A structural model for the matrix of genetic correlations between countries in international evaluations.

I. Delaunay, Vincent Ducrocq, Didier Boichard

A STRUCTURAL MODEL FOR THE MATRIX OF GENETIC CORRELATIONS BETWEEN COUNTRIES IN INTERNATIONAL EVALUATIONS

I. Delaunay, V. Ducrocq and D. Boichard

INRA, Station de Génétique Quantitative et Appliquée 78352 Jouy en Josas Cédex, France

INTRODUCTION

Since 1996, Interbull (http://www.interbull.org) has been routinely calculating international breeding values of dairy bulls from their national proofs. With the steady increase in international exchanges of breeding material, these international evaluations have become a decisive tool in selection schemes. In November 2001, 25 countries were subscribing to the service. The current Interbull procedure includes the de-regression of national estimated breeding values (EBVs) followed by the analysis of these deregressed proofs to estimate genetic correlation between countries and to evaluate each bull in each participating country, using a MACE (Multitrait Across Country Evaluation - Schaeffer, 1994) approach.

In the Holstein breed, 27 populations were jointly evaluated (Jordani, 2001), requiring the knowledge of 27 heritabilities and 351 genetic correlations. These were computed using a Restricted Maximum Likelihood approach and an Expectation-Maximisation algorithm (EM-REML - Sigurdsson *et al.*, 1995). Computational constraints related to REML and EM prevents the simultaneous estimation of all genetic correlations. Instead, these were obtained combining at most 10 populations at a time and in many cases (127 out of 351), indirect or adhoc approximations were used (see the Interbull website). The main reasons for poor estimations of genetic correlations were : i) weak genetic links between countries, ii) genetic correlations close to the border of the parameter space (they were all between 0.76 and 0.96) and iii) extremely slow convergence of the EM algorithm. The problem of imprecise estimates of genetic correlations is likely to worsen with new countries joining Interbull. Another estimation strategy must be found to maintain confidence and credibility.

Among other alternatives, Rekaya *et al.* (1999) proposed a structural (parsimonious) model for genetic covariances: in their model, each covariance was described as the sum of an average value for all covariances and a multiple regression on measures of genetic, management and climate similarities between countries.

The approach proposed here is another structural model, free of assumptions on the exact nature of the characteristics responsible for correlations less than unity. The model should allow the estimation of parameters close to (or even on) the border of the parameter space (e.g., correlations very close or equal to 1). This is required for the use of more efficient estimation algorithms. In particular, it has been found that the AI (Average Information) - REML algorithm (Johnson and Thomson, 1995, Gilmour *et al.*, 1995), a second order maximisation algorithm, is much faster than EM-REML and leads to asymptotic standard errors of the estimates, but can fail when iterative solutions go out of the parameter space.

This work is part of PROTEJE (Production Traits European Joint Evaluation – Canavesi *et al.*, 2002), a European project designed to study potential improvements to the current Interbull methodology.

MATERIAL AND METHODS

Structural model. The main problem with Rekaya's structural model (Rekaya *et al.*, 1999) is the arbitrariness of the similarity measures used. The objective definition of a relevant and flexible management or a genetic similarity measure seems hopeless. The basis of the reparameterisation proposed here is to let the data define the most discriminant measures. To illustrate the approach, consider four countries and three unobservable (sets of) characteristics (x_{i1}, x_{i2}, and x_{i3}) for each country i that condition genetic correlations between countries. These characteristics can be represented as a point in a 3-dimensional space. Let $P_i= [x_{i1} \quad x_{i2} \quad x_{i3}]'$. The proposed structural model defines each genetic correlation (between countries i and j) as a function of the distance between points (P_i and P_j). If the euclidian norm is chosen, $d_{ij} = d(P_i, P_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$. To ensure that genetic correlations are less or equal to 1, define :

$$\rho_{ij} = \exp \{-d_{ij}\} \qquad [1]$$

The model being overparameterised (there is an infinite set of 4 points with the same distance between them), one can fix one point at the origin (P_1= [0 0 0]') , the second one along the first axis (P_2= [x_{21} 0 0]') and so on: P_3= [x_{31} x_{32} 0]', P_4= [x_{41} x_{42} x_{43}]'.

