Estimation of Genetic Correlations Between Countries[1]

Agust Sigurdsson and Georgios Banos
INTERBULL Centre, Box 7023, S-750 07 Uppsala, Sweden.

[1]Paper presented at the INTERBULL open meeting, september 7-8 1995, Prague, Czech Republic

Introduction

Estimation of genetic parameters in a multi-country scenario is a special case of parameter estimation as different traits are observed on different groups of animals. The error covariance between traits is thus assumed to be zero. Currently, international evaluation of dairy bulls are based on national evaluation results. As individual observations for daughters are not used, parameter estimation with conventional methods is not possible.

A procedure based on Expectation Maximization algorithm to produce restricted maximum likelihood estimates of international (co)variance components, using national evaluation results from different countries, was tested with simulation.

The International sire model

The transformed version of the MME for the multi-trait international sire model can be written:

$$\begin{bmatrix} X'R^{-1}X & 0 & X'R^{-1}Z \\ 0 & Q'A^{-1}Q \otimes G^{-1} & -Q'A^{-1}\otimes G^{-1} \\ Z'R^{-1}X & -A^{-1}Q \otimes G^{-1} & Z'R^{-1}Z+A^{-1}\otimes G^{-1} \end{bmatrix} \begin{bmatrix} \hat{c} \\ \hat{g} \\ Q\hat{g}+\hat{s} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ 0 \\ Z'R^{-1}y \end{bmatrix} \quad (1)$$

where y=de-regressed proofs; ĉ=fixed country effect; ĝ=random genetic group effects; ŝ=random bull effects; Q is a matrix that assigns bulls to phantom parent groups; A^{-1} is the inverse of the male relationship matrix; R^{-1} is a diagonal matrix with diagonals equal to the number of daughters of each bull in each country times the inverse of the residual variance for that particular country; G^{-1} is the sire genetic (co)variance matrix of order equal to number of countries; and X,Z are incidence matrices. The present study presents possible ways of obtaining reliable estimates for the G matrix and the residual variances.

REML procedure

Let $\hat{u} = Q\hat{g}+\hat{s}$ and define

$$T = \begin{bmatrix} Q'A^{-1}Q & -Q'A^{-1} \\ -A^{-1}Q & A^{-1} \end{bmatrix} = \begin{bmatrix} T_{gg} & T_{gu} \\ T_{ug} & T_{uu} \end{bmatrix} \quad (2)$$

and

$$W = \begin{bmatrix} X'R^{-1}X & 0 & X'R^{-1}Z \\ 0 & T_{gg} \otimes G^{-1} & T_{gu}\otimes G^{-1} \\ Z'R^{-1}X & T_{ug} \otimes G^{-1} & Z'R^{-1}Z+T_{uu}\otimes G^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} W^{cc} & W^{cg} & W^{cu} \\ W^{gc} & W^{gg} & W^{gu} \\ W^{uc} & W^{ug} & W^{uu} \end{bmatrix} \quad (3)$$

where bulls are ordered within country (trait). For the multi-country international sire model the EM-REML equation for G can be written:

$$G_{ij}^{(k+1)}=\frac{[\hat{u}_i^{(k)}T_{uu}\hat{u}_j^{(k)}-\hat{g}_i^{(k)}T_{gg}\hat{g}_j^{(k)}+tr(T_{uu}W_{ij}^{uu(k)})+2tr(T_{gu}W_{ij}^{ug(k)})+tr(T_{gg}W_{ij}^{gg(k)})]}{q} \quad (4)$$

and an approximation for the error variance:

$$R_{ii}^{(k+1)}=G_{ii}^{(k+1)}\lambda_i \quad (5)$$

for i =1,...,c and j =i,...c; where c is number of countries; q is total number of bulls; k is the iteration round; tr( ) is the trace operator; λ is the assumed environmental to sire variance ratio σ_e^2/σ_s^2 according to the heritability estimated or assumed in each country.

Simulation

The suitability of the above methodology was tested with stochastic simulation. Individual performance records were simulated for two dairy cattle populations (A, B) covering 10 generations. First lactation records were simulated according to genetic standard deviation (σ_g) 320 kg and heritability (h^2) 0.25 for both populations and genetic correlation (r_g) of .90 between the two populations. Systematic exchange of bull parents between countries was allowed resulting in well connected populations. Animal model genetic evaluation systems were implemented in both countries and within and across country selection was being practiced. After 10 generations the data for each population numbered 66000 cows and 600 bulls where 54 bulls had daughters in both countries; 60 bulls had full-sib brothers in the other country; and the remaining bulls had at least 1 half-sib brother in the other population.

