COMPUTING FOR MULTI-TRAIT SINGLE-STEP GENOMIC EVALUATION OF AUSTRALIAN SHEEP

Karin Meyer, Andrew Swan and Bruce Tier

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

SUMMARY

The impact of parameterising to genetic principal components and dimension reduction on computational requirements is examined for a subset of traits considered in single step evaluation of sheep in Australia. Together with judicious treatment of dense blocks due to genomic relationships in the mixed model equations, such models can reduce computational requirements many-fold.

INTRODUCTION

Genetic evaluation utilizing genomic information is in the process of being adopted in many livestock improvement schemes. In particular, the so-called 'single-step' procedure allows for joint evaluation of all animals – genotyped or not – utilising all pedigree information available at the same time (Misztal *et al.* 2009). It can be thought of as an extension of previous, best linear unbiased prediction schemes, replacing the pedigree based numerator relationship matrix between animals, $\mathbf{A}$, by it's counterpart, $\mathbf{H}$, which combines the genomic relationship matrix among genotyped animals, $\mathbf{G}$, with relationships derived from the pedigree. Computing the inverse of $\mathbf{H}$ requires large matrix products and direct inversion of $\mathbf{G}$ and the corresponding submatrix of $\mathbf{A}$, and challenges thus posed have attracted considerable attention (e.g. Aguilar *et al.* 2011).

Computational requirements to estimate breeding values are heavily dependent on the number non-zero (NNZ) elements in the coefficient matrix, $\mathbf{C}$, of the mixed model equations (MME) to be solved. In a multivariate analysis comprising q traits, each non-zero element of the inverse of the relationship matrix can contribute up to q^2 elements to this matrix. Equivalent and reduced rank models have been proposed which can reduce this number (Meyer and Kirkpatrick 2005; Meyer 2009), but have seen little practical use. Let animals be grouped according to their genomic information status, with $\mathbf{H}_{22}$ the submatrix of $\mathbf{H}$ for genotyped individuals. Typically, $\mathbf{H}_{22}$ and the corresponding block of $\mathbf{H}^{-1}$ are dense, i.e. contain few zero elements. Hence, the NNZ elements in $\mathbf{C}$ arising from elements of $\mathbf{H}^{-1}$ becomes more important than previously, where the inverse relationship matrix $\mathbf{A}^{-1}$ was sparse throughout. Furthermore, existence of dense blocks in the MME together with substantial amounts of random access memory (RAM) available in modern hardware readily allow matrix manipulation routines from highly optimized software libraries to be exploited. We examine the utility of equivalent or reduced rank models together with the use of multi-threaded library routines for dense matrix calculations for an application of single-step genetic evaluation to Australian sheep data.

EQUIVALENT MODELS AND BEYOND

Consider a linear mixed model for q traits, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ with $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$ and $\mathbf{e}$ the vectors of observations, fixed and random effects, and residuals, and $\mathbf{X}$ and $\mathbf{Z}$ the pertaining incidence matrices. Let $\mathbf{u}$ represent animals' additive genetic effects, ordered by animals within traits so that $\mathrm{Var}(\mathbf{u}) = \boldsymbol{\Sigma} \otimes \mathbf{H}$, with $\boldsymbol{\Sigma}$ the genetic covariance matrix among traits. For $\mathrm{Var}(\mathbf{e}) = \mathbf{R}$, the diagonal block in $\mathbf{C}$ for $\mathbf{u}$ is then $\mathbf{C}_{uu} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Sigma}^{-1} \otimes \mathbf{H}^{-1}$. The first part, $\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}$, is block-diagonal for animals, with blocks of size $q \times q$. If $\boldsymbol{\Sigma}^{-1}$ has no zero elements, $\boldsymbol{\Sigma}^{-1} \otimes \mathbf{H}^{-1}$, however, contributes q^2 non-zero elements to $\mathbf{C}_{uu}$ for each non-zero off-diagonal element of $\mathbf{H}^{-1}$.

