AN ALTERNATIVE FOR MIXED MODEL ANALYSES OF LARGE, MESSY DATA SETS (MTDFREML)

L. D. Van Vleck

R. K. Splan

AN ALTERNATIVE FOR MIXED MODEL ANALYSES OF LARGE, MESSY DATA SETS (MTDFREML)

L. D. Van Vleck
USDA, ARS, R. L. Hruska U.S. Meat Animal Research Center, Lincoln, NE

R. K. Splan
University of Nebraska, Lincoln

ABSTRACT

Portable Fortran based programs (MTDFREML) were developed using a derivative-free algorithm to obtain REML estimates of (co)variance components. Computations are based on Henderson's mixed model equations for multiple-trait models with missing observations on some traits and incorporation of relationships among relatives. Many fixed and random factors are allowed with number of levels dependent on computer memory. Data sets with more than 40,000 genetic effects have been analyzed. Options allow solving MME at convergence. Constraints are automatically imposed. Expectations, standard errors of contrasts of solutions for fixed effects and prediction error variances of solutions for random effects can be obtained. Dimensions can be changed to match data with computer capability. A Fortran compiler is necessary. No fee is charged but the University of Waterloo must certify a license has been obtained for sparse matrix subroutines (SPARSPAK) used in the program. As an example, birth weights of 4891 progeny of 389 sires nested within 12 breeds and of 2893 dams nested within 3 breeds of dam were analyzed to estimate components of variance due to sires and dams and to estimate differences among breeds of sires. For MTDFREML the analysis was trivial but for PROC MIXED the analysis was impossible unless dams were dropped from the model.

Key words: Variance component estimation, Mixed model equations, Sparse matrix methods

1. Introduction

Many applications for mixed model analyses have too much data or too many levels of fixed or random factors to use commercial statistical packages such as PROC MIXED of SAS (1996). Animal breeders typically need to estimate variance components, predict genetic values, and estimate contrasts of fixed effects from such data. Until recently a specific computer program usually was written for each analysis. With increases in computing power and memory several quite general packages have been developed by animal breeders. The purpose of this paper is to describe a little of the history as well as the capabilities of one such package used by animal breeders -- MTDFREML (multiple trait derivative-free restricted maximum likelihood). A more complete history of development of the program is given in proceedings of the conference in honor of Shayle R. Searle (Van Vleck, 1996). The computational aspects are given in detail in the manual that accompanies the program statements (Boldman, et al., 1995).

2. Motivation

An example of a relatively common problem for animal breeders with outcomes when attempted with MTDFREML and PROC MIXED of SAS (1996) will be described.

Genetic evaluations for bulls for weights of progeny are done separately for each breed of beef cattle at one of four universities in the United States. These evaluations provide best evaluations for comparison of bulls within a breed. Many breeders, however, want to compare bulls of different breeds. The USDA Meat Animal Research Center (USMARC) has characterized many breeds in their Germ Plasm Evaluation project. Their data are not used in the national cattle evaluations, but some bulls used at USMARC have been used elsewhere and thus have within-breed evaluations (commonly called EPDs, expected progeny differences). Notter and Cundiff (1991) developed a system for adjusting the within-breed evaluations so that bulls can be compared across breeds. The basis of the across-breed adjustment is the direct comparison of records of progeny of bulls used at USMARC, i.e., breed of sire differences. Table 1 shows the numbers of records that are available for the 12 breeds compared, and the number of sires of each breed for three traits. Table 2 further illustrates the data structure including the number of breeds of cows mated to the bulls and the number of dams of each dam breed for the birth weight trait. In addition, the model includes classification factors for sex of calf and age of dam and a linear covariate for calendar day of birth.

Unquestionably, these are "messy" data although not a large data set by animal breeding standards. A mixed model with two random classification effects is indicated: sires within-breed of sire and dams within-breed of dam. The problem then is to jointly estimate components of variance due to sire effects, dam effects and residual effects and also breed of sire effects. The other effects can also be estimated but are only "noise" for the goals of the analysis.

