Recommendations for Estimation of Variance Components for International Sire Evaluation

A.-M. Tyrisevä[1], K. Meyer[2], F. Fikse[3], V. Ducrocq[4], J. Jakobsen[5], M. H. Lidauer[1], E. A. Mäntysaari[1]
[1]*MTT Agrifood Research Finland, Biotechnology and Food Research, Biometrical Genetics, Jokioinen, Finland;*
[2]*Animal Genetics and Breeding Unit, University of New England, Armidale, Australia;*
[3]*Dept. Animal Breeding and Genetics, SLU, Uppsala, Sweden;*
[4]*UMR1313 Animal Genetics and Integrative Biology, 78350 Jouy en Josas, France;*
[5]*Interbull Centre, Dept. Animal Breeding and Genetics, SLU, Uppsala, Sweden*

Abstract

This study assessed the impact of alternative parameterizations for the estimation of variance components on practical predictions of breeding values with MACE. Interbull MACE Holstein evaluations for somatic cell count (April 2009) and protein yield (August 2007) were considered. The MACE model was expressed in terms of a random regression model, which facilitates exploitation of principal component and factor analytic approaches. Both methods allow a reduction of the number of parameters to be estimated and benefit from the more parsimonious variance structure. Genetic parameters from different approaches were very similar, when the optimal fit was used. Over-fitting did not affect the estimates, but increased estimation time, whereas fitting too few parameters affected bull rankings in different countries.

Introduction

Principal component (PC) and factor analytic (FA) approaches to model covariance matrices facilitate a reduction in the number of parameters to be estimated and are thus an attractive proposition to ease the computational burden of variance component estimation for multiple-trait across country evaluation (MACE, e.g., Mäntysaari, 2004, Leclerc *et al.*, 2005). Both methods decompose genetic covariance matrices into the pertaining matrices of eigen-values and -vectors, i.e. principal components. For highly correlated traits, some principal components explain virtually no genetic variation (i.e. have eigenvalues close to zero) and can be omitted, so that only the leading principal components are fitted in the model.

The bottom-up PC approach, a sequential PC method proposed by Mäntysaari (2004), is designed for large-scale, over-parameterized models, but has so far been tested on a simulated data set only. The direct PC and FA approaches, suggested by Kirkpatrick and Meyer (2004), were also designed for a large, multi-trait framework and showed their potential in large beef cattle data sets (e.g., Meyer, 2007a). The aim of this study was to assess the usefulness of the bottom-up PC, direct PC and FA approaches for MACE.

Material and Methods

Random regression MACE

The random regression (RR) MACE model for sire *i* is expressed as follows:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{V}\mathbf{v}_i + \boldsymbol{\varepsilon}_i, \qquad (1)$$

where $\mathbf{y}_i$ is the vector of n_i de-regressed, national breeding values for bull *i*, $\mathbf{b}$ is the vector of *t* country effects, $\mathbf{v}_i$ is the vector of *t* regression coefficients for bull *i*, and $\boldsymbol{\varepsilon}_i$ is the corresponding vector of n_i residuals. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are incidence matrices assigning observations to the respective effects. $\mathbf{G}$, the $t \times t$ VCV matrix of sire breeding values $\mathbf{u}_i$, is decomposed into the matrices of eigenvalues ($\mathbf{D}$) and eigenvectors ($\mathbf{V}$); $\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ with $\mathbf{D} = Var(\mathbf{v}_i)$. Further, the residual variance is $Var(\boldsymbol{\varepsilon}_i) = diag(g_{jj}\lambda_j / EDC_{ij})$ with g_{jj} the sire variance for country *j*, $\lambda_j = (4 - h_j^2)/h_j^2$ with h_j^2 the heritability in country *j* and EDC_{ij} the effective daughter contribution for bull *i* in country *j*. In this study, genetic groups were not included for the variance component estimations, but they were included for the predictions of the breeding values.

PC approaches

The direct PC approach fits the leading principal components directly, requiring multiple analyses to determine the appropriate rank (Kirkpatrick and Meyer, 2004). The bottom-up PC approach, in turn, starts analysis with a sub-set of countries and adds new countries sequentially, assessing in each step whether or not a new country increases the rank (Tyrisevä *et al.*, 2009). Now, $\mathbf{G}_1 = \mathbf{V}_1\mathbf{D}_1\mathbf{V}_1^T$, where $\mathbf{D}_1$ contains the r largest eigenvalues and $\mathbf{V}_1$ the r corresponding eigenvectors, with $r < t$.

