Principal Components Analysis for Conformation Traits in International Sire Evaluations

M. del P. Schneider[1] and W.F. Fikse[2]

[1]*UR337 Station de Génétique Quantitative et Appliquée, Département de Génétique Animale, INRA, 78352 Jouy en Josas, France*
[2]*Interbull Centre, Dept Animal Breeding and Genetics, SLU, P.O. Box 7023, SE-75007 Uppsala, Sweden*

Abstract

Reparameterization using an approximate factor analysis was done to exploit patterns in the genetic correlation matrix of 19 Holstein conformation traits. For the udder traits teat length and rear teat placement, only 3 and 5 principal components explained 98% and 99% of the total variation, respectively; whereas 9 to 13 were necessary for overall conformation traits. The absolute deviations of the reparemeterized genetic correlations from the initial genetic correlations were lower than 0.03 when 98% of the total variation was considered. The countries with the most frequent large contributions (>0.30) to the first eigenvectors were Australia, New Zealand, Switzerland, Estonia and Poland.

1. Introduction

The increasing number of participating countries in the international genetic evaluations, and the lack of genetic links among some of them make the estimation of genetic correlations a difficult task. The application of methods that would result in the reduction of the number of parameters to be estimated is therefore very important. Principal components (**PC**) and factor analysis are two potential approaches that can be used to identify patterns in the data, and to summarize the data with little loss of information. Leclerc *et al.* (2005) has shown that a PC and an approximate factor analysis (**AFA**) gave good approximations of estimated genetic correlations for milk production in international sire genetic evaluations.

The objective of this study was to explore, using an approximate factor analysis approach, patterns in the genetic correlation matrix of conformation traits and to investigate whether these patterns could be exploited for estimation of genetic correlations.

2. Material and Methods

Genetic correlations (**rG**) of 19 Holstein conformation traits from the May 2006 Interbull routine genetic evaluation were used (Interbull, 2006).

The rG matrix of each conformation trait was reparameterized using an approximate factor analysis. First, principal components for the rG matrix were obtained by means of a canonical decomposition of the rG matrix: $\mathbf{rG} = \mathbf{U \cdot D \cdot U'} = \mathbf{V \cdot V'}$, with $\mathbf{U \cdot U'} = \mathbf{I_M}$ and $\mathbf{U' \cdot U} = \mathbf{I_N}$, where $\mathbf{D}$ is the diagonal matrix of eigenvalues, $\mathbf{V}$ is the corresponding set of orthogonal eigenvectors, $\mathbf{V}$ is $\mathbf{U \cdot D}^{1/2}$, and $\mathbf{I}_i$ is the identity matrix of dimension i. 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 rank of $\mathbf{rG}$ can be reduced by setting to zero the smallest eigenvalues in the $\mathbf{D}$ matrix, and deleting the corresponding eigenvectors from $\mathbf{U}$: $\mathbf{P^*} = \mathbf{U^* \cdot D^* \cdot U^{*'}} = \mathbf{V^* \cdot V^{*'}}$, with $\mathbf{U^* \cdot U^{*'}} = \mathbf{I_N}$ and $\mathbf{U^{*'} \cdot U^*} = \mathbf{I_M}$, where $\mathbf{P^*}$ is a matrix with rank N<M, $\mathbf{D^*}$ is the matrix of modified eigenvalues, $\mathbf{U^*}$ contains the corresponding set of orthogonal eigenvectors, and $\mathbf{V^*}$ is the $\mathbf{U^* D^*}^{1/2}$.

The P* matrix does not necessarily have ones on the diagonal anymore, and to obtain the reparemeterized genetic correlation matrix

(**rG***), matrix **P*** can be rescaled ($rG^*_{ij} = P^*_{ij} / \sqrt{P^*_{ii} \times P^*_{jj}}$, referred to as PC approach by Leclerc *et al.* (2005)), or a diagonal matrix **F** can be added to make the diagonal elements of the **rG*** matrix equal to one. The elements of **F** can be considered the specific factors in the factor analysis. These elements were defined as constraints **F = I –**

diag (**P***), and not estimated at the same time as **P***, as is done in the formal factor analysis. With the AFA approach, there is a reduction of the number of parameters to be estimated: N(2M–N+1)/2 (Meyer and Kirkpatrick, 2005). However, **rG*** remains full rank. More details are given in Leclerc *et al.* (2005).

Table 1. Number of principal components necessary to retain 0.980% and 0.990 % of the total variation and maximum, minimum, average and average absolute deviations of the reparameterized genetic correlations with an approximate factor approach from the genetic correlations (rG), for the principal component which retains 98% of the total variation, for 19 Holstein conformation traits.

