Using transformation algorithms to estimate (co)variance components by REML in models with equal design matrices

K.V. Konstantinov* and G.J. Erasmus
Department of Animal Science, University of the Orange Free State, P.O. Box 339, Bloemfontein, 9300 Republic of South Africa

Received 21 April 1992; accepted 12 December 1992

The reduction of computational demands on Restricted Maximum Likelihood (REML) procedures by a diagonalization approach is extended to multiple traits by the use of canonical transformations. A computing strategy is developed for use on large data sets employing two different REML algorithms for the estimation of (co)variance components. Results from a simulation study indicate that (co)variance components can be estimated efficiently at a low cost on a wide range of computer systems.

Die vermindering van rekenaarbehoeftes vir Beperkte Maksimum Aanneemlikheid(REML)-prosedures deur 'n diagonalisasie benadering word uitgebrei na meervoudige eienskappe deur die gebruik van kanoniese transformasies. 'n Berekeningstrategie is ontwikkel vir gebruik op groot datastelle deur die aanwending van twee REML-algoritmes vir die bepaling van (ko)variansie-komponente. Resultate met 'n simulasiestudie dui aan dat (ko)variansie-komponente doeltreffend en teen lae koste op 'n wye reeks rekenaarstelsels beraam kan word.

Keywords: Canonical transformations, diagonalization, REML, variance components.

* To whom correspondence should be addressed.

Introduction

In practical animal breeding, multiple traits are usually measured on each individual to collect as much information as possible about its productivity. These traits are often correlated. In genetic studies, multivariate estimation of (co)-variance components and genetic values for sire selection has recently received considerable attention. Although statistically appealing, it is computationally challenging to apply multivariate analyses to large data sets. The difficulty with multivariate analyses is the increasing number of equations to be solved. However, there are situations where canonical transformations can be used to transform all correlated variables to uncorrelated canonical variables. The proviso is that all traits are to be recorded on all animals. Typical examples of such a situation may be milk production traits in dairy cattle and wool traits in sheep. Several authors have presented algorithms for Restricted Maximum Likelihood (REML) estimation of (co)variance components with equal design matrices (Meyer, 1985; Schaeffer, 1986; Jensen & Mao, 1988). There is, however, room for more research in this field in an effort to decrease computational demands.

The objectives of this study were: (i) to extend the diagonalization approach (Konstantinov & Erasmus, 1992) to multivariate analysis using canonical transformations, (ii) to develop a computing strategy suitable for large data sets, and (iii) to determine its properties using two different computer systems.

The Model

Consider **q** traits and let **Y** be **nxq** matrix of observations on **n** individuals each with records on **q** traits. Now, let the model for each trait be denoted as

$$y_i = X_i b_i + Z_i u_i + e_i \qquad (i = 1, ..., q) \qquad (1)$$

where b_i is a ***pxl*** vector of fixed effects, u_i is an ***sxl*** vector of random effects and e_i is an ***nxl*** vector of random residuals. The matrices X_i and Z_i are known with dimensions **nxp** and **nxs**, respectively, and relate observations in y_i to classes in b_i and u_i. If these matrices are equal for all traits, i.e. $X_i = X$ and $Z_i = Z$ for all **i**, then the multivariate model can be written as (Thompson, 1973):

$$y = (I_q * X)b + (I_q * Z)u + e \qquad (2)$$

where '*' denotes direct product operation (Searle, 1982), $b' = [b_1', b_2', ..., b_q']$, $u' = [u_1', u_2', ..., u_q']$, and $e' = [e_1', e_2', ..., e_q']$. The expectation and (co)variance matrices are: $E(y) = (I_q * X)b$, $E(u) = 0$, $E(e) = 0$, $Var(y) = G*ZAZ' + R*I_n$, $Var(u) = G*A$, and $Var(e) = R*I_n$, where **G** and **R** are (co)variance matrices of the **q** traits for the random and residual factors, respectively, and **A** is the numerator relationship matrix. Under these assumptions the matrix formulations of the mixed model equations (MME) of Henderson (1973) are:

$$\begin{bmatrix} R^{-1}*X'X & R^{-1}*X'Z \\ R^{-1}*Z'X & R^{-1}*Z'Z + G^{-1}*A^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} (R^{-1}*X')y \\ (R^{-1}*Z')y \end{bmatrix} \qquad (3)$$

