Multiple-Trait Across-Country Evaluations Using Singular (Co)Variance Matrix and Random Regression Model

Esa A. Mäntysaari
MTT Agrifood Research Finland, Animal Production, SF 31600 Jokioinen

1. Introduction

Current Interbull approach for across country evaluation (MACE) assumes records on each country to represent an expression of different trait. In practice this leads into large number of equations since each sire evaluated will, conceptually, have a breeding value for all traits, i.e. countries, although it might have daughters only in one. To be able to solve the across country evaluations a (co)variance matrix of sire effects of the size number of traits in the evaluation has to be known. When current Interbull Holstein run for production traits includes 26 countries, the (co)variance matrix of sire effects involves 325 correlations. Such a large matrix with relatively high correlations is deemed to be close to singular or even non-positive definite and only small inconsistencies can cause higher order partial correlations to be unexpected (van der Beek, 1999).

Several approaches to estimate across country (co)variances have been suggested. Variance component estimation by Sigurdsson *et al.* (1995) was based on straightforward REML by EM algorithm. Klei and Weigel (1998) proposed use of reduced set of equations, where bulls had equations only for countries where they had been used. Instead of EM, Madsen *et al.* (2000) advocated use of AI-REML algorithm, because of its' favorable convergence characteristics. Despite of the method used, the (co)variances have been estimated using subsets of the data. Each subset includes countries that have the best possible connections through common sires. Estimates from multiple REML are then combined into a single (co)variance matrix. If combining is done by simple averages, the final matrix has to be bended to be positive definite. Another alternative could be to use all subset estimates, and combine them iteratively using "summing of expanded part matrices" (Mäntysaari , 1999).

Weigel and Rekaya (1999) suggested to replace the country based trait definitions by production clusters that would be determined by type of production, herd size, climate etc. Miné ry *et al.* (2003) proposed to set up a structure for sire country/trait correlations, so that they could be estimated without running all possible subset REML analyses. However, their approach would still imply 26 by 26 sire (co)variance matrices in the final MACE run.

The objective of the current paper is to present a MACE model based on random regression (RR) coefficients. An advantage of the RR model is that it can be easily modified to use a rank deficient sire (co)variance matrix. On practical evaluations such model requires less computational power. An other advantage comes from relaxed model assumptions. In reduced rank RR model the genetic correlations between country-traits are generated by covariance functions, and are therefore more flexible in modeling high correlations and partial correlations.

2. Method

Define a simple multi trait (MT) model for international evaluations:

$$\mathbf{y}_i=\mathbf{X}_i\mathbf{b}+\mathbf{Z}_i\mathbf{u}_i+\varepsilon_i, \qquad [1]$$

where $\mathbf{y}_i$ is a vector of daughter yield deviations for bull i, $\mathbf{X}_i\mathbf{b}$ vector of country effects, or genetic groups, $\mathbf{u}_i$ is a vector of *t* transmitting abilities of bull i, $\mathbf{Z}_i$ is a matrix that joins the observations to corresponding traits in $\mathbf{u}_i$, and $\boldsymbol{\varepsilon}_i$ is a vector of residuals. If bull has observations for all traits $\mathbf{Z}_i=\mathbf{I}$, or if observation is only for trait k then $\mathbf{Z}_i$ is k'th

row of corresponding identity matrix. In MACE the $var(\mathbf{u}_i)=\mathbf{G}_0$ and $var(\varepsilon_i)=diag(g_{kk}\lambda_k/n_{ik})$, where g_{kk} is the sire variance of the k'th country/trait (diagonal of $\mathbf{G}_0$), $\lambda_k=(4-h_k^2)/h_k^2$ as defined by country providing the data, and n_{ik} is the number of daughters the bull *i* has in country-trait k.

Random regression model

The model [1] can be considered as single trait RR model without modifications. Now the design matrix $\mathbf{Z}_i$ is matrix of covariables, and the vector of sire breeding values are taken as RR coefficients. In [1] the $var(\varepsilon_i)$ was defined as a diagonal matrix, in RR model the same can be visioned as uncorrelated residuals with heterogenous variance.

