Reduced Rank Estimation of (Co)-variance Components For International Evaluation Using AI-REML

Per Madsen[1], Just Jensen[1], and Thomas Mark[2]

[1]Danish Institute of Agricultural Sciences, Research Centre Foulum, P.O. Box 50, DK-8830 Tjele, Denmark.
[2]Danish Agricultural Advisory Centre, Udkaersvej 15, Skejby, DK-8200 Aarhus N, Denmark.

Introduction

In multiple across country evaluations (MACE) of dairy bulls, (Shaeffer & Zhang, 1993), different national evaluations are regarded as different traits. This has the advantage of taking differences in trait definitions and national evaluation systems into account as well as possible genotype by environment interactions due to e.g. climate and/or differences in productions-systems and -circumstances. The disadvantage is that a genetic (co)-variance matrix of dimension corresponding to the number of traits included is required. The expression of similar traits in different countries tends to be highly correlated. As the number of countries increases, the corresponding covariance matrix will tend to have one or more non-positive eigenvalues, indicating that it is not of full rank. Previous methods for the estimation of (co)-variance matrices for MACE have been REML based on the EM algorithm of Dempster et al. (1977). This algorithm is generally known to have very slow convergence properties, and furthermore it is difficult to estimate standard errors of the estimates obtained with the EM-algorithm. The multivariate AI-REML algorithm of Jensen et al. (1997) generally has very good convergence properties, and in most cases, it can provide standard errors of resulting estimates.

In many cases, it is desired to include more than one trait per country. This may; for example, be the case when analysing somatic cell count and clinical mastitis where all countries have somatic cell count evaluations but only a small subset of countries has evaluations on clinical mastitis. This introduces residual co-variances between traits that are recorded in the same country since they typical are recorded on the same set of animals. Current methods for estimation of (co)-variance matrices for MACE have assumed zero residual co-variances. In the algorithm of Jensen et al. (1997), a complete specification of existence or non-existence of residual co-variances is possible.

Current methods generally assume that the ratio between sire and residual variances within countries are known, and tend to perform best if only a well-connected subset of the data is analysed (Sigurdsson et al. 1996, Klei & Weigel, 1998). In analysing all data from all bulls evaluated in each country, it should be possible to estimate sire variances for each country at the same time as estimating sire co-variances and existing residual co-variances. In the present approach sire and residual variances in each country is always estimated and it is therefore conjectured that more efficient and unbiased estimated will be obtained by including data from all bulls evaluated.

The purpose of this paper is to present an extension of the algorithm of Jensen et al. (1997) for use in the estimation of (co)-variances in across country models. This includes:

- Development of an AI-REML algorithm that handles records with different residual variance, due to different number of daughters per sire in the analysis.
- Modify the algorithm in order to estimate (co)-variance matrices of reduced rank.
- Modify the across country model to accept residual co-variances among traits if more than one trait per country is included in the analysis.

Methods

Model

The MACE model (Schaeffer & Zhang, 1993):

$$\mathbf{y = Cc + ZQg + Zs + e}$$

where **y** is a vector of de-regressed proofs, **c** is a vector of fixed country effects, **g** is a vector of fixed phantom group effects, **s** is a vector of random bull effects, and **e** is a vector of random residuals. **C, Z,** and **Q** are design matrices relating de-regressed proofs to countries and bulls, and relating bulls to phantom groups respectively.

