Application of a Structural Model to Estimate Genetic Correlations between Countries

S. Minéry[1], W.F.Fikse[2] and V. Ducrocq[3]

[1] *Département de Génétique, Institut de l'Elevage, INRA-SGQA, 78352 Jouy en Josas Cédex, France*

[2]*Interbull Centre, Department of Animal Breeding & Genetics, Swedish University of Agricultural Sciences, Box 7023, S - 750 07, Uppsala, Sweden*

[3] *INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy en Josas Cédex, France*

Introduction

One of the major problems of international genetic evaluations is the estimation of genetic covariances. The use of structural models has been suggested to exploit patterns in the genetic correlations matrix and to reduce the number of parameters to estimate.

The structural model project of Delaunay *et al.* (2002) is a part of PROTEJE. It proposed to use the countries themselves to characterize genetic correlations between them, instead of external information as in the structural model of Rekaya *et al.* (2001). The genetic correlations between two countries are defined as an exponential function of the Euclidian distance between these countries. In this structural model, (k+1) countries can be represented in a k-dimensional space. The reduction of the number of dimensions of the space allows to reduce the number of parameters that has to be inferred. For example, in a 3-dimensional space only 72 coordinates need to be estimated to compute 325 genetic correlations between 26 countries.

The structural model of Delaunay *et al.* (2002) was successfully tested on simulated data, and on the current genetic correlations matrix used by Interbull. The aim of the present study was to apply the structural model for estimation of genetic correlations on field data, and to compare these results with the corresponding estimates obtained with an "unstructured" model. An additional aim was to study the possible use of the coordinates of different structural models related to the same space, to calculate directly genetic correlations, without having to estimate them.

Material and Methods

Data available were deregressed national breeding values of bulls and their effective daughter contributions (EDC) used for Holstein milk yield international genetic evaluation of February 2003 (26 populations, here after referred to as countries). Different subsets of countries were considered.

The sire model currently used in international genetic evaluations was applied (Schaeffer, 1994). Each observation was weighed by the EDC of the bull. Genetic groups were considered as fixed effects and based on selection path*year of birth*origin. Small groups were merged, first by origin, and then by year of birth. Minimum group size was 500 bulls to avoid problems with inference of small group sizes when group effects are assumed fixed.

Two different models were used for the genetic covariances: the structural model (SM) and an "unstructured" model, called classical model (CM). In both cases, the residual variances and the genetic variances within countries were estimated. In the classical model the genetic covariances across countries were estimated. In the structural model the coordinates for each axis were estimated for each country.

The choice of countries to define axes for the structural model was based on results of cluster analyses of Weigel and Zwald (2002). The aim was to select axes countries that represented distinct production environment. The Netherlands defined the origin of the space, USA the first axis, New Zealand the second axis and Hungary the third axis, unless mentioned otherwise.

AI-REML algorithm implemented in the program AIREMLUPG (Druet *et al.,* 2003) was used for parameters estimation. It also provided the asymptotic standard errors of the estimates. For the structural model, the AI-REML algorithm used a simplified average information matrix, ignoring non-zero terms of the second derivative of the genetic (co)variance matrix that occur when covariances are a non-linear function of the parameters. (Gilmour *et al.,* 1995). For the classical model, the update of the genetic covariance matrix combined the AI with an EM, if a pure AI update yielded parameters outside the parameters space (Jensen *et al.,* 1997). For the structural model, the update was based on a line search procedure in which the step size was repeatedly divided by two until the likelihood increased (Dennis and Schnabel, 1983).

The structural and the classical models were compared using the estimated genetic correlations, minus two times the logarithm of the likelihood function and two information criteria that take into account the number of parameters to estimate: Akaike's Information Criterion (AIC) and Schwarz' Bayesian Information Criterion (BIC). The total number of deregressed national breeding values was used to compute BIC. BIC put more penalties on the number of parameters than AIC (Wolfinger, 1993).

Results and Discussion

1. Comparison of structural and classical models (SM and CM).

Genetic correlations between five countries (the four axes countries and Denmark) were estimated with three different models: the classical model (CM5), a structural model with four dimensions (SM45) and a parsimonious structural model, with three dimensions (SM35). Total number of observations was 32730, for 31009 bulls and 33 genetic groups.

SM45 and SM35 gave the same estimates of the genetic correlations and coordinates (not shown), and had the same -2logL (Table 1). Thus, information criteria were lower for SM35, since it had one parameter less.

Table 1. Number of parameters, -2logL and information criteria for CM5, SM45 and SM35.

	CM5	SM45	SM35
No. parameters	20	20	19
-2log L	**424709.3**	424717.0	424717.0
AIC	**424749.3**	424757.0	424755.0
BIC	424917.1	424924.8	**424914.4**

Table 2. Estimated genetic correlations from SM35 and their standard errors (above diagonal), deviations (SM35-CM5) in estimated genetic correlations and deviations in standard errors (below diagonal).

