Principal Component Approach in Describing the Across Country Genetic Correlations

A.-M. Tyrisevä[1], M. H. Lidauer[1], V. Ducrocq[2], P. Back[1], F. Fikse[3] and E. A. Mäntysaari[1]
[1]MTT Agrifood Research Finland, Biotechnology and Food Science, Biometrical Genetics, Jokioinen, Finland;
[2]INRA, Station de Génétique Quantitative et Appliquée, Jouy-en-Josas, France;
[3]Interbull Centre, Dept. Animal Breeding and Genetics, SLU, Uppsala, Sweden

1. Introduction

Current multiple-trait across country evaluation (MACE) for Holstein includes 25 countries/traits. Estimation of the variance-covariance (VCV) matrix is difficult since the size of the matrix is too large to be estimated in a single analysis and the model is over-parameterized due to high genetic correlations between the different countries (e.g., van der Beek 1999).

Different strategies have been proposed to solve problems in variance component estimation when VCV matrix is close to singularity. Mäntysaari (2004) suggested a method involving an ascending sequence of variance component estimation, rank reduction and addition of a new trait/country until all traits/countries are included in the analysis. Leclerc *et al.* (2005) presented a method where the rank of the Interbull VCV matrix for well-connected countries is first reduced using principal components (PC) or factor analysis (FA). Then covariances between the reduced rank VCV matrix of the well-connected countries and the other countries are computed and finally the remaining missing variance components for the other countries are derived. Another appealing approach is the direct estimation of the leading PC of the VCV matrix (Meyer 2008). However, the proposed strategies require a prior knowledge of the number of the leading PC in the VCV matrix.

The objective of the paper is to investigate the effects of rank reduction in MACE analyses and its effect on the breeding values. Opportunities and constraints of PC and FA approach will be assessed.

2. Material and methods

2.1MACE as a random regression model

The classical MACE sire model (CM) including *t* countries can be written as:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \qquad [1]$$

where $\mathbf{y}_i$ is a vector of national de-regressed breeding values for bull i ($i=1, \ldots, n$); $\mathbf{b}$ is a vector of country effects, $\mathbf{u}_i$ is a vector of t international breeding values of bull i with $\text{var}(\mathbf{u}_i) = \mathbf{G}_0$, $\boldsymbol{\varepsilon}_i$ is a vector of residuals with $\text{var}(\boldsymbol{\varepsilon}_i) = diag\{g_{kk}\lambda_k / n_{ik}\}$, and $\mathbf{X}_i$ and $\mathbf{Z}_i$ are incidence matrices of appropriate order. $\mathbf{G}_0$ is the $t \times t$ genetic VCV matrix, g_{kk} is the diagonal element of $\mathbf{G}_0$ for country k, $\lambda_k = \left(4 - h_k^2\right)/h_k^2$, and n_{ik} is the effective daughter contribution for bull i in country k. Genetic groups are accounted by incorporating random phantom parent groups into the additive genetic relationship matrix.

Decomposing $\mathbf{G}_0 = \mathbf{S}_0\mathbf{V}_0\mathbf{D}_0\mathbf{V}_0^{\mathrm{T}}\mathbf{S}_0$, where $\mathbf{S}_0$ is a diagonal matrix with genetic standard deviations, and $\mathbf{V}_0\mathbf{D}_0\mathbf{V}_0^{\mathrm{T}}$ is the eigenvalue decomposition of the genetic correlation matrix with $\mathbf{V}_0$ being the dense matrix of eigenvectors and $\mathbf{D}_0$ the diagonal matrix of eigenvalues, allows setting up an equivalent random regression model (RRM):

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{S}_0\mathbf{V}_0\boldsymbol{\nu}_i + \boldsymbol{\varepsilon}_i, \qquad [2]$$

where $\boldsymbol{\nu}_i$ is a vector of t regression coefficients for bull i with $\text{var}(\boldsymbol{\nu}_i) = \mathbf{D}_0$. Then, the breeding values of bull i can be back transformed as: $\mathbf{u}_i = \mathbf{S}_0\mathbf{V}_0\boldsymbol{\nu}_i$.