What to do? Although PROC MIXED of SAS (1996) was not designed for a problem such as this, several options were tried with indicated outcomes: 1) both sires and dams included as random classification effects–failed to run, 2) sire as the only random classification effect and dams ignored–converged quickly with estimates of sire and residual components of variance and also breed of sire contrasts, 3) as a partial test of capacity, dams were included as the only random classification effect and sires ignored–failed to run. Specifying the model as sire within-breed worked but as expected the model for dam within-breed did not run.

The MTDFREML program was expected to run with all such models and did run. For the sire model, the solutions for variance components and estimable breed of sire contrasts were the same with PROC MIXED and with MTDFREML. Not unexpectedly, with MTDFREML breed of sire differences were not greatly different for models ignoring random effects of dams or ignoring both sire and dam effects. The sampling errors of the breed of sire contrasts are, however, markedly underestimated if sires are ignored and slightly underestimated if sires are considered random effects and dams are ignored (Barkhouse, et al., 1998).

3. Derivative-free REML

A survey of the audience revealed few but those with training in animal breeding had heard of derivative-free REML (DFREML). The idea of REML appeared in Patterson and Thompson (1971). The history of DFREML for animal breeders dates to1986 when Smith and Graser (1986) published a paper in the Journal of Dairy Science. A year later a more complete exposition was published by Graser, Smith and Tier (1987) in the Journal of Animal Science. Thus, most animal breeders could have been exposed to the idea. Smith had been a student in a course with Shayle Searle that used Searle's red bound notes (1979) that were a precursor to the variance components book (Searle, et al., 1992). Smith also was in a computer science course that covered computing the log determinant of the coefficient matrix of least squares equation and the residual sums of squares by Gaussian elimination (e.g., Stewart, 1994). The estimation on variance component notes included results of Harville (1977) that can be used to write the likelihood in terms of the mixed model equations. These two ideas are critical to the development of derivative-free algorithms for obtaining REML solutions for variance components. Karin Meyer (1988, 1989) popularized the method by distributing Fortran statements for a general mixed model program utilizing sparse matrix techniques with Gaussian elimination. She also demonstrated how to extend the method to multiple-trait analyses (Meyer, 1991).

4. The connection with mixed model equations

To show in simple terms what is required to calculate the likelihood given the data to consider the usual general linear mixed model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where

$\mathbf{y}$ is the vector of observations,
β is the vector of fixed effects with $\mathbf{X}$ the matrix associating fixed effects with observations,
$\mathbf{u}$ is the vector of random effects with $\mathbf{Z}$ the matrix associating random effects with observations and
$\mathbf{e}$ is a vector of residual effects. The general first and second moments are:

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{X}\beta\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \text{ and}$$

$$V\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{ZGZ'} + \mathbf{R} & \mathbf{ZG} & \mathbf{R}\\ \mathbf{GZ'} & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix}$$

Henderson's mixed model equations which are well-known to be a rather simple modification of least squares equations can be written for the general case as:

$$\begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \mathbf{G^{-1}} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{pmatrix}.$$

The single trait ($R = I\sigma^2$) version of these was invented by Henderson about 1948 (Henderson, preface, 1984). The proofs that the $\hat{\beta}$ are BLUE and $\hat{\mathbf{u}}$ are BLUP came later (Henderson et al., 1959 and Henderson, 1963). Properties are listed in Henderson (1975).

Let the MME be $\mathbf{Cs} = \mathbf{r}$. The key features of these equations for large sets of messy data are the large number of equations and the sparsity of the coefficient matrix, **C**. For the example with BWT, the number of equations was 3320 and the fraction of non-zero cells of **C** was only 1.16%. Making use of this sparsity is the key to the computational efficiency of DFREML.