Factor analytic approach

The FA approach is closely related to the PC approach, but it divides additive genetic effects and their variances into a common and a trait specific part. Now, the RR MACE model is re-parameterized as follows:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i(\mathbf{L}\boldsymbol{\delta}_i + \boldsymbol{\tau}_i) + \boldsymbol{\varepsilon}_i, \qquad (2)$$

where $\boldsymbol{\delta}_i$ is the vector of common factors, with $Var(\boldsymbol{\delta}_i) = \mathbf{I}$, and $\boldsymbol{\tau}_i$, the vector of country specific effects, with $Var(\boldsymbol{\tau}_i) = \mathbf{F} = diag\left\{\sigma^2_{\tau ij}\right\}$. $\mathbf{L}$ denotes the matrix of factor loadings. The sire effects are now expressed as: $\mathbf{u}_i = \mathbf{L}\boldsymbol{\delta}_i + \boldsymbol{\tau}_i$, and the VCV matrix as: $\mathbf{G} = \mathbf{L}\mathbf{I}\mathbf{L}^T + \mathbf{F}$.

The resulting **G** in the FA model is of full rank, provided that none of the trait-specific variances is zero. However, the model is very parsimonious since the matrix of factor loadings can be of reduced rank as in the PC approach. Usually, the number of common factors is notably smaller than the number of PCs in the reduced rank model.

Data sets

Somatic cell count (SCC) from April 2009 and protein yield from August 2007 MACE Interbull Holstein evaluations were used as test data sets. The SCC data comprised 100 551 bulls from 23 countries and that for protein yield 103 676 bulls from 25 countries. Number of bulls per country ranged from less than 200 to more than 20 000 for both traits. The majority of bulls had daughters in one country only. The number of common bulls – defined as bulls with daughters in a pair of countries without restrictions on the country of origin – varied from one to 1 526 for SCC with a mean of 240, and from zero to 1 194 for protein yield with a mean of 158. A sire model with a sire-maternal grandsire pedigree was fitted with 107 728 animals for SCC and 106 003 animals for protein yield.

In addition, later data sets were analysed for both traits, to test the estimation time of variance components under the direct PC approach and with the earlier calculated parameters as starting values. For SCC, April 2010 evaluation data with 24 countries, and for protein yield, April 2009 evaluation data with 26 countries were tested. Now, the number of bulls for SCC was 110 049 with 114 290 bulls in the pedigree and for protein yield 109 845 bulls with 115 049 bulls in the pedigree.

Models

For both traits, ranks used in the direct PC analyses were estimated using the bottom-up PC approach (Tyrisevä *et al.*, 2009). For SCC, the appropriate rank was 15 (PC15) and that for protein yield was 20 (PC20). Further, for both traits, additional analyses under the models fitting too low or too high a rank were performed for comparison: rank 10 (PC10) and rank 23 (PC23) for SCC and rank 15 (PC15) and rank 25 (PC25) for protein yield.

The correct fit in the FA analyses was determined following the suggestions by Meyer and Kirkpatrick (2008). The appropriate fits were 7 (FA7) for SCC and 9 (FA9) for protein yield.

While estimating new variance components for the updated data sets, input parameters were those obtained from the 2009 analysis for SCC and from the 2007 analysis for protein yield. The starting value for a new country was defined as the average variance of the countries already included in the model and covariances for a new country were derived from an average correlation of the countries

already in the model. Ranks used were the same as in the previous analyses.

Variance components were estimated by restricted maximum likelihood, using an average information algorithm as implemented in WOMBAT (Meyer, 2007b). The numbers of parameters for the SCC 2009 were: 186, 241, 277 and 164 for PC10, PC15, PC23 and FA7, respectively. The numbers of parameters for the protein yield 2007 were: 271, 311, 326 and 215 for PC15, PC20, PC25 and FA9, respectively.

Estimated breeding values

The prediction of breeding values in (1) and (2) followed Tyrisevä *et al.* (2008). The correlations of the estimated breeding values (EBVs) between different approaches under the optimal fit were studied, as well as the correlations between the direct PC under the optimal and too low a rank. Further, correlations between EBVs from the direct PC analysis under the optimal rank and EBVs from Interbull are presented for comparison. The same comparisons were performed in four subgroups defined as: A) bulls used only in their own country, B) bulls used in their own country and abroad, C) bulls used only abroad, and D) imported bulls.

Breeding values were obtained using a preconditioned conjugated gradient iteration on data algorithm as implemented in MiX99 (Vuori *et al.*, 2006).

