Principal Component and Factor Analytic Models In International Sire Evaluation

A.-M. Tyrisevä[*], K. Meyer[†], W.F. Fikse[‡], V. Ducrocq[§], J. Jakobsen[¶], M.H. Lidauer[*] and E.A. Mäntysaari[*]

Introduction

Various studies have addressed the challenge of variance component estimation for multiple-trait across country evaluation (MACE) and attempted to ease the burden of the estimation process. Several of these have focused on using the decomposition of the genetic covariance matrices into the pertaining matrices of eigenvalues and -vectors, namely principal component (PC) and factor analytic (FA) approaches (e.g., Leclerc et al., 2005; Mäntysaari, 2004). For highly correlated traits, some eigenvalues have only a very small effect on the genetic variation. This is utilized by ignoring the PCs with negligible effects. For the PC approach this results in dimension reduction. The FA model also includes trait specific variances. This results in a full rank (co)variance (VCV) matrix unless some of the latter are zero. Leclerc et al. (2005) studied both PC and FA approaches for a sub-set of well-linked base countries, performing dimension reduction for this sub-set and estimating the contribution of the remaining countries to these PCs or factors. Mäntysaari (2004) introduced a bottom-up PC approach: this begins with a sub-set of countries, adding in the remaining countries sequentially. By examining in each step whether or not the new country increases the rank of the genetic VCV matrix, it only fits PCs with non-negligible eigenvalues and thus avoids over-parameterized models. Direct estimation of the important genetic principal components only has been proposed by Kirkpatrick and Meyer (2004). However, this requires the appropriate rank to be known or to be estimated. Similarly, we can estimate a VCV matrix imposing a FA structure directly. The bottom-up approach has recently been tested for variance component estimation for MACE with promising results (Tyrisevä et al., 2009). Both direct PC and FA approaches have been applied to beef cattle data sets, and have demonstrated their potential to be used for large, multi-trait data sets (e.g., Meyer, 2007a). The objectives of this study are to assess the impact of alternative parameterizations (PC and FA) for the estimation of variance components on practical predictions of breeding values with MACE.

Material and methods

Random regression MACE The MACE model for i^{th} sire can be expressed in terms of a random regression (RR) model as:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\mathbf{V}\boldsymbol{\nu}_i + \boldsymbol{\epsilon}_i, \quad (1)$$

[*]Biotechnology and Food Research, MTT Agrifood Research Finland, 31600 Jokioinen, Finland
[†]Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia
[‡]Department of Animal Breeding and Genetics, SLU, Box 7023, S-75007 Uppsala, Sweden
[§]UMR 1313 INRA, Génétique Animale et Biologie Intégrative, 78352 Jouy-en-Josas Cedex, France
[¶]Interbull Centre, Department of Animal Breeding and Genetics, SLU, Box 7023, S-75007 Uppsala, Sweden

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, $\boldsymbol{\nu}_i$ is the vector of t regression coefficients for bull i, $\boldsymbol{\epsilon}_i$ is the corresponding vector of n_i residuals, and $\mathbf{X}_i$ and $\mathbf{Z}_i$ denote the pertaining incidence matrices. Decompose $\mathbf{G}$, the $t \times t$ VCV matrix of sire effects $\mathbf{u_i}$, as $\mathbf{G} = \mathbf{VDV}^T$ with $\mathbf{D}$ and $\mathbf{V}$ the matrices of the eigen-values and -vectors, respectively, gives Var($\boldsymbol{\nu}_i$) $= \mathbf{D}$. Further, for g_{jj} the sire variance and h_j^2 the heritability in country j, Var($\boldsymbol{\epsilon}_i$) $= diag(g_{jj}\,\lambda_j/EDC_{ij})$, with $\lambda_j = (4 - h_j^2)/h_j^2$ and EDC_{ij} the effective daughter contribution for bull i in country j.

PC approach This representation facilitates parameter reduction when $\mathbf{G}$ is positive semi-definite. If the PCs with the smallest eigenvalues have no influence, they can be ignored without impairing the accuracy of the estimation. $\mathbf{G}$ can then be replaced by $\mathbf{G}_1 = \mathbf{V}_1\mathbf{D}_1\mathbf{V}_1^T$, where $\mathbf{D}_1$ contains the r largest eigenvalues and $\mathbf{V}_1$ the r corresponding eigenvectors, with $r < t$.