What must be computed to calculate the log likelihood, log L or rather -2log L? The results in Harville (1977) and Searle (1979) can be used to show (Meyer, 1991) that

$$-2\log L = \text{constant} + \log|\mathbf{R}| + \log|\mathbf{G}| + \log|\mathbf{C}^*| + \mathbf{y'Py}$$

which are all terms used in Henderson's MME. The $\mathbf{C}^*$ is a full-rank subset of **C** and $\mathbf{y'Py}$ reduces to a generalized residual sum of squares.

For a simple model for one trait that animal breeders call an animal model with **u**, the vector of genetic values:

$$\mathbf{R} = \mathbf{I}\,\sigma_e^2 \text{ and } \mathbf{G} = \mathbf{A}\sigma_g^2$$

where

A is the matrix of numerator relationships among the animals with genetic values in **u**. Then using rules from Searle (1982):

$$\log|\mathbf{R}| = N \log(\sigma_e^2) \text{ and } \log|\mathbf{G}| = \log|\mathbf{A}| + q \log(\sigma_g^2)$$

where

N = the number of observations for the trait,
q = the number of animals with genetic values in u (which can include ancestors in A that do not have records but that is another story of Henderson, 1976a,b; Quaas, 1976),
σ_e^2 = the residual variance and
σ_g^2 = the variance of additive genetic values.

5. The connection with the sparspak subroutines and the simplex algorithm

The idea of the derivative-free method is to search for estimates of $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ that will maximize log L or equivalently minimize the more convenient form, -2log L. A generally efficient way to do this is with the Simplex algorithm (Nelder and Mead, 1965) which is described as the amoeba algorithm in Press et al. (1989). The simplex algorithm finds the minimum for a non-linear function; hence the use of -2log L. A detailed description is given in the manual for MTDFREML (Boldman et al., 1993). This algorithm can be thought of as the second of two important *black* boxes (something in and a good result out without having to understand much of what is happening).

The simplex algorithm directs the search by controlling step sizes and directions of changes in the variance and covariance components that make up **R** and **G**. At each up-date of the simplex, the four non-constant parts of -2log L are computed. The extension of $\log|\mathbf{R}|$ and $\log|\mathbf{G}|$ to more complex models is not difficult so these two parts remain easy to compute (except for forcing the (co)variance components to remain in the allowable parameter space). The other two parts are computationally intensive. The key to efficient calculations is to take advantage of the sparseness of **C***. The "black" box that was used in MTDFREML was a set of sparse matrix subroutines (SPARSPAK) developed at the University of Waterloo, Canada (George et al., 1980). SPARSPAK is copyrighted and requires a license to use. That package is built around sparse Choleski factorization. An initial read of half-stored non-zero coefficients of **C** is followed by a one-time reordering of the equations to minimize steps in sparse matrix factorization of **C** in subsequent rounds. Then in each round, triangular Choleski factor, **L**, is calculated such that $\mathbf{LL'} = \mathbf{C}$. The log determinant of **L** is $\sum^{NE} \log(\ell_{ii})$ where ℓ_{ii} is the i^{th} diagonal of L and NE is the number of equations. Then $\log|\mathbf{C}| = 2\sum^{NE} \log(\ell_{ii})$. At the end of the factorization for each round, **s** is calculated with a sparse Choleski down and up solve. The calculation of $\mathbf{y'Py}$ is $\sum^{N} \mathbf{y}_i'\mathbf{R}_i^{-1}\mathbf{y}_i - \mathbf{s}'\mathbf{r}$ where $\mathbf{y}_i$ is the vector of traits for the i^{th} animal and $\mathbf{R}_i$ is the variance-covariance matrix of residuals for the number of traits measured on animal i (see manual for more detail). The uncorrected total sum of squares would be calculated during the read of the data using the current update of components of **R**. The time consuming part is to calculate **L** from up-dated **C**. After **L** is calculated, then $\log|\mathbf{L}|$ and $\mathbf{y'Py}$ require little time to compute. What can be noted is that these computations are general for multiple-trait models with only $\Sigma\mathbf{y}_i'\mathbf{R}_i^{-1}\mathbf{y}_i$ slightly complicated to compute.