Results and Discussion

Genetic correlations

The estimates of genetic correlations from the PC and FA approaches under the optimal fit were almost identical, except for some differences in minimum values (Table 1). Further, the non-post-processed Interbull estimates were in a good accordance with them.

There was a notable difference in the general level of genetic correlations between SCC and protein yield. Those for the protein yield were on a lower level (mean values: SCC 0.88, protein yield 0.69) and some of the correlations were extremely low (Table 1). The low genetic correlations of protein yield were associated with low number of records and weak ties with the other countries. Further, Jakobsen *et al.* (2009) observed that different trait definitions and national genetic evaluation models, as well as genotype by environment interactions cause low to moderate genetic correlations between countries. Currently, Interbull performs a post-processing step (Interbull, 2010) that raises the genetic correlations to a level that corresponds to the common knowledge of their level. Thus, the difference of the post-processed genetic correlations with the estimates from the other approaches for protein yield is relatively large (Table 1).

Variance component analyses for SCC 2009 required 5, 3, 16, 3, 7 and >30 days for FA6, FA7, FA8, PC10, PC15 and PC23, respectively. Variance component analyses for protein yield 2007 took 14.5, 3.5, 31.5, 21.5, 5 and 16.5 days for FA7, FA9, FA11, PC15, PC20, and PC25, respectively. Estimation times of the optimal fits are underlined. It seems clear that the PC and FA analyses had notable problems to find the maximum, if the fit was selected wrongly. Updating the variance components for SCC, computing time was reduced from 7 to 5 days (Table 2) and remained the same for protein yield (Table 3).

EBVs

EBVs from the PC and FA approaches under the optimal fit were identical or almost identical, with few exceptions (Table 4). Further, correlations among EBVs from the full and optimal rank models were unity (Tyrisevä *et al.*, 2010; in preparation). EBV correlations from different approaches were in places lower than 0.99 for countries with few records and low number of common bulls with the other countries. As expected, the differences were largest in the subgroup C (bulls used only abroad), in which the prediction of the breeding values was totally based on the links through the pedigree (Table 4).

EBVs started to differ, when the predictions between the optimal and too low a fit were compared (Table 4). Based on the current study, the estimation of the variance components is more sensitive to the use of correct fit than prediction of the EBVs (Tyrisevä *et al.*, 2010; in preparation). Further, the Interbull's post-processing step of the parameters created larger differences in the EBV correlations than too strong a rank reduction (Table 4).

Times required for solving the MACE system were at most 6 min for SCC and 7 min for protein yield, both with Interbull predictions. Solving times for the models under the optimal fit were at their shortest 4 min for SCC, and 5 min for protein yield.

Recommendations

The study shows that the principal component and factor analytic approaches are useful in variance component estimation and breeding value prediction for the international sire evaluation. Both methods facilitate more parsimonious models that are useful in MACE, where over-parameterized models and the problems associated with them are common. Solving time was clearly shortest for the models fitting optimal number of parameters. Further, no data sub-setting was needed for such models.

Using too low a fit affects the accuracy of variance component estimation, as well as the accuracy of breeding value predictions. Based on our results, it is, however, unlikely that minor deviations from the correct fit have practical influences. Over-parameterization had no influence on the accuracy, but it notably increased the estimation time.

The bottom-up PC approach can be utilized to determine the correct rank and the direct PC approach for routine analyses. Ranks must be re-estimated only when major changes occur in the data sets. Further, there is no need to start the bottom-up PC process from the beginning, but utilize the last parameters and rank as input for a new analysis (Table 2, 3). Currently, there is no such an option for the FA approach.

This study clearly reveals some problems associated with disconnectedness and variability of data sets. We therefore recommend that participating countries pay attention to the quality of their national genetic evaluation models. Estimation difficulties associated with the quality of the data causes low genetic correlations, lower accuracy of the breeding value predictions and longer solving time of the bottom-up PC (Table 3). The bottom-up PC run for protein yield clearly demonstrated this. The majority of the first 15 countries introduced in the analysis were well-connected, large countries. They contributed 88% of the total data, but the computing time was less than 9% of the total time used. On the other hand, the computing time for the bottom-up PC for SCC was almost the same as for the direct PC (Table 2), hinting that there were no large problems embedded in the data sets with this trait.

References

Interbull, 2010. Genetic correlation estimation procedure. URL: http://www-interbull.slu.se/documents/Genetic_correlation_estimation_procedure_2007t2.pdf

Jakobsen, J.H., Dürr, J.W., Jorjani, H., Forabasco, A., Loberg, A. & Philipsson, J. 2009. Genotype by environment interactions in international genetic evaluations of dairy bulls. In: *Proc 18th Assoc. Advmt. Anim. Breed. Genet., Roseworthy, Australia,* 133-142.

