# A Method To Estimate Correlations Among Traits In Different Countries Using Data On All Bulls

*L. Klei* [1] *and K. A. Weigel* [2]

[1] *Holstein Association USA, Inc. Brattleboro, Vermont, USA*

[2] *University of Wisconsin, Madison, Wisconsin, USA*

## Introduction

The international evaluation of bulls using multiple across country evaluations (MACE) requires estimates of the genetic (co)-variances among countries. By considering the trait of interest in each country as a different trait the estimation of MACE parameters resembles the estimation of parameters for a multiple trait evaluation. However, the estimation of MACE parameters is atypical in that many animals have evaluations in only one country and ties among countries are in general very limited. Sigurdsson and Banos (1995) noted that this can cause problems while estimating variance components using restricted maximum likelihood (REML). This was explained by that in their algorithm many bulls have indirect proofs for some of the countries, i.e. based solely on pedigree indices and correlated information from a different country. In order to circumvent this problem they suggested the use of a subset of well connected bulls. Well-connected being those bulls that have proofs in more than one country as well as bulls that are members of full-sibs groups that have members with proofs in more than one country.

The purpose of this paper is to show an expectation maximization (EM) algorithm for REML that allows for the use of information on all bulls in data from several countries. Results of a simulation will be shown to compare correlations estimated from all data versus parameters estimated from well-connected subset of the same data.

## Method and Material

The method presented in this paper is based on the idea that estimation of (co)-variances among countries only requires bull equations for bulls within a country when he contributes additional information from that country. This means that he either has an own observation in a country or he has a descendant with an observation in a country. All other bulls will have evaluations based on parents. All information about this is contained within the parent information. This approach requires the development of an estimation procedure similar to the one described by Klei (1995).

As an example, let *[m]* be a representation of the countries in which a bull has information. I.e. for two countries, *[10]* indicates information in country one only, *[01]* indicates information in country two only, while *[11]* indicates information in country one and two. Also let $q_{[m]}$ be the number of bulls in each category, then the total number of bull in the evaluation $q$ equals

$$\sum q_{[m]}$$

This approach requires

$$\sum_{[m]} sum_{[m]} q_{[m]}$$

equations where $sum_{[m]}$ is the number of ones in the representation (i.e. $sum_{[11]}$ = 2). Sigurdsson and Banos (1995) described a more traditional approach in which equations were assigned to each bull in each country. This approach requires $n_c \times q$ equations, where $n_c$ is the total number of countries. This value is always larger or equal to

$$\sum_{[m]} sum_{[m]} q_{[m]}$$

## Model

The common MACE model is (Schaeffer and Zhang, 1993):

$$y = Cc + ZQg + Zs + e$$

in which

| | | |
|---|---|---|
| $y$ | : | vector of de-regressed proofs |
| $c$ | : | vector of country effects |
| $g$ | : | vector of phantom group effects |
| $s$ | : | vector of random bull effects |
| $e$ | : | vector of random residuals |
| $C$ | : | matrix that assigns de-regressed proof to a country |
| $Z$ | : | matrix that assigns de-regressed proof to a sire |
| $Q$ | : | matrix that assigns group effects to bulls |

In general, applications of MACE assign phantom groups to unknown parents within country. The result of this is that phantom groups are totally confounded with country effects and therefore the MACE model can also be written as:

$$y = ZQf + Zs + e$$

with the following distribution properties of the random variables:

$$\begin{pmatrix} y \\ s \\ e \end{pmatrix} \sim MVN \left\{ \begin{bmatrix} \begin{pmatrix} ZQf \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} ZGZ^T + R & GZ^T & R \\ & G & 0 \\ symm. & & R \end{pmatrix} \end{bmatrix} \right\}$$

where:

| | | |
|---|---|---|
| $f$ | : | vector of phantom group + country effects |
| $G$ | : | (co)-variance matrix among the elements of $s$ |
| $R$ | : | (co)-variance matrix among the elements of $y$ |