2.2 Reduced rank

Principal components (PC). When $\mathbf{G}_0$ is close to singular, only r ($r<t$) eigenvalues in $\mathbf{D}_0$ affect on the (co)variances in $\mathbf{G}_0$. Hence, $\mathbf{G}_0$ can be approximated by the singular matrix $\mathbf{G}_{PC} = \mathbf{S}\mathbf{V}_1\mathbf{D}_1\mathbf{V}_1^{\mathrm{T}}\mathbf{S}$ of rank r, where $\mathbf{D}_1$ contains the r largest eigenvalues, $\mathbf{V}_1$ the corresponding eigenvectors, and $\mathbf{S}$ is a diagonal scaling matrix that returns the original variances. Replacing $\mathbf{V}_0$ in model [2] by $\mathbf{V}_1$ will yield for each bull r random regression coefficients ($\mathbf{v}_i^*$) that can be used as $\mathbf{u}_i \cong \mathbf{S}\mathbf{V}_1\mathbf{v}_i^*$.

Factor analysis (FA). Alternatively, $\mathbf{G}_0$ can be decomposed into two variance terms, one term describing the VCV structure common to all countries and another the country specific variances: $\mathbf{G}_0 = \mathbf{L}_0\mathbf{W}_0\mathbf{L}_0^{\mathrm{T}} + \mathbf{F}_0$. The FA gives a parsimonious structure to VCV matrix and therefore is of interest. Hence, the FA model for MACE is:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{L}_0\boldsymbol{\delta}_i + \boldsymbol{\tau}_i + \boldsymbol{\varepsilon}_i, \qquad [3]$$

where $\boldsymbol{\delta}_i$ is a vector of r regression coefficients for bull i with $\mathrm{var}(\boldsymbol{\delta}_i) = \mathbf{W}_0$, $\boldsymbol{\tau}_i$ is a vector of country specific breeding values with $\mathrm{var}(\boldsymbol{\tau}_i) = \mathbf{F}_0 = diag\{\sigma^2_{\tau_{ij}}\}$, and $\mathrm{cov}(\boldsymbol{\delta}_i,\boldsymbol{\tau}_i)=0$. The back transformed breeding values for bull i are: $\mathbf{u}_i = \mathbf{L}_0\boldsymbol{\delta}_i + \boldsymbol{\tau}_i$.

The first term of the FA model allows a reduced order fit. An EM-algorithm (Mäntysaari 1999) was applied for an approximated decomposition of $\mathbf{G}_0$ into $\mathbf{G}_{FA} = \mathbf{L}\mathbf{W}\mathbf{L}^{\mathrm{T}} + \mathbf{F}_0$, with $\mathbf{L}\mathbf{W}\mathbf{L}^{\mathrm{T}}$ of order r.

2.3 Test application

Data used for a test application was August 2007 MACE Interbull Holstein evaluation for protein yield including all 25 countries. CM evaluation was compared to four RRM evaluations where only the largest 20, 17, 15 and 10 PC were considered. CM evaluation was also compared to FA RRM with order 10. Only results highlighting key features are shown.

3. Results

3.1 Approximated genetic VCV matrix

Variances remained unchanged in the tested approaches ensuring that variance differences were unaffected. However, applying FA yielded a clear separation between the common variance (variances of the loadings) and country specific variances (Table 1). The latter explained up to 17% (for Czech Republic) of the total variance.

Table 1. Genetic variance for Holstein protein MACE when rank/order is reduced to 10 either by means of principal components (PC) or by means of factor analysis.

