Principal Components and Factorial Approaches for Estimating Genetic Correlations in International Sire Evaluation

H. Leclerc,^{1,2} W. F. Fikse,¹ and V. Ducrocq³

¹Interbull Centre, Department of Animal Breeding and Genetics, SLU, Box 7023, Uppsala 75007, Sweden

²Union Nationale des Coopératives agricoles d'Élevage et d'Insémination Animale, Paris 75595, France

³Station de Génétique Quantitative et Appliquée, INRA, Jouy-en-Josas 78352, France

ABSTRACT

The increasing number of participating countries and the lack of genetic links among some of them lead to statistical and computational difficulties in estimating the genetic (co)variance matrix needed for international sire evaluation of milk yield. Reparameterization using principal components or factorial approaches is proposed to exploit patterns in the genetic correlation matrix in order to reduce the number of parameters to be estimated without much loss of information. A 2-step approach was used. First, the genetic matrix between 8 or 9 “base” countries was used to determine a reduced number of principal components or factors. Then, the contributions of the remaining countries to these principal components or factors were computed. The resulting genetic correlations for the 18 countries were compared with the “reference” genetic correlations obtained with a classical model. The impact of using reparameterized genetic correlation matrices on breeding value prediction was investigated for both approaches. A better agreement between predicted breeding values and stability of their rankings was found when an approximate factor analysis was used, whatever the number of factors considered. The estimation of genetic correlations among 18 countries using an approximate factorial approach with 5 factors taken into account led to a reduction of the number of parameters to estimate from 171 to 80. The average absolute deviation of the correlations estimated with an approximate factorial approach from the “reference” genetic correlations was 0.014, which is considered very satisfactory in light of the computational ease.

(Key words: genetic correlation matrix, principal component analysis, factor analysis, international evaluation)

Abbreviation key: A_FA = approximate factor analysis, CM = classical model, Interbull = International

Bull Evaluation Service, MACE = multiple-trait across-country evaluation, PBV = predicted breeding value, PC = principal components, rG_BB = genetic correlations among base countries, rG_BO = genetic correlations between base and other countries, rG_OO = genetic correlations among other countries.

INTRODUCTION

The continuous exchange of genetic material such as semen and embryos at the international level has led to the development of across-country bull comparisons. The International Bull Evaluation Service (Interbull) provides international predicted breeding values (PBV) for each bull on the scale of each participating country using a multiple-trait across-country evaluation (MACE; Schaeffer, 1994) approach with national genetic evaluation results as input data. Knowledge of genetic parameters 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 estimating the genetic (co)variance matrix.

Recently (November 2004), Interbull introduced a new procedure for the estimation of genetic correlations between countries to avoid assignment of arbitrary correlations. In a first step, genetic correlations for all country pairs are estimated with an approximate expectation-maximization-REML procedure (Klei and Weigel, 1998). In a second step, estimated correlations are combined with prior information (correlations) using a Bayesian approach. Thus, estimates are regressed to an expected correlation, the regression depending on the strength of genetic ties between countries. In addition, genetic correlations are required to fall within certain ranges. The expected correlation and the ranges are based on similarities between, for example, production system and climate between countries. For numerically smaller breeds, prior information from other breeds is also utilized. For the largest breed (Holstein) with 27 participating populations, 351 genetic correla-

Received February 3, 2005.

Accepted May 5, 2005.

Corresponding author: Hélène Leclerc; e-mail: helene.leclerc@jouy.inra.fr.

tions between countries need to be estimated, which is a challenging task.

In MACE, performances in all participating countries are considered different traits, whereas the underlying trait, for example, milk production, is quite 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 more zero or very small eigenvalues (e.g., 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. Structural models were proposed to exploit patterns in the genetic correlation matrix (Rekaya et al., 2001; Delaunay et al., 2002). Delaunay et al. (2002) suggested using data itself to characterize genetic correlations between countries, instead of external information as in the structural model of Rekaya et al. (2001). Promising results were obtained on simulated data and on milk yield field data (Delaunay et al., 2002; Delaunay, 2003; Minéry et al., 2003), but disappointing results were found with poorly connected countries and for moderately correlated traits (Leclerc et al., 2004). A proposed alternative is to use principal components (PC) or factor analysis approaches (Madsen et al., 2000; M. E. Goddard, 2004, personal communication; Mäntysaari, 2004; Meyer and Kirkpatrick, 2005). The main difference between the PC and factor analysis is that 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 whereas factor analysis does not. Nevertheless, both approaches are of interest because they lead to a more parsimonious model.