FA approach For the FA approach, we divide $\mathbf{u}_i$ into vectors of common factors, $\boldsymbol{\delta}_i$, with Var($\boldsymbol{\delta}_i$) $= \mathbf{W} = \mathbf{I}$, and country specific effects, $\boldsymbol{\tau_i}$, with Var($\boldsymbol{\tau}_i$) $= \mathbf{F} = diag\{\sigma^2_{\tau\,ij}\}$, i.e. $\mathbf{u}_i = \mathbf{L}\boldsymbol{\delta}_i + \boldsymbol{\tau}_i$, where $\mathbf{L}$ denotes the matrix of factor loadings. This gives

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b} + \mathbf{Z}_i\Big(\mathbf{L}\boldsymbol{\delta}_i + \boldsymbol{\tau}_i\Big) + \boldsymbol{\epsilon}_i \tag{2}$$

The FA approach models $\mathbf{G}$ as the sum of two terms: the common (co)variances and the trait-specific variances, i.e. $\mathbf{G} = \mathbf{LWL}^T + \mathbf{F}$. As for the PC approach, the rank of the matrix of loadings is reduced, i.e. $r < t$ of the PCs explain the common covariances. The resulting model, however, will not be of reduced rank due to the country specific variances. The advantage of the FA model is that $\mathbf{G}$ is decribed very parsimoniously.

Data Data consisted of protein yield deregressed breeding values from the August 2007 Interbull MACE Holstein evaluation. Data comprised 100 551 bulls in 25 countries. The majority of the bulls were used in one country only with the number of bulls per country ranging from 145 (French Red Holstein) to 23 380 (USA), 4 678 on average. The number of common bulls varied from zero to 1 194 with a mean of 178. Here, common bulls were bulls with daughters in both countries, without restrictions on the country of origin. Analyses fitted a sire model, with information on sire and maternal grand-sire pedigrees increasing the number of sires to 106 003.

Models included in the comparison Estimates of $\mathbf{G}$ were determined in a previous study (Tyrisevä et al., 2010; in preparation), where the appropriate fit were chosen based on Akaike's information criterion (AIC) and the comparison of estimates from analyses using successive number of PCs (see Meyer and Kirkpatrick, 2008). For the direct PC approach, rank 19 (PC19) was selected as best (Tyrisevä et al., 2009). Results using too low a rank (PC15) and full rank (PC25) are presented for comparison. Similarly, for the FA approach, a model fitting 9 factors (FA9) was chosen. The number of parameters was 271 for PC15, 305 for PC19, 326 for PC25 and 215 for FA9. Variance components were estimated by restricted maximum likelihood, using an average information algorithm as implemented in WOMBAT (Meyer, 2007b).

Analysis of estimated breeding values (EBVs) The prediction of breeding values in (1) and (2) followed (Tyrisevä et al., 2008). Correlations between EBVs from PC and FA approaches under the optimal fits and correlations between EBVs from PC 15, 19 and 25 were studied. Further, correlations between EBVs from PC19 and FA9 and from PC15 and PC19 for each

Table 1: Quantiles, minima, maxima and means of genetic correlations for protein yield.

Approach	Min	1st Quant.	Median	Mean	3rd Quant.	Max
Direct PC, rank 15	-0.05	0.56	0.71	0.68	0.81	0.95
Direct PC, rank 19	0.09	0.56	0.71	0.69	0.82	0.94
Direct PC, rank 25	0.08	0.56	0.71	0.69	0.82	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

country were considered for four subgroups: A) bulls used only in their own country, B) bulls used in their own country and abroad, C) bulls not used in the country of EBV estimation, 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

Variances Estimates of genetic variances from PC19, PC25 and FA9 were almost identical, except for some differences between approaches for French Red Holstein (FFR) (PC19: 80.4 ± 9.06, PC25: 80.6 ± 9.16, FA9: 76.9 ± 8.60). The differences in estimates and their high standard errors can be attributed to the low number of the bulls (145) in this population. For FA9, there was substantial variation in the amount of the country specific variance. On average, the proportion of the total genetic variance attributed to country specific effects was 5%, with proportions highest for Australia (19%) and Latvia (31%). In 9 of the 25 countries (Switzerland, Great Britain, New Zealand, Czech Republic, Slovenia, Israel, French Red Holstein, South Africa and Japan) the genetic variance was totally explained by the common variance. Computing time were shortest for the optimal fit (e.g., FA7: 14.5 days, FA9: 3.5 days, FA11: 31.5 days, and PC15: 21.5 days, PC19: 9 days, PC25: 16.5 days).

