Prospects of Principal Components and Factor Analysis for Estimation of Genetic Correlations among Countries for Milk Production Traits

H. Leclerc[1,2], W.F. Fikse[1], and V. Ducrocq[3]

[1] *Interbull Centre, Department of Animal Breeding & Genetics, SLU, Box 7023, S – 750 07, Uppsala, Sweden*

[2] *Union Nationale des Coopératives agricoles d'Elevage et d'Insémination Animale, 75 595 Paris Cedex 12, France*

[3] *INRA, Station de Génétique Quantitative et Appliquée, 78 352 Jouy-en-Josas Cedex, France*

Introduction

The estimation of genetic correlations between countries is a prerequisite for international evaluation. Unfortunately, the increase in the number of participating countries and the lack of genetic links between some of them leads to statistical and computational difficulties.

In Mutiple-trait Across Country Evaluation (MACE), performances in all countries are considered as different traits whereas the underlying trait, e.g. milk production, is often similar for all countries. Thus, the expression of this trait in different countries tends to be highly correlated, and the genetic covariance matrix tends to have one or several very small eigenvalues (Van der Beek, 1999). A reduction of the number of parameters to be estimated requires the assumption of an underlying structure of the correlation matrix. One of the proposed alternatives is to use principal components (PC) or factor analysis approaches (Madsen *et al.,* 2000; Mäntysaari, 2004; Goddard, 2004, personal communication; Meyer and Kirkpatrick, 2005).

PC analysis assumes that all of the genetic variance is explained by a reduced number of principal components common to all countries, whereas factorial approaches assume that only a part of this variance is shared with other countries, the remaining part being country specific. As a result, PC analysis leads to a rank reduction of the genetic (co)variance matrix while factor analysis does not. Nevertheless, both approaches are of interest because they lead to more parsimonious models.

By only considering the first N principal com-ponents or factors ($N<M$, M is the number of original variables), it is possible to summarize the information in the data with limited loss of information. In this context, Madsen *et al.* (2000) and Mäntysaari (2004) proposed to estimate (co)variance components with a classical model and to reparametrize the corresponding matrix, discarding the smallest eigenvalues of the full rank matrix. A second approach which directly estimates a reduced rank (co)variance matrix was proposed by Meyer and Kirkpatrick (2005).

The aim of this study was to assess the prospects of using PC and factor analysis approaches for the estimation of the international genetic correlation matrix for milk yield data.

Material & Methods

The data were deregressed national breeding values of Holstein bulls and their effective daughter contributions (EDC) used in the Interbull routine evaluation of August 2003 for milk yield from 18 member countries (Australia, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Hungary, Ireland, Italy, New-Zealand, Poland, Spain, Switzerland, The Netherlands, United Kingdom and the United States), with a number of common bulls ranging from 6 (Finland – Poland) to 772 (Canada – the United States).

The sire model currently used in international genetic evaluations (Schaeffer, 1994) was applied.

An "unstructured" model, called here classical model (CM) was used to estimate "reference" genetic correlations among the 18 countries. A bending procedure (Jorjani *et al.,* 2003) was applied to ensure that the genetic correlation matrix was positive definite. Two approaches were used to reparametrize the genetic correla-tions matrix. The PC approach was based on the canonical decomposition of the correlation matrix $\mathbf{rG} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}' = \mathbf{V} \cdot \mathbf{V}'$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues,

$\mathbf{U}$ is the corresponding set of orthogonal eigenvectors, and $\mathbf{V}=\mathbf{U}\cdot\mathbf{D}^{1/2}$. The rank of $\mathbf{rG}$ (equal to the number of countries *M*) was reduced to *N* by setting to zero the smallest eigenvalues in $\mathbf{D}$, and deleting the corresponding eigenvectors from $\mathbf{U}$, such as $\mathbf{P}^{*}=\mathbf{U}^{*}\cdot\mathbf{D}^{*}\cdot\mathbf{U}^{*'}=\mathbf{V}^{*}\cdot\mathbf{V}^{*'}$. Then, $\mathbf{P}^{*}$ was rescaled to make it a correlation matrix $\mathbf{rG}^{*}$, with ij^{th} elements computed as $rG^{*}_{ij}=P^{*}_{ij}\Big/\sqrt{P^{*}_{ii}\times P^{*}_{jj}}$. The number of parameters to estimate is $N(2M-N+1)\big/2$ (Meyer and Kirkpatrick, 2005).

