation of the genomic relationship matrix or the corresponding part of the NRM nor their inverses. Thus it becomes increasingly appealing as the number of genotypes grows, especially if a moderately sized subset of markers suffices to model genetic effects. However, explicit imputation of genotypes for non-genotyped animals is required instead. Fernando *et al.* (2016a) describe efficient computing strategies to fit such models and emphasize their scope for parallel processing.

WOMBAT provides a simple implementation for iterative solution of the MME for the HM, using a PCG algorithm with diagonal preconditioner. This includes a set-up step to impute missing genotypes. As above, sparse parts of **C** are held in core, but storage of potentially large, dense submatrices is mostly avoided by calculating the product **Cr** in multiple steps; details are given in Meyer (2017). This module is selected by run option `--s3step`, again with additional parameter file options available, e.g. to select the imputation scheme for non-genotyped animals or the degree of in-core storage to apply.

Genomic relationship matrices and $\mathbf{H}^{-1}$

To assist with calculations required to obtain genomic relationship matrices (**G**), their inverses or $\mathbf{H}^{-1}$, WOMBAT provides a separate module invoked with run option `--hinv`. It has a number of options (set in the parameter file) to choose which calculations to carry out and which results to write out and their format. These include the choice of method to calculate **G**, a weighted average between genomic and pedigree based components and inclusion of 'meta-founders' (Legarra *et al.*, 2015). Currently additional options are being implemented to approximate the inverse of **G**, either in 'APY' form (Misztal *et al.*, 2014) or using the partial Cholesky or RQ decompositions proposed by Hancock (2017) and Fernando *et al.* (2016b), respectively.

Availability

Executables for WOMBAT together with the user manual and worked examples are available for free download via **http://didgeridoo.une.edu.au/km/wombat.php**.

Conclusions

Our software package WOMBAT has been available to the scientific community for over a decade. A number of features added in recent years that are less well known are described to encourage their uptake and to inspire 'better' estimates for multivariate analyses through penalized estimation or appropriate pooling of results from analyses by parts. Special features for genomic data are provided to increase the scope of the package.

Acknowledgements

AGBU is a joint venture of NSW Department of Primary Industries and the University of New England. Work was supported by Meat and Livestock Australia grant L.GEN.1704.

List of References

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney & D. Sorensen, 1999. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, Third edition.

Blackford, L., J. Demmel, J. Dongarra, I. Duff, S. Hammarling, G. Henry, M. Heroux, L. Kaufman, A. Limsdaine, A. Petitet, R. Pozo, K. Remington & R. C. Whaley, 2002. An updated set of Basic Linear Algebra Subprograms (BLAS). ACM Trans. Math. Softw. 28(2):135–151.

Fernando, R. L., J. C. M. Dekkers & D. J. Garrick, 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50.

Fernando, R. L., H. Cheng, B. L. Golden & D. J. Garrick, 2016a. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genet. Sel. Evol. 48(1):96.

Fernando, R. L., H. Cheng & D. J. Garrick, 2016b. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. Genet. Sel. Evol. 48:80.

Hancock, T., 2017. Approximate GBLUP for efficient routine evaluations. Proc. Ass. Advan. Anim. Breed. Genet. 22: Paper no. 21.

Legarra, A., O. F. Christensen, I. Aguilar & I. Misztal, 2014. Single step, a general approach for genomic selection. Livest. Sci. 166:54–65.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar & I. Misztal, 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics 200(2):455–468.

Mäntysaari, E. A., 1999. Derivation of multiple trait reduced random regression (RR) model for the first lactation test day records of milk, protein and fat. In: *Proceedings of the 50th Annual Meeting of the European Association of Animal Production*. Europ. Ass. Anim. Prod.

Masuda, Y., T. Baba & M. Suzuki, 2014. Application of supernodal sparse factorization and inversion to the estimation of (co)variance components by residual maximum likelihood. J. Anim. Breed. Genet. 131(3):227–236.

Meyer, K., 2006. "WOMBAT" – digging deep for quantitative genetic analyses using restricted maximum likelihood. CD-ROM Eighth World Congr. Genet. Appl. Livest. Prod. Communication No. 27–14.

Meyer, K., 2011. Performance of penalized maximum likelihood in estimation of genetic covariances matrices. Genet. Sel. Evol. 43:39.

Meyer, K., 2013. A penalized likelihood approach to pooling estimates of covariance components from analyses by parts. J. Anim. Breed. Genet. 130(4):270–285.

Meyer, K., 2016a. Simple penalties on maximum likelihood estimates of genetic parameters to reduce sampling variation. Genetics 203(4):1885–1900.

Meyer, K., 2016b. Technical note: A successive over-relaxation pre-conditioner to solve mixed model equations for genetic evaluation. J. Anim. Sci. 94(11):4530–4535.

Meyer, K., 2017. A look at computations for multivariate single-step genomic evaluation fitting the 'hybrid model'. Proc. Ass. Advan. Anim. Breed. Genet. 22: Paper no. 98.

Meyer, K. & M. Kirkpatrick, 2010. Better estimates of genetic covariance matrices by 'bending' using penalized maximum likelihood. Genetics 185(3):1097–1110.

Misztal, I., A. Legarra & I. Aguilar, 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97(6):3943–3952.

Thompson, R., S. Brotherstone & I. M. S. White, 2005. Estimation of quantitative genetic parameters. Phil. Trans. Roy. Soc. B 360:1469–1477.