For MACE, since the actual daughter observations are unknown, $R$ is assumed to be a diagonal matrix with diagonal elements for sire k in country i (Sigurdsson and Banos, 1995):

$$r_{ik,ik} = \frac{4 - h_i^2}{h_i^2} g_{ii} \Big/ n_{ik}$$

where:

| | | |
|---|---|---|
| $g_{ii}$ | : | genetic variance in country i |
| $h_i^2$ | : | heritability in country i |
| $n_{ik}$ | : | number of daughters on which the proof of bull k in country i was based |

The associated mixed model equations are:

$$\begin{pmatrix} Q^T Z^T R^{-1} ZQ & Q^T Z^T R^{-1} Z \\ symm. & Z^T R^{-1} Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{f} \\ \hat{s} \end{pmatrix} = \begin{pmatrix} Q^T Z^T R^{-1} y \\ Z^T R^{-1} y \end{pmatrix}$$

## EM-REML algorithm

The EM algorithm for REML was described by Dempster et al. (1977). This algorithm was modified by Klei (1995) to efficiently handle the special data structure occurring in multiple country evaluations. The latter method can easily be modified to accommodate the special needs of MACE in which only the genetic (co)-variance components are estimated and the residuals are assumed to be functions of these.

Define $\hat{\eta}_k$ to be the set of known solutions for bull k in the vector of bull solutions $\hat{s}$. In the example, for a bull in set $q_{[01]}$ these would be the solution for country two. It is also beneficial to define a matrix $H_{[m]}$, which is a picker matrix obtained by deleting a row from $I_{n_c}$ here in the [m] representation the corresponding element of [m] is 0. In the example,

$$H_{[10]} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \qquad H_{[01]} = \begin{pmatrix} 0 & 1 \end{pmatrix}, \qquad \text{and} \qquad H_{[11]} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Also define:

$$\hat{S}_{[m]k} = \hat{\eta}_{[m]k} \hat{\eta}_{[m]k}^T + \operatorname{var}\left( \hat{\eta}_{[m]k} - \eta_{[m]k} \right)$$

the sum of squares of BLUPs and prediction error (co)-variances of the BLUPs for sire k in group [m]. These can be found by obtaining solutions to the mixed model equation and inversion of the left hand sides of these equations. In this study FSPAK (Perez-Enciso et al., 1994) was used to obtain these values.

Define also:

$$\hat{S}_{[m].} = \sum_{k=1}^{q_{[m]}} \hat{S}_{[m]k} \text{ and } \hat{S}_{..} = \sum_{[m]} H_{[m]} \hat{S}_{[m].} H_{[m]}^T$$

It can then be derived that an update for the estimate of $G_o$ in round (t+1) can be obtained through iteration on:

$$G_o^{(t+1)} = G_o^{(t)} + G_o^{(t)} \sum_{[m]} \left\{ H_{[m]} (G_o^{(t)^{-1}} \hat{S}_{[m].} G_o^{(t)^{-1}} - q_{[m]} G_o^{(t)^{-1}} \right) H_{[m]}^T \right\} G_o^{(t)} / q$$

until $G_o^{(t+1)} = G_o^{(t)}$ and then use the $G_o^{(t+1)}$ as the value in the next iteration round. The EM-algorithm can be described as:

E-step : compute BLUP and PEV for *s*.

M1-step : compute $\hat{S}_{[m].} = \sum_{k=1}^{q_{[m]}} \hat{S}_{[m]k}$

M2-step : iterate on the update until $G_o^{(t+1)} = G_o^{(t)}$ and repeat the steps until convergence.

Note that the update reduces to:

$$G_o^{(t+1)} = \hat{S}_{..} / q$$

when all animals have equations in all countries. This is the more conventional update for the genetic (co)-variance components.

