FACTOR-ANALYTIC MODELS TO REDUCE COMPUTATIONAL REQUIREMENTS IN INTERNATIONAL GENETIC EVALUATION OF BEEF CATTLE

Karin Meyer

Animal Genetics and Breeding Unit*, University of New England, Armidale, NSW 2351

SUMMARY

There has been considerable recent interest in factor-analytic models for the analysis of genotype by environment type problems. It is shown that, compared to 'standard' multivariate analyses, such models can substantially reduce computational requirements for international genetic evaluation of beef cattle, fitting an animal model and treating performance in different countries as separate traits.

INTRODUCTION

Factor-analytic (FA) structures provide a powerful means to model multivariate covariance matrices for 'similar' traits parsimoniously, and are readily implemented in our standard mixed model framework for estimation and prediction. FA models are used increasingly, in particular in the analysis of data from plant breeding trials subject to genotype by environment (G×E) interactions; see Meyer (2009) for a recent review. This paper demonstrates that FA models can provide an attractive alternative to standard, multi-trait models for international genetic evaluation of beef cattle.

MATERIAL AND METHODS

Alternative models. Consider an animal model analysis of q traits where each individual has a record for a single trait ('country' or 'location') only. Assume covariance matrices for genetic effects have a factor-analytic structure, i.e. can be written as $\mathbf{\Sigma} = \mathbf{\Gamma\Gamma}' + \mathbf{\Psi}$, with $\mathbf{\Psi}$ diagonal. We then have 3 equivalent models.

i) Multivariate (MV): In a multi-trait analysis we estimate genetic effects ($\mathbf{u}$) in each country directly invoking the standard, linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$, with $\mathrm{Var}(\mathbf{u}) = \mathbf{\Sigma} \otimes \mathbf{A}$ and $\mathbf{A}$ denoting the numerator relationship matrix (NRM). This requires $\mathbf{\Sigma}$ to have full rank.

ii) 'Extended' factor analytic (XFA): If $\mathbf{\Sigma} = \mathbf{\Gamma\Gamma}' + \mathbf{\Psi}$, we can partition genetic effects into m common ($\mathbf{c}$) and q specific ($\mathbf{s}$) factors, $\mathbf{u} = (\mathbf{\Gamma} \otimes \mathbf{I})\mathbf{c} + \mathbf{s}$, with $\mathrm{Var}(\mathbf{c}) = \mathbf{I} \otimes \mathbf{A}$ and $\mathrm{Var}(\mathbf{s}) = \mathbf{\Psi} \otimes \mathbf{A}$. Fitting these separately gives XFA model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^\star\mathbf{c} + \mathbf{Zs} + \mathbf{e}$, with $\mathbf{Z}^\star = \mathbf{Z}(\mathbf{\Gamma} \otimes \mathbf{I})$ (Thompson et al. 2003).

iii) Principal components (PC): If specific effects are assumed absent, i.e. $\mathbf{\Psi} = \mathbf{0}$, the XFA model reduces to a PC model. An equivalent model to MV and XFA is obtained if we decompose $\mathbf{\Gamma\Gamma}' + \mathbf{\Psi} = \mathbf{\Gamma}^\star\mathbf{\Gamma}^{\star\prime}$ (to form $\mathbf{Z}^\star = \mathbf{Z}(\mathbf{\Gamma}^\star \otimes \mathbf{I})$) and consider all factors, i.e $m = q$. For $m < q$, we obtain a reduced rank model, i.e., the PC model can accommodate $\mathbf{\Sigma}$ which does not have full rank.

Transformations. Solutions for genetic effects from one model are readily transformed to those from another. Fitting an XFA or full rank PC model, $\hat{\mathbf{u}} = (\mathbf{\Gamma} \otimes \mathbf{I})\hat{\mathbf{c}} + \hat{\mathbf{s}}$ or $\hat{\mathbf{u}} = (\mathbf{\Gamma}^\star \otimes \mathbf{I})\hat{\mathbf{c}}$. Conversely, fitting a MV model, $\hat{\mathbf{c}} = (\mathbf{\Gamma}'\mathbf{\Sigma}^{-1} \otimes \mathbf{I})\hat{\mathbf{u}}$ and $\hat{\mathbf{s}} = (\mathbf{\Psi\Sigma}^{-1} \otimes \mathbf{I})\hat{\mathbf{u}}$ (Smith et al. 2001).