Let v_i be the genetic standard deviation in country i. All the elements of the matrix of genetic (co)variances G are functions of the v_i's and d_{ij}'s (with, obviously d_{ii} = d(P_i, P_j) = 0). For all i, j:

$$g_{ij}= v_i v_j \exp \{-d_{ij}\}$$

As with the original parameterisation, there are n(n+1)/2 unknowns for a matrix G of dimension n: n genetic variances and n(n-1)/2 genetic coordinates x_{ij}. The point estimates have an appealing intuitive interpretation. If these points are close for two countries, the corresponding genetic correlation is close to 1. If the two points are located at the same place, the correlation is 1, which is now within the parameter space. Although we don't have a formal proof at this stage, it was observed that correlation matrices defined using [1] are always (semi-) positive definite, while negative definite matrices cannot be described in this way.

More parsimonious models can be easily derived: if only 2 (respectively 1) unknown characteristics are sufficient to explain differences in genetic correlations, expression [1] can still be applied, with points P_i chosen on a plane : P_1= [0 0]'; P_2= [x_{21} 0]'; P_3= [x_{31} x_{32}]'and P_4= [x_{41} x_{42}]' (respectively on a line: P_1= [0]'; P_2= [x_{21}]'; P_3= [x_{31}]'and P_4= [x_{41}]'). The proper dimension can be assessed using likelihood ratio tests. Reduced rank genetic matrices are obtained imposing constraints on the P_i's. An example is when P_i is set equal to P_j, for some i,j. A limitation however is that negative correlations cannot be represented. This is not a problem when very similar traits are studied (as for lactation yield) but is more troublesome when correlated traits (yield and persistency for example) will be analysed together.

AI-REML. The joint estimation of the genetic variances and the P_i's can be implemented using an AI-REML algorithm. The main difficulty lies on the fact that the G matrix is not a linear function of the parameters to be estimated and the strict calculation of the matrix of second derivative of the log restricted likelihood is extremely complex. Instead, Gilmour *et al.* (1995) proposed to use a simplified average information matrix, ignoring non-zero terms of the second derivatives of G. This strategy is implemented in the ASREML software (Gilmour *et al.*, 2000)

Illustration : simulated data sets. A hypothetical dairy cattle "international" population was created. Data covered a period of 5 generations in four countries. For each generation and each country, milk yield of 2560 cows (total: 51200) was simulated according to the model:

$$y_{ijk} = b_k + a_{ij} + e_{ijk} \qquad [2]$$

where y_{ij} is the record of a cow i in herd k of country j; b_k is a herd effect (25 herds per country and generation) with b_k~N(0,1); a_{ij} is the true additive genetic value of cow i in country j with a_{ij}= ½ a_{sire} + ½ a_{dam} + ϕ_{ij} ; ϕ_{ij} being the cow's mendelian sampling ; $\mathbf{a}=\{a_{ij}\}\sim N(0, G\otimes A)$ and e_{ijk} is the residual with $\mathbf{e}\sim(0, I\otimes I\,\sigma^2_e)$.

Parents of the next generation were selected after a simulated international evaluation (MACE) at each generation. The male population was structured with, at each generation, 40 young sires with 96 daughters each, 16 proven sires with 400 daughters each and 8 sires of sons. In this illustrative example, connectedness was ensured through the use of each bull in each country. genetic and residual variances were σ^2_a =0.25 and σ^2_e =0.75 for all countries. Finally, the correlation matrix was chosen (here) to have a uniform pattern: $\rho_{ij} = \rho$ =0.9 (case A) and 0.99 (case B) for all countries i and j,. REML estimates of all parameters were obtained assuming an animal model and using the AI-REML algorithm (with the ASREML software) either with an "unstructured model" or structural models where the P_i's were in a 1, 2 or 3 dimensional space.