The eigenvectors of the genetic (co)variance matrix define the so-called principal components, which are independent linear functions of the variables considered. Their statistical significance is assessed through the amplitude of the corresponding eigenvalues. The first principal component (or factor in factor analysis) accounts for as much of the variability in the data as possible, and each succeeding component accounts for the maximum amount of variation remaining. The last components, with variance (eigenvalues) close to zero, have a minimal contribution to the variability in the data. By reducing the number of dimensions, that is, only considering the first N principal components or factors ($N < M$, M is the number of original variables), it is possible to summarize the information in the data without much 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 reparameterize 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 the present study was to assess the prospects and consequences of using PC and factor analysis approaches for the estimation of the international genetic correlation matrix for milk yield data. The objectives were 2-fold: 1) to estimate the international genetic correlation matrix for milk yield using approximate PC or factorial approaches for 18 participating countries and to compare results to "reference" genetic correlations obtained with a classical model, and 2) to measure the impact of reparameterization with various numbers of principal components or factors, on international PBV and bull rankings. The proposed approaches could replace the REML estimation in the first step of the current Interbull procedure. It was not intended at this stage to develop an exact procedure to directly estimate principal components or factors from all the data. As shown by Meyer and Kirkpatrick (2005), such an approach should work but its development and application are very demanding compared with what is presented here.

MATERIALS AND METHODS

Data

Data were deregressed national breeding values (Jairath et al., 1998) of Holstein bulls and their effective daughter contributions (Fikse and Banos, 2001) used in the Interbull routine evaluation of August 2003 for milk yield. Data were edited to include only national evaluations for bulls born after 1984. All observations were included in the estimation of the genetic correlations, in contrast with the current Interbull practice that builds on the use of subsets of well-connected bulls (Interbull, 2005). The sire variances considered were estimated within country using a single-trait expectation-maximization REML procedure (Sullivan, 1999). Heritabilities of milk yield were provided by the participating countries.

Data from 18 Interbull member countries were used for this study (Australia, Belgium, Canada, Czech Republic, Denmark, Finland, France, Germany, Hungary, Ireland, Italy, New Zealand, Poland, Spain, Switzerland, The Netherlands, United Kingdom, and United States). The data set is characterized in Table 1. The selected countries represented a wide range of production systems and environments. Genetic ties between countries were variable, with the number of bulls with daughters in 2 countries (hereafter called common

Table 1. Number of bulls per country¹ with national evaluation (on the diagonal) and the number of bulls with an evaluation in both countries for each country pair (above the diagonal).

	AUS	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	HUN	IRL	ITA	NLD	NZL	POL	USA
AUS	3596	73	346	60	119	214	82	139	10	193	235	116	105	154	277	417	44	501
BEL		424	96	61	106	184	50	104	9	152	138	68	84	117	286	76	48	109
CAN			5446	83	205	282	98	201	12	215	665	213	108	225	286	334	70	772
CHE				519	65	104	42	83	5	96	111	48	62	83	101	59	30	91
CZE					1653	419	89	184	19	247	228	211	103	220	298	150	98	371
DEU						11,782	162	229	23	331	329	197	156	350	616	214	134	521
DNK							4737	92	16	89	129	97	72	140	148	89	50	124
ESP								905	21	172	252	153	122	230	222	134	80	267
FIN									615	26	27	14	15	24	22	15	6	20
FRA										8371	252	127	139	241	265	152	74	443
GBR											3118	207	211	284	372	285	81	445
HUN												1329	97	187	185	152	75	306
IRL													878	126	163	125	46	135
ITA														4590	298	183	70	348
NLD															6819	376	89	657
NZL																3009	50	437
POL																	3765	91
USA																		18,121

¹Country codes: AUS = Australia, BEL = Belgium, CAN = Canada, CHE = Switzerland, CZE = Czech Republic, DEU = Germany, DNK = Denmark, ESP = Spain, FIN = Finland, FRA = France, GBR = United Kingdom, HUN = Hungary, IRL = Ireland, ITA = Italy, NLD = The Netherlands, NZL = New Zealand, POL = Poland, USA = United States.

bulls) ranging from 6 (Finland–Poland) to 772 (Canada–United States).

Model for Observations

Milk production in different countries were considered different traits that are genetically correlated. A multiple-trait sire model was used to combine deregressed national genetic values from several countries into international sire evaluations (MACE; Schaeffer, 1994):

$$\mathbf{y}_i = \mu_i \mathbf{I} + \mathbf{Z}_i \mathbf{Q} \mathbf{g}_i + \mathbf{Z}_i \mathbf{s}_i + \mathbf{e}_i$$

where $\mathbf{y}_i$ is the vector of deregressed national genetic merit from country i , μ_i is the fixed effect of country mean, $\mathbf{g}_i$ is the vector of genetic groups effects treated as random for the prediction of the bull's international breeding values and as fixed for the estimation of genetic correlations between countries, $\mathbf{s}_i$ is the vector of random sire genetic merit for country i , $\mathbf{e}_i$ is the vector of random residual, $\mathbf{Z}_i$ is the sire incidence matrix, and $\mathbf{Q}$ is the matrix assigning sires in $\mathbf{s}$ to group effects in $\mathbf{g}$. The joint distribution of the random variables was assumed to be multivariate normal, with $\text{var}(\mathbf{s}) = \mathbf{A} \otimes \mathbf{G}$, $\text{var}(\mathbf{g}) = \mathbf{I} \otimes \mathbf{G}$, (only for prediction of breeding values) and $\text{var}(\mathbf{e}_i) = \mathbf{D}_i \sigma_{e_i}^2$. $\mathbf{G}$ is the genetic (co)variance matrix between traits measured in different countries, $\mathbf{A}$ is the additive genetic relationship matrix relating bulls to their sires and maternal grand-sires, $\mathbf{D}_i$ is a diagonal matrix with effective daughter contributions (Fikse and

Banos, 2001) as elements (weighting factors), and $\sigma_{e_i}^2$ is the residual variance for country i .