National evaluations of bulls were first de-regressed within country and sire variances were estimated with the approximate EM-REML procedure. The within country variances were estimated based on all data and data where bull proofs from generation 1 to 5 were omitted. All relationship information was, however, included in both cases. The same procedure for estimating genetic correlations was then tested on a joint data comprising de-regressed evaluations from both countries. The impact of excluding national evaluations of imported bulls on estimated genetic correlations, as well as the effects of bias in import evaluations and weak genetic ties between the populations were studied. Bias was introduced by multiplying the estimated breeding values of exchange bulls in the importing country by (1+b) where b was picked from the standard normal distribution with a mean of 0.05-0.15 and range ±0.05.

Weak data connectedness was reached by reducing genetic relationship among bulls from different populations. This was done by simulating the pedigree structure in such a way that only 50 % of the bulls had half-brothers in the other country but the same number of bulls with multiple proofs and full-sib brothers, as before, was kept. Further, 15% and 20% respectively of sires and maternal-grand-sires of the bulls lacking sib in the other country was randomly replaced by phantom group. When ties were weak, genetic correlations were also estimated based on subsets of 'well connected' data including bulls with evaluations in both countries, full-sib families with members in both countries, and ancestors.

Within country variances

Genetic standard deviations estimated within country with the approximate EM-REML procedure are shown in table 1. The estimates based on all data (All) were in close agreement with the true base population variances. If national proofs from generation 1 to 5 were omitted but all relationship information included (Cut) the base population genetic variance was underestimated by 5 %. Genetic standard deviation estimate, based on the pooled geometric mean of standard deviations of national and

de-regressed bull evaluations was 13% lower than the true genetic variance.

Table 1. Estimated within country genetic standard deviation for the simulated data by REML[1], as geometric mean of national proofs and de-regressed proofs (GEO)[2] and the true base population genetic standard deviation in respective country (TRUE)[3] (SE≈1.5).

Country	REML		GEO	TRUE
	All	Cut		
A	162.3	153.5	140.8	161.5
B	160.8	151.6	141.0	161.0

[1]REML: Estimates from the approximate EM-REML procedure using all data (All) or omitting national proofs from generation 1-5 but including all pedigree information (Cut)
[2]GEO: Pooled average of geometric mean of standard deviations of national proofs and de-regressed proofs calculated within year
[3]TRUE: True base population genetic standard deviation

Genetic correlations

Estimates of genetic correlation between the two simulated populations, considering several investigation factors are shown in table 2.

Table 2. Impact of including/excluding biased/unbiased national evaluations of imported bulls and effect of weak ties between the two populations on estimated genetic correlation (SE ≈ .02).

Import proofs		Relationship	Data	Estimated r_G
Included	Mean Bias	ties	considered	(true r_G = .9)
Yes	0 %	Strong	All	.900
No	0 %	Strong	All	.780
Yes	5 %	Strong	All	.899
Yes	10 %	Strong	All	.893
Yes	15 %	Strong	All	.882
Yes	0 %	Weak	All	.820
Yes	0 %	Weak	Well connected[1]	.896

[1]Well connected data: Bulls evaluated in both countries, full sib families with members in both countries, and ancestors.

When all data with strong ties was included in the analysis the estimated genetic correlation was the same as the true value. This attests to the suitability of the method under ideal circumstances. When national evaluations of imported bulls were excluded the genetic correlation was underestimated by 12 %. So even in this well balanced case, with strong genetic ties between populations, direct ties appeared essential for estimating the genetic correlation.

When 5 to 15 % average bias was introduced into national evaluations of imported proofs in one of the countries and genetic correlations were estimated including these biased proofs, results were slight underestimates. The effect was not overwhelming and even with a 15 % bias in import proofs only a 2 % bias was detected in the genetic correlation estimate.

When genetic correlations were based on all data available and genetic ties between populations were weak, the result was underestimate by 8 %. However, when the estimation was based on a well connected subset of the otherwise loosely connected data, the genetic correlation estimate was close to the true value. In the latter case, the well connected subset consisted of bulls evaluated in both countries, full-sibs with members in both countries, and ancestors.

Application with field data

The methodology was tested on real data consisting of Holstein dairy bull records from five countries. The countries chosen were Germany (DEU), France (FRA), Italy (ITA), the Netherlands (NLD) and the United States of America (USA)

Estimates of within country standard deviation from the approximate EM-REML procedure are listed in table 3 for the five countries studied. Estimates based on the geometric mean of within year variance of national and de-regressed evaluations are also listed. The latter method was the past choice in international evaluations. The sire standard deviation computed from reported population variances by each country are also listed in the table.

Table 3. Estimated within country sire standard deviation for milk yield by REML[1], geometric mean of national evaluations and de-regressed proofs (GEO)[2] and half the population genetic standard deviation reported by each country (POP)[3]. Units are kilograms for all countries.