## Simulation

Data were simulated as bivariate normal for two populations with correlation of either .70 or .95. Heritability parameters were .10 or .35, and the maximum percentage of sires of sires and sires of cows which could be imported from the other population was fixed at either 15 or 50. Population size was 24,000 cows and 300 progeny test bulls for each population in each generation. Twenty sires of sons were chosen in each population per generation, in addition to 50 "proven" sires of cows. Each progeny test bull had 80 progeny in the current generation, and could have 300 additional progeny in the next generation (in either population) if selected as a "proven" sire of cows. Six generations of selection were carried out. Following simulation, breeding values were estimated within each population using a univariate animal model. For each of the eight parameter combinations 5 samples were generated and analyzed.

## Methods of Evaluation

Estimated breeding values and progeny counts on bulls were used to compute de-regressed proofs according to the method described by Rozzi and Schaeffer (1996). De-regressed data were subsequently analyzed using the following five different methods:

I. All data, assigning an equation to a bull in a country where he has information as defined previously.

II. All data, assigning an equation to each bull in each country.

III. Well connected subset, assigning an equation to a bull in a country where he has information, and keeping the variances equal to the within country variances estimated from a single country evaluation on all data.

IV. Well connected subset, assigning an equation to a bull in a country where he has information, and estimating the variances from the subset data.

V. Well connected subset, assigning an equation to each bull in each country, and estimating the variances from the subset data.

For all analyzes the same programs were used. Iterations were stopped when all the parameters showed a relative change to the previous round of less than $10^{-7}$ or when 1000 rounds of iteration were reached.

## Results and Discussion

Table 1 shows the average number of rounds needed to reach convergence. In this table methods that are immediately comparable, I with II and IV with V, show that the method of assigning equations based on information in a country converges more rapidly than when each animal has an equation in each country. This can be explained by the large number of nuisance parameters (bulls without information in a country) when one equation per bull per country approach is used. When all data is being used many samples did not reach the required convergence within the maximum number of rounds. This table also shows that it is easier to estimate correlations well within the parameter space. Estimates close to the edge (.95) were slower to converge. As expected it was also easier to estimate parameters when heritabilities were high. This can be explained by the increase in information about the genetic component in the de-regressed proofs.

Table 1. Average number of rounds of iteration for the different parameter combinations and methods of analysis.

| | | | Method | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | I | | II | | III | | IV | | V | |
| $r_g$ | $h^2$ | % exch. | ave. | | ave. | | ave. | | ave. | | ave. | |
| .70 | .10 | 15 | 269 | | 1000 | *5*[a] | 103 | | 119 | | 173 | |
| | | 50 | 99 | | 559 | | 59 | | 75 | | 112 | |
| | .35 | 15 | 170 | | 975 | *2* | 59 | | 81 | | 113 | |
| | | 50 | 74 | | 371 | | 34 | | 51 | | 86 | |
| .95 | .10 | 15 | 503 | | 1000 | *5* | 182 | | 193 | | 296 | |
| | | 50 | 605 | *2* | 1000 | *5* | 400 | *1* | 386 | *1* | 489 | *1* |
| | .35 | 15 | 228 | | 1000 | *5* | 82 | | 159 | | 242 | |
| | | 50 | 92 | | 501 | | 53 | | 88 | | 139 | |

[a]indicates the number of samples for which a 1000 rounds of iteration was reached before the stopping criterion of $10^{-7}$ .

| | |
|---|---|
| $h^2$ | : heritability |
| $r_g$ | : true genetic correlation |
| % exch. | : percent exchange of genetic material |
| Method | : evaluation method, for description see main text. |

In situations where countries have a large number of genetic ties (high exchange percentage) estimation of parameters becomes easier. The only exception being the case of a heritability of .10 and a correlation of .95. This might be explained by the high estimated correlation (.97, Table 3) for this situation. In this case the closeness to the edge of the parameter space might have affected convergence.