Genetic groups for missing ancestors were formed based on selection path (sire, maternal grand-sires, and maternal grand-dam), year of birth and origin (defined following country borders). Small groups were merged together first by birth year and then by origin to achieve a minimum group size of 500 bulls with unknown parents for the estimation of genetic correlations across country and 15 bulls for the prediction of international breeding values.

Methods

Classical model for genetic correlations. An unstructured model (classical model; **CM**) was used to estimate “reference” genetic correlations among the 18 countries. The restricted likelihood function was defined in terms of sire variances and genetic correlations, and was maximized using an average information REML algorithm (Johnson and Thompson, 1995). Because of memory constraints, the number of countries included per run was limited to 10. A complete matrix with correlations (18×18) was obtained by taking the average correlation for each country pair from correlations estimated in 30 runs each including between 4 and 10 countries. The number of estimates for each country pair was between 1 and 10, that is, a total number of estimates of 497 for 153 correlations. Each run included on average 3.6 well-connected populations (among Australia, Canada, France, Germany, Italy,

The Netherlands, New Zealand, United Kingdom, and United States) and at least one of the largest Holstein populations: France, Germany, The Netherlands, and United States. The estimation of genetic correlations between countries with weak ties was facilitated by utilizing indirect links provided by the 9 well-connected countries. The “reference” correlation matrix was obtained after applying a weighted bending procedure to make the matrix positive definite (Jorjani et al., 2003), because genetic correlations were not estimated simultaneously. The estimation of genetic correlations with a classical model was a compulsory step in this study because these estimates were used as “reference” correlations. For 6 populations (Estonia, Israel, Slovenia, and the 3 red populations from Denmark, France, and Switzerland), the genetic ties were too weak to estimate proper “reference” genetic correlations, particularly with countries other than the well-connected ones. Thus, a fair comparison between the studied approaches could not be made for these populations, which were therefore not considered in this study. It is important to note that these exclusions do not preclude the use of the proposed approaches for any country.

Reparameterization of the genetic correlation matrix. The genetic correlation matrix was reparameterized using 2 approaches. The PC approach was based on the canonical decomposition of the correlation matrix $\mathbf{rG}$, which can be written as:

$$\mathbf{rG} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}' = \mathbf{V} \cdot \mathbf{V}',$$

with $\mathbf{U} \cdot \mathbf{U}' = \mathbf{I}_M$ and $\mathbf{U}' \cdot \mathbf{U} = \mathbf{I}_M$

where $\mathbf{D}$ is the diagonal matrix of eigenvalues, $\mathbf{U}$ is the corresponding set of orthogonal eigenvectors, $\mathbf{V}$ is $\mathbf{U} \cdot \mathbf{D}^{1/2}$, and $\mathbf{I}_i$ is the identity matrix of dimension i . The rank of the correlation matrix can be reduced by setting to zero the smallest eigenvalues in the matrix $\mathbf{D}$, and deleting the corresponding eigenvectors from $\mathbf{U}$:

$$\mathbf{P}^* = \mathbf{U}^* \cdot \mathbf{D}^* \cdot \mathbf{U}^{*'} = \mathbf{V}^* \cdot \mathbf{V}^{*'},$$

with $\mathbf{U}^* \cdot \mathbf{U}^{*'} = \mathbf{I}_N$ and $\mathbf{U}^{*'} \cdot \mathbf{U}^* = \mathbf{I}_M$

where $\mathbf{P}^*$ is a matrix with rank $N < M$, $\mathbf{D}^*$ is the diagonal matrix of modified eigenvalues, $\mathbf{U}^*$ contains the corresponding set of orthogonal eigenvectors, and $\mathbf{V}^*$ is $\mathbf{U}^* \cdot \mathbf{D}^{*1/2}$. The symmetric matrix $\mathbf{P}^*$ was then rescaled to make it a correlation matrix $\mathbf{rG}^*$, with diagonal elements equal to one. The elements of the $\mathbf{rG}^*$ matrix

(rG_{ij}^*) were computed as $rG_{ij}^* = \frac{P_{ij}^*}{\sqrt{P_{ii}^* \times P_{jj}^*}}$, where rG_{ij}^* is

element (i, j) of matrix $\mathbf{rG}^*$, and P_{ij}^* is element (i, j) of matrix $\mathbf{P}^*$. Then, $\text{rank}(\mathbf{rG}^*) = \text{rank}(\mathbf{P}^*) = N$. At first

glance, it might appear that the reparameterization has increased the number of parameters to estimate, unless N is substantially smaller than M . However, eigenvectors are orthogonal and the restriction $\mathbf{U}^* \cdot \mathbf{U}^{*'} = \mathbf{I}_N$ implies that for the i th column of $\mathbf{V}^*$ only $M - i + 1$ elements need to be estimated. Therefore, the number of parameters to estimate the genetic variance covariance matrix is

$$\frac{N(2M - N + 1)}{2}$$

(Meyer and Kirkpatrick, 2005).