An Expectation Maximisation (EM) algorithm of Dempster *et al.* (1977) for estimation of the ij^{th} element of **G** and **R** is:

$$\hat{g}_{ij}^{(k+1)} = [\hat{u}_i'^{(k)} \hat{u}_j^{(k)} + tr(A^{-1}C_{ij}^{(k)})]/s \qquad (4)$$

$$\hat{r}_{ij}^{(k+1)} = [\hat{e}_i'^{(k)} \hat{e}_j^{(k)} + trB_{ij}^{(k)}]/n \qquad (5)$$

where C_{ij} is a submatrix of a generalized inverse of the coefficient matrix in eqn. (3) corresponding to the i^{th} and j^{th} subvectors in u, B_{ij} is a submatrix of **WCW'**, corresponding to the ij^{th} pair of traits, where $W = [X : Z]$, **s** is the number of classes in u and **n** is the total number of observations.

In the estimation of (co)variance components, we are not interested in solutions for the fixed effects and therefore they

can be absorbed to obtain the system:

$$(\mathbf{R}^{-1} * \mathbf{Z'SZ} + \mathbf{G}^{-1} * \mathbf{A}^{-1})u = (\mathbf{R}^{-1} * \mathbf{Z'S})y \quad (6)$$

where $\mathbf{S} = \mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-}\mathbf{X'}$ is a projection matrix.

Canonical transformation

It is clear from eqn. (3) that if $\mathbf{R} = \mathbf{I}_n$ and **G** is diagonal matrix, the multitrait analysis can be divided into **q** single trait analyses. Such a transformation, called a canonical transformation, was first suggested for animal breeding problems by Thompson (1976) and has been applied to data in several publications (Meyer, 1985; Schaeffer, 1986; Taylor *et al.*, 1985; Jensen & Mao, 1988). The different authors presented different algorithms for transformation of **G** and **R** to the canonical scale. In the present study the algorithm of Golub & Van Loan (1983) was chosen.

Suppose that y_k is a vector of **q** variates measured on animal **k**. The (co)variance matrix of y is $\mathbf{Var}(y) = \mathbf{V} = \mathbf{G} + \mathbf{R}$. Transformation of y_k is needed to z_k (say, $z_k = \mathbf{U}y_k$) such that $\mathbf{URU'} = \mathbf{I}$ and $\mathbf{UGU'} = \mathbf{D}$, where **U** is a transformation matrix, **I** is an identity matrix and **D** is a diagonal matrix. Because there are no covariances among the elements of z_k, each element of z_k can be analysed by a single-trait variance-component method. The appropriate transformation is obtained as follows:

Given $\mathbf{G} = \mathbf{G'}$ and $\mathbf{R} = \mathbf{R'}$, with **R** positive definite:

(i) Compute Cholesky decomposition of $\mathbf{R} = \mathbf{LL'}$ and set $\mathbf{B} = \mathbf{L}^{-1}\mathbf{GL'}^{-1}$.

(ii) Use Householder's transformations and the QL algorithm to calculate eigenvalues and eigenvectors of **B**. Store eigenvectors in a matrix, say **Q**.

(iii) Set $\mathbf{U} = \mathbf{L'}^{-1}\mathbf{Q}$.

Let $\hat{G}_c$ and $\hat{R}_c$ be the estimates of the random factor and residual (co)variance matrices on the canonical scale. The inverse transformation is then used to obtain **G** and **R** on the original scale. That is:

$$\hat{G} = \mathbf{U}^{-1}\hat{G}_c\mathbf{U'}^{-1} \quad (7)$$

$$\hat{R} = \mathbf{U}^{-1}\hat{R}_c\mathbf{U'}^{-1} \quad (8)$$

EM–REML algorithm

After canonical transformation, the mixed model equations in (3) contain **q** diagonal blocks, corresponding to the **q** canonical variates. The equation in the i^{th} block is:

$$(\mathbf{Z'SZ} + \alpha_i^{-1}\mathbf{A}^{-1})\mathbf{u}_{ci} = \mathbf{Z'Sy}_{ci} \quad (9)$$

where α_i is the i^{th} diagonal element of $\mathbf{G}_c$. Let $\mathbf{C}_c = (\mathbf{Z'SZ} + \alpha^{-1}\mathbf{A}^{-1})^{-1}$. Then for the $(\mathbf{k} + \mathbf{1})$ round of iteration, estimators of the elements of $\mathbf{G}_c$ and $\mathbf{R}_c$ are:

$$\hat{g}_{cii}^{(k+1)} = [\hat{u}'_{ci}\mathbf{A}^{-1}\hat{u}_{ci} + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{C})] / \mathbf{s} \quad (10)$$

$$\hat{g}_{cij}^{(k+1)} = [\hat{u}'_{ci}\mathbf{A}^{-1}\hat{u}_{cj}] / \mathbf{s} \qquad \text{for } \mathbf{i} \neq \mathbf{j} \quad (11)$$

$$\hat{r}_{cii}^{(k+1)} = [\hat{e}'_{ci}\hat{e}_{ci} + \mathbf{r(X)} + \mathbf{s} - \alpha_i^{-1}\mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}_c)] / \mathbf{n} \quad (12)$$

$$\hat{r}_{cij}^{(k+1)} = [\hat{e}'_{ci}\hat{e}_{cj}] / \mathbf{n} \qquad \text{for } \mathbf{i} \neq \mathbf{j} \quad (13)$$

where $\hat{e}'_{ci}\hat{e}_{cj} = \mathbf{y'Sy} - \hat{u}'_{ci}\mathbf{Z'Sy}_{cj} - \alpha\hat{u}'_{ci}\hat{u}_{cj}$ for all **i** and **j**.

Formulae (10) and (13) are presented in Jensen & Mao (1988). The estimation of the covariances is simplified, because of the block diagonal structure of the multitrait mixed model equations on the canonical scale. A further simplification is possible if the diagonalization algorithm described by Konstantinov & Erasmus (1992) is employed. The computing strategy using this approach is as follows:

(i) Obtain initial values for **G** and **R**, which could be 'guesses' or literature values, for the first round of iteration.

(ii) Perform absorption and subsequent diagonalization of the MME coefficient matrix as described by Konstantinov & Erasmus (1992).

(iii) Determine the transformation to canonical scale, **U**, such that **G** is diagonal and **R** is identity matrices, using steps (i) to (iii) in the algorithm described.

(iv) Transform the right-hand side matrix, **P'Z'SY** and the matrix of the sums of squares and cross-products **Y'SY** to the canonical scale as follows:

$$\mathbf{P'Z'SY}_c = \mathbf{P'Z'SYU'} \quad (14)$$

$$\mathbf{Y'}_c\mathbf{SY}_c = \mathbf{UY'SYU'} \quad (15)$$

where **P** is the transformation matrix used to diagonalize the MME coefficient matrix.

(v) Obtain estimates for $\mathbf{g}_{ij}$ and $\mathbf{r}_{ij}$ for $\mathbf{i}, \mathbf{j} = 1, \ldots, \mathbf{q}$ according to eqns. (10) to (13).

(vi) Obtain estimates for **G** and **R** on the original scale according to eqns. (7) and (8).

Repeat steps (iii)–(vi) until convergence is reached.

Numerical example

Consider the records for two traits for 704 progeny of six bulls, distributed in 10 herds with a data structure as summarized in Table 1. The six bulls are assumed related according to the following pedigree:

Bull	Sire	MGS
2	1	0
3	1	2
4	1	2
5	3	4
6	3	4
7	5	6

where MGS = maternal grand sire and '0' = unknown parent. In order to compute the Cholesky decomposition of the numerator relationship matrix **A**, a small modification to the algorithm of Quaas (1976) was made as shown by Henderson (1976). For starting values the following (co)variance matrices were used:

$$\mathbf{G} = \begin{bmatrix} 51000.0 & 7100.0 \\ 7100.0 & 5000.0 \end{bmatrix} \text{ and } \mathbf{R} = \begin{bmatrix} 300000.0 & 70000.0 \\ 70000.0 & 80000.0 \end{bmatrix}$$