The simplex algorithm then determines up-dates of **R** and **G** until convergence. Generally convergence is declared when the variance of the -2log L's in the simplex is small, e.g., .00001. The simplex contains as many -2log L's as the number of variances and covariances plus one. The simplex algorithm is not guaranteed to converge to a global minimum (maximum of log L) so restarts are necessary to spread the search to determine if convergence is to the global minimum. Usually single-trait models without a covariance term such as a genetic covariance between direct and maternal effects converge with the first start. Multiple-trait analyses may

require many restarts. An ad hoc check for global convergence is whether -2log L changes in front of the second decimal and whether the proportions of total variance change from one restart to the next.

6. Advantages and disadvantages

This list will be sketchy and in some cases what is listed as an advantage may in some cases also be a disadvantage. Several advantages that could be listed under the heading of magic will show that. No quadratics other than the total sum of squares are computed. That is a computational advantage but knowing what quadratics are computed sometimes helps in understanding estimates of variance components. As stated in the name, not even a set of first derivatives is needed. Consequently, expectations with traces involving inverse elements of $\mathbf{C}^*$ are not needed as with the EM algorithm. Similarly, matrices of second derivatives as needed for Newton-Raphson (N-R) type algorithms need not be computed. For non-mathematical researchers those are major hurdles and with DF are not needed. The disadvantage is that for algorithms where the information matrix or the expected value of the information matrix can be computed, convergence is usually much more rapid than with DF algorithms.

One problem with most REML algorithms is that solutions try to escape the parameter space, i.e., variances go negative or some eigenvalues of $\mathbf{R}_0$ or $\mathbf{G}_0$ (the variance-covariance matrices for an animal with more than one trait) become negative. Rules for what to do are not standard. The EM algorithm often creeps to the boundary. The N-R based algorithms usually require the estimate to be set to zero which complicates re-entry of a component. The DF algorithm assigns a -2log L that is much too large which forces contractions in the simplex algorithm until the estimates are in the parameter space. Each of these contractions does not require much time but usually contractions indicate poorly behaved data, i.e., many rounds to reach convergence.

Some advantages of DFREML involve equivalent models and multiple-trait models. For example in multiple-trait models with repeated measures for one trait and not for another trait, the environmental covariance between traits can be forced into what animal breeders call a covariance between permanent environmental (environmental effects being repeated in subsequent records) effects for the trait with repeated measures and also for the trait without repeated measures. Another example is for estimation of the correlation between effects of varieties expressed in several locations–the genotype by environment problem. The program can handle this as a multiple-trait analysis with the trait being defined as the location where the "trait" is measured. Still another example is in determining the correlation between the expression of a sire's genes for sex-limited traits, e.g., scrotal circumference in male progeny and age at puberty in female progeny. A simpler application would be to determine the genetic correlation between weaning weight in male and in female progeny.

Some computing considerations also can be considered advantages or disadvantages. The program is portable in the form of Fortran statements. But a compiler is then needed and Fortran seems not to be a priority for programmers anymore. The program is flexible in that vector sizes can be changed easily to match computer memory and models for the traits. Thus, large data sets can be handled with "large" depending on the number of factors and levels and computer memory. For obtaining solutions to MME the number of equations has exceeded 100,000. For variance component estimation, problems with 50,000 or more equations have been managed.

The program can handle multiple-trait problems with different models for different traits and with no requirement that all traits be measured on each animal. Of special interest to animal breeders is that models with two genetic effects (commonly a direct effect of the animal and a maternal effect from the dam of the animal) can be modeled with a covariance between the two effects. Most such models require the genetic relationships among the animals be used which somewhat increases the number of non-zero elements in the coefficient matrix.