Following the same ideas as Klei & Weigel, (1998) where unknown parents were assigned to phantom parent groups on a within country basis, the MACE model becomes:

$$\mathbf{y = ZQf + Zs + e}$$

with the following distributional properties:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{s} \\ \mathbf{e} \end{pmatrix} \sim \mathbf{MVN}\left\{ \begin{pmatrix} \mathbf{ZQf} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{ZGZ'+R} & \mathbf{GZ'} & \mathbf{R} \\ \text{Symm.} & \mathbf{G} & \mathbf{0} \\ & & \mathbf{R} \end{pmatrix} \right\}$$

where **f** is a vector of fixed phantom group + country effects, **G** is the (co)-variance matrix among elements in **s,** and **R** is the residual (co)-variance matrix

Since the actual daughter records are usually not known at the international level, the residual co-variance matrix is assumed to be block diagonal with blocks for each bull of same size as the number of countries plus traits where the bulls have evaluations. For example, if bull k has records from two countries. Trait one is evaluated in both countries but trait two is only evaluated in country one, then the corresponding block in the residual (co)-variance matrix, **R**, for bull k will have the following form:

$$\mathbf{R}_k = \begin{pmatrix} \mathbf{r}_{1.1,1} & \mathbf{r}_{1.1,2} & \mathbf{0} \\ \mathbf{r}_{1.2,1} & \mathbf{r}_{1.2,2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{r}_{2.1,1} \end{pmatrix}$$

where the first subscript refers to country and the last two to traits within country respectively. The elements $\mathbf{r}_{c.i,j}$ are $\boldsymbol{\sigma}_{c.i,j}/\min(\mathbf{n}_{C.i},\mathbf{n}_{C.j})$, where $\boldsymbol{\sigma}_{c.i,j}$ is the residual co-variance between trait **i** and **j** in country **c** and $\mathbf{n}_{C.i}$ is the effective number daughters for **i** trait in country **c**. This assumes that the trait with the smallest number of records is measured on a subset of the animals recorded for the other trait.

Algorithm

The algorithm used is an extension of the multivariate algorithm presented by Jensen et al. (1997). This involves computing first and second derivatives of the likelihood function and estimating the (co)-variance matrices in a Newton-Raphson like algorithm. Use of sparse matrix methodology is necessary for efficient implementation of such a procedure (Misztal & Perez-Enciso, 1993). Algorithms based on second differentials are not guaranteed to stay within the parameter space and therefore the algorithm gradually can switch to an EM-algorithm in order to guarantee that the likelihood function increases in each iteration.

The above-mentioned methods were extended to include models for records with heterogeneous residual variance (different number of daughters per bull). In each iteration the resulting (co)-variance matrices is checked to be positive definite. If one or more non-positive eigenvalues are detected the corresponding covariance matrix was replaced with: $\mathbf{G_0} = \mathbf{H\Delta H'}$, where $\boldsymbol{\Delta}$ is a diagonal matrix of modified eigenvalues and **H** is the corresponding set of orthonormal eigenvectors. In practice non-positive eigenvalue were replaced with a small positive number $\boldsymbol{\tau}$**.** The effect of this is that the corresponding dimension is dropped in the (co)-variance matrix but in general, all variances can still be estimated.

A full description of the algorithm will be presented in Madsen & Jensen (2000).

Materials

The algorithm was tested on three datasets:

The first dataset were in principle the same as described by Mark et al. (2000) and included real data (evaluations) on somatic cell count (SCC) on 21963 bulls from nine countries. This dataset forms a well-connected subset as defined by Sigurdsson et al. (1996). Thus, for this dataset, a nine by nine sire covariance matrix should be estimated and all residuals were assumed to be uncorrelated.

The second dataset consisted of data from the same nine countries but extended with results from all evaluated bulls with at least 50 daughters in one of the nine countries. In total

data on 34978 bulls were included. Furthermore, the dataset included evaluations on clinical mastitis (CM) from three of the nine countries.

To test the possibility of estimating within country residual co-variances a subset containing SCC and CM for the three countries were build. For this third dataset, a 6 by 6 sire (co)-variance matrix should be estimated. SCC and CM were assumed to have residual co-variances within the county.

Results and Discussion

Number of rounds to convergence, dimension and rank of the resulting covariance matrix for the three datasets are shown in Table 1.