Kirkpatrick, M. & Meyer, K. 2004. Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. *Genetics 168,* 2295-2306.

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. 2007a. Multivariate analyses of carcass traits for Angus cattle fitting reduced rank and factor analytic models. *J. Anim. Breed. Genet. 124,* 50-64.

Meyer, K. 2007b. "WOMBAT - A tool for mixed model analyses in quantitative genetics by REML". *J. Zheijang Univ. Sci. B 8,* 815-821.

Meyer, K. & Kirkpatrick, M. 2008. Perils of parsimony: Properties of reduced-rank estimates of genetic covariance matrices. *Genetics 180,* 1153-1166.

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

Tyrisevä, A.-M., Lidauer, M.H., Ducrocq, V., Back, P., Fikse, W.F. & Mäntysaari, E. A. 2008. Principal component approach in describing the across country genetic correlations. *Interbull Bulletin 38,* 142–145.

Tyrisevä, A.-M., Meyer, K., Fikse, F., Ducrocq, V., Jakobsen, J., Lidauer, M.H. & Mäntysaari, E.A. 2009. Comparison of different variance component estimation approaches for MACE – direct and bottom-up PC. *Interbull Bulletin 40,* 72-76.

Vuori, K., Strandén, I. & Lidauer, M.H. 2006. MiX99 – effective solver for large and complex linear mixed models. In: *Proc. 8th WCGALP, Belo Horizonte, Brazil,* CD-ROM, No 27: 33.

Table 1. Descriptive statistics of estimates of genetic correlations from different approaches.

Approach	Min	1st quartile	Median	Mean	3rd quartile	Max
Somatic cell count 2009						
Direct PC, rank 15	0.62	0.84	0.88	0.87	0.92	0.97
Bottom-up PC, rank15	0.65	0.85	0.89	0.88	0.92	0.98
Factor analysis, fit 7	0.63	0.84	0.89	0.87	0.92	0.97
Non-post-processed Interbull	0.61	0.86	0.90	0.89	0.93	0.98
Post-processed Interbull	0.74	0.86	0.90	0.89	0.93	0.97
Protein yield 2007						
Direct PC, rank 20	0.08	0.56	0.71	0.69	0.82	0.94
Bottom-up PC, rank 20	0.05	0.57	0.71	0.68	0.81	0.94
Factor analysis, fit 9	0.13	0.57	0.71	0.69	0.82	0.94
Non-post-processed Interbull	0.02	0.59	0.74	0.70	0.83	0.94
Post-processed Interbull	0.75	0.79	0.85	0.84	0.87	0.93

Table 2. Variance component estimation times of the principal component approaches for somatic cell count.

Bottom-up SCC 2009	No of countries	No of country	Time, d:hr:min	No of rank reduction	Rank	Time, d:hr:min	Total d:hr:min
	7	25	0:02:45	3 + 2 + 4	5	0:00:28	0:03:13
	8	14	0:00:56	3	6	0:00:09	0:01:05
	9	13	0:01:29	5	7	0:00:22	0:01:51
	10	6	0:01:19	5	8	0:00:37	0:01:56
	11	6	0:01:58	5	9	0:00:56	0:02:54
	12	3	0:01:50	21	9	0:05:05	0:06:55
	13	11	0:03:59	5	10	0:01:21	0:05:20
	14	26	0:12:18	8	11	0:02:49	0:15:07
	15	22	0:13:43	6	11	0:03:11	0:16:54
	16	8	0:05:26	4	11	0:02:21	0:07:47
	17	9	0:06:01	4	11	0:02:17	0:08:18
	18	10	0:06:31	8	12	0:03:58	0:10:29
	19	12	0:10:09	6	12	0:04:04	0:14:13
	20	11	0:11:20	5	13	0:03:51	0:15:11
	21	13	0:14:34	7	14	0:06:25	0:20:59
	22	15	1:01:54	6	14	0:07:39	1:09:33
	23	9	0:13:13	7	15	0:07:59	0:21:12
							7:18:57
Direct PC							
SCC 2009	23			86	15		7:00:02
SCC 2010	24			70	15		5:01:57

Table 3. Variance component estimation times of the principal component approaches for protein yield.