The program statements are free and a detailed manual is available. But because the SPARSPAK subroutines (Chu et al., 1984) are integral they must be licensed for a relatively small annual fee for single computer or for somewhat larger institutional fee for an unlimited number of computers . Currently an agreement is being formalized with the University of Waterloo on how the distribution can be made.

One of the major advantages of the MTDFREML package relates to SPARSPAK. Ordinarily a Choleski factor can be obtained only for a full rank matrix. The **C** matrix except in very simple models is never full rank. To make such a matrix full rank would require imposing constraints before factorization. With messy data and many effects in the model, the necessary constraints are not always easy to determine. Steve Kachman (Boldman et al., 1993), however, modified the SPARSPAK code to impose necessary constraints during factorization to determine L. Those constraints are remembered for every up-date of **G** and **R** which insures -2log L does not fluctuate due to the constraints that are imposed. Thus, the code for MTDFREML needs to be distributed with SPARSPAK subroutines already modified and user-ready.

One complication of not knowing what constraints are imposed is in determining what appropriate estimable functions are or if even if what would seem to be an estimable function is really estimable. Solutions corresponding to the constrained to zero equations are also true zero in the solution vector but that is not always a certain way to determine estimability. Therefore, an option was added to determine expected values of solutions.

The logic for calculation of expectations is this. Because $\mathbf{s} = \mathbf{C}^{-}\mathbf{r}$ even if **s** is not obtained from a generalized inverse (in fact, **s** is obtained from a generalized **L**), then $E[\mathbf{s}] = \mathbf{C}^{-} E(\mathbf{r})$ which can be rewritten in terms of the MME and blocks of the generalized inverse of **C** as

$$\mathbf{E[s]} = \begin{pmatrix} \mathbf{C}^{xx} & \mathbf{C}^{xz} \\ \mathbf{C}^{zx} & \mathbf{C}^{zz+} \end{pmatrix} \begin{pmatrix} \mathbf{X'R}^{-1}\mathbf{X} \\ \mathbf{Z'R}^{-1}\mathbf{X} \end{pmatrix} \beta$$

Examination of this expression reveals that the coefficients of β for the i^{th} element of **s** are obtained by multiplying the i^{th} row of $\mathbf{C}^{-}$ by each of the columns of $\begin{pmatrix} \mathbf{X'R}^{-1}\mathbf{X} \\ \mathbf{Z'R}^{-1}\mathbf{X} \end{pmatrix}$.

The i^{th} row of $\mathbf{C}^{-}$, $\mathbf{c}_i$, is obtained quickly with a Choleski down- and up-solve of $\mathbf{Cc}_i = \mathbf{I}_i$ where $\mathbf{I}_i$ is the i^{th} column of **I**. The multiplication of $\mathbf{c}_i$ by the lefthand columns of **C** requires reading of the prepared data file and sparse matrix multiplication with $\mathbf{c}_i$, i.e., $\mathbf{E[s}_i\mathbf{]} = \mathbf{c}_i' \begin{pmatrix} \mathbf{X'R}^{-1}\mathbf{X} \\ \mathbf{Z'R}^{-1}\mathbf{X} \end{pmatrix} \beta$.

Each expectation is relatively time consuming but ordinarily only a few expectations are needed to find the pattern.

The same general idea is used to obtain variances of linear contrasts and is attributed to Harville (1974) who used the term "mixed model conjugate normal equations". Again for $\mathbf{Cs} = \mathbf{r}$, $\mathbf{s} = \mathbf{C}^{-}\mathbf{r}$ whether **s** is solved for directly or not. For a linear contrast $\mathbf{k's}$, $V(\mathbf{k's}) = \mathbf{k}'V(\mathbf{s})\,\mathbf{k} = \mathbf{k'C}^{-}\mathbf{k}$ where V(**s**) is the variance of prediction errors. Now equate $\mathbf{C}\phi = \mathbf{k}$ so that $\phi = \mathbf{C}^{-}\mathbf{k}$, whether solved for directly or not. In fact, do a Choleski down- and up-solve to obtain ϕ using **k** as the right hand side vector (usually mostly zeroes except for, for example, a 1 and a -1 for a linear contrast between two levels of a factor). From these identities $\mathbf{k}'\phi$ is the necessary calculation to obtain $\mathbf{k'C}^{-}\mathbf{k} = V(\mathbf{k's})$ which is the variance of prediction error for the linear contrast $\mathbf{k's}$. Thus, variances (and standard errors) of linear contrasts can be obtained quickly and for any linear function of the solution vector. Care must be taken to insure estimability as the computations do not require the contrast to be estimable.