An equivalent model is obtained by expanding $\mathbf{Z}\mathbf{u}$ to $\mathbf{Z}(\mathbf{Q} \otimes \mathbf{I})(\mathbf{Q}^{-1} \otimes \mathbf{I})\mathbf{u} = \mathbf{Z}^{\star}\mathbf{u}^{\star}$, with $\mathbf{I}$ an

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

identity matrix. This gives $\text{Var}(\mathbf{u}^\star) = \mathbf{Q}^{-1}\boldsymbol{\Sigma}\mathbf{Q}^{-T} \otimes \mathbf{H} = \boldsymbol{\Sigma}^\star \otimes \mathbf{H}$ and $\mathbf{C}^*_{uu} = \mathbf{Z}^{\star\prime}\mathbf{R}^{-1}\mathbf{Z}^\star + (\boldsymbol{\Sigma}^\star)^{-1} \otimes \mathbf{H}^{-1}$. Choosing $\mathbf{Q}$ so that $\boldsymbol{\Sigma}^\star$ is diagonal reduces the NNZ elements contributed by each non-zero element of $\mathbf{H}^{-1}$ to q. The trade-off for this is that $\mathbf{Z}^\star$ has up to q non-zero elements per observation compared to, typically, a single element of unity in $\mathbf{Z}$. This gives rise to some extra non-zero elements in other parts of $\mathbf{C}^\star$, especially in the off-diagonal block for fixed × random effects, $\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}^\star$. Suitable matrices $\mathbf{Q}$ can be obtained from the eigen-decomposition $\boldsymbol{\Sigma} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'$, either the matrix of eigenvectors, $\mathbf{Q} = \mathbf{E}$, or the matrix of 'factor loadings', $\mathbf{Q} = \mathbf{E}\boldsymbol{\Lambda}^{-1/2}$. The latter can be rotated to lower triangular form, $\mathbf{Q} = \mathbf{E}\boldsymbol{\Lambda}^{-1/2}\mathbf{T}$ (with $\mathbf{T}\mathbf{T}' = \mathbf{I}$) which reduces the NNZ elements in $\mathbf{Q}$ to $q(q+1)/2$ and thus the number of multiplications to set up the MME and the NNZ in $\mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}^\star$.

Furthermore, this parametrization can directly yield substantial, additional computational savings by invoking a 'reduced rank' model, if $\boldsymbol{\Sigma}$ has $q - r$ negligible eigenvalues, which generally holds for larger values of q. This involves estimating only the first r principal components (i.e. elements of $\mathbf{u}^\star$) for each animal which, at convergence, are combined to give the q corresponding elements of $\mathbf{u}$. This is achieved by simply considering only the first r columns of $\mathbf{Q}$, which reduces the number of equations in the model as well as the NNZ elements.

MATERIAL AND METHODS

Data consisted of 5.24 million records for 5 traits recorded on 1.77 million animals in the LAMBPLAN terminal sire breeds evaluation (Brown *et al.* 2007), representing the most commonly recorded traits in these breeds, namely birth, weaning and post-weaning weights, and post-weaning eye muscle and fat depth. Including parents without records there were 1,995,755 animals of which 10,698 (N) where genotyped for 48,599 single nucleotide polymorphisms. To build $\mathbf{H}^{-1}$, genomic relationships were computed following Yang *et al.* (2010). This yielded 63,793,942 NNZ elements in $\mathbf{H}^{-1}$ (halfstored), compared to 6,584,393 elements in the corresponding pedigree based matrix $\mathbf{A}^{-1}$.

As in the routine LAMBPLAN evaluation, records were pre-corrected for the effects of birth-rearing type, age at measurement and age of dam, and body weight as a covariate for eye muscle and fat depth. The model of analysis then comprised contemporary groups as fixed effects, animals' additive genetic effects, dams' permanent environmental effects for the body weights (653,067 levels), and genetic groups (93 levels) as random effects. The latter were fitted 'explicitly' – assigning proportions of membership for each animal – as augmenting the pedigree by phantom parents in single-step applications can be problematic (Misztal *et al.* 2013).