Reduced rank random regression model

For full rank RR model the variance of sire values $\mathbf{G}_0$ was defined as *t* by *t* matrix. Consider a eigenvalue decomposition:

$$\mathbf{G}_0=\mathbf{V}_0\mathbf{D}_0\mathbf{V}_0^T$$

When $\mathbf{G_0}$ is close to singular, the smallest eigenvalues in $\mathbf{D}_0$ can be set to zero and deleted, and the corresponding columns in $\mathbf{V}_0$ removed. Define **D** as a diagonal matrix with r significant eigenvalues and **V** as a *t* by *r* matrix of corresponding eigenvectors. If t > r, the singular matrix:

$$\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{V}^T$$

approximates the original $\mathbf{G_0}$, but has a rank r. Now we can replace the RR model by its approximate reduced rank model:

$$\mathbf{y}_i=\mathbf{X}_i\mathbf{b}+\mathbf{Z}_i\mathbf{V}\mathbf{V}^T\mathbf{u}_i+\varepsilon_I \quad [2]$$
$$\Leftrightarrow\mathbf{y}_i=\mathbf{X}_i\mathbf{b}+\Phi_I\,\nu_i+\varepsilon_i. \quad [3]$$

In [2] the ν_i correspond to r largest genetic principal components that describe the breeding value of t traits of bull i. The columns of Φ_i represent coefficients of corresponding eigenfunctions. In contrast to sparse $\mathbf{Z}_i$ the covariable matrix is Φ_i dense. If r=t the model [1] MT solutions for breeding values can be back solved as $\mathbf{u}_i=\mathbf{V}\boldsymbol{\nu}_I$, and for if r < t, the back scaled breeding values $\mathbf{V}\boldsymbol{\nu}_t$, are approximately $\mathbf{u}_i$.

Computational technique to solve the RR REML

Estimation of MACE breeding values using [2] should be straightforward with any RR BLUP software that is capable for weighted analysis. In this study goal was to test the suitability of reduced rank RR model in estimation of across country (co)variances. An unquestioned choice to estimate the singular **G** is to first use [2] but keeping the matrix **V** equal to identity matrix. Then, after convergence, decompose **G** into **D** and **V**. More challenging is to start the analysis with subset of traits and subsequently add each trait to the analysis until all traits are modeled with r random regression coefficients.

The latter approach was implemented using standard single trait RR REML program and Unix shell scripts. Only modification to REML program was exclusion of the update of residual variance. The iteration cycle started from initialization step with q traits, q=4, and the first traits were analyzed without rank reduction. Then the iteration continued with transformation steps.

initial step: start with q=4; r=q;

1. Form weights $w_{ik} = \hat{g}_{kk}\lambda_k / n_{ik}$ for each sire
2. Solve new $\hat{\mathbf{G}}$ using RR REML
3. If REML program in step 2 needs more than 1 iteration, return to step 1.

iteration steps:

1. Start with q traits and rank r=q. Decompose current $\hat{\mathbf{G}}$ and form $\hat{\mathbf{V}}$ and $\hat{\mathbf{D}}$. Both are of order q x q
2. Add new trait into analysis, i.e., q=q+1, Correspondingly, add rows of zeros into $\hat{\mathbf{V}}$ and $\hat{\mathbf{D}}$. Let $\hat{\mathbf{V}}_{q,r+1} = 1.0$ and $\hat{\mathbf{D}}_{r+1,r+1} = 1.0$

3. Form weights $w_{ik} = \hat{g}_{kk}\lambda_k / n_{ik}$. Solve new $\hat{\mathbf{D}}$ using RR REML

4. Back solve $\hat{\mathbf{G}} = \hat{\mathbf{V}}\hat{\mathbf{D}}\hat{\mathbf{V}}^T$, Decompose $\hat{\mathbf{G}}$ and form $\hat{\mathbf{V}}$ and $\hat{\mathbf{D}}$. Determine rank r

5. If REML program in step 3 required iterations, return to step 3

6. If more traits to add, return to step 2

3. Example simulation

Example data were simulated using a simple sire model and 10 countries. First 50 unrelated (grand) sires were generated using Holstein (co) variance matrix that was copied from Interbull web site at 1st September 2003 (http://www-interbull.slu.se/eval-/appen023.html). Next, for each sire s, n_{sk}~binomial(25,0.2)) of sons were generated on each of 10 countries. For each son i, a daughter yield deviation for trait *k* was formed as:

$$y_{iks}=\mu_k+sire_{ks}+\lambda_k\sigma_k^2\varepsilon_{ik}/n_{ik},$$

where λ_k was based on heritability in country, a convenient μ_k was formed as $\mu_k=10*\lambda_k$, and ε_{ik}~N(0,1). Values for λ_k were generated uniformly [16,..., 20], corresponding h^2= 0.19,...,0.24. Simulation lead into weak unbalancedness where 12 out of 500 sire*country classes had less than two observations, and the average number of sons per sire per country was 4.96. The number of daughters per son n_{ik} was generated uniformly [3,...,18].