The second approach was an approximate factor analysis, hereafter referred to as **A_FA**. Here, $\mathbf{P}^*$ 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. The elements of $\mathbf{F}$ can be considered the specific factors in the factor analysis. However, these elements were not estimated as in a formal factor analysis (i.e. at the same time as $\mathbf{P}^*$) but defined by the constraints $\mathbf{F} = \mathbf{I} - \text{diag}(\mathbf{P}^*)$. 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 and not $N < M$.

In the remainder of this study, the estimates obtained with CM will be considered the “reference” genetic correlations. PC_i and A_FA_i will represent the reparameterized genetic correlation matrices using the PC or the A_FA approach respectively, considering the i largest eigenvalues and their corresponding eigenvectors.

Definition and Use of the Principal Components

A reduced number K of the 18 countries with strong links and representative for the production systems prevailing worldwide were chosen to define the principal components. These countries will be referred to as base countries. Two different sets of base countries were compared: the first one with mainly large connecting populations and the second one with mainly populations having large contribution to the first eigenvectors of the correlation matrix used in the Interbull routine evaluation of August 2003.

For these K base countries, the correlations estimated with CM were compared with genetic correlations obtained after reparameterization considering various numbers of eigenvalues and corresponding eigenvectors estimated with PC or A_FA approaches. To quantify the influence of the changes in the genetic correlation matrix, PBV of bulls were computed for each country scale using the “reference” genetic correlations and genetic correlations obtained with both reparameteriza-

tion approaches for the first set of base countries. Bulls were ranked on PBV on each country scale for all tested alternatives and the rankings were compared for all bulls on the one hand and for the top 100 bulls on the other hand.

The lower (or upper) triangle of the full genetic correlation matrix for the 18 countries can be divided into 3 parts: the correlations among base countries ($\mathbf{rG_BB}$), the correlations between base and other countries ($\mathbf{rG_BO}$), and the correlations among other countries ($\mathbf{rG_OO}$). Genetic correlations among all 18 countries were computed using the following approach for each set of base countries. First, $\mathbf{rG_BO}$ was estimated using a classical model keeping the $\mathbf{rG_BB}$ fixed. Only a limited number of the other countries (1 or 2 in practice) were included at a time. For the $\mathbf{rG_BB}$, the correlations obtained after PC or A_FA reparameterizations were used. This way, genetic correlations between the base countries on the one hand and the other countries on the other hand were relative to the same eigenspace defined by the base countries ($\mathbf{V}^*$). Then, the elements of the N principal components for the other countries were obtained by regressing the $\mathbf{rG_BO}$ correlations on the N eigenvectors ($\mathbf{V}^*$) of the base countries:

$$\mathbf{H}^{*'} = (\mathbf{V}^{*'} \cdot \mathbf{V}^*)^{-1} \cdot \mathbf{V}^{*'} \cdot \mathbf{rG_BO}$$

The reparameterized genetic correlation matrix ($\mathbf{rG}^*$) between all the 18 countries can be computed as:

• $\mathbf{rG}^*$, which is the $\mathbf{P}^*$ matrix after rescaling to make the diagonal elements equal to one, where $\mathbf{P}^* = \mathbf{W} \cdot \mathbf{W}'$

with $\mathbf{W} = \begin{bmatrix} \mathbf{V}^* \\ \mathbf{H}^* \end{bmatrix}$ of rank N , for the PC approach, and

• $\mathbf{rG}^* = \mathbf{W} \cdot \mathbf{W}' + \mathbf{F}$ with $\mathbf{W} = \begin{bmatrix} \mathbf{V}^* \\ \mathbf{H}^* \end{bmatrix}$ of rank N and $\mathbf{F} = \mathbf{I} - \text{diag}(\mathbf{W} \cdot \mathbf{W}')$ for the A_FA approach.

Thus, $\mathbf{rG_OO}$ does not need to be estimated but is computed from the regression of correlations $\mathbf{rG_BO}$ on the N eigenvectors.

RESULTS

Base Countries

Two alternatives to select base countries were studied. For the first one, 9 countries were taken into account, hereafter referred to as base9. They included 7 large countries having a large number of common bulls (i.e., Australia, France, Germany, Italy, New Zealand, United Kingdom, and United States). These were expected to be the most representative for the production systems encountered in the 18 countries considered in this study. Two countries (Hungary and Czech Republic) were added, based on their large contribution to the

first eigenvectors of the CM correlation matrix for all 18 countries (Table 2). For the second alternative, 8 countries were used (base8). Five countries having sufficient links with the other countries were chosen based on their large contribution to the first eigenvectors (i.e., Czech Republic, Germany, Hungary, New Zealand, and Switzerland). Three countries providing many connections to all the other ones were added (i.e., The Netherlands, United Kingdom, and United States).