Bottom-up PC Protein yield 2007	No of countries	No of iterations, country addition	Time, d:hr:min	No of iterations, rank reduction	Rank	Time, d:hr:min	Total time, d:hr:min
	7	5	0.00:46	4	7	0:00:26	0:01:12
	8	9	0:01:48	4	8	0:00:41	0:02:29
	9	8	0:02:21	5	9	0:01:12	0:03:33
	10	8	0.03:24	6	10	0:02:02	0:05:26
	11	11	0:06:05	5	11	0:02:25	0:08:30
	12	14	0:10:24	6	11	0:03:49	0:14:13
	13	13	0:10:53	6	12	0:03:50	0:14:43
	14	13	0:14:09	6	13	0:05:00	0:19:09
	15	12	0:16:34	5	14	0:05:28	0:22:02
	16	77	6:03:56	8	15	0:11:04	6:15:00
	17	12	1:06:04	6	16	0:10:40	1:16:44
	18	17	2:10:31	13	16	1:03:47	3:14:18
	19	12	1:13:49	6	17	0:13:00	2:02:49
	20	21	3:14:19	12	17	1:07:15	4:21:34
	21	14	1:22:37	5	18	0:13:08	2:11:45
	22	28	5:11:05	7	19	0:22:04	6:09:09
	23	15	3:14.23	11	19	1:15:25	5:04:48
	24	15	3:17:11	6	20	1:24:00	4:17:35
	25	14	4:05:09	12	20	2:04:03	6:09:12
							46:23:11
Direct PC							
Protein yield 2007	25			24	20		5:13:27
Protein yield 2009	26			42	20		5:16:46

Table 4. Correlations between estimated breeding values from different approaches and from optimal and too low a rank under the direct PC approach in the subgroup C, i.e. bulls used only abroad. For Interbull predictions, post-processed parameters were used as starting values.

Country	Somatic cell count 2009				Protein yield 2007			
	PC15[a] BUP[b]	PC15 FA7[c]	PC15 INTB[d]	PC10 PC15	PC20 BUP	PC20 FA9	PC20 INTB	PC15 PC20
Canada	0.999	1.000	0.999	0.998	1.000	1.000	0.998	0.999
Germany	1.000	1.000	0.999	0.999	1.000	1.000	0.997	1.000
Denmark, Finland, Sweden	0.999	1.000	0.999	0.998	1.000	1.000	0.998	1.000
France	1.000	1.000	0.999	0.999	1.000	1.000	0.997	1.000
Italy	1.000	1.000	1.000	0.999	1.000	1.000	0.998	1.000
The Netherlands	1.000	1.000	0.999	0.999	1.000	1.000	0.998	1.000
USA	0.999	1.000	0.999	0.999	1.000	1.000	0.996	1.000
Switzerland	0.994	0.999	0.998	0.995	0.997	1.000	0.997	0.999
Great Britain	1.000	1.000	0.999	0.999	1.000	1.000	0.997	1.000
New-Zealand	0.999	0.999	0.993	0.997	0.997	0.999	0.984	0.995
Australia	0.999	0.999	0.998	0.997	1.000	1.000	0.995	1.000
Belgium	0.996	0.999	0.997	0.997	0.996	1.000	0.994	0.997
Ireland	0.999	0.999	0.998	0.999	1.000	0.999	0.993	0.997
Spain	0.999	1.000	0.999	0.999	1.000	1.000	0.996	1.000
Czech Republic	0.999	0.998	0.998	0.994	0.999	0.999	0.981	0.993
Slovenia	-[e]	-	-	-	0.978	0.979	0.883	0.994
Estonia	0.985	0.992	0.984	0.990	0.999	0.997	0.986	0.995
Israel	0.980	0.989	0.984	0.990	0.988	0.993	0.975	0.985
Swiss Red Holstein	0.995	0.996	0.994	0.994	0.998	0.999	0.994	0.998
French Red Holstein	0.979	0.996	0.994	0.997	0.972	0.988	0.987	0.998
Hungary	0.998	0.999	0.999	0.995	1.000	1.000	0.996	0.999
Poland	-	-	-	-	1.000	0.999	0.987	0.998
South Africa	0.997	0.999	0.996	0.995	0.997	0.998	0.956	0.992
Japan	0.998	0.997	0.992	0.998	1.000	1.000	0.997	0.999
Latvia	-	-	-	-	0.993	0.977	0.918	0.982
Danish Red Holstein	0.993	0.994	0.988	0.998	-	-	-	-

[a]Direct PC, rank 15
[b]Parameters from the bottom-up PC analysis
[c]Factor analysis, fit 7
[d]Interbull
[e]Country was not participated in the evaluation