Similarly, the variance of prediction error for **u** can be obtained. In that case, $\mathbf{k}_i$ is a vector of zeroes but with the element corresponding to the i^{th} element in u being set to unity: then $\mathbf{k}_i'\phi = V(\hat{u}_i - u_i)$.

Standard errors of variance components or of functions such as heritability are not yet available with MTDFREML for general multiple-trait models. For single-trait models or multiple-trait models with all traits measured on all animals, a visiting scientist, Dr. Joerg Dodenhoff, utilized the computing strategies involved with average information REML (AIREML) to obtain the average of the matrix of second derivatives and its expected value to obtain asymptotic standard errors (Dodenhoff et al., 1998). A version for cases with missing observations on some traits is needed.

Some other disadvantages of DFREML include sensitivity to starting values. If the starting values are on the order of a magnitude too large or small, the step size in the simplex algorithm may not be great enough to reach a set of parameter estimates close to the minimum of -2log L. One solution to the problem that usually works is to force starting values for variance components to sum to less than the unadjusted phenotypic variance.

In some cases rounding error seems to prevent reaching convergence. Extremely small changes in variance components then result in relatively large changes in -2log L or at least in V(-2log L) of those retained in the simplex. In most such cases the estimates and -2log L will not have changed in any important way in spite of a V(-2log L) that is larger than desired.

Starting values of zero for covariances will not be changed as the simplex makes proportional changes when updating. Starting values close to zero will change very slowly. A related problem is that starting with the wrong sign on a covariance will result in many rounds to change the sign.

A major disadvantage is for three or more traits in an analysis. With many covariance terms the number of possible combinations of estimates of variances and covariances is large. Restarts are essential as global convergence may require several restarts as well as many rounds to reach each local convergence (Boldman and Van Vleck, 1990; Groeneveld and Kovac, 1990). A rule of thumb is that time to convergence increases according to t^5 where t is the number of traits (Misztal, 1994). For the example, a single trait took 7 minutes on a relatively ancient 486/33, two traits took hours, and three traits took days.

7. Development credits

The development of MTDFREML was a joint effort largely financed by a USDA post doctoral associate position. The series of post-docs spent some of their time on various aspects of the program. The chief architect of the program and the person who made the program easy to use was Keith Boldman, the first post-doc. Lisa Kriese succeeded him and added several parts as did I. Curt Van Tassell, the final post-doc, cleaned up some minor bugs and made compilation much easier (Boldman, et al., 1995). As an aside, Curt also developed a similar program based on a Gibbs Sampling Bayesian algorithm (Van Tassell and Van Vleck, 1995). Steve Kachman, as already mentioned, made the important contribution of modification of the Choleski algorithm to work with singular coefficient matrices. Joerg Dodenhoff added steps to compute asymptotic standard errors for the variance components, a feature that was not a part of the original package.

8. Some similar packages

This paper has described some of the features of only one such package developed by animal breeders for computation of REML estimates. The MTDFREML program may be one of the easiest to use but is not necessarily as fast or as efficient as some others which are listed here.