The aim was to find alternatives using PC or A_FA approaches to estimate genetic correlations as accurately as possible with a minimum number of components or factors to reduce the number of parameters to estimate. For base9, the comparison of reparameterized genetic correlations obtained with the PC approach with the CM genetic correlations showed relatively large average differences (more than 0.030) when up to 4 principal components were taken into account (Table 3). If 5 or more PC were considered (i.e., at least 98% of the variance was explained), the genetic correlations became on average closer to the CM estimates (average absolute deviation of correlations was lower than 0.020), even though the maximum deviations of correlations were still large (0.056 for PC5, 0.036 for PC6). For the A_FA approach, the reparameterized genetic correlations were closer to the CM genetic correlations (Table 3). When only 2 eigenvalues and their corresponding eigenvectors (i.e., 90.3% of the variance explained) were considered, the average absolute deviation of correlations was less than 0.030. When 5 or more PC were considered, the maximum deviation of correlations was less than 0.030. In addition, an average of diagonal elements in the $\mathbf{F}$ matrix less than 0.030 was found when at least 5 PC were taken into account. The $\mathbf{F}$ components represent the components specific to each country. For base8, the comparison of reparameterized genetic correlations obtained for PC and A_FA approaches for various numbers of eigenvalues and corresponding eigenvectors with the CM genetic correlations showed similar patterns (results not shown).

The comparison of breeding values predicted with base9 reparameterized genetic correlations or CM genetic correlations showed product-moment correlations that were always larger than 0.99 for PC approach taking into account at least 5 components (Figure 1). Increasing the number of PC improved the correlations between PBV: the minimum was 0.9779 for PC3 and Czech Republic, whereas the maximum was 0.9991 for PC6 and United Kingdom. For A_FA considering at least 3 eigenvalues and their corresponding eigenvectors, the correlations were always larger than 0.999 (Figure 1) for all 9 countries (except for A_FA3 and Czech Republic). Similar correlations were obtained for

Table 2. Eigenvalues, relative and cumulative proportions of the variances explained of the classical model genetic correlation matrix for 18 countries, and countries with large contributions (>0.30) to these eigenvectors.

Rank	Eigenvalue	Relative proportion	Cumulative proportion	Countries ¹ with large weight in eigenvector definition (> 0.30)	
				Positive sign	Negative sign
1	15.397	0.855	0.855		
2	0.789	0.044	0.899	AUS, NZL	
3	0.562	0.031	0.930	NZL	POL, CZE
4	0.432	0.024	0.954	POL, CHE	CZE
5	0.228	0.013	0.967	DEU, NLD	HUN, CHE
6	0.196	0.011	0.978	BEL	FIN
7	0.095	0.005	0.983	USA, AUS	BEL, HUN
8	0.073	0.004	0.987	NZL, DNK	AUS, GBR
9	0.061	0.003	0.991	NLD	DEU, AUS, ESP
10	0.051	0.003	0.994	FRA, ITA, IRL	HUN, DNK
11	0.045	0.003	0.996	NLD	CAN
12	0.031	0.002	0.998	DEU, FRA	USA, GBR
13	0.022	0.001	0.999	ITA, NLD, CAN	USA, FRA
14	0.015	0.001	1.000	BEL, FIN	NLD, IRL
15	0.001	0.000	1.000	CHE	POL
16	0.000	0.000	1.000	NZL, GBR, ESP	IRL, DNK
17	0.000	0.000	1.000	ESP	ITA, GBR
18	0.000	0.000	1.000	USA	DNK

¹Country codes: AUS = Australia, BEL = Belgium, CAN = Canada, CHE = Switzerland, CZE = Czech Republic, DEU = Germany, DNK = Denmark, ESP = Spain, FIN = Finland, FRA = France, GBR = United Kingdom, HUN = Hungary, IRL = Ireland, ITA = Italy, NLD = The Netherlands, NZL = New Zealand, POL = Poland, USA = United States.

PBV rankings (results not shown). Standard deviations of the changes in ranks for breeding values predicted using CM correlations and breeding values predicted using a reparameterized genetic correlation matrix for both PC and A_FA approaches enabled judging of the amplitude of reranking (results not shown). For the PC approach, the standard deviations decreased with the addition of an extra PC. Surprisingly, this was not always the case for the A_FA approach. However, the standard deviations obtained for the A_FA analyses were between 2.8 and 6 times smaller than for the PC analyses.

For the analysis of top bulls, the best agreement between the CM and PC approach was obtained when 6

PC were chosen (PC6; Figure 2). However, the results obtained were on average relatively disappointing given the limited changes observed when all PBV were considered (PC6; Figure 1); for instance, for PC6, the agreement between top lists based on CM and A_FA ranged from 81 bulls for the Czech Republic to 93 bulls for New Zealand. For the A_FA approach, the results depended on the country, for example, the best result (largest number of bulls in common between ranking based on CM and A_FA) for France, Germany, and United States were obtained with A_FA3, whereas this was with A_FA6 for New Zealand. However, whatever the number of factors considered (with a minimum number of 3), the stability of PBV rankings was always

Table 3. Maximum and average absolute deviations of reparameterized genetic correlations (r_G) with a principal components (PC) approach and a approximate factorial (A_FA) approach computed for various numbers of eigenvalues and corresponding eigenvectors (between one and 8) from genetic correlations estimated with a classical model after bending (CM) for 9 countries (Australia, Czech Republic, France, Germany, Hungary, Italy, New Zealand, United Kingdom, and United States).