Estimated correlations for the different situations using the five methods are in Table 3. From this table it can be seen that subset analysis (Method III through V) always gave lower estimates for the low heritability. For high heritability the subset analyzes always give higher estimates. Many of these differences were not significant, possibly due to the small number of samples. Largest differences were observed in the situation of moderate correlation and high heritability. From this it appears that subset analysis can be used when estimating correlation that are larger than .90. Due to the upper limit of the parameter space the risk of severe bias is reduced.

In cases where estimated correlations are smaller than .90 it appears to be beneficial to verify subset analysis periodically with a complete data analysis.

Table 2. Standard deviation of the difference in correlation estimates from Method I (full data) and Method III (subset data) for the different parameter combinations.

| $r_g$ | $h^2$ | % exch. | standard deviation |
|---|---|---|---|
| .70 | .10 | 15 | .06 |
| | | 50 | .01 |
| | .35 | 15 | .03 |
| | | 50 | .01 |
| .95 | .10 | 15 | .01 |
| | | 50 | .01 |
| | .35 | 15 | .01 |
| | | 50 | .00 |

$h^2$ : heritability
$r_g$ : true genetic correlation
% exch. : percent exchange of genetic material
Method : evaluation method, for description see main text

Contrary to expectation (Sigurdsson and Banos, 1995), Methods I and II gave similar results. An explanation is that termination of iterations in this study was based on relative changes in parameters and not on a maximum number of rounds of iteration. It appears that the right parameters will be found when using Method II, however, that this method might be more time consuming than Method I. Time advantages from Method I result from fewer rounds of iteration and less time per round. This is indicated by the situation of high correlation (.95), low heritability (.10) and low exchange (15). In that situation none of the samples converged within the required number of rounds for Method II and as a result the estimates are .01 lower than those obtained from Method I.

Instead of looking at the average difference between subset and full data analysis, it is also of interest to observe the difference between individual samples. Table 2 shows the variation that is observed in the estimates obtained for these two situations. From this table it can be seen that only in cases of low exchange and low correlation large difference can be expected in estimates from individual samples.

Table 3. Mean correlation estimate and standard error (se) for the different parameter combinations and methods of analysis.

| | | | Method | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | I | | II | | III | | IV | | V | |
| $r_g$ | $h^2$ | % exch | mean | se | mean | se | mean | se | mean | se | mean | se |
| .70 | .10 | 15 | .72 | .04 | .72 | .04 | .71 | .03 | .71 | .03 | .71 | .03 |
| | | 50 | .74 | .02 | .74 | .02 | .72 | .02 | .72 | .02 | .72 | .02 |
| | .35 | 15 | .69 | .04 | .69 | .04 | .72 | .03 | .72 | .03 | .72 | .03 |
| | | 50 | .70 | .01 | .70 | .01 | .73 | .01 | .73 | .01 | .73 | .01 |
| .95 | .10 | 15 | .93 | .01 | .92 | .01 | .91 | .01 | .91 | .01 | .91 | .01 |
| | | 50 | .97 | .01 | .97 | .01 | .96 | .02 | .96 | .02 | .96 | .02 |
| | .35 | 15 | .94 | .01 | .94 | .01 | .95 | .01 | .95 | .01 | .95 | .01 |
| | | 50 | .94 | .00 | .94 | .00 | .94 | .00 | .94 | .00 | .94 | .00 |

$h^2$ : heritability
$r_g$ : true genetic correlation
% exch. : percent exchange of genetic material
Method : evaluation method, for description see main text

Table 4 shows the estimates of the within country variances when using the various methods. Since results for Methods I and II and Methods IV and V were similar, only result from Methods I, III, and IV are shown. From this table it can be concluded that subset analysis gives severely upward biased estimate of the within country variances. Estimates of within country variances from Method III were comparable to Method I. Since not all data on which selection was based were included in the single country analysis (Method III), estimates were smaller than those from Method I, as expected (Meyer and Thompson, 1984).