Karin Meyer, who divides her time between Edinburgh, Scotland and Armidale, Australia developed the first general DFREML package and has continued to modify and extend that package. There now is an option to use AIREML which generally converges in many fewer rounds than DFREML but which requires elements of the inverse of the coefficient matrix which may limit the allowable number of equations. Eildert Groeneveld in Germany has a general purpose package for solving MME called PEST. He also has a version to estimate variance components called VCE. Jöst Jensen and Per Madsen in Denmark have the DMU program which utilizes AIREML. Arthur Gilmour in Australia has Arthur's REML, ASREML, which is based on AIREML. Joerg Dodenhoff, from Germany and currently at Iowa State University, while at the University of Nebraska working with me and Steve Kachman developed an AIREML package fashioned after MTDFREML for a general single-trait model with up to three genetic (direct, maternal, and grand maternal) effects (Dodenhoff et al., 1998). Many other similar packages have probably been written but this list shows that various options are now available even before SAS develops a package for large sets of messy data.

9. Summary

This paper has described a computational package for relatively large messy data sets based on a derivative-free algorithm for obtaining REML estimates of variances and covariances. Some history of DFREML and the underlying principles were described. A computing algorithm was outlined based on sparse matrix Choleski factorization of the non-full rank matrix of coefficients of Henderson's mixed model equations. Capabilities of the program include: estimation of (co)variance components, estimation of contrasts of fixed effects and their standard errors, calculation of prediction error variances of random effects, and calculation of expected values of solutions. The program was developed by USDA-ARS with collaboration from Steve Kachman at the University of Nebraska. Some similar packages developed in other countries were also listed.

References

Barkhouse, K. L., L. D. Van Vleck, and L. V. Cundiff. 1998. Effect of ignoring random sire and dam effects on estimates and standard errors of breed comparisons. J. Anim. Sci. 76:2279-2286.

Boldman, K. G. and L. D. Van Vleck. 1990. Effect of different starting values on parameter estimates by DF-REML and EM-REML in an animal model with maternal effects. J. Anim. Sci. 68(suppl. 2):71 (abstr.).

Boldman, K. G. and L. D. Van Vleck. 1991. Derivative-free restricted maximum likelihood estimation in animal models with a sparse matrix solver. J. Dairy Sci. 74:4337-4343.

Boldman, K. G., L. A. Kriese, L. D. Van Vleck, and S. D. Kachman. 1993. *A Manual for Use of MTDFREML*. A set of programs to obtain estimates of variances and covariances (DRAFT). USDA-ARS, Roman L. Hruska U.S. Meat Animal Research Center, Clay Center, NE.

Boldman, K. G., L. A. Kriese, L. D. Van Vleck, C. P. Van Tassell and S. D. Kachman. 1995. *A Manual for Use of MTDFREML*. A set of programs to obtain estimates of variances and covariances (DRAFT). Revised. USDA-ARS, Roman L. Hruska U.S. Meat Animal Research Center, Clay Center, NE.

Chu, E., A. George, J. Liu, and E. Ng. 1984. SPARSPAK: Waterloo sparse matrix package user's guide for SPARSPAK-A. CS-84-36, Dept. Computer Sci., Univ. Waterloo, Waterloo, ON, Canada.

Dodenhoff, J, L. D. Van Vleck, S. D. Kachman, and R. M. Koch. 1998. Parameter estimates for direct, maternal, and grand maternal genetic effects for birth weight and weaning weight in Hereford cattle. J. Anim. Sci. 76:2521-2527.

George, A., J. Liu, and E. Ng. 1980. User guide for SPARSPAK: Waterloo sparse linear equations package. CS-78-30, Dept. Computer Sci., Univ. Waterloo, ON, Canada.

Graser, H. -U., S. P. Smith, and B. Tier. 1987. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. J. Anim. Sci. 64:1362-1370.

Groeneveld, E. and M. Kovac. 1990. A note on multiple solutions in multivariate restricted maximum likelihood covariance component estimation. J. Dairy Sci. 73:2221-2229.

Harville, D. A. 1974. Some useful representations for constrained mixed-model estimation. J. Am. Stat. Assn. 74:200.