Table 1. Estimation statistics for test datasets

Data-Set	# of iterations	Sire (co)-variance matrix		
		Dim.	Rank	% var. expl.
1	50	9	7	99.7
2	86	9	6	99.6
3	37	6	5	99.9

The amount of work in one round of AI-REML is roughly the same as in one round of EM-REML since, usually, the major part of the time is used on computing the sparse inverse of the MME. The results shows that large (co)-variance matrices of reduced rank can be estimated with a relatively low number of iterations. The stopping criterion used was that the norm of the gradient vector should be less than 10^{-4}. Generally, this corresponds to a much smaller round to round change in the parameter vector. In comparison, the analysis of Mark et al. (2000) took more than 1000 rounds for convergence, although from a different set of priors.

Results from analysis of dataset 1 and 2 on SCC from nine countries are shown in Table 2. For comparison, estimates from Mark et al. (2000) using the algorithm of Klei & Weigel (1998) are included. The results from dataset 1 (the well-connected subset) clearly show that very similar estimates were obtained from the two methods. Standard errors (s.e.) of genetic correlations between SCC in different countries range between .01 and .46. If country four is excluded, the range is between .01 and .03. The reason for the large s.e. of estimates related to country four is that only 54 bulls with evaluations from this country is included in the "well-connected" subset.

The correlations estimated on dataset 2 were generally lower than the estimates from dataset 1. Especially correlations involving countries 1, 3, and 4 were lower. The reason could be the same bias problem as shown by Sigurdsson et al. (1996), and Klei & Weigel, (1998) using all data. Another reason could be differences/bias in the national evaluation system. The possible bias problem will be addressed in Madsen & Jensen (2000). Also here, very accurate ($0.01 < \text{s.e.} < 0.03$) results were found, except for country four.

The estimated genetic and residual correlations for the third dataset are in Table 3. The genetic correlations between SCC and CM within country are between .49 and .58. Residual correlations within country range from .11 to .35. This shows that residual correlations can be estimated. Methods assuming these correlations to be zero may lead to biased inferences. The bias-/unbiasedness of these estimators will also be addressed in Madsen & Jensen (2000).

All three analysis showed that a reduction in rank of the (co)-variance matrices were possible. This may lead to reductions in computational cost as the dimensions corresponding to non-positive eigenvalues can be dropped from the international evaluation. Furthermore reducing the rank of the (co)-variance matrices will lead to models that are more parsimonious. In the current implementation decisions on reducing the rank is made purely on mathematical grounds. It may be possible to develop a "statistical" rule such that all eigenvalues that explains less than a certain threshold of all genetic variance is dropped. This would lead to models with even fewer parameters to be estimated than the current implementation.

Earlier methods were tested on simulated data with varying degree of connectedness and varying heritability of the traits analysed. Similar tests will be included in Madsen & Jensen (2000) for the current algorithm.

Conclusions

- The AI-REML algorithm was able to estimate sire covariance matrices for multiple traits recorded in several countries and generally converges in much fewer rounds of iteration that previously used procedures based on the EM algorithm.
- The algorithm was able to estimate (co)-variance matrices of reduced rank.
- The algorithm provides standard errors of the resulting estimates.
- The procedure can take non-zero residual co-variances into account. Sire and residual (co)-variances is estimated for each trait and in each country where the trait is recorded.

References

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Proc. R. Stat. Soc. B 39*, 1.

Jensen, J., Mantysaari, E.A., Madsen, P. & Thompson, R. 1997. Residual Maximum Likelihood Estimation of (Co) Variance Components in Multivariate Mixed Linear Models using Average Information. *J. Indian Soc. Agr. Stat. 49*, 215.

Klei, L. & Weigel, K.A. 1998. A method to estimate correlations among traits in different countries using data on all bulls. Proceedings of the 1998 Interbull meet., Rotorua, New Zealand. *Interbull Bulletin 17*, 8-14.