Table 4. Mean and standard error (se) of the within country variances for the different parameter combinations using different methods of evaluation.

| | | | | A | | B | |
|---|---|---|---|---|---|---|---|
| $r_g$ | $h^2$ | % exch. | Method | mean | se | mean | se |
| .70 | .10 | 15 | I | 11.68 | .25 | 12.10 | .14 |
| | | | III | 11.63 | .24 | 12.03 | .15 |
| | | | IV | 11.85 | .56 | 12.93 | .41 |
| | | 50 | I | 11.04 | .06 | 11.14 | .11 |
| | | | III | 11.23 | .08 | 10.95 | .08 |
| | | | IV | 11.17 | .05 | 11.35 | .28 |
| | .35 | 15 | I | 33.98 | .52 | 33.02 | .25 |
| | | | III | 33.82 | .53 | 32.81 | .26 |
| | | | IV | 35.60 | 1.32 | 37.93 | 3.28 |
| | | 50 | I | 33.85 | .35 | 33.14 | .37 |
| | | | III | 33.43 | .30 | 32.46 | .41 |
| | | | IV | 35.25 | 1.40 | 37.37 | .93 |
| .95 | .10 | 15 | I | 11.67 | .14 | 11.79 | .25 |
| | | | III | 11.55 | .13 | 11.66 | .27 |
| | | | IV | 11.84 | .50 | 11.66 | .57 |
| | | 50 | I | 11.82 | .20 | 11.45 | .18 |
| | | | III | 11.48 | .19 | 11.14 | .16 |
| | | | IV | 11.30 | .44 | 10.92 | .42 |
| | .35 | 15 | I | 33.44 | .58 | 33.12 | .76 |
| | | | III | 33.18 | .64 | 32.60 | .69 |
| | | | IV | 35.71 | 2.37 | 36.26 | 1.66 |
| | | 50 | I | 32.97 | .74 | 32.71 | .29 |
| | | | III | 32.00 | .72 | 31.57 | .39 |
| | | | IV | 31.35 | .83 | 31.79 | .72 |

$h^2$ : heritability
$r_g$ : true genetic correlation
% exch. : percent exchange of genetic material
Method : evaluation method, for description see main text.

## Conclusions

The approach in which equations are assigned to bulls within country influences the rate of convergence of the estimating procedure. This effect increases with the number of animals with an equation in a country where there is no additional information.

Correlation estimates based on a selected subset of animals have a tendency to be biased downward for situations with low heritability and tend to be biased upward in situations of high heritability.

Estimates of within country variances should be based on complete data. Within country variances based on single country data tend to be smaller than those based upon a multiple country evaluation.

Decisions to terminate iteration should be based on a measure indicating change in parameters.

Method III appears to be a viable alternative to complete data analysis when correlations are expected to be high or when exchange rates of germplasm are high.

## Acknowledgments

This research was made possible with financial assistance of the National Association of Animal Breeders and World Wide Sires.

## Literature Cited

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

Klei, L. 1995. Evaluating Holstein sires for conformation traits using data from the United States and The Netherlands. *Ph.D. Dissertation* Cornell University. Ithaca, New York.

Meyer, K. & Thompson, R. 1984. Bias in variance and covariance components estimators due to selection on a correlated traits. *Z. Tierz. Züchtungsbiol. 101*, 34.

Perez-Enciso, M., Misztal, I. & Elzo, M.A. 1994. FSPAK: An interface for public domain sparse matrix routines. In *Proc. 5th. World Congress on Genetics Applied to Livestock Production.* Vol. 22:87.

Rozzi, P. & Schaeffer, L.R. 1996. New deregression procedure used on type traits. Paper presented at *Interbull workshop*. Nov. 25-26. Verden, Germany.

Schaeffer, L.R. & Zhang, W. 1993. Multi-traits, across country evaluation of dairy sires. *Proc. of the open session of the Interbull annual meeting.* Aug 19-20. Aarhus, Denmark.

Sigurdsson, A. & Banos, G. 1995. Estimation of genetic correlations between countries. *Proc. of the open session of the Interbull annual meeting.* Sept. 7-8. Prague, Czech Republic.