Based on these starting values, the transformation to the canonical scale for the first round of iteration yields eigenvalues of the sire (co)variance matrix $\lambda_1 = 0.05921610$ and $\lambda_2 = 0.18088861$ with the transformation matrix and its inverse

$$U = \begin{bmatrix} -0.0003362 & 0.0037817 \\ 0.0020188 & -0.0011856 \end{bmatrix} \quad U^{-1} = \begin{bmatrix} 163.852 & 522.640 \\ 278.999 & 46.467 \end{bmatrix}$$

This yields estimates on the canonical scale $g^c_{11} = 0.113880$, $g^c_{12} = -0.139806$, $g^c_{22} = 0.261649$, $r^c_{11} = 0.734939$, $r^c_{12} = -0.665665$ and $r^c_{22} = 1.969377$ which, after being transformed back to the original scale, yield the following estimates for the first round of iteration:

$$\hat{G} = \begin{bmatrix} 50900.012 & 7240.134 \\ 7240.134 & 5258.250 \end{bmatrix} \quad \hat{R} = \begin{bmatrix} 318872.266 & 73254.481 \\ 73254.481 & 82259.810 \end{bmatrix}$$

As a convergence criterion, the norm of the matrix of differences of the parameter estimates was used as suggested by Jensen & Mao (1988) and Schaeffer (1986), i.e.

$$\left\| \hat{G}_{(k+1)} - \hat{G}_{(k)} \right\| \quad \text{and} \quad \left\| \hat{R}_{(k+1)} - \hat{R}_{(k)} \right\|$$

Both norms were required to be less than 10^{-8}. Under these conditions convergence was reached after 62 rounds of iteration with the following estimates:

$$\hat{G} = \begin{bmatrix} 51142.982 & 7118.611 \\ 7118.611 & 5188.666 \end{bmatrix} \quad \hat{R} = \begin{bmatrix} 318864.396 & 73272.382 \\ 73272.382 & 82963.886 \end{bmatrix}$$

Simulation study

The most time-consuming part of the algorithm presented is the Householder tridiagonalization and subsequent diagonalization of the MME coefficient matrix using the QL algorithm. In order to investigate the time required for both transformations with a large data set, a simulation study was conducted. Three traits were simulated and the following model was fitted for all traits:

$$y_{ijkl} = \mu + h_{il} + s_{jl} + e_{ijkl} \qquad (16)$$

where y_{ijkl} is the l^{th} record of the k^{th} progeny of the j^{th} sire, within the i^{th} herd; h_{il} is a fixed effect for the l^{th} record within the i^{th} herd; s_{jl} is a random effect of sire j on his progeny's l^{th} record, and e_{ijkl} is the random error associated with each record. The simulation consisted of first through third record of 127960 progeny of 1000 sires, born over a period of 20 years and distributed over 100 herds. Within each year, 5 'proven' and 45 'young' sires were used. The 'proven' sires were allowed to have from 3 to 5 progeny within each of 20 randomly allocated herds. Sires 1 through 100 were assumed unrelated and they represented the so-called 'base' population, and were randomly assigned as sires and MGS to bulls 101 through 450. For bulls 451 through 1100, sires and MGS were assigned at random from every year's batch of bulls, starting from year 1. For example, bulls 451 through

Table 1 Data for the numerical example: numbers of records and totals for sire × herd subclasses

	Sire								
	2			3			4		
Herd	n	T1	T2	n	T1	T2	n	T1	T2
1	5	22169.04	842.04	20	72608.52	3032.53	15	63574.51	2132.78
2	20	87038.94	3077.38	16	59858.82	2346.82	14	58570.57	2114.60
3	15	58397.78	2320.47	18	67806.75	2712.51	8	32169.70	1230.42
4	19	79004.69	2756.54	11	43257.35	1931.04	14	54870.16	1827.52
5	15	64660.39	2293.50	17	61505.48	2440.74	7	28741.89	1079.30
6	5	23128.68	687.05	14	55088.63	2091.93	13	53240.39	1952.38
7	11	45877.57	1586.72	19	69520.09	2733.26	6	24767.62	961.54
8	13	56002.25	1934.09	13	49343.14	1910.94	15	61699.61	2515.58
9	20	82026.43	3115.39	5	19064.53	752.01	15	64513.32	2319.18
10	15	64230.84	2611.55	14	51085.83	2096.56	16	65635.60	2376.24