The second approach was an approximate factor analysis, hereafter referred to as A_FA. $\mathbf{P}^{*}$ was computed as for the PC approach, but instead of rescaling it to obtain $\mathbf{rG}^{*}$, a diagonal matrix $\mathbf{F}$ was added such that the diagonal elements of the genetic correlation matrix were equal to unity. These elements were not estimated as in a formal factor analysis (i.e. at the same time as $\mathbf{P}^{*}$) but defined by the constraint $\mathbf{F}=\mathbf{I}-diag(\mathbf{P}^{*})$. Therefore, the number of parameters to estimate was the same as for the PC approach. However, the rank of the correlation matrix $\mathbf{rG}^{*}$ was still *M*.

In the following analyses, the estimates obtained with CM will be considered as the "reference" genetic correlations. PC*i* and A_FA*i* will represent the reparametrized genetic correlation matrices using the PC or the A_FA approach respectively, considering the *i* largest eigenvalues and their corresponding eigenvectors.

Results and Discussion

1. Base countries

A reduced number of the 18 countries with strong links and representative for the production systems prevailing world-wide were chosen to define the principal components. These countries will be referred to as base countries. Two sets of base countries were compared. Base9 included seven large connecting countries (AUS, DEU, FRA, GBR, ITA, NZL, USA) and two countries with large contributions to the first eigenvectors of the 18 countries CM correlation matrix (HUN, CZE) (Table 1). Base8 included five countries having large contributions to the first eigenvectors (CHE, CZE, DEU, HUN, NZL) and three countries improving links with all the others (GBR, NLD, USA).

Table 1. Eigenvalues, relative proportion of the explained variances of the 18×18 CM genetic correlation matrix, and countries with large contributions (>0.30) to these eigenvectors.

	Eigenv.	Proportion	Countries >0.30 in eigenvector def.	
			positive sign	negative sign
1	15.397	0.855		
2	0.789	0.044	AUS, NZL	
3	0.562	0.031	NZL	POL, CZE
4	0.432	0.024	POL, CHE	CZE
5	0.228	0.013	DEU, NLD	HUN, CHE
6	0.196	0.011	BEL	FIN
7	0.095	0.005	USA, AUS	BEL, HUN
8	0.073	0.004	NZL, DNK	AUS, GBR
9	0.061	0.003	NLD	DEU, AUS, ESP
10	0.051	0.003	FRA, ITA, IRL	HUN, DNK
11	0.045	0.003	NLD	CAN
12	0.031	0.002	DEU, FRA	USA, GBR
13	0.022	0.001	ITA, NLD, CAN	USA, FRA
14	0.015	0.001	BEL, FIN	NLD, IRL
15	0.001	0.000	CHE	POL
16	0.000	0.000	NZL, GBR, ESP	IRL, DNK
17	0.000	0.000	ESP	ITA, GBR
18	0.000	0.000	USA	DNK

For base9, the comparison of the CM genetic correlations with genetic correlations obtained for a PC approach considering various numbers of eigenvalues and corresponding eigenvectors showed large average differences. Five princi-pal components should be considered to obtain an average absolute deviation of correlations lower than 0.030 (Table 2). For the A_FA approach, only 2 eigenvalues were needed to obtain an average absolute deviation of corre-lations lower than 0.030. Similar patterns were obtained with base8.

Table 2. Maximum and average absolute deviations of reparametrized genetic correlations (*rG*) with PC approach and A_FA approach computed for various numbers of eigenvalues and corresponding eigenvectors (between one and 8) from CM genetic correlations for the 9 countries of base9.

		1 Eig.	2 Eig.	3 Eig.	4 Eig.	5 Eig.	6 Eig.	7 Eig.	8 Eig.
	Eigenvalues' cumulative proportion of	83.8	90.3	95.1	96.9	98.0	98.7	99.3	99.7
PC	Maximum deviation *rG* PC-CM	0.361	0.335	0.147	0.076	0.056	0.036	0.025	0.022
	Average absolute deviation *rG* PC-CM	0.185	0.103	0.050	0.031	0.020	0.013	0.007	0.003
A_FA	Maximum deviation *rG* A FA-CM	0.124	0.132	0.058	0.036	0.019	0.013	0.012	0.011
	Average absolute deviation *rG* A FA-CM	0.045	0.026	0.012	0.009	0.006	0.005	0.003	0.001
	Average F	0.162	0.097	0.049	0.031	0.020	0.013	0.007	0.003