RESULTS AND DISCUSSION

AI-REML converged quickly (6 to 7 iterations). As expected (same number of parameters), the maximum log-likelihood value and the estimated genetic correlations were identical with the unstructured model and the structural model in dimension 3 when the common correlation between countries was 0.9 (tables 1 and 2). However, a more parsimonious model (structural model in dimension 2) gave results that are not statistically different from the unstructured model: a likelihood ratio test would accept the structural model in dimension 2. When the P_i's for all countries i and j, were forced to be on a line, the model was very bad. This is not a surprise here: all correlations between equal, one expects the points to be approximately equidistant, which is difficult on a straight line !

When ρ=0.99, the smallest three eigenvalues of G are equal to 0.01 (instead of 0.1 for ρ=0.9): the correlation matrix is much closer to the border of the parameter space. Then, the unstructured model failed completely: the iterative steps of AI-REML took the correlation matrix out of the parameter space and could not find its way back. In contrast, the structural models converged and the model in dimension 1 appears acceptable (table 2).

Table 1. Maximum log-likelihood value

	$\rho = 0.9$	$\rho = 0.99$
Unstructured Model	-24887.7	Did not converge
Structural model – dimension 3	-24887.7	-24543.5
Structural model – dimension 2	**-24888.2**	-24543.7
Structural model – dimension 1	-24914.1	**-24544.9**

Table 2. Estimates of the coordinates of the P_i's and of genetic correlations

	Coordinates of				lower triangle of the genetic correlation matrix		
	P_1	P_2	P_3	P_4			
$\rho = 0.9$; Structural model	0	0.117	0.057	0.102	0.890		
dimension 3	0	0	0.094	0.042	0.895	0.894	
	0	0	0	0.073	0.876	0.918	0.904
$\rho = 0.9$; Structural model	0	0.109	0.040	0.131	0.896		
dimension 2	0	0	0.097	0.073	0.901	0.887	
					0.861	0.926	0.909
$\rho = 0.99$; Structural model	0	0.000	-0.008	-0.006	1.000		
dimension 2	0	0	0.000	0.000	0.992	0.992	
					0.994	0.994	0.998
$\rho = 0.99$; Structural model	0	0.004	-0.007	-0.020	0.996		
dimension 1					0.993	0.989	
					0.980	0.984	0.987

CONCLUSION

The two examples presented here display some of the potential advantages of the structural model proposed : it cancels several problems associated with the estimation of genetic correlations required for international evaluations : estimates on (or very close to) the border of the parameter space can be obtained. Second order REML algorithms do not fail any more. Convergence is faster and asymptotic standard errors of estimates are available. More parsimonious models (including those leading to reduced ranked genetic correlation matrices) can be proposed and tested. Similarities between countries are easily visualised.

Within the PROTEJE project (Canavesi *et al.*, 2002) and in connection with Interbull, we will now study more deeply this type of structural models on more complex situations. In particular the influence of weak connections between countries must be assessed.

REFERENCES

Canavesi, F., Boichard, D., Ducrocq, V. *et al.* (2002) *These proceedings.*

Gilmour, A. R., Thomson, R. and Cullis, B. R. (1995*) Biometrics* **51** : 1440-1450.

Gilmour, A. R., Cullis, B. R., Welham, S.J.,Thomson, R. (2000) ASREML Reference Manual.

Johnson, D.L. and Thomson, R. (1995) *J. Dairy Sci.* **75** : 449-456.

Jorjani, H. (2001) *Interbull Bull.* **27** : 84-89.

Rekaya R., Weigel K. and Gianola, D. (1999) *Interbull Bull.* **22** : 25-30.

Schaeffer., L.R. (1994) *J. Dairy Sci.* **77** : 2671-2678.

Sigurdsson, A., Banos, G. and Philipsson J. (1996) *Acta Agric. Scand.* **46** : 129-136.