	Sire								
	5			6			7		
Herd	n	T1	T2	n	T1	T2	n	T1	T2
1	14	56100.34	1887.45	12	52098.71	1915.03	17	68285.66	2606.86
2	13	49726.02	1677.48	8	33330.61	1197.40	17	67152.82	2697.66
3	9	34352.20	1279.98	8	32542.49	1168.55	6	20642.85	760.06
4	5	20399.52	819.34	12	48014.98	1875.83	19	79823.12	3006.25
5	11	42984.14	1678.76	7	27563.95	1068.99	12	51049.62	2105.27
6	20	79168.45	2884.40	5	20916.99	796.87	16	63918.02	2557.09
7	6	25670.19	883.32	11	45710.53	1734.48	15	57405.08	2294.63
8	9	34144.61	1345.47	20	83088.58	3073.69	18	72571.82	2959.69
9	6	21730.02	816.15	15	58748.61	2299.39	6	23267.50	964.72
10	9	34477.63	1275.83	13	50640.71	1874.98	9	37591.49	1467.95

500 were sons of five randomly chosen sires 101 through 150 and grandsons of the rest. The sire and the residual effects were generated by random sampling from a normal distribution with zero mean and variance of 1. The following (co)-variance matrices were used as 'true' values:

$$\mathbf{G} = \begin{bmatrix} 8000.0 & 300.0 & 800.0 \\ & 600.0 & 80.0 \\ \textit{symmetric} & & 200.0 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 360000.0 & 75000.0 & 80000.0 \\ & 40000.0 & 25000.0 \\ \textit{symmetric} & & 30000.0 \end{bmatrix}$$

Programming and computers

Two versions of the program were developed for conducting the analysis. Both were written in standard FORTRAN 77. Two types of computers were employed. The first was a 386 PC with 'Cyrix' high-performance math coprocessor, and the second was a CONVEX C-120 with special vector capabilities. Both computers employed FORTRAN compilers with the highest possible degree of optimisation. The first version of the program was compiled with a 'LAREY' FORTRAN compiler with scalar optimisation, and the second was designed and compiled with vector optimisation on the CONVEX. Only default system options were used. For (co)variance components estimation two algorithms were used, viz. EM and the algorithm presented by Meyer (1985). The same convergence criterion as for the numerical example was used.

Results and Discussion

The REML algorithm presented by Meyer (1985) converges more rapidly than the EM–REML algorithm. Convergence was obtained after 26 and 138 rounds of iteration, respectively, with the following estimates of the (co)variance matrices:

$$\hat{G} = \begin{bmatrix} 7737.934 & 316.583 & 660.998 \\ & 644.779 & 64.920 \\ \textit{symmetric} & & 330.996 \end{bmatrix}$$

$$\hat{R} = \begin{bmatrix} 357044.447 & 73884.212 & 79996.391 \\ & 36617.216 & 22027.696 \\ \textit{symmetric} & & 28321.521 \end{bmatrix}$$

Although it was not the objective of this study to compare the speed of convergence of different REML algorithms, Meyer's algorithm (Meyer, 1985) was employed simply to show that more complicated algorithms can be handled with ease using canonical transformations and the diagonalization approach.

The CPU time required for Householder's tridiagonalization and the subsequent diagonalization of the MME coefficient matrix using the QL algorithm were 107.69 and 4.75 min, and 29.28 and 1.78 min for 386 PC and CONVEX, respectively.

Tridiagonalization and subsequent diagonalization require $2n^{3}/3$ and $30n$ operations, respectively (Golub & Van Loan, 1983). This task is more demanding than the $n^{3}/2$ operations needed to invert a symmetric matrix. However, once the transformations are completed, the time required for the iterations is trivial. Only 12 and 63 s were needed for convergence with Meyer's and the EM algorithm, respectively on the CONVEX. The corresponding times for the 386 PC were 23 and 121 s, respectively.