The comparison of breeding values predicted on each country scale with base9 reparametrized genetic correlations and CM genetic correla-tions (Figure 1) showed product moment corre-lations that were larger than 0.990 for PC approaches including at least 5 components. Increasing the number of principal components improved the correlations between predicted breeding values. For A_FA, the correlations were larger than 0.999 for 8 of the 9 countries when at least 3 factors were considered. For the analysis of the top 100 bulls, the larger the number of principal components considered, the better the agreement between the CM and PC approaches (Figure 2). However, the num-ber of common top bulls was on average disap-pointing low given the very high product moment correlations. For the A_FA approach, whatever the number of factors considered, the stability of predicted breeding values rankings was always good with at least 91 of the top 100 bulls remaining in this category.

2. All countries

Based on the previous results and the substantial reduction of the number of parameters to estimate (80 for the variance-covariance matrix instead of 171 with a CM), A_FA5 was chosen to estimate the 18×18 correlation matrix for both sets of base countries. The lower (or upper) triangle of the full genetic correlation matrix for the 18 countries was divided into 3 parts: correlations among base countries (rG_BB), correlations between base and other countries (rG_BO), and correlations among other countries (rG_OO).

Genetic correlations among all countries were computed using the following approach (see Leclerc *et al.,* 2005 for details). First, one or two countries at a time were added to the base countries and their correlations (i.e. elements of rG_BO) were estimated keeping rG_BB fixed. Then, the rG_BO estimates were regressed on the 5 vectors defining $\mathbf{V}^*$ for the base countries. The estimated regression coefficients were used to define the part of $\mathbf{V}^*$ corresponding to the other countries. From this, $\mathbf{rG}^*$ was created ($\mathbf{rG}^* = \mathbf{P}^* + \left(\mathbf{I} - diag(\mathbf{P}^*)\right)$ with $\mathbf{P}^* = \mathbf{V}^* \cdot \mathbf{V}^{*'}$). The lower diagonal block of $\mathbf{rG}^*$ is rG_OO.

Figure 1. Product moment correlations between breeding values predicted with CM genetic correlations and genetic correlations based PC*i* or A_FA*i* approaches (◊ PC3, □ PC4, △ PC5, ○ PC6, and ♦ A_FA3, ■ A_FA4, ▲ A_FA5, ● A_FA6) for 9 countries.

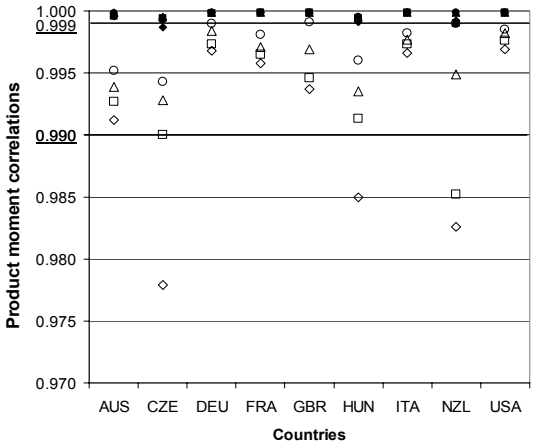
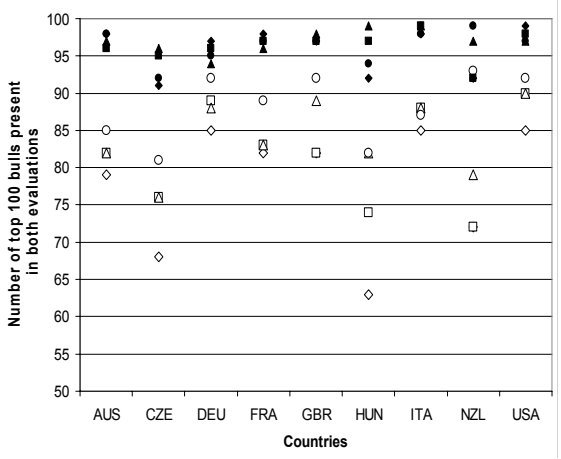


Figure 2. Number of top 100 bulls with breeding values predicted with CM genetic correlations also in the top 100 with BV predicted with PC*i* or A_FA*i* approaches for 9 countries (◊ PC3, □ PC4, △ PC5, ○ PC6, A_FA3, ■ A_FA4, ▲ A_FA5, ● A_FA6).



