Comparison of Different Variance Component Estimation Approaches for MACE – Direct and Bottom-up PC

A.-M. Tyrisevä[1], K. Meyer[2], F. Fikse[3], V. Ducrocq[4], J. Jakobsen[5], M.H. Lidauer[1] and 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]INRA, Station de Génétique Quantitative et Appliquée, Jouy-en-Josas, France ;
[5]Interbull Centre, Dept. Animal Breeding and Genetics, SLU, Uppsala, Sweden

1. Introduction

Multiple-trait across country evaluation (MACE) is used for international genetic evaluation of dairy bulls. MACE treats records in different countries as different traits. Thus, a sire will get a breeding value for each participating country. Whenever a country makes changes to their national evaluation model, the genetic variance-covariance (VCV) matrix needs to be re-estimated.

Estimation of the VCV matrix is a difficult task. For the Holstein production evaluation, which includes 26 traits, it is not possible to estimate the VCV matrix in a single analysis with the currently available estimation methods and the given time constraints. Hence, the complete matrix is built from analyses of sub-sets. This readily results in a non-positive definite matrix and a bending procedure (Jorjani *et al.,* 2003) needs to be applied to obtain a positive definite matrix. In addition, the VCV matrix is usually over-parameterized as genetic correlations between countries are generally high.

Mäntysaari (2004) showed that the classical MACE model can be described as a random regression (RR) MACE model. This provides the opportunity to model the complete VCV parsimoniously by fitting only the principal components (PC) with non-negligible variance, which yields a VCV matrix of reduced rank. In turn, this can reduce the dimension of the system of equations to be solved in MACE, and thus the computational effort required (Tyrisevä *et al.*, 2008).

The aim of this study is to compare two approaches available for reduced rank estimation of the VCV matrix: the direct PC method proposed by Kirkpatrick and Meyer (2004) and the bottom-up PC procedure suggested by Mäntysaari (2004).

2. Material and Methods

2.1 Random regression MACE and rank reduction

As has been shown (Mäntysaari 2004), classical MACE and random regression MACE are equivalent models. The genetic VCV matrix of sire effects, var($\mathbf{u}_i$) = **G**, can be decomposed as:

$$\mathbf{G} = \mathbf{SCS} \text{ and } \mathbf{C} = \mathbf{VDV}^{\mathrm{T}},$$

where **S** is a diagonal matrix of standard deviations, **C** is the genetic correlation matrix, **V** is a matrix of eigenfunctions and **D** the diagonal matrix of eigenvalues of **C**. Then the classical MACE model for *t* countries

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

equals the RR MACE model

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{SV}\boldsymbol{\nu}_i + \boldsymbol{\varepsilon}_i \qquad [2]$$

For both models $\mathbf{y}_i$ is a $n\times1$ vector of national de-regressed breeding values for bull *i*, **b** is a $t\times1$ vector of country effects and $\boldsymbol{\varepsilon}_i$ is a vector of residuals with var(ε_i) = **I**. $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the incidence matrices. In [2], $\boldsymbol{\nu}_i$ is a vector of *t* regression coefficients for bull *i* with $\mathrm{var}(\boldsymbol{\nu}_i) = \mathbf{D}$.

If **G** is close to singular, only *r* eigenvalues ($r < t$) can be considered, and **G** can be replaced with $\mathbf{G}_{1} = \mathbf{S}\mathbf{V}_{1}\mathbf{D}_{1}\mathbf{V}_{1}{}^{\mathrm{T}}\mathbf{S}$. $\mathbf{D}_1$ contains the *r* significant eigenvalues and $\mathbf{V}_1$ the *r* corresponding eigenfunctions. See Tyrisevä *et al.* (2008) for further details.