	NLD	USA	NZL	HUN	DNK
NLD		0.927	0.779	0.857	0.955
		(0.006)	*(0.016)*	*(0.015)*	*(0.007)*
USA	0.002		0.724	0.886	0.968
	-0.001		*(0.012)*	*(0.020)*	*(0.011)*
NZL	-0.005	0.017		0.685	**0.747**
	0.011	*0.006*		*(0.026)*	*(0.019)*
HUN	-0.001	-0.001	-0.005		0.872
	-0.004	*0.004*	*0.008*		*(0.020)*
DNK	-0.001	-0.006	**-0.056**	-0.006	
	-0.010	*-0.005*	*0.007*	*-0.007*	

Table 3. Number of bulls with records in the country (on the diagonal) and number of common bulls (above the diagonal).

	DNK	FIN	FRA	NLD	USA	NZL	AUS	HUN
DNK	4538	15	82	135	110	85	77	91
FIN		607	26	21	19	15	10	14
FRA			8037	253	420	146	180	120
NLD				6461	622	357	257	172
USA					17458	431	484	289
NZL						2997	412	147
AUS							3544	110
HUN								1276

AUS: Australia; DNK: Denmark; FIN: Finland; FRA: France; HUN: Hungary; NLD: The Netherlands; NZL: New Zealand.

Genetic correlations estimated by SM35 were very close to those estimated with CM5, except the one between Denmark and New Zealand that differed by almost 0.06 (Table 2). This correlation was based on the lowest number of common bulls in this subset of countries, which could explain why it was less stable than the others (Table 3). The higher -2logL observed for SM35 was compensated by the reduction of parameters as shown by lower BIC criteria.

CM5 and SM45 gave different estimates of genetic correlations, although SM45 had the same number of parameters. CM5 had a better fit than SM45 as indicated by the lower -2logL. One explanation could be that SM45 imposed more constraints on the genetic correlations. Delaunay *et al.* (2002) had

already mentioned this problem for a spatial representation of the correlations. For example, three countries A, B and C could be represented in 2 dimensions. Correlations between A and B could be transformed into a distance D_{AB}, using the definition of the correlation in the structural model. Similarly, the other correlations determine D_{AC} and D_{BC}. But if the sum of D_{AC} and D_{BC} is less than D_{AB}, then a spatial representation of these genetic correlations is impossible. The triangle (A,B,C) can not be formed. This was observed for the genetic correlations estimated with CM5; they could not be converted into coordinates in a 4-dimensional space.

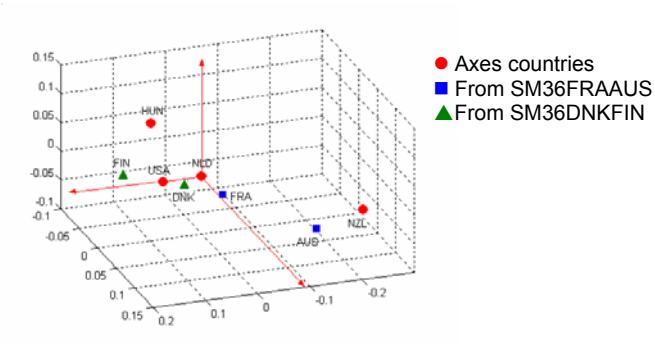

Figure 1. Combination (CALC) of the coordinates obtained from SM36FRAAUS and SM36DNKFIN.

2. *Use of the estimated coordinates to calculate the correlations.*

Two sets of countries were analyzed with the structural model. The first set included the 4 axes countries plus France and Australia (SM36FRAAUS: 39773 observations). The second set included the same axes countries plus Denmark and Finland (SM36DNKFIN: 33337 observations). The two sets of coordinates obtained were combined. Coordinates between countries in this common space were used to calculate distances and genetic correlations (CALC) (Figure 1). Thus, correlations between France-Denmark, France-Finland, Australia-Denmark and Australia-Finland were obtained without having to estimate them.

The eight countries were also analyzed jointly with the structural model (SM38: 44918 observations) and with a classical model (CM8).

Genetic correlations calculated (CALC) and estimated with SM38 were nearly the same (Table 4). In SM38, the number of ancestors was larger than in each SM36, which could have created additional genetic links between countries and could explain the differences with CALC.

Table 4. Genetic correlations (standard errors) calculated from SM36 (CALC) or estimated with SM38 and CM8.

		DNK	FIN
FRA	CALC	0.930	0.825
	SM38	0.930 (*0.012*)	0.833 (*0.024*)
	CM8	0.941 (*0.003*)	0.798 (*0.009*)
AUS	CALC	0.810	0.729
	SM38	0.812 (*0.013*)	0.733 (*0.025*)
	CM8	0.823 (*0.007*)	0.643 (*0.013*)

Table 5. -2logL and information criteria for CALC, SM38 and CM8.