Factor rotation. Fitting m common factors, $\mathbf{\Gamma}$ of size $q \times m$ has $m(m-1)/2$ elements given by orthogonality constraints. $\mathbf{\Gamma}$ is not unique and can be rotated, i.e. replaced by $\mathbf{\Gamma T}$ with $\mathbf{T}$ an orthogonal matrix. A particular rotation yields $\mathbf{\Gamma}^+$ with all above diagonal elements equal to zero, which can be interpreted as the matrix derived from the first m columns of the Cholesky factor of $\mathbf{\Sigma}$. Such 'trian-

*AGBU is a joint venture of NSW Department of Primary Industries and University of New England

gular' $\mathbf{\Gamma}^+$ yields a less dense $\mathbf{Z}^\star$. Let $\boldsymbol{\gamma}'_j$ denote the $j-$th row of $\mathbf{\Gamma}^+$. For an individual with a record in location j, $\boldsymbol{\gamma}'_j$ represents the respective part of the design matrix, contribution to the coefficient matrix ($\mathbf{C}$) in the mixed model equations (MME) is proportional to $\boldsymbol{\gamma}_j\boldsymbol{\gamma}'_j$. As elements $j+1$ to m of $\boldsymbol{\gamma}_j$ are zero, for a single record per individual, the corresponding $m\times m$ diagonal block of $\mathbf{C}$ then consists of a dense sub-block for rows and columns 1 to j, while the remaining $m-j$ rows and columns are diagonal (with non-zero diagonal elements due to the NRM). For instance, for $j=2$ only row and column 1 and 2 are linked by a non-zero element, and only for $j=q$ are all m^2 elements in the diagonal block non-zero.

Data. Case I comprised simulated records for a half-sib structure, considering $q = 8, 12$ or 16 countries, with 100 'global' sires used in all countries, 900 'local' sires used in a particular country only, and 50 progeny per sire and country. Both sires and progeny were assumed to have a single record, with Global sires belonging to country 1. The model of analysis was a simple animal model, fitting country means as the only fixed effects. Data were simulated assuming heterogeneous variances, moderate heritabilities and high genetic correlations between countries, using population values of covariance matrices in solving the MME.

Case II considered 865 129 weaning weight records for Australian and New Zealand Angus calves in 1 336 herds, pre-corrected for the effects of age at weaning, birth type and dam age, and $N = 998\,479$ animals in the pedigree. Records in individual herds were assigned to 10 different 'countries', considering herds as they occurred in the data, resulting in 69 875 (country 7) to 104 903 (country 2) records per trait. The model of analysis fitted direct and maternal additive genetic effects and maternal permanent environmental effects (324 613 dams) as random effects, and contemporary groups (between 9 243 and 12 415 per 'country') as fixed effects. Covariance matrices used for the mixed model analyses were constructed from standard variance components for this trait and breed, with genetic correlations assumed to be about 0.8 (0.78 to 0.82).

Case III involved weaning weight records for Hereford calves in 9 countries, as collated for a global international evaluation feasibility study (Graser 2008). There were 4 281 659 records, 2 678 762 dams with progeny in the data, and 6 648 388 animals in the pedigree. For efficiency, countries were renumbered in descending order of the number of records (c.f. Meyer 2009), with 2.67, 0.55, 0.52, 0.22 and 0.16 million for countries 1 to 5, and less than 50 000 for the remainder. The model of analysis was as for case II. Covariance components utilised were similar to those in the Hereford pilot study, but ignored three small within country direct-maternal genetic covariances.

Analyses. Estimates of fixed effects and predictions for random effects were obtained by solving the corresponding MME iteratively, using a pre-conditioned conjugate gradient (PCG) algorithm, as implemented in the iterative solutions module of WOMBAT (Meyer 2007) which holds the complete MME in core, using sparse matrix storage techniques. All calculations were carried out using double precision (8 Byte) floating point variables. XFA and PC models were parameterised using the rotation of factors to 'triangular' $\mathbf{\Gamma}^+$. Strategy A used a simple, diagonal pre-conditioning matrix, obtained as the reciprocals of the diagonal elements of $\mathbf{C}$, and stored in core. Strategy B employed a block pre-conditioner for random effects. For MV models this involved the inverse of the $q\times q$ diagonal block for each level, treated as dense. For XFA and PC models blocks were of size $m+q$ and m, respectively, i.e. jointly considered common and specific factors (if fitted) for an individual. Inversion of these blocks exploited their known sparsity structure. Dense blocks of all inverses were written to disk and re-read for each PCG iterate. The algorithm was assumed to have converged when the ratio of the sum of squared deviations in solutions between iterates dived by the sum solutions squared was less than 10^{-14}. Computations for analyses requiring less than 4 Gb of memory (RAM) were carried out on a single user, 64-bit machine with a dual-core processor rated at a speed