2.2 Bottom-up PC approach

The bottom-up PC approach adds traits (=countries) sequentially into the analysis until all traits are included (Mäntysaari 2004). Whenever a new trait has been added to the analyses, Akaike's information criterion (AIC) is used to decide whether a model with lower rank is more likely. The stages of the algorithm are as following:

1. start REML analyses with *n* countries (7 countries in this study)
2. repeatedly omit one PC and run REML until AIC starts to deteriorate
3. select the best model to go forward (matrices of eigenvalues and eigenfunctions in step 2 are saved and those for the best model are used)
4. add a new country and run REML
5. reduce by one PC and run REML (only one reduction needs to be tested after step 1 since the other fits have already been tested in the earlier steps of the algorithm)
6. compare models in steps 4 and 5 and select the model with the best AIC to go forward
7. repeat steps 4-6 until all countries have been added and the best final model has been selected
8. update weights

2.3 Direct PC approach

In direct PC approach the genetic VCV matrix is decomposed into the matrices of eigenvalues and eigenfunctions, and only the leading PCs are fitted to model the genetic VCV matrix (Kirkpatrick and Meyer 2004). Therefore, compared to the bottom-up PC approach, the direct PC approach differs in the matrix used for the eigenvalue decomposition.

2.4 Correct rank of matrix

The bottom-up approach is designed to estimate the appropriate rank of the VCV matrix. The appropriate rank can be inferred by a series of model comparisons for analysis with different ranks. The direct PC approach requires, in turn, the number of PCs to be fitted (i.e., rank) to be specified *a priori*.

Based on the results of Meyer and Kirkpatrick (2008), selecting too low a rank can lead to picking up the wrong subset of principal components in the direct PC approach. Further, removing more than 0.5% of the sum of the eigenvalues in VCV matrix affected genetic correlations as found in the current study (results not shown). Thus, it is important to find the appropriate rank when rank reduction is desired in RR MACE.

The ability of the bottom-up PC approach to find the correct rank was validated by determining the appropriate rank using methods suggested by Meyer and Kirkpatrick (2008). For this, the VCV matrix for protein yield estimated by Interbull in the routine evaluation was decomposed, and the magnitude of the eigenvalues was studied to make a reasonable guess of the appropriate rank. After this, several direct PC analyses for ranks bracketing this value were carried out, and the resulting Maximum Log Likelihood values, AIC, sum of the eigenvalues and magnitude of the leading eigenvalues were studied in order to establish the correct rank.

2.5 Test application

Data sets used for testing were August 2007 and April 2009 MACE Interbull Holstein evaluations for protein yield and SCC, respectively. In total, 25 countries/traits for protein yield and 23 countries/traits for SCC were included.

Bottom-up PC runs were performed for both protein yield and SCC. For protein yield, direct PC approach runs with ranks 15, 17, 19 and 20 were performed. For SCC, the direct PC analysis was only carried out for rank 15, the final rank of bottom-up approach.

The current algorithm for the bottom-up PC is driven by shell scripts tailored for the variance component estimation software WOMBAT (Meyer 2007b). The analyses for direct PC approach were also carried out with

WOMBAT. An average information REML algorithm was used for both direct and bottom-up PC analyses.

3. Results

3.1 Bottom-up PC approach

The estimated rank from the bottom-up PC approach for the VCV matrix for protein yield was 20. Comparison of Maximum Log Likelihood and AIC values as proposed by Meyer (2007a), as well as comparison of the leading principal components (Table 2) suggested that rank 20 was a good choice, although the differences between ranks 19 and 20 were small.

3.2 Comparison of genetic correlations

In general, genetic correlations obtained using different approaches were rather similar, especially for SCC. Selected estimates of genetic correlations for protein yield and SCC are plotted in Figures 1 and 2, respectively. Non-post-processed Interbull estimates (Interbull) were used for comparison.

On average, correlation estimates from the bottom-up approach were somewhat lower than those estimated by the direct PC approach with equal rank. However, the differences in the minimum, maximum and mean values were small between the approaches (Table 1).

The level of the genetic correlations for SCC was high throughout; the minimum value being no lower than 0.61 (an Interbull estimate). Surprisingly, lower genetic correlation estimates were obtained for protein yield. Estimates for some between-country correlations were as low as 0.10. This was in contrast to our hypothesis of high genetic correlations between the countries for protein yield. For both traits, lowest correlation estimates were from Interbull estimates.