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

¹Number of eigenvalues and corresponding eigenvectors used to reparameterize the 9×9 genetic correlation matrix.

²Elements in diagonal matrix $\mathbf{F} = \mathbf{I} - \text{diag}(\mathbf{P}^*)$.

Figure 1. Product-moment correlations between breeding values predicted with genetic correlations obtained for a classical model after bending and genetic correlations reparameterized based on principal components (PC_{*i*}) or approximate factorial (A_FA_{*i*}) approaches, with *i* representing the number of eigenvalues and their corresponding eigenvectors used to reparameterize the genetic correlations matrix (◇ PC3, □ PC4, △ PC5, ○ PC6, and ◆ A_FA3, ■ A_FA4, ▲ A_FA5, ● A_FA6) for 9 countries (AUS = Australia, CZE = Czech Republic, DEU = Germany, FRA = France, GBR = United Kingdom, HUN = Hungary, ITA = Italy, NZL = New Zealand, USA = United States).

Figure 2. Number of top 100 bulls with breeding values predicted with genetic correlations obtained for a classical model after bending also in the top 100 with PBV estimated with principal components (PC_{*i*}) or approximate factorial (A_FA_{*i*}) approaches, with *i* representing the number of eigenvalues and their corresponding eigenvectors used to reparameterize the genetic correlations matrix (◇ PC3, □ PC4, △ PC5, ○ PC6, and ◆ A_FA3, ■ A_FA4, ▲ A_FA5, ● A_FA6) for 9 countries (AUS = Australia, CZE = Czech Republic, DEU = Germany, FRA = France, GBR = United Kingdom, HUN = Hungary, ITA = Italy, NZL = New Zealand, USA = United States).

good for the A_FA approach with at least 91 of the top 100 bulls remaining in this category.

All Countries

Based on the previous results, the alternative A_FA5 was chosen for the comparison of the correlation matrix for 18 countries and for both sets of base countries. For each set of base countries, the rG_BB correlations considering 5 factors were fixed. Coordinates of the 9 or 10 other countries in these eigenspaces were obtained from the correlations estimated with a classical model including the base countries and one or 2 other countries at a time. Then, the RG_OO correlations were estimated from the regression of the rG_BO correlations on the 5 fixed factors. 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 4). Base8 yielded a lower maximum deviation (−0.096) than base9 (−0.127). From CM estimates, 15.8% of the genetic correlations computed from A_FA5 deviated by more than 0.030 for base9, but only 12.5% for base8 (Table 4). 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 (Table 4). For the correlations between base countries on the one hand and other countries on the other hand (rG_BO), only 8.6% deviated by more than 0.030 from the CM for base9 and 11.4% for base8 (Table 4). For rG_OO genetic correlations, deviations were larger. However, base8 presented a lower average absolute deviation of correlations than base9 (0.024 instead of 0.035; Table 4). For base9, 47.2% of the genetic correlations were lower with A_FA5 than with CM, whereas the corresponding figure for base8 was only 22.2% (Table 4). The deviations of correlations larger than 0.050 concerned the following country pairs: Denmark–Poland, Denmark–Switzerland, Finland–Poland, Ireland–Poland, and Poland–Switzerland for base9, and Finland–Ireland and Finland–Poland for base8. The number of common bulls for these country pairs was 50 or less.

DISCUSSION

The main objective of the 2 studied approaches was to reduce the number of parameters to estimate without reducing the accuracy of genetic correlations estimates and PBV and exploiting the structure of the genetic correlation matrix. The PC or A_FA approaches considering 5 principal components or factors for 18 countries require 80 parameters to estimate the genetic variance-covariance matrix instead of 171 with a classical model.

Table 4. Average, maximum, and distribution of deviations of genetic correlations (rG) computed for a approximate factorial approach (A_FA) taking into account the 5 largest eigenvalues and their corresponding eigenvectors from correlations obtained with a classical model after bending (CM) for all countries and groups of countries (among base countries, between base and other countries, and among other countries) on 18 countries.

Number of countries defining the base ²		All ¹		Among base		Between base and other		Among other	
		Base9	Base8	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 absolute deviation 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 ≤ 0.05	0.7	0.7	0.0	0.0	1.2	1.3	0.0	0.0
	0.01 < x ≤ 0.03	12.4	5.2	19.4	21.4	11.1	2.5	8.3	0.0
	-0.01 ≤ x ≤ 0.01	52.3	51.0	77.8	75.0	58.0	63.8	13.9	15.6
	-0.03 ≤ x < -0.01	19.6	31.4	2.8	3.6	22.2	22.5	30.6	62.2
	-0.05 ≤ 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

¹Statistics on rG among all countries, rG among base countries, rG between base and other countries and rG among other countries

²Base9 with base countries: Australia, Czech Republic, France, Germany, Hungary, Italy, New Zealand, United Kingdom, and United States; Base8 with base countries: Czech Republic, Germany, Hungary, The Netherlands, New Zealand, Switzerland, United Kingdom, and United States.