Trait	NCou[1]	λ^2 98%	λ^2 99%	rG[3] Max	rG[3] Min	rG[3] Abs Avg	F[4] Max	F[4] Min	F[4] Avg
Stature	20	5	7	0.019	-0.020	0.005	0.045	0.005	0.017
Angularity	18	9	11	0.024	-0.030	0.007	0.045	0.003	0.019
Body depth	20	7	9	0.018	-0.020	0.005	0.041	0.001	0.017
Chest width	20	9	10	0.011	-0.020	0.004	0.045	0.002	0.014
Rump angle	20	4	6	0.013	-0.020	0.005	0.030	0.004	0.018
Rump width	20	7	9	0.020	-0.020	0.005	0.034	0.005	0.017
Foot angle	19	10	11	0.018	-0.020	0.004	0.030	0.001	0.014
Rear leg rear view	18	8	9	0.016	-0.020	0.005	0.039	0.001	0.013
Rear leg set	20	8	10	0.025	-0.030	0.006	0.072	0.002	0.018
Udder support	20	8	9	0.014	-0.020	0.004	0.032	0.002	0.014
Fore udder	20	8	10	0.016	-0.020	0.004	0.031	0.002	0.015
Rear udder height	20	8	10	0.021	-0.020	0.006	0.041	0.006	0.017
Udder depth	19	4	6	0.015	-0.020	0.004	0.029	0.003	0.016
Front teat placement	20	5	7	0.019	-0.020	0.005	0.061	0.004	0.019
Teat length	19	3	5	0.018	-0.010	0.005	0.030	0.002	0.018
Rear teat placement	14	3	5	0.011	-0.020	0.005	0.029	0.007	0.019
Overall conformation	19	11	13	0.018	-0.020	0.005	0.046	0.001	0.016
Overall udder	19	9	10	0.018	-0.020	0.004	0.054	0.001	0.014
Overall feet and legs	18	10	11	0.022	-0.030	0.006	0.038	0.000	0.016

[1] Ncou = Number of participating countries in the May 2006 international genetic evaluation.
[2] Number of eigenvalue and corresponding eigenvector used to reparameterize the genetic correlation matrix.
[3] The average deviation for the genetic correlations was -0.001 for all conformation traits.
[4] Elements in diagonal matrix F (F = I – diag(P*)).
Avg = average deviation, Abs Avg = absolute average deviation, Max = maximum deviation, Min = minimum deviation.

3. Results and Discussion

For the udder traits teat length and rear teat placement, only 3 and 5 PCs were enough to explain 98% and 99% of the total variation, respectively (Table 1). Overall conformation traits (overall conformation, overall udder and overall feet and legs) needed 9 to 13 PC. For the linearly scored feet & legs traits 8 to 10 PC

were needed to explain 98% of the variation. No clear pattern was observed for the body and udder traits.

There was considerable variation in the number of PC among the traits. This variation could reflect differences in trait definition, models, genetic by environment interactions (i.e. tied *vs.* free stalls for feet & legs traits).

The higher number of components for the overall traits reflects the variability in trait definition.

Deviations of the reparemeterized genetic correlations with the AFA approach from the rG, for various numbers of eigenvalues and corresponding eigenvectors, are presented in Table 1 and 2. In Table 1, deviations for the PC which retained 98% of the total variation are shown. Note that the number of PC and the number of participating countries in the genetic evaluation varies among the traits. For rG, the maximum and minimum deviations were <0.03 in all cases, and the absolute deviations ranged from 0.004 to 0.007. Therefore, genetic correlations obtained with the reparameterized rG matrix were very close to the "initial" rG. The average values of F (elements added to the diagonal) were <0.020. The F components represent the components specific for each country.

Two extreme examples, teat length and overall conformation, are presented in Table 2. Three and 11 PC explained 98% of the total variation, respectively. For teat length only one PC was needed to obtain an average absolute deviation <0.03, whereas 4 PC were needed for overall conformation. For the F components, 2 and 9 PC were necessary to get an average deviation <0.03, for teat length and overall conformation, respectively.

Table 2. Maximum, minimum, average and average absolute deviations of the reparameterized genetic correlations with an approximate factor approach from the genetic correlations (rG), for teat length (TL) and overall conformation (OC).