Table 1. Minimum, maximum and mean values of estimated genetic correlations for protein 2007 and SCC 2009.

Trait and approach	Min	Max	Mean
Protein 2007			
Direct PC 20	0.08	0.94	0.69
Bottom-up PC 20	0.04	0.94	0.68
Interbull	0.02	0.94	0.70
SCC 2009			
Direct PC 15	0.73	0.97	0.89
Bottom-up PC 15	0.65	0.98	0.88
Interbull	0.61	0.98	0.89

4. Discussion

For both traits the estimated rank was smaller than the size of the VCV matrix. This reduced the number of genetic parameters to be estimated from 325 to 311 for protein yield and from 276 to 211 for SCC. Use of correct rank for protein yield also resulted in the fastest running time for direct PC. Analysis of direct PC 15 required 20 days, whereas that for direct PC 20 needed only 5.5 days.

All approaches seemed to perform well for analyses of SCC, even though the number of common bulls was low or zero in some cases (Figure 2). One has to bear in mind that there are, however, still links through the pedigree in these cases.

Compared to SCC, approaches used for protein yield behaved differently in the problems associated with the data structure. Low numbers of common bulls and some challenging countries/populations were at least to some extent associated with the lower correlation estimates. Further, they were clearly associated with the larger differences between results from the two methods. The bottom-up PC approach had a tendency to lead to the lowest estimates in these cases.

5. Conclusions

In practice, direct and bottom-up PC approaches performed equally well and enabled the use of more parsimonious models through random regression MACE. Bottom-up PC can be utilized for direct estimation of the rank of the VCV matrix to be used.

References

Jorjani, H., Klei, L. & Emanuelson, U. 2003. A Simple Method for Weighted Bending of Genetic (Co)variance Matrices. *J. Dairy Sci. 86*, 677-679.

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

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

Table 2. Comparison of Log Likelihood, AIC and eigenvalues from reduced rank variance component analysis with different rank for the protein yield.

Approach	No of para-meters	Deviation of the highest Max LogL	Deviation of the highest -½AIC	Sum of eigen-values	Eigenvalues 1	2	3	4	5	6	7	8
Direct PC15	271	-105	-68	1695.5	1325.7	78.9	69.8	43.6	36.6	30.9	22.3	19.7
Direct PC17	290	-36	-18	1695.4	1329.7	76.7	65.0	44.5	35.2	30.4	21.3	17.8
Direct PC19	305	-2	0	1694.6	1330.9	76.1	60.3	47.4	33.2	28.8	21.4	17.2
Direct PC20	311	0	-3	1694.6	1331.0	76.1	60.1	47.2	33.0	28.6	21.3	17.3