The number of parameters for a “real” factor analysis would be $\frac{N(2M - N + 1)}{2} + M$ (Smith et al., 2001); that is, M parameters more than the approaches studied here. For the 27 populations currently in the international genetic evaluation of production traits for Holsteins, the number of parameters could be reduced by a factor of about 3, relative to CM, using a PC or an A_FA approach taking 5 principal components or factors into account relatively to a classical model. The PC approach has the main advantage of reducing the rank of the correlation matrix, which decreases the computational requirements both for estimation of genetic parameters and prediction of international breeding values, whereas the A_FA approach results in a full rank matrix. Moreover, the PC approach enables to summarize information into only a few components. As these components often have biological/geographical interpretation, nonparticipating countries to Interbull could choose combinations of these components based on their own particular situation.

The comparison of PBV showed relatively large differences depending on whether the PC approach or A_FA approach was used for the reparameterization. Ranking of breeding values is sensitive to between-countries correlations and seemingly small changes in correlation estimates can have large consequences on ranking. Changes in the top bulls’ rankings are especially critical considering the impact of these top bulls in genetic exchange.

For the comparison of genetic correlations for 18 countries, only the approximate factorial approach was

considered, as it presented satisfactory characteristics of stability of PBV, even though it did not reduce the rank of the genetic correlation matrix. Compared with a full rank approach, too much loss of accuracy was observed for PC to advise the use of this approach for the estimation of international genetic correlations. All eigenvalues contributing more than 1% to the total variance, that is, 5 in the A_FA approach, were taken into account for the comparison of genetic correlations estimated between 18 countries.

The accuracy of genetic correlation estimates obtained from the studied approaches in comparison with CM was not always related to the number of eigenvalues and corresponding eigenvectors considered in the analysis. One of the possible explanations is that the increase of the number of factors considered increased the number of parameters to estimate. As a result, the precision of the estimated parameters corresponding to the factors with a small contribution to the variance is lower than the precision of parameters corresponding to the first factors. Above a certain optimum, this loss of precision becomes larger than the gain due to better “coverage” of the real full matrix.

The choice of countries to define the base appeared to be important as shown by the rather large difference between rG_{OO} correlations estimated from base9 and correlations estimated from base8. With the set of countries corresponding to base9, we observed poor results for some countries like Poland and Switzerland. This means that these countries were not well represented in the “base”. These 2 countries had a large contribution to the first eigenvectors (Table 2), but we first decided

Table 5. Average, maximum, and distribution of deviations of genetic correlations (r_G) estimated with a classical model after bending (CM), a approximate factorial approach taking into account the 5 largest eigenvalues and their corresponding eigenvectors (A_FA5) for a base formed with 9 countries or with 8 countries from the corresponding correlations used by Interbull (IB) for milk yield in the evaluation of August 2003 (<http://www-interbull.slu.se/eval/framesida-prod.htm>) on 18 countries.

		CM	A_FA5 Base9 ¹	A_FA5 Base8 ²
Average deviation r_G Model-IB		-0.035	-0.045	-0.046
Average absolute deviation r_G Model-IB		0.050	0.056	0.054
Maximum deviation r_G Model-IB		-0.255	-0.268	-0.255
Frequency	0.05 < x	1.3	1.3	0.0
of correlation	0.03 < x ≤ 0.05	9.2	5.9	5.2
deviations	0.01 < x ≤ 0.03	15.0	12.4	9.8
Model-IB (%)	-0.01 ≤ x ≤ 0.01	23.5	17.6	23.5
	-0.03 ≤ x < -0.01	13.1	13.7	15.0
	-0.05 ≤ x < -0.03	6.5	14.4	10.5
	x < -0.05	31.4	34.6	35.9

¹Base9 with base countries: Australia, Czech Republic, France, Germany, Hungary, Italy, New Zealand, United Kingdom, and United States.

²Base8 with base countries: Czech Republic, Germany, Hungary, The Netherlands, New Zealand, Switzerland, United Kingdom, and United States.

not to take them into account to define the base countries because of, respectively, weak ties with the other countries and small population size. On the other hand, some countries, such as France, Italy, or the United States, were used even though they did not appear to have extreme contributions to the first eigenvectors. In the definition of base8, the choice of countries was balanced between large contribution to the first eigenvectors (i.e., Czech Republic, Germany, Hungary, New Zealand, and Switzerland) and large connecting countries improving genetic ties between all countries (i.e., The Netherlands, United Kingdom, and United States). The number of base countries was reduced to 8 countries, which reduced the number of runs needed to estimate the correlations among the 18 countries.