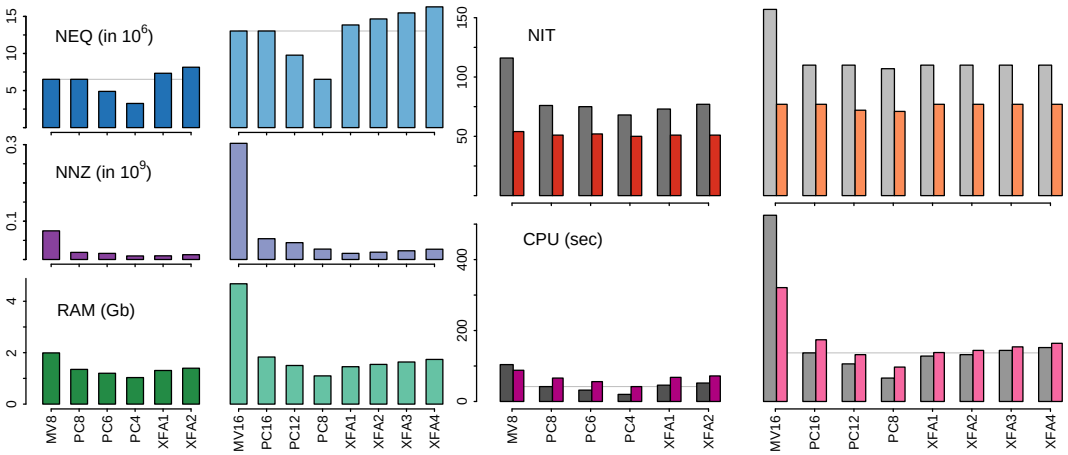

Figure 2. Results for simulated data for 8 (dark) and 16 (light) traits (see text for definitions)

of 2.2 GHz. The remaining analyses were performed on a multi-user, 64-bit machine with dual-core processors rated at 2.6 Ghz.

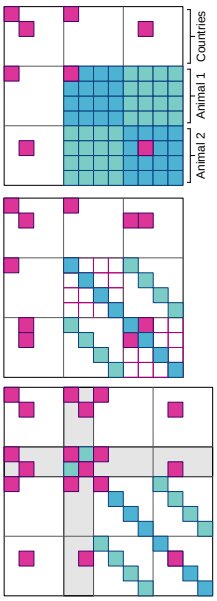

Figure 1. Toy case

RESULTS

Figure 1 shows the sparsity pattern of $\mathbf{C}$ for the 3 alternative models for a toy example comprising $q = 4$, means for each trait (rows/columns 1 to 4) and effects for 2 animals, with records in country 1 and 2, respectively, and animal 1 the sire of animal 2. For MV (top), each element of the NRM inverse contributes q^2 coefficients, resulting in dense animal $\times$ animal blocks. In contrast, for PC (middle, $m = q$) or XFA (bottom, $m = 1$), each such element contributes only m or $m + q$ coefficients. This is off-set, in part at least, by a denser design matrix $\mathbf{Z}^\star$, i.e. more contributions from the 'data part'. However, for 'triangular' $\mathbf{\Gamma}^+$, contributions are limited. For XFA, there are extra equations (gray background) and additional 'data' terms linking equations for $\mathbf{c}$ and $\mathbf{s}$, but diagonal blocks of $\mathbf{C}$ are sparser still.

Case I. Characteristics of the MME and computational requirements to solve them for the simulated data are summarised in Figure 2. To alleviate scale problems, values for $q = 8$ (except NIT) are plotted at twice their actual value. While MVq and PCq involved the same number of equations (NEQ), the number of non-zero off-diagonal elements in one triangle of $\mathbf{C}$ (NNZ) and, consequently, the memory required (RAM) differed substantially. Using a simple, diagonal pre-conditioning matrix (left bars) required marked more PCG iterates (NIT) than a block-diagonal matrix (right bars), but, except for MV, required less time in total (CPU). Increasing the number of common factors m in XFAm analyses augmented NEQ linearly, but appeared to have relatively little impact on computational requirements.