	Classical model, Rank 25	Rank 10		
			Factor analysis	
		PC	Variance of loadings	Country specific variance
Canada	122.32	122.32	119.54	2.78
France	87.24	87.24	85.33	1.91
Italy	85.56	85.56	76.25	9.31
USA	334.89	334.89	328.98	5.91
New-Zealand	21.25	21.25	18.47	2.78
Belgium	45.43	45.43	38.66	6.77
Czech Republic	87.80	87.80	74.86	12.94
Slovenia	7.67	7.67	6.80	0.87
Japan	67.24	67.24	67.04	0.20
Latvia	21.81	21.81	21.55	0.26

Original genetic correlations between countries vary from 0.75 to 0.93, being lowest between New-Zealand and other countries (Table 3). The PC approach shows hardly any differences in genetic correlations, when rank was reduced to 20, and acceptable differences (average increase of 2%) with rank reduced to 17. When rank was further reduced to 15 and 10 genetic correlations increased almost uniformly, being on average 3% (rank 15) and 6% (rank 10) higher than the original correlations (Tables 3 and 4). However, with the FA approach and an order down to 10, genetic correlations between countries remained unchanged (Table 4).

3.2 Effect of PC and FA approaches on breeding values

Rank/order reduction to 15 or 10 had no influence on the breeding values of bulls that have been used only in their own country.

Studying the group of bulls that had daughters both in own and foreign countries revealed a decrease in correlations between breeding values from CM and breeding values from RRM model with rank reduced to 15 or 10 (Table 5). The decrease in correlations was clearest for bulls from New-Zealand, whose production system differ substantially from the other countries, as well as in countries like Slovenia, where only few bulls are of own origin.

3.3 Solving the reduced rank MACE model

Rank reduction by PC method yielded a clear reduction in the number of equations and computing time needed for solving the model. The CM included 2.7 million equation, required 123 Mb of memory and 11 minutes of solving time, when solving with a preconditioned conjugated gradient iteration on data algorithm. Corresponding values for the reduce rank MACE model with rank 10 were 1.1 million equations, 61 Mb memory, and 3 minutes of solving time, respectively. In contrast, FA will not yield any reduction in the number of equations and computing time.

4. Discussion

Rank reduction by means of PC tended to increase genetic correlations between countries. Using only the 17 largest PC, which explained 98% of the total variance, did not affect breeding values. Further reduction to rank 15 or 10 showed changes in breeding values in the countries with weak ties with the other countries. It is difficult to assess the significance of these changes since the accuracy for the same (co)variances in the original $\mathbf{G}_0$ matrix may be low. For Italy and Slovenia, genetic correlations increased on average by 6.8% to 0.924 and 8.2% to 0.927, respectively, when rank was reduced to 15. However, for Italy the correlation between breeding values from CM evaluation and PC reduced rank 15 evaluation was 0.9999, while for Slovenia corresponding correlation was as low as 0.93. It can be speculated whether rank reduction to as low as 15 is justified. The low correlations between full and reduced rank breeding values for countries like Slovenia may be related to genetic correlations with other countries which might be too low in the current MACE model.

Reduced order and more parsimonious model when using FA was clearly superior to PC and suggests that order might be less than 10 when applying FA. This may open an opportunity for direct estimation of reduced rank parameters, which will sufficiently approximate $\mathbf{G}_0$. As found here, FA does not reduce the size of the problem for the breeding value estimation. However, this might be of minor importance.

5. Conclusions

Results suggest that a considerable reduction in rank/order of the genetic VCV matrix for multiple-trait across country evaluation is possible. This supports an attempt to step-wisely or directly estimate the reduced rank/order genetic VCV matrix with reliable genetic correlation estimates among all participating countries.

References

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

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. 2008. Parameter expansion for estimation of reduced rank covariance matrices (Open access publication). *Genet. Sel. Evol. 40*, 3-24.

Mäntysaari, E.A. 1999. Derivation of multiple trait reduced rank random regression (RR) model for the first lactation test day records of milk, protein and fat. *Proc 50th EAAP*, 22.-26. 8. 1999. Zurich, Switzerland.