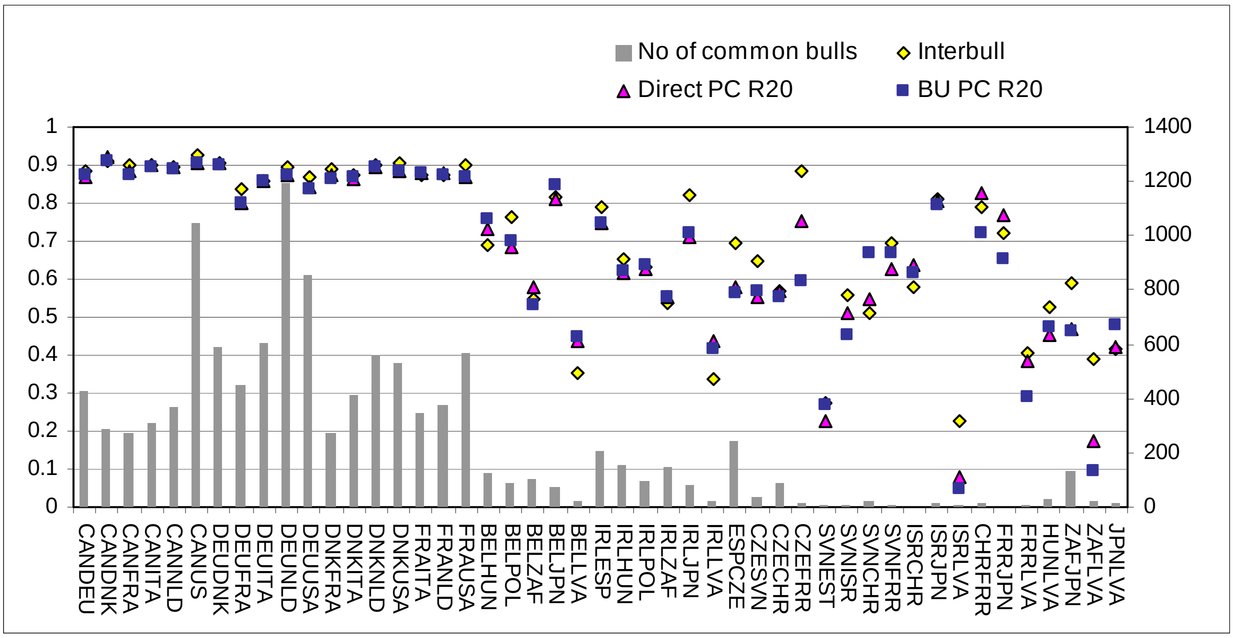

Figure 1. Examples of genetic correlations for protein 2007 using both bottom-up and direct PC approaches together with the non-post-processed Interbull correlations for comparison.

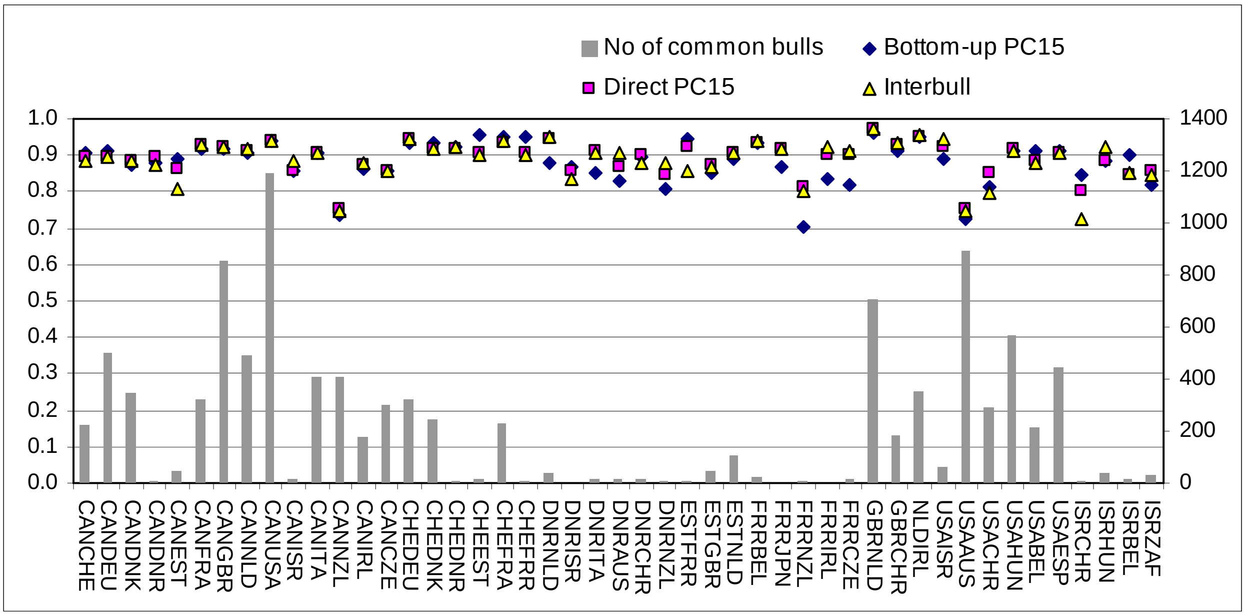

Figure 2. Examples of genetic correlations for SCC 2009 using both bottom-up and direct PC approaches together with the non-post-processed Interbull correlations for comparison.