Analyses fitted standard multivariate (MV) and the principal components (PC) models described above. Dense diagonal blocks in $\mathbf{C}$ (or $\mathbf{C}^\star$) for genotyped animals were stored in two-dimensional arrays, a single matrix of size $qN \times qN$ for MV and r blocks of size $N \times N$ for PC model analyses. Similarly, if fitted, genetic groups were held in a single dense block. No distinction between MV and PC was made for this effect, as the transformation yielded sufficient additional coefficients between levels for different traits from the data part, $\mathbf{Z}^{\star\prime}\mathbf{R}^{-1}\mathbf{Z}^\star$, for the corresponding off-diagonal blocks to be almost dense. The remaining non-zero coefficients in the coefficient matrix were held in compressed sparse row format. A preconditioned conjugate gradient (PCG) algorithm (e.g. Tsuruta *et al.* 2001) with partial Cholesky decomposition preconditioner was used to solve the MME. Cholesky factors and solutions for the dense blocks were obtained using LAPACK routines `DPOTRF` and `DPOTRS` (Anderson *et al.* 1999), respectively. The product of the coefficient matrix and a vector required in each PCG iterate was formed using routines `DSYMV` from the BLAS library (Blackford *et al.* 2002) and the Intel sparse matrix equivalent, `MKL_DCSRSYMV`.

Computations were carried under Linux on a machine with 256GB of RAM and 16 Intel Xeon CPU E5-2630 cores, rated at 2.4Ghz with a cache size of 20MB. BLAS and LAPACK routines used were loaded from the Intel Math Kernel Library (MKL), version 11.1.

Table 1. Computing requirements for equivalent models for 5 traits

			Without genetic groups				With genetic groups			
			Pedigree		Genomic		Pedigree		Genomic	
			MV	PC	MV	PC	MV	PC	MV	PC
No. of equations			12,182,223				12,182,688			
NNZ[a]	Sparse	after data	50.7	66.2	50.7	66.2	223.8	316.0	223.8	316.0
		after random	179.2	89.1	178.9	89.1	352.3	338.9	352.0	338.9
	Dense	genotyped	–	–	1430.6	286.1	–	–	1430.6	286.1
Total			191.4	101.3	1621.6	387.4	364.6	351.2	1794.8	637.2
Memory (GB)			4.3	3.3	25.6	7.8	7.8	7.7	28.8	11.6
No. of PCG iterates			684	693	682	690	1357	1387	1339	1389
Time[b]	single		22.1	19.1	90.5	28.9	44.8	46.4	165.0	64.4
	multi		20.5	20.8	65.3	31.7	42.2	42.9	122.6	61.1

[a]No. of non-zero elements in coefficient matrix (in million) [b]in minutes, for single- and multi-threaded MKL routines

RESULTS

Computational requirements for analyses fitting equivalent models are summarized in Table 1, comparing models with and without the use of genomic information. Values given for NNZ elements pertain to one triangle of the symmetric coefficient matrix. As expected, there were marked differences in the NNZ elements between MV and PC models, with more elements arising from the 'data part' but substantially less non-zero elements due to covariances between random effects for the PC models, especially for single-step analyses. Fitting genetic groups increased the NNZ elements substantially and almost doubled the number of PCG iterates required. PC models proved highly advantageous, with overall computing times reduced 2- to 3-fold when genomic relationships were considered. While CPU time summed over threads when using multi-threaded MKL routines (not shown) seemed to indicate pronounced parallel processing, differences in elapsed time to single-thread runs were surprisingly small, suggesting 'processor spin' rather than actual simultaneous execution.

Corresponding results for a 10-trait scenario, obtained by doubling the data, for single-step models with genetic groups are given in Table 2. Considering more traits amplified differences between models and improved multi-thread performance, especially for the Cholesky decomposition of the diagonal block(s) for genotyped animals in the preconditioning step. Reducing the number of principal components fitted decreased the number of equations in the model and NNZ elements in the coefficient matrix. Results clearly illustrate the increasing advantage of PC over MV models with the number of traits and number of negligible eigenvalues in the genetic covariance matrix among traits.