Madsen, P. & Jensen, J. 2000. Estimation of Reduced Rank (Co)-variance matrices Using AI-REML. (in preparation).

Mark,T.W., Fikse, F., Sigurdsson, A. & Philipsson, J. 2000. Feasibility of International Genetic Evaluations of Dairy Sires for Somatic Cell Count and Clinical Mastitis. *Paper presented at the 2000 Interbull meet., Bled.* 6 pp.

Misztal, I. & Perez-Enciso, M. 1993. Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation-maximazation. *J. Dairy Sci. 76*, 1479.

Schaeffer, L.R. & Zhang, W. 1993. Multitrait, across country evaluation of dairy sires. Proceedings of the open session of the Interbull annual meet., Aarhus, Denmark. *Interbull Bulletin 8.* 21pp.

Sigurdsson, A., Banos, G. & Philipsson, J. 1996. Estimation of genetic (co)variance components for international evaluation of dairy bulls. *Acta Agric. Scand. Sect. A, Anim. Sci. 46*, 129.

Table 2. Estimated genetic correlations for somatic cell count (SCC) in nine countries. Asymptotic standard error of estimates in parentheses

Dataset[1)]	Country	Country							
		2	3	4	5	6	7	8	9
1		.93 (.01)	.80 (.03)	.96 (.46)	.94 (.01)	.96 (.01)	.94 (.01)	.81 (.03)	.94 (.01)
2	1	.77 (.01)	.79 (.01)	.69 (.42)	.88 (.01)	.93 (.01)	.88 (.01)	.74 (.03)	.82(.01)
M		.90	.80	.94	.93	.96	.93	.82	.93
1			.85 (.03)	.96 (.17)	.96 (.01)	.94 (.01)	.98 (.01)	.81 (.03)	.89 (.01)
2	2		.69(.03)	.97 (.20)	.93 (.01)	.86 (.01)	.96 (.01)	.81 (.03)	.90 (.01)
M			.85	.88	.95	.94	.97	.81	.87
1				.84 (.09)	.88 (.02)	.87 (.02)	.87 (.02)	.97 (.03)	.87 (.02)
2	3			.65 (.16)	.82 (.03)	.85 (.03)	.80 (.03)	.91 (.03)	.76 (.02)
M				.79	.89	.89	.88	.98	.88
1					.96 (.14)	.97 (.14)	.97 (.16)	.84 (.13)	.92 (.13)
2	4				.90 (.15)	.81 (.35)	.91 (.19)	.82 (.18)	.92 (.14)
M					.94	.96	.94	.83	.91
1						.96 (.01)	.96 (.01)	.88 (.03)	.92 (.01)
2	5					.95 (.01)	.95 (.01)	.88 (.02)	.90 (.01)
M						.98	.96	.89	.92
1							.97 (.01)	.88 (.03)	.92 (.01)
2	6						.95 (.01)	.87 (.02)	.87 (.01)
M							.97	.91	.93
1								.85 (.03)	.91 (.01)
2	7							.85 (.03)	.90 (.01)
M								.87	.90
1									.87 (.02)
2	8								.87 (.02)
M									.91

[1)] 1 estimates from dataset 1, 2 estimates from dataset 2, M estimates from Mark et al. (2000).

Table 3. Estimated genetic (above diagonal) and residual (below diagonal) correlation's between somatic cell count (SCC) and clinical mastitis (CM) in three countries (dataset 3). Asymptotic standard error of estimates in parentheses

Trait	Country	SCC			CM		
		1	2	3	1	2	3
S	1		.80 (.21)	.96 (.03)	.49 (.05)	.82 (.51)	.38 (.07)
C	2			.83 (.21)	.19 (.34)	.58 (.09)	.18 (.28)
C	3				.63 (.07)	.68 (.36)	.53 (.04)
C	1	.32 (.05)				.09 (.81)	.92 (.09)
M	2		.35 (.11)				-.10 (.47)
	3			.11 (.12)			