The comparison of the genetic correlations estimated with the classical model and the approximate factor analysis approach showed large differences with the correlations actually used by Interbull for milk yield in the evaluation of August 2003 (Table 5). The average absolute deviation of Interbull correlations from CM was 0.050, 0.054 with A_FA5 base8, and 0.056 with A_FA5 base9. For all models, around one-half of the genetic correlations deviated by more than 0.030 from the Interbull correlations, and about one-third of the genetic correlations were much lower (by more than -0.050) compared with the Interbull ones (Table 5). The CM and A_FA5 correlations could have been closer to Interbull correlations used in August 2003, if procedures applied had been similar to Interbull ones, that is, same data (selected subsets of well-connected bulls which tend to yield higher genetic correlation estimates; Sigurdsson et al., 1996) and a postprocessing step, which results in an average increase of genetic

correlations of 0.07 (W. F. Fikse, 2004, personal communication). Postprocessing mainly affects country pairs with fewer common bulls. The postprocessing step is subjective because it is difficult to know whether a low correlation is caused by estimation problems (mainly for country pairs with poor links) or by true differences such as differences in national models and trait definitions, genotype × environment interaction, and data quality problems. The postprocessing step used until August 2004 lacked transparency and the property of being repeatable. The new postprocessing step implemented since November 2004 formalizes the use of prior information.

Estimation of genetic correlations using the proposed approximate PC and factorial approaches can be easily extended to all participating countries because only correlations between the base countries and the remaining countries are needed. Small and poorly connected populations like Finland have been used without encountering any particular problems in the estimation of genetic correlations.

CONCLUSIONS

The proposed methods (PC and A_FA approaches) are pragmatic and give very good approximations of estimated genetic correlations. Moreover, the number of parameters to estimate is reduced compared with a classical model and they do not necessitate new programs or algorithms.

Nevertheless, the PC and A_FA approaches in the present study are only approximate methods to estimate genetic correlations. It would be useful to compare approximate correlations with the ones obtained with

a direct estimation approach like the one implemented by Meyer and Kirkpatrick (2005). In this way, it would be possible to judge the impact of the approximation. At a minimum, the proposed approaches have the advantage of being straightforward and immediately applicable.

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

REFERENCES

- Delaunay, I. 2003. Méthodologie des évaluations génétiques internationales chez les bovins laitiers. Ph.D. Thesis. Inst. Nat. Agron. Paris-Grignon, Paris, France.
- Delaunay, I., V. Ducrocq, and D. Boichard. 2002. A structural model for the matrix of genetic correlations between countries in international evaluation. Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France. Commun. No. 01-14.
- Fikse, W. F., and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. *J. Dairy Sci.* 84:1759-1767.
- INTERBULL. 2005. Genetic correlation estimation procedure. Online. Available http://www-interbull.slu.se/documents/Genetic_correlation_estimation_procedure.pdf. Accessed Jan. 3, 2005.
- Jairath, L., J. C. M. Dekkers, L. R. Schaeffer, Z. Liu, E. B. Burnside, and B. Kolstad. 1998. Genetic evaluation for herd life in Canada. *J. Dairy Sci.* 81:550-562.
- Johnson, D. L., and R. Thompson. 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and an average information. *J. Dairy Sci.* 78:449-456.
- Jorjani, H., L. Klei, and U. Emanuelson. 2003. A simple method for weighted bending of genetic (co)variance matrices. *J. Dairy Sci.* 86:677-679.
- Klei, L., and K. A. Weigel. 1998. A method to estimate correlations among traits in different countries using data on all bulls. *Interbull Bull.* 17:8-14.
- Leclerc, H., S. Minéry, W. F. Fikse, V. Ducrocq, and T. Druet. 2004. Estimation of genetic correlations using a structural model for milk yield and type traits: Limits and opportunities. *Interbull Bull.* 32:65-69.
- Madsen, P., J. Jensen, and T. Mark. 2000. Reduced rank estimation of (co)variance components for international evaluation using AI-REML. *Interbull Bull.* 25:46-50.
- Mäntysaari, E. A. 2004. Multiple-trait across-country evaluation using singular (co)variance matrix and random regression model. *Interbull Bull.* 32:70-74.
- Meyer, K., and M. Kirkpatrick. 2005. Restricted maximum likelihood estimation of genetic principal components and smoothed covariance matrices. *Genet. Sel. Evol.* 37:1-30.
- Minéry, S., W. F. Fikse, and V. Ducrocq. 2003. Application of a structural model to estimate genetic correlations between countries. *Interbull Bull.* 31:175-179.
- Rekaya, R., K. A. Weigel, and D. Gianola. 2001. Application of a structural model for genetic covariances in international dairy sire evaluations. *J. Dairy Sci.* 84:1525-1530.
- Schaeffer, L. R. 1994. Multiple-country comparison of dairy sires. *J. Dairy Sci.* 77:2671-2678.
- Sigurdsson, A., G. Banos, and J. Philipsson. 1996. Estimation of genetic (co)variance components for international evaluation of dairy bulls. *Acta Agric. Scand. A Anim. Sci.* 46:129-136.
- Smith, A. B., B. R. Cullis, and R. Thompson. 2001. Analysing variety by environment data using multiplicative mixed models. *Biometrics* 57:1138-1147.
- Sullivan, P. G. 1999. REML estimation of heterogeneous sire (co)variances for MACE. *Interbull Bull.* 22:146-148.
- Van Der Beek, S. 1999. Exploring the (inverse of the) international genetic correlation matrix. *Interbull Bull.* 22:14-20.