DISCUSSION

We have described a simple reparameterisation of the standard multivariate mixed model – estimating genetic effects for principal components rather than the traits of interest – and illustrated its potential to reduce computational requirements, especially when parts of the inverse of the relationship matrix are dense. In addition, this parameterisation directly lends itself to dimension reduction by eliminating the principal components which explain virtually no genetic variation, which becomes increasingly important with the number of traits considered. Even a relatively small reduction, say 10 to 20% can have a huge impact, typically without affecting the accuracy of genetic evaluation notably.

Calculations shown for the small subset of traits in LAMBPLAN considered here held the MME in core. In practice, this is unlikely to be feasible and an 'iteration on data' type strategy needs to be employed instead (Tier and Graser 1991). However, the NNZ in the coefficient matrix is likely

Table 2. Computing requirements for full and reduced rank models for 10 traits

			MV10	PC10	PC9	PC8	PC7	PC6
No. of equations (in million)			24.37	24.37	22.37	20.37	18.38	16.38
NNZ[a]	Sparse	after groups	901.8	1355.4	1208.4	1015.2	840.3	683.6
		after random	1420.1	1401.2	1249.6	1051.9	872.4	711.1
	Dense	genotyped	5722.4	572.3	515.1	457.8	400.6	343.4
		groups	0.313	0.331	0.271	0.215	0.164	0.121
	Total		7167.1	1989.1	1787.3	1530.2	1291.4	1070.9
Memory (GB)			104.0	28.9	26.4	23.1	20.2	17.3
No. of PCG iterates			1797	1938	1969	1913	1891	1517
Time[b]	single	Precondition	293.3	3.25	3.00	2.57	2.40	1.95
		Total	959	202	188	155	139	126
	multi	Precondition	25.0	0.6	0.5	0.4	0.4	0.3
		Total	551	147	140	127	116	88

[a]No. of non-zero elements in coefficient matrix (in million) [b]in minutes, for single- and multi-threaded MKL routines

to be at least equally important in such schemes. In addition, if sufficient memory is available, they are readily combined with in-core storage of dense blocks and experience gained here with library routines for matrix computations should be directly transferable.

CONCLUSIONS

Computational strategies described are expected to play an essential rôle in making multi-trait, single-step genetic evaluation for Australian livestock computationally feasible.

ACKNOWLEDGEMENTS

Work was supported by Meat and Livestock Australia grants B.BFG.0050, B.SGN.0027 and B.SGN.0028.

REFERENCES

Aguilar I., Misztal I., Legarra A. and Tsuruta S. (2011) *J. Anim. Breed. Genet.* **128**:422.

Anderson E., Bai Z., Bischof C., Blackford S., Demmel J., Dongarra J., Du Croz J., Greenbaum A., Hammarling S., McKenney A. and Sorensen D. (1999) *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, Third edition.

Blackford L., Demmel J., Dongarra J., Duff I., Hammarling S., Henry G., Heroux M., Kaufman L., Limsdaine A., Petitet A., Pozo R., Remington K. and Whaley R.C. (2002) *ACM Trans. Math. Softw.* **28**:135.

Brown D.J., Huisman A.E., Swan A.A., Graser H.U., Woolaston R.R., Ball A.J., Atkins K.D. and Banks R.G.a. (2007) *Proc. Ass. Advan. Anim. Breed. Genet.* **17**:187.

Meyer K. (2009) *Proc. Ass. Advan. Anim. Breed. Genet.* **18**:442.

Meyer K. and Kirkpatrick M. (2005) *Genet. Sel. Evol.* **37**:1.

Misztal I., Legarra A. and Aguilar I. (2009) *J. Dairy Sci.* **92**:4648.

Misztal I., Vitezica Z.G., Legarra A., Aguilar I. and Swan A.A. (2013) *J. Anim. Breed. Genet.* **130**:252.

Tier B. and Graser H.U. (1991) *J. Anim. Breed. Genet.* **108**:81.

Tsuruta S., Misztal I. and Strandén I. (2001) *J. Anim. Sci.* **79**:1166.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., Nyholt D.R., Madden P.A., Heath A.C., Martin N.G., Montgomery G.W., Goddard M.E. and Visscher P.M. (2010) *Nature Genet.* **42**:565.