	CALC	SM38	CM8
-2logL	572423.7[a] 572422.7[b]	572419.9	**572389.0**
AIC	-	572487.9	**572477.0**
BIC	-	**572783.8**	572859.9

a. calculated with residual and sire variances obtained from the two SM36
b. calculated with residual and sire variances obtained from SM38

For CALCa the residual and sire variances obtained from the two SM36 analyses were used to compute -2logL while for CALCb those estimated with SM38 were used (Table 5). Thus, the difference between CALCa and CALCb could be attributed to the different residual and sire variances. CALCb and SM38 had the same residual and sire variances, but -2logL were different due to the differences in the genetic correlations. BIC favoured SM8 in comparison to CM8 because of the decrease of the number of parameters to estimate from 44 with CM38 to 34.

3. General Discussion

Selection of countries to define the axes, and number of dimension for the structural model are important issues. Information criteria could be used to determine the best combination of axes countries. When information criteria values are the same, the structural model with axes countries that defined the larger volume in the space gave more precise estimated coordinates (see Minéry (2003) for an illustration). This volume could also be considered to select the best axes countries.

Furthermore, it seems reasonable to include at least one well-connected country among the axes countries, like the USA. Genetic correlations involving poor connected countries are estimated with too little precision for such countries to be chosen as axes countries.

This study shows that the structural model could allow a drastic decrease of the number of runs to estimate the correlations between countries. Only coordinates in the space defined by the axes countries would need to be estimated. But before using this method, it would be necessary to ensure that new participating countries are correctly represented in the space defined by the axes countries chosen.

Most of the correlations estimated with SM or CM were similar or lower than those used by Interbull in February 2003. This is partly due to differences in data selection, procedures used by Interbull to estimate correlations for poor connected countries, and difference in choice of residual variances (Interbull uses heritabilities provided by the countries).

Estimation of the parameters with the structural model needed 0.6 Gb of memory and took some hours on a Intel Xeon 2.8 GHz computer for 5 countries, to 2.1 Gb and a half day for 8 countries. It is feasible to apply in test run. The number of iterations with the structural model was usually higher than with the classical model. Some improvements of the algorithm can be done with respect to the rules of convergence and the way to change the step size.

Finally, the structural model could be tested on other traits, like conformation traits which are less correlated across countries than production traits. The best combinations of countries to define axes would not be necessarily the same as for milk yield. The main question will be to know how much the differences between genetic correlations estimated with the structural model and with the classical model are acceptable compared to the benefit of reducing the number of parameters.

Conclusion

The structural model of Delaunay *et al.* (2002) applied on Holstein milk yield international data and using an AI-REML algorithm was able to explain genetic covariance between countries. These examples show that reduction of the number of parameters is possible with the structural model. This reduction compensated the constraints of the spatial representation. These results are promising and much more drastic reduction of the number of parameters could be planned in the future. Moreover, the use of the coordinates of the countries to calculate the genetic correlations could reduce the number of runs (thus the time) needed to determine the all genetic (co)variance matrix.

Determination of the best axes countries and the optimal number of axes needs to be investigate further, using information criteria and regarding the volume of the space defined by the axes countries. An efficient procedure to select the best axes countries with the optimal number of axes countries should be defined and tested.

References

Delaunay, I., Ducrocq, V. & Boichard, D. 2002. A structural model for the matrix of genetic correlations between countries in international evaluations. *Proc. 7th WCGALP*. CD-ROM comm. n° 01-14.

Dennis, J.E. & Schnabel, R.B. 1983. *Numerical methods for unconstrained optimization and nonlinear equations.*

Prentice-Hall, Englewood Cliffs, New Jersey, USA. 378 p.

Druet, T., Jaffrézic, F., Boichard, D. & Ducrocq, V. 2003. Modeling lactation curves and estimation of genetic parameters for first lactation test-day records of French Holstein cows. *J. Dairy Sci. 86,* 2480-2490.

Gilmour, A.R., Thomson, R. & Cullis, B.R. 1995. Average Information REML: An efficient algorithm for variance parameters estimation in linear mixed models. *Biometrics 51,* 1440-1450.

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. *Jour. Ind. Soc. Ag. Statistics 49,* 215-236.

Minéry, S. 2003. Application of a structural model to estimate genetic correlations between countries. *Report of MSc*. 47 p.

Rekaya, R., Weigel, K. & Gianola, D. 2001. Application of a structural model for genetic covariances in international dairy sire evaluations. *J. Dairy Sci. 84,* 1525-1530.

Schaeffer, L.R. 1994. Model for international evaluation of dairy sires. *J. Dairy Sci. 77,* 2671-2678.

Wolfinger, R. 1993. Covariance structure selection in general mixed models. *Comm. Statist.-Simula. 22(4),* 1079-1106.