Conclusions

The results from this study show that large data sets can be analysed with ease and at a low cost using canonical transformation and the diagonalization approach. However, there are two aspects which need attention when dealing with multiple trait estimation of (co)variance components. The first is that in the analysis of highly correlated traits there is an ever-existing possibility that, owing to sampling, the estimated (co)variance matrix of the random factor **G** could be non-positive definite. As pointed out by Hill & Thompson (1978), the probability of this occurring increases with the number of variances included simultaneously. The algorithm of Golub & Van Loan (1983) requires **R** to be positive definite, and hence the transformation to canonical scale is possible only if this is true. How to deal with negative definite estimates remains an open question. Meyer (1985) suggested that at convergence negative roots should be set to zero before transforming estimates to the original scale. Another possibility would be to impose a constraint on the parameter space while iterating. Such an algorithm was proposed by Schaeffer (1986). This algorithm only estimates genetic and residual variances on canonical scale. The variance estimates at convergence are then back-transformed to obtain both variances and covariances on the original scale by the use of the inverse of the initial transformation matrix (**U**). Such a procedure would yield estimates that are dependent on the initial values chosen for **G** and **R** as shown by Jensen & Mao (1988).

The second problem is the choice of convergence criteria. Jensen & Mao (1988) suggested conservative stopping criterion instead of fixed number of iterations. For the numerical example in this study we, however, forced the norms of the matrix differences to zero. The number of iterations then increased from 62 to 77 and the CPU time from 3 to 4 s for the EM algorithm on the 386 PC. This result suggests that, using the algorithm presented, the precision of a conservative stopping criterion can be increased to a very high level.

References

DEMPSTER, A.P., LARID, N.M. & RUBIN, D.B., 1977. Maximum Likelihood from incomplete data via EM algorithm. *Jl R. statist. Soc., B* 39, 1.

GOLUB, G.H. & VAN LOAN, C.F., 1983. Matrix computations. The John Hopkins University Press, Baltimore, Maryland.

HENDERSON, C.R., 1973. Sire evaluation and genetic trends. *Proc. Anim. Breed. Genet. Symp.*, in Honor of Dr J.L. Lush. Am. Soc. Anim. Sci., Champaign, IL.

HENDERSON, C.R., 1976. Inverse of a matrix of relationships due to sires and maternal grand sires in an inbred population. *J. Dairy Sci.* 59, 1585.

HILL, W.G. & THOMPSON, R., 1978. Probabilities of nonpositive definite between group or genetic covariance matrices. *Biometrics* 34, 429.

JENSEN, J. & MAO, I.L. 1988. Transformation algorithms in analysis of single trait and multitrait models with equal design matrices and one random factor per trait: A review. *J. Anim. Sci.* 66, 2750.

KONSTANTINOV, K.V. & ERASMUS, G.J., 1992. Application of Householder's transformations and the QL algorithm to REML estimation of variance components. *S. Afr. J. Anim. Sci.* 22, 92.

MEYER, KARIN, 1985. Maximum Likelihood estimation of variance components for multivariate model with equal design matrices. *Biometrics* 41, 153.

QUAAS, R.L., 1976. Computation of the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 32, 949.

SCHAEFFER, L.R., 1986. Estimation of variances and covariances within allowable parameter space. *J. Dairy Sci.* 69, 187.

SEARLE, S.R., 1982. Matrix algebra useful for statistics. John Wiley & Sons, NY.

TAYLOR, J.F., BEAN, B., MARSHAL, C.E. & SULLIVAN, J.J., 1985. Genetic and environmental components of semen production traits of artificial insemination of Holstein bulls. *J. Dairy Sci.* 68, 2703.

THOMPSON, R., 1973. The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics* 22, 527.

THOMPSON, R., 1976. Estimation of quantitative genetic parameters. In: Proc. of the Int. Conf. on Quantitative Genetics. Eds. Pollak, F., Kempthorne, O. & Bailey, T.B., Iowa State Univ. Press, Ames. pp. 639–657.