			rG								F^3					
	Cum. prop.[2]		Max		Min		Avg		Abs Avg		Max		Min		Avg	
λ^1	TL	OC	TL	OC	TL	OC	TL	OC	TL	OC	TL	OC	TL	OC	TL	OC
1	0.963	0.756	0.034	0.311	-0.052	-0.290	-0.002	-0.013	0.009	0.049	0.100	0.600	0.010	0.126	0.037	0.244
2	0.975	0.821	0.025	0.155	-0.023	-0.153	-0.001	-0.010	0.006	0.041	0.068	0.328	0.010	0.116	0.025	0.179
3	0.982	0.863	0.018	0.140	-0.015	-0.113	-0.001	-0.007	0.005	0.031	0.030	0.301	0.002	0.040	0.018	0.137
4	0.987	0.898	0.012	0.089	-0.015	-0.108	-0.001	-0.006	0.003	0.022	0.027	0.280	0.001	0.030	0.013	0.102
5	0.990	0.926	0.009	0.060	-0.009	-0.059	-0.001	-0.004	0.003	0.015	0.020	0.143	0.001	0.018	0.010	0.074
6	0.993	0.943	0.006	0.028	-0.007	-0.060	0.000	-0.003	0.002	0.012	0.014	0.091	0.001	0.015	0.007	0.057
7	0.995	0.954	0.005	0.025	-0.005	-0.034	0.000	-0.003	0.001	0.010	0.012	0.083	0.001	0.014	0.005	0.046
8	0.997	0.963	0.003	0.022	-0.005	-0.034	0.000	-0.002	0.001	0.008	0.011	0.082	0.001	0.002	0.003	0.037
9	0.998	0.971	0.003	0.022	-0.004	-0.036	0.000	-0.002	0.001	0.007	0.010	0.061	0.001	0.001	0.002	0.029
10	0.999	0.978	0.002	0.018	-0.003	-0.026	0.000	-0.001	0.001	0.006	0.004	0.048	0.000	0.001	0.001	0.022
11	0.999	0.984	0.001	0.018	-0.002	-0.023	0.000	-0.001	0.000	0.005	0.004	0.046	0.000	0.001	0.001	0.016

[1] Number of eigenvalue and corresponding eigenvector used to reparameterize the genetic correlation matrix.
[2] Cum. Prop = cumulative proportion explained variance by the eigenvalues/eigenvectors.
[3] Elements in diagonal matrix F (F = I – diag(P*)).
Avg = average deviation, Abs Avg = absolute average deviation, Max = maximum deviation, Min = minimum deviation.

Table 3 summarizes how many times a given country had large contributions (>0.30) to the first eigenvectors. For instance, Australia (AUS) had 13, 4, and 7 times large contributions (>0.30) to the second ($\lambda 2$), third ($\lambda 3$), and fourth ($\lambda 4$) eigenvector, respectively. The countries with the most frequent large contributions were Australia, New Zealand (NLZ), Switzerland (CHE/CHR), Estonia (EST) and Poland (POL). These groups of countries were relatively different from the rest. Australia and New Zealand have different production systems compared to Northern Hemisphere countries. Estonia has relatively weak links and the estimated genetic correlations (rG) have relatively high standard errors.

With the AFA approach the number of parameters to be estimated is reduced. For example, if only 3 PC are kept for teat length,

the number of parameters will be reduced from 190 to 54.

Stature data (*i.e.* deregressed breeding values) were further analyzed with the approach proposed by Leclerc *et al.* (2005) (results not shown). Five PC (*i.e.* countries) were chosen to be the "base countries" together with two "link provider" major countries to ensure good connectedness of the data. Correlations among the base countries and between the base and other countries were estimated, and correlations among other countries were computed. Deviations of the reparameterized genetic correlation matrix with the AFA approach from the initial rG were large for some set of countries, much larger in comparison with the results by Leclerc for milk production. Surprisingly, the addition of a country with large deviations as base country did not reduce the deviations. Choosing a "good" base countries group turned out to be a difficult task, and procedures for objective choice of base countries need to be devised before this approach can be considered for routine application.

Table 3. Number of times a country had large contributions (>0.30) to the second, third and fourth eigenvectors (λ).

Country	λ_2	λ_3	λ_4
AUS	13	4	7
BEL	2	0	3
CAN	1	0	0
CHE	5	2	1
CHR	5	5	6
CZE	1	6	1
DEU	0	0	4
DFS	3	1	3
ESP	0	1	2
EST	5	10	8
FRA	4	1	1
GBR	3	1	1
HUN	3	1	4
ITA	3	2	3
JPN	3	0	3
NLD	0	1	1
NZL	7	7	5
POL	8	10	4
USA	2	2	4
ZAF	1	1	2

λ_2 = second eigenvector, λ_3 = third eigenvector, λ_4 = fourth eigenvector.

4. Conclusions

Results showed that by keeping a number of PC which explained 98% of the total variation, absolute deviations < 0.03 could be obtained.

Udder support needing equally many PCs to retain 98% of the variation as two of the feet & legs traits was somewhat surprising. This type of analyses provide worthwhile information for breed societies, *e.g.* the World Holstein-Friesian Federation, in their efforts to harmonize type classification across countries: it tells which traits are similarly defined across countries, and which countries have deviating trait definitions.

Results were used for the estimation of genetic correlations in a preliminary study using stature data. The selection of the base country group turned out to be difficult. The large variability found among the type traits, will require selection and grouping of countries specific for each trait.

AFA is an approximate method to estimate genetic correlations; more research is still needed in the field.

References

Interbull, 2006. Interbull Routine Genetic Evaluation for Conformation Traits February 2006. http://www.interbull.slu.se/conform/framesida-conf.htm. Accessed February 15, 2006.

Leclerc, H., Fikse, W.F. & Ducrocq, V. 2005. Principal components and factorial approaches for estimating genetic correlations in international sire evaluation. *J. Dairy Sci. 88,* 3306-3315.

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