The simulated data was first analyzed with full rank RR model, corresponding ordinary MT REML (MT). Next the rank reduction algorithm in section 2.3 was applied but starting directly from 10 traits (RR10). And finally, rank reduction was started using the first 4 countries, and then sequentially adding one trait at the time until all traits were in the analysis (RR4). Rank reduction was based on eigenvalues of correlation matrices. Small eigenvalues were dropped as long as their sum did not contribute more than 0.05% of the variation.

4. Results

The MT REML algorithm converged unexpectedly well. At the first tests, the efficiency of parameter estimation was so good, that the data generation was repeated using unrealistically small daughter group sizes. Still EM REML did perform without problems: presumable the balancedness of the design and missing selection made analysis easy.

The weights in the analysis were updated only on main cycles. Full convergence required 8 cycles and all together 240 EM iterates. In RR10 the REML cycles were started from converged MT values. At the first iteration of first cycle the rank of $\hat{\mathbf{G}}$ was reduced from 10 to 7 and the iteration stopped because of convergence.

When RR4 was started with four first countries in the analysis the algorithm went downward to rank=2 before new variables were added. Each time a new trait was included algorithm required 4-5 new main cycles, and between 109-120 REML iterates. After convergence the $\hat{\mathbf{G}}$ matrix had rank=6. Two modifications were tested. In the first, the rank was restricted always to stay higher than 4; and in the second, the traits were reordered so that countries entered in to analysis in increasing order of "average correlations" to others. No clear improvement was noticed in either scheme, however results presented here are from the restricted rank analysis.

Figure 1 illustrates the differences on estimates of the country-trait sire variances. In the simulation the Interbull (co)variance matrix was scaled to have unity for country 1. The true sire variances are given by the var(u). Basically there was no differences between estimation methods on recovering the sire variances.

Figure 2 illustrates the differences on estimates of genetic correlations between countries calculated by different approaches. The graph displays the correlations of country 10 with other countries. The line marked **G** illustrates the true values on generated sire effects. The difference between true value and estimated values is notable, but also expected.

All estimated values are much closer to each other than to the true values. MT and RR10 go exactly over each other.

5. Conclusions

The multiple trait MACE model can be considered equivalent to RR MACE model. The latter has an advantage that the sire effects can be modeled using eigenfunctions with desired order. This removes the restriction that the correlations between countries have to be below one. Another advantage is computational efficiency. When bulls would be evaluated for 13 RR coefficients instead of 26 traits almost only half of the equations would be needed. Moreover, there would be room to increase the complexity, i.e., add more countries or more traits per country.

Estimation of sire variances using reduced rank RR models seemed to work well. Approximation using rank=6 gave almost exactly the same solutions as did the full rank MT REML. However, estimates of correlations were slightly different depending on iteration scheme adopted. Results here can not be used to make inferences about the accuracy of estimates since the MT REML tends to underestimate the correlations when number of traits is this high.

In variance component estimation the reduced rank RR allows analysis of more countries simultaneously and thus sub setting of the countries might not be needed. Reduced RR model does not automatically work in independent sub analyses that have been proposed with structural model (Minéry *et al.,* 2003). This because each time new country data is added, it can also affect on estimates from previous country-traits. It might be possible to include always the best connected countries into each subset and then include each country at the time to estimate country-trait eigenfunction coefficients.

Acknowledgement

Nordisk Avelsvärdering supported the traveling costs of the author.

6. References

Klei, B. & Weigel, K.A. 1998. A method to estimate correlations among traits in different countries using data on all bulls. *Interbull Bulletin 17,* 8-14.

Madsen, P., Jensen, J. & Mark, T. 2000. Reduced Rank Estimation of (Co)variance components for International Evaluation using AI-REML. *Interbull Bulletin 25,* 46-50.

Minéry, S., Fikse, W.F. & Ducrocq, V. 2003. Application of a structural model to estimate genetic correlations between countries. *Interbull Bulletin 31,* 175-179.

Mäntysaari, E.A. 1999. Derivation of multiple trait reduced rank random regression (RR) model for the first lactation test day records of milk, protein and fat. *Proc 50th EAAP* 22.-26.8.1999. Zurich, Switzerland.