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Stat. Assoc. 72:320.

Henderson, C. R. 1963. Selection index and expected genetic advance. In: Statistical Genetics in Plant Breeding. NAS-NRC publication 982.

Henderson, C. R. 1975. Best linear unbiased prediction under a selection model. Biometrics 31:423-447.

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

Henderson, C. R. 1976b. Inverse of a matrix of relationships due to sires and maternal grandsires in an inbred population. J. Dairy Sci. 59:1585-1588.

Henderson, C. R. 1984. *Application of linear models in animal breeding.* U. Guelph, Guelph, ON, Canada.

Henderson, C. R., O. Kempthorne, S. R. Searle, and C. M. vonKrosigk. 1959. Estimation of environmental and genetic trends from records subject to culling. Biometrics 15:192-218.

Meyer, K. 1988. DFREML. A set of programs to estimate variance components under individual animal model. J. Dairy Sci. 71(suppl. 2):33.

Meyer, K. 1989. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. Genet. Sel. Evol. 21:317.

Meyer, K. 1991. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. Genet. Sel. Evol. 23:67-83.

Misztal, I. 1994. Comparison of software packages in animal breeding. Proc. 5th World Cong. Genetic Applied to Livest. Prod. 22:3-10.

Nelder, J. A. and R. Mead. 1965. A simplex method for function minimization. Computer J. 7:308.

Notter, D. R., and L. V. Cundiff. 1991. Across-breed expected progeny differences: Use of within-breed expected progeny differences to adjust breed evaluations for sire sampling and genetic trend. J. Anim. Sci. 69:4763-4776.

Patterson, H. D. and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrics 58:545.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1989. *Numerical recipes.* Cambridge University Press, New York.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics 32:949-953.

SAS Institute. 1996. SAS/STAT software: Changes and enhancements. SAS Institute, Cary, NC.

Searle, S. R. 1979. Notes on variance component estimation: A detailed account of maximum likelihood and kindred methodology. Paper BU-673-M, Biometrics Unit, Cornell Univ.

Searle, S. R. 1982. *Matrix Algebra Useful for Statistics.* John Wiley and Sons, Inc., New York.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance Components.* John Wiley and Sons, Inc., New York.

Smith, S. P. and H. -U. Graser. 1986. Estimating variance components in a class of mixed models by restricted maximum likelihood. J. Dairy Sci. 69:1156-1165.

Stewart, G. W. 1994. Gauss, statistics, and Gaussian elimination. Proc. 26th Symp. on the Interface. Computing Science and Statistics 26:1.

Van Tassell, C. P., and L. D. Van Vleck. 1995. *A Manual for Use of MTGSAM.* A set of Fortran programs to apply Gibbs Sampling to animal models for variance component estimation. USDA-ARS, Roman L. Hruska U.S. Meat Animal Research Center, Clay Center, NE.

Van Vleck, L. D. 1996. Development of a flexible, portable, efficient, free software program for estimation of (co)variance components for multiple models (MTDFREML). In: Proc., Conf. in honor of Shayle R. Searle, Cornell Univ., Aug. 9-10, 1996.

Table 1. Records on progeny of 12 sire breeds used to estimate breed of sire differences

Sire breed	No. sires	No. progeny BWT	WWT	YWT
A	67	856	826	762
B	68	676	619	576
C	25	181	170	168
D	28	422	358	312
E	28	422	368	332
F	20	387	338	334
G	63	583	506	468
H	15	174	155	154
I	24	365	336	334
J	16	435	415	347
K	7	199	191	189
L	27	189	176	173
Total	389	4891	4458	4149

Table 2. Other factors in model to estimate breed of sire differences

Dam breeds	Mated to sire breeds	Years	Number Dams	Progeny
A	Not A	15	1118	1875
B	Not B	15	1287	1413
X	A,B,D	3	488	603
3	12	15	2893	4891