The average absolute deviations of the 153 genetic correlations obtained for A_FA5 from those estimated with CM for the 18 countries were 0.016 and 0.014 for base9 and base8 respectively (Table 3). The rG_BB correlations estimated with A_FA5 were accurate, with no deviation of correlations larger than 0.030 from CM for both sets of base countries. The main difference between both sets of base countries

Table 3. Average, maximum and distribution of deviations of genetic correlations (*rG*) computed for A_FA5 approach from CM correlations for all countries and groups of countries (among base countries (BB), between base and other countries (BO) and among other countries(OO)).

		All[1]		In rG BB		In rG BO		In rG OO	
Definition of the base		base9[2]	base8[3]	base9	base8	base9	base8	base9	base8
No. estimates (/153 *rG* CM)		153	153	36	28	81	80	36	45
Average deviation *rG* A FA-CM		-0.010	-0.011	0.003	0.003	-0.005	-0.009	-0.031	-0.023
Average abs. deviat. *rG* A_FA-CM		0.016	0.014	0.006	0.006	0.012	0.012	0.035	0.024
Maximum deviation *rG* A_FA-CM		-0.127	-0.096	0.019	0.018	0.047	-0.054	-0.127	-0.096
Frequency of correlation deviations A_FA – CM (%)	$0.05 < x$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	$0.03 < x \leq 0.05$	0.7	0.7	0.0	0.0	1.2	1.3	0.0	0.0
	$0.01 < x \leq 0.03$	12.4	5.2	19.4	21.4	11.1	2.5	8.3	0.0
	$-0.01 \leq x \leq 0.01$	52.3	51.0	77.8	75.0	58.0	63.8	13.9	15.6
	$-0.03 \leq x < -0.01$	19.6	31.4	2.8	3.6	22.2	22.5	30.6	62.2
	$-0.05 \leq x < -0.03$	11.8	9.8	0.0	0.0	7.4	8.8	33.3	17.8
	$x < -0.05$	3.3	2.0	0.0	0.0	0.0	1.3	13.9	4.4

[1] Statistics on *rG* among all countries, *rG* among base countries, *rG* between base and other countries and *rG* among other countries
[2] Base9 with base countries: AUS, CZE, DEU, FRA, GBR, HUN, ITA, NZL, USA.
[3] Base8 with base countries: CHE, CZE, DEU, GBR, HUN, NLD, NZL, USA.

is related to the rG_OO genetic correlations, of which 47.2% deviated by more than 0.030 from CM estimates for base9, and only 22.2% for base8. Thus, the choice of countries to define the base appeared to have a large impact on results. The deviations of genetic correlations larger than 0.050 concerned country pairs with less than 50 common bulls.

Note that genetic correlations obtained with CM are estimates, and are not necessarily the the correct ones.

Conclusion

The PC and A_FA approaches proposed here are pragmatic and give very good approximations of estimated genetic correlations. The number of parameters to estimate is reduced in comparison with a CM and neither approach requires new programs and/or algorithms.

Both approached can be easily extended to all participating countries since only correlations between the base countries and the remaining countries are needed. Small and/or poorly connected populations such as Finland were used here without encountering any particular problems in the estimation of genetic correlations.

Compared to the reference situation represented by the CM correlation matrix in this study, the impact on breeding values of using correlations estimated from the PC approach was larger than with the A_FA approach. However, only the PC approach makes it possible to obtain a reduced rank correlation matrix and/or to have a limited number of lists/scales.

References

Jorjani, H., Klei, L. & Emanuelson, U. 2003. A simple method for weighted bending of genetic (co)variance matrices. *J. Dairy Sci. 86*, 677-679.

Leclerc, H., Fikse, W.F. & Ducrocq, V. 2005 Principal components and approximate factorial approaches for estimating genetic correlations among countries in international dairy sire evaluation. *J. Dairy Sci, in press*.

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.

Mäntysaari, E.A. 2004. Multiple-Trait Across-Country Evaluations Using Singular (Co)Variance Matrix and Random Regression Model. *Interbull Bulletin 32*, 70-74.

Meyer, K. & Kirkpatrick, M. 2005. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Genet. Sel. Evol. 37*, 1-30.

Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. *J. Dairy Sci. 77*, 2671-2678.

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