Case II. Corresponding results for the Angus data are shown in Figure 3. A number of combinations PCm_1/m_2 or XFAm_1/m_2 are examined, with m_1 and m_2 the number of common factors fitted for direct and maternal genetic effects. For NIT and CPU, values pertain to strategy B for MV10 and strategy A otherwise. With multiple random effects, differences between MV and either of the FA

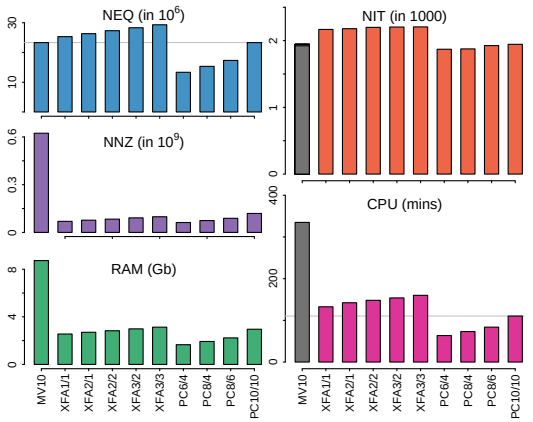

Figure 3. Results for Angus analyses (see text)

models were even more pronounced than for case I. Again, the number of common factors had little effect on NIT, but increased time per iterate and thus CPU proportionally. CPU time required for full rank PC was less than for any of the XFA analyses and only a third of that for the MV model. Resources required for reduced rank analyses are even less. E.g., omitting 2 direct and 4 maternal genetic factors reduced CPU to 75% of that for a full rank model.

Case III. Results for the large scale analyses for Herefords are given in Table 1. 'Best' values for NNZ (in million) pertain to countries numbered as described above, 'worst' used the reverse order. Using $\mathbf{\Gamma}$ with *mq* non-zero elements rather than $\mathbf{\Gamma}^+$, NNZ for PC9/9 would have been as large as $1\,005.6 \times 10^6$. A MV analysis holding MME in core was not feasible for this case. A corresponding analysis using an 'iteration on data' type strategy required more than 60 hours CPU time (Tier 2008; *pers.comm.*). Again PC analyses appeared advantageous over XFA implementations, and reducing the rank of fit even slightly had a dramatic impact on computational requirements.

DISCUSSION

Phocas et al. (2005) showed that animal model analyses would be preferable for international genetic evaluation of beef cattle. Computational requirements for such analyses can be large. However, as results clearly demonstrate, judicious choice of an equivalent model can greatly aid with this task. Parameterisation to a FA model, combined with a rotation of the factor matrix to triangular form, is especially beneficial for G×E scenarios where individuals have records in a single location only. Moreover, separation of genetic effects into common and specific factors is appealing, as estimates have a direct interpretation as global breeding values and local deviations. Results suggest that PC models are most advantageous computationally, even if little or no rank reduction is feasible. As outlined, there is a direct relationship between solutions from the different alternative models, i.e. estimates of 'global' breeding values are readily determined from those from a PC analysis.

Table 1. Global Hereford evaluation study

Model	NEQ	NNZ		RAM	NIT	CPU
		worst	best	(Gb)		(h)
XFA1/1	157.7	407.3	407.3	14.7	3051	29.8
XFA2/2	171.0	489.4	460.3	16.2	3110	32.0
PC9/9	144.4	899.8	386.7	13.9	3002	27.2
PC8/7	124.4	725.8	354.9	12.0	2690	20.6
PC7/6	111.1	599.9	294.1	10.5	2598	17.6

REFERENCES

Graser, H.-U. (2008) World Hereford Conf., Copenhagen, Denmark, June, 30 - July, 1.
Meyer, K. (2007) *J. Zhejiang Uni. SCIENCE B* **8**:815.
Meyer, K. (2009) *Genet. Select. Evol.* **41**:21.
Phocas, F., Donoghue, K. and Graser, H.-U. (2005) *Genet. Select. Evol.* **37**:361.
Smith, A. B., Cullis, B. R. and Thompson, R. (2001) *Biometrics* **57**:1138.
Thompson, R., Cullis, B. R., Smith, A. B. and Gilmour, A. R. (2003) *Austr. New Zeal. J. Stat.* **45**:445.