Sigurdsson, A., Banos, G. & Philipsson, J. 1995. Estimation of international (co)variance components. *Acta Agric. Scand. Sect. A, Anim. Sci. 46,* 129-136.

van der Beek, S. 1999. Exploring the (inverse of the) genetic international correlation matrix. *Interbull Bulletin 22,* 14-20.

Weigel, K.A. & Rekaya, R. 1999. Clustering herds across country boarders for international genetic evaluation. *Interbull Bulletin 22,* 31-37.

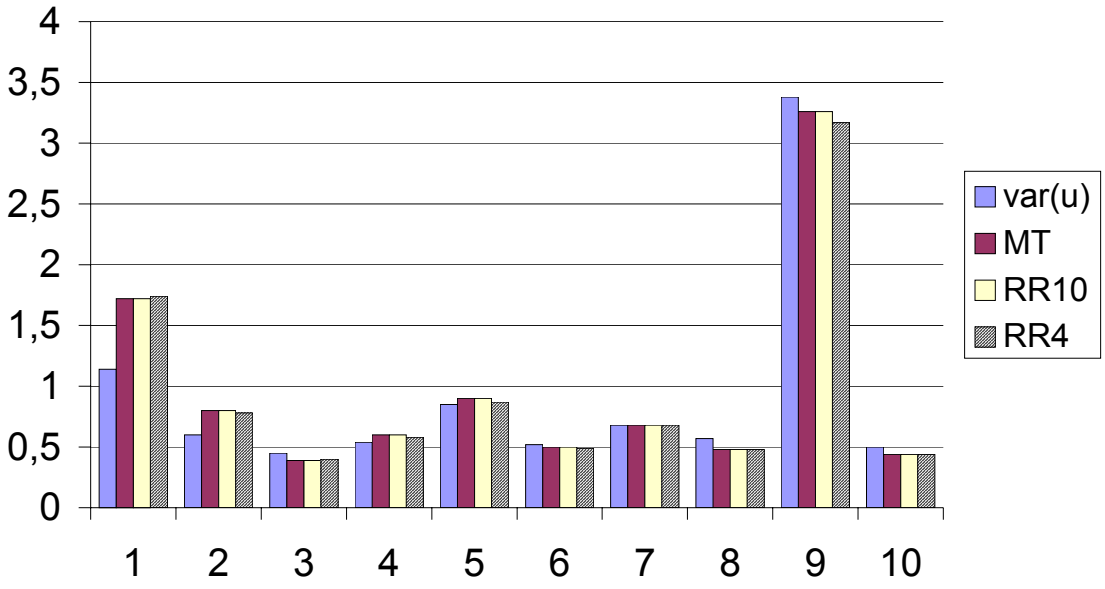

Figure 1. True simulated sire variances (var(u)), and the estimates from multi trait REML (MT), reduced rank random regression from all countries (RR10), and from inclusion of traits sequentially from 4 to 10 (RR4).

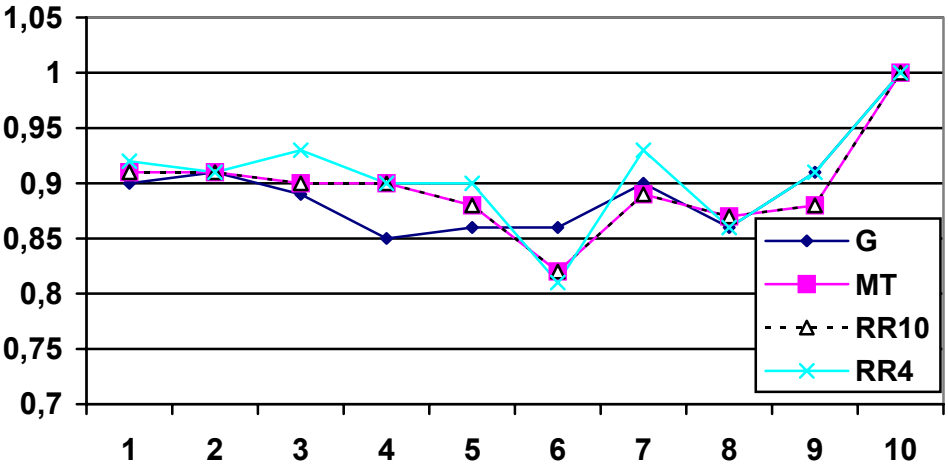

Figure 2. Genetic correlations between each of the traits and the trait 10 from true simulated sire (co)variances (**G**), and the estimates from multi trait REML (MT), reduced rank random regression from all countries (RR10), and from inclusion of traits sequentially from 4 to 10 (RR4).