Genetic correlations Except for minimum values, there were hardly any differences in estimates of genetic correlations from PC15, PC19, PC25 and FA9 (Table 1). Furthermore, the non-post-processed Interbull estimates, included in the Table 1 for comparison, were almost identical with the estimates from the other approaches.

EBVs Using the PC19 model reduced the number of equations in the mixed model by 24% compared to PC25. No reduction in the number of equations was gained from the FA9 model. Times required for solution ranged from 5min (PC19) to 7min (FA9). The EBVs from PC19 and PC25 were identical. This was expected, given the almost identical genetic correlations between the approaches. EBV correlations between PC15 and PC19 and between PC15 and PC25 were lower than those between PC19 and PC25, demonstrating that the use of too low a rank affected the estimation. Further, few differences between EBVs from FA9 and PC19 were noticeable. Exceptions for which EBV correlations were less than 0.99 were Slovenia, FFR and Latvia. These were the countries with the lowest number of records and weak ties with the other countries. The mean number of common bulls between FFR and the other countries was as low as 9, and those for Latvia and Slovenia were 29 and 32, respectively. In all studied subgroups, EBVs from FA9 and PC19 were unity or close to unity, except in the subgroup C for Slovenia, FRR and Latvia ($<$

0.99). Correlations between EBVs from PC15 and PC19 tended to be lower than those between FA9 and PC19, but they were still very high (> 0.99), except in subgroup C for Israel and Latvia. Representative examples of the above correlations are presented in Table 2. Results agreed well with a previous study, in which the input parameters used for the prediction of EBVs were provided by Interbull (Tyrisevä et al., 2008).

Table 2: Correlations between EBVs: PC analyses with optimal and too low a rank, and for PC and FA approaches for optimal fit.

	Complete data		Subgroup C	
Country	PC15 PC19	FA9 PC19	PC15 PC19	FA9 PC19
Canada	0.999	1.000	0.999	1.000
Italy	1.000	1.000	1.000	1.000
Netherlands	1.000	1.000	1.000	1.000
USA	1.000	1.000	1.000	1.000
Czech Republic	0.999	1.000	0.999	1.000
Australia	1.000	1.000	1.000	1.000
Belgium	0.997	1.000	0.997	1.000
Ireland	0.997	0.999	0.997	0.999
Slovenia	0.994	0.979	0.994	0.979
Israel	0.985	0.993	0.985	0.993
French Red Hol.	0.999	0.988	0.999	0.988
Latvia	0.982	0.977	0.982	0.977

Conclusion

The RR representation of MACE facilitates exploitation of PC or FA approaches for variance component estimation and prediction of breeding values for international sire evaluation. Both PC and FA allow a reduction of the number of parameters to be estimated, and both methods benefit from the more parsimonious variance structure. Genetic parameters from different approaches were very similar, when the optimal number of PCs/factors was fitted. Overfitting did not affect the estimates of genetic correlations and breeding values, but increased the estimation time, whereas fitting too low number of parameters affected bull rankings in different countries.

References

Leclerc, H., Fikse, W. F., and Ducrocq, V. (2005). *J. Dairy Sci.*, 88:3306–3315.

Mäntysaari, E. A. (2004). *Interbull Bull.*, 32:70–74.

Meyer, K. (2007a). *J. Anim. Breed. Genet.*, 124:50–64.

Meyer, K. (2007b). *J. Zheijang Univ. Sci. B*, 8:815–821.

Meyer, K. and Kirkpatrick, M. (2008). *Genetics*, 108:1153–1166.

Tyrisevä, A.-M., Lidauer, M. H., Ducrocq, V., Back, P., Fikse, W. F., and Mäntysaari, E. A. (2008). *Interbull Bull.*, 38:142–145.

Tyrisevä, A.-M., Meyer, K., Fikse, W. F., Ducrocq, V., Jakobsen, J., Lidauer, M. H., and Mäntysaari, E. A. (2009). *Interbull Bull.*, 40:72–76.

Vuori, K., Strandén, I., and Lidauer, M. H. (2006). In *Proc 8th WCGALP*, CD-ROM, 27:33.