Country[4]	REML	GEO	POP
DEU	253	239	271
FRA	323	315	320
ITA	287	264	290
NLD	263	255	260
USA	344	309	325

[1]REML: Estimates from the approximate EM-REML procedure
[2]GEO: Pooled average of geometric mean of national evaluations and de-regressed proofs calculated within year
[3]POP: Estimated population sire standard deviation supplied by individual country
[4]DEU=Germany;FRA=France;ITA=Italy;NLD=the Netherlands;USA=United States of America

REML estimates were in all cases higher than the approximation based on geometric mean. This agrees with the results from the simulation. For France, Italy and the Netherlands the REML estimates were very similar to the population parameters reported by each country. For Germany the REML estimate was 6% lower and for USA 6% higher than the population parameter. Population parameters, however, were differently derived in each country and may in some cases not reflect the true genetic standard deviation of the base population.

Genetic correlation estimates considering data sets that excluded or included import evaluations as well as well connected data subsets are in table 4. Estimates between Italy and the Netherlands and between Italy and USA are used for illustration but the pattern was similar for all two- country combinations.

Table 4. Effects of including or excluding national evaluations of imported bulls and using only well connected subset of the data on estimated genetic correlation (r_G)

Pair	Import proofs included	Data considered	Estimated r_G
ITA-NLD	No	All	.64
	Yes	All	.87
	Yes	Well connected[1]	.94
ITA-USA	No	All	.77
	Yes	All	.95
	Yes	Well connected	.96

[1]Well connected data: Bulls evaluated in both countries, full sib families with members in both countries, and ancestors.

Similarly to simulated data, the importance of direct ties in estimation of genetic correlation was clearly reflected. If direct ties (national evaluations of common bulls) were excluded from the analysis the

genetic correlations were heavily affected. Also in agreement with the simulation, genetic correlation estimation based only on bulls with multiple evaluations, full-sib families with members in both countries and ancestors were higher than estimates based on all data. Differences between the latter varied across country pairs, depending on the degree of data connectedness. It is not possible to conclude that the highest obtained value is the true genetic correlation but according to the simulation study, it ought to be closer to it. The two country pairs in table 4 were picked as they represented the two extremes regarding number of bulls in common. Italy and the Netherlands had the least and Italy and USA the most bulls in common among all country pairs. The estimated genetic correlation between Italy and the Netherlands based on the well connected subset was 7% higher than based on all data. The corresponding difference in the case of Italy and USA was only 1%.

Genetic correlations were estimated between all five countries considering a well connected subset of the available data. This was done both in two-country scenarios considering all possible combinations and for all five countries simultaneously. Results are listed in table 5. The table also shows product moment correlations of national evaluations of bulls evaluated in both countries. Estimated genetic correlations were, as expected, higher than proof correlations. The difference was consistent for the country combinations studied e.g.. the highest genetic correlation estimate was associated with the highest proof correlation. Only trivial differences between the bi- and multi-country genetic correlation estimates were observed. Multi-country estimates should have smaller sampling errors due to more data. These estimated genetic correlations did not appear to be sensitive to the number of countries included in the analysis.

Table 5. Estimated genetic correlations from a bi-country (r_G-bi) and multi-country (r_G-multi) approximate EM-REML and national evaluation correlations (r_{PROOF})[1].

Combination[2]	r_G-bi	Bulls in subset[3]	r_G-multi based on 6560 bulls	r_{PROOF}
DEU-FRA	.91	1110	.91	.81
DEU-ITA	.91	700	.91	.83
DEU-NLD	.97	983	.96	.90
DEU-USA	.89	1042	.90	.83
FRA-ITA	.95	887	.95	.86
FRA-NLD	.91	1391	.92	.88
FRA-USA	.97	2042	.97	.92
ITA-NLD	.94	620	.92	.86
ITA-USA	.96	1347	.96	.90
NLD-USA	.93	1039	.92	.87

[1] r_{PROOFS}: Product moment correlation of national evaluations for bulls with proofs in both countries. Calculated within birthyear of bulls and presented as pooled average.
[2] DEU=Germany;FRA=France;ITA=Italy;NLD=the Netherlands;USA=United States of America
[3] Total number of bulls in each bi-country subset

In general all genetic correlation estimates listed in table 5. were high and ranged from .89, in a bi-variate analysis of German and USA data, to .97 between France and USA. Looking at individual estimates, some particular patterns were detected. France, Italy and USA were all very highly correlated and the estimate between Germany and the Netherlands was also high. The estimated genetic correlations between these two groups of countries were on the other hand lower, particularly the correlations with Germany.

Conclusions

Estimation of genetic parameters within and across country, based on national bull evaluation results, is possible with the approximate EM-REML procedure presented in this paper. Direct ties between populations in form of bulls evaluated in more than one country must exist to assure estimability of genetic correlations. Even when imported proofs are biased the benefit of including them in the analysis outweighs the negative impact of bias on genetic correlation estimation. Further, if the genetic ties between the populations are weak, data for genetic correlation estimation should be restricted to a well connected subset comprising bulls with evaluations in several countries and full-sib families with members in more than one country. Further theoretical research is needed to quantify the minimum amount of ties needed for genetic correlation estimation with international data.