Mäntysaari, E.A. 2004. Multiple-trait across country evaluations using singular (co)variance matrix and random regression model. *Interbull Bulletin 32*, 70-74.

Table 3. Genetic correlations between a sample of ten countries (upper triangle) and changes when reducing rank to 15 by means of principal components (lower triangle).

	Canada	France	Italy	USA	New-Zealand	Belgium	Czech Republic	Slovenia	Japan	Latvia
Canada		0.90	0.90	0.93	0.75	0.85	0.85	0.86	0.93	0.86
France	0.04		0.88	0.89	0.75	0.85	0.85	0.86	0.90	0.86
Italy	0.05	0.04		0.92	0.75	0.85	0.85	0.86	0.89	0.86
USA	0.05	0.02	0.07		0.75	0.85	0.85	0.86	0.92	0.86
New-Zealand	0.01	0.01	0.03	0.02		0.75	0.75	0.77	0.75	0.77
Belgium	0.01	0.01	0.03	0.02	0.00		0.85	0.87	0.85	0.87
Czech Republic	0.02	0.01	0.02	0.02	0.01	0.00		0.87	0.85	0.86
Slovenia	0.04	0.03	0.04	0.04	0.03	0.04	0.04		0.86	0.89
Japan	0.04	0.03	0.04	0.03	0.02	0.02	0.01	0.04		0.87
Latvia	0.05	0.03	0.04	0.03	0.03	0.02	0.01	0.08	0.02	

Table 4. Changes in genetic correlations between a sample of ten countries when rank/order of genetic correlation matrix was reduced to 10; either by means of factor analysis (upper triangle) or by means of principal components (low triangle).

	Canada	France	Italy	USA	New-Zealand	Belgium	Czech Republic	Slovenia	Japan	Latvia
Canada		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
France	0.05		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Italy	0.07	0.06		0.00	0.00	0.00	0.00	0.00	0.00	0.00
USA	0.06	0.05	0.07		0.00	0.00	0.00	0.00	0.00	0.00
New-Zealand	0.04	0.03	0.05	0.04		-0.01	0.00	0.00	0.00	0.00
Belgium	0.04	0.05	0.05	0.03	0.05		0.01	0.01	0.00	0.00
Czech Republic	0.05	0.02	0.05	0.05	0.04	0.03		0.00	0.00	0.00
Slovenia	0.06	0.06	0.06	0.05	0.06	0.10	0.09		0.00	0.00
Japan	0.05	0.07	0.04	0.03	0.04	0.03	0.05	0.06		0.00
Latvia	0.07	0.07	0.08	0.06	0.07	0.10	0.09	0.10	0.06	

Table 5. Correlation of breeding values for bulls from full rank and reduced rank/order MACE evaluation. Here, bulls have daughters both in own and foreign countries.

Country	PC Full vs. rank15	PC Full vs. rank10	FA Full vs. order10	No of bulls	Proportion of domestic daughters, %	Domestic daughters/bull on average	Foreign daughters/bull on average	Ratio of domestic/ foreign daughters per bull, %
Canada	0.9999	0.9989	0.9999	1907	21	1071	4017	26.7
France	0.9999	0.9978	0.9999	1371	38	3875	6313	61.4
Italy	0.9999	0.9970	0.9999	1062	16	2116	11455	18.5
USA	0.9998	0.9993	1.0000	3207	22	1059	3818	27.7
New-Zealand	0.9802	0.9767	0.9999	1237	17	1607	7946	20.2
Belgium	0.9991	0.9416	0.9999	587	1	157	17479	0.9
Czech Republic	0.9996	0.9981	0.9999	1251	3	226	8164	2.8
Slovenia	0.9323	0.8654	0.9994	75	1	193	30207	0.6
Japan	0.9975	0.9931	0.9999	232	5	1209	24635	4.9
Latvia	0.9277	0.9061	0.9998	119	<1	73	20550	0.4