Technical note: Genetic principal component models for multitrait single-step genomic evaluation

K. Meyer, A. Swan, and B. Tier

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

ABSTRACT: A reparameterization of the multivariate linear mixed model in genetic evaluation to principal components is described. This yields an equivalent model with a sparser coefficient matrix in the mixed model equations and, thus, reduced computational requirements to solve them. It is especially advantageous for analyses incorporating genomic relationship information with many nonzero elements in the inverse of the relationship matrix. Moreover, the framework lends itself directly to dimension reduction and, thus, further computational savings by omitting genetic principal components with negligible eigenvalues. The potential impact on computational demands is illustrated for an application to single-step genomic evaluation of Australian sheep.

Key words: computing, dimension reduction, equivalent model, genomic evaluation, multitrait, single step

INTRODUCTION

Genetic evaluation using genomic information has been adopted or is currently being introduced in many livestock improvement schemes worldwide. A substantial proportion of these uses the so-called “single-step” procedure (Misztal et al., 2009), which allows for joint evaluation of all animals, genotyped or not (see Legarra et al., 2014, for a recent review). The single-step approach can be viewed as an extension of previous BLUP models, replacing the pedigree-based numerator relationship matrix between animals, $A$, by its counterpart, $H$, which combines the genomic relationship matrix among genotyped animals, $G$, with relationships derived from the pedigree. As $G$ and the corresponding part of $H$ are dense, computational requirements to predict breeding values are increased by orders of magnitude. We describe a parameterization of the multitrait model (also applicable for related, multicoefficient models) to genetic principal components (PC). This results in a drastic reduction in the number of nonzero (NNZ) elements in the coefficient matrix of the mixed model equations (MME) arising from covariances among genetic effects while somewhat increasing the number in the remaining parts. We demonstrate that this can substantially reduce the computational requirements in the context of single-step genomic analyses, especially if the genetic covariance matrix has eigenvalues close to zero and dimension reduction is feasible.

EQUIVALENT MODELS AND BEYOND

Consider a linear, mixed model for $q$ traits:

$$y = X\beta + Zu + e,$$  \[1\]

with $y$, $\beta$, $u$, and $e$ the vectors of observations, fixed and random effects, and residuals and $X$ and $Z$ the pertaining incidence matrices. The model given in Eq. [1] can include additional random effects or represent other models with multiple coefficients per individual, such as random regression models. For clarity of presentation, we consider a simple multitrait animal model in the following, in which $Z$ comprises $q$ columns per animal. Let $u$ represent animals’ additive genetic effects, ordered by animals within traits. This gives

$$\text{Var}(u) = \Sigma \otimes H,$$  \[2\]
Multitrait genomic best linear unbiased prediction

\[
\begin{align*}
\text{Var}(e) &= \text{var}(Q \Sigma), \text{which typically has a single element of unity (except for residual covariances between traits recorded on the same animal. If \(u\) is ordered by traits within animals, this is block-diagonal for animals, with blocks of size \(q \times q\). The second part, \(\Sigma^{-1} \otimes H^{-1}\), generates genetic links between animals and traits. Assuming \(\Sigma^{-1}\) has no special structure (i.e., no elements of zero), each nonzero off-diagonal element of \(H^{-1}\) contributes \(q^2\) new, nonzero elements to \(C_U\).}

It has long been recognized that reformulating a linear mixed model without changing means or variances can yield “equivalent” models for many problems that are computationally more tractable than the original (Henderson, 1985). Expand Eq. [1] to

\[
y = X\beta + Z(Q \otimes I)(Q^{-1} \otimes I)u + e = X\beta + Z^*u^* + e, \quad [4]
\]

with \(I\) denoting an identity matrix, \(Z^* = Z(Q \otimes I)\) and \(u^* = (Q^{-1} \otimes I)\). This gives an equivalent model with \(\text{var}(u^*) = Q^{-1}\Sigma Q^{-T} \otimes H = \Sigma^* \otimes H\) (superscript “\(^-T\)” denotes the transpose of the matrix inverse). The corresponding MME are

\[
\begin{bmatrix}
X'R'X & X'R'Z' \\
Z'R'X & Z'R'Z' + \Sigma^{-1} \otimes H^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
u^*
\end{bmatrix}
= \begin{bmatrix} X'R'y \\ Z'R'y \end{bmatrix} \quad [5]
\]

Let \(C^*_u\) denote the coefficient matrix in Eq. [5]. Judicious choice of \(Q\) can yield a matrix \(\Sigma^*_u\), which is diagonal. This reduces the NNZ elements contributed to the coefficient matrix by each nonzero (off-diagonal) element of \(H^{-1}\) from \(q^2\) to \(q\). The trade-off is that \(Z^*\) has up to \(q\) nonzero elements per observation compared with \(Z\), which typically has a single element of unity (except for random regression models). The additional elements in the incidence matrix give rise to extra nonzero elements in the other parts of \(C^*_u\) involving \(Z^*\). The equivalent model is computationally advantageous if the latter do not exceed the reduction above. This is generally the case if there is at least a modest number of genotyped animals and \(H^{-1}\) has corresponding dense parts.

Suitable matrices \(Q\) arise from the eigendecomposition of the genetic covariance matrix, \(\Sigma = E\Lambda E'\). An obvious choice is the matrix of eigenvectors, \(Q = E\), which gives \(\Sigma^* = \Lambda\) and results in elements of \(u^*\) equal to genetic PC scores (Kirkpatrick and Meyer, 2004). The elements of the diagonal matrix \(\Lambda\) are the eigenvalues of \(\Sigma\). Any PC with eigenvalues close to zero contribute negligible information and can, therefore, be omitted. Such dimension reduction changes the equivalent to a reduced rank model. Assume eigenvalues are in descending order with the last \(q - r\) values deemed indistinguishable from zero. This is often the case for larger values of \(q\), that is, analyses considering more than a few traits. The reduced rank model is then obtained by setting \(Q\) to the submatrix of \(E\) comprising columns 1 to \(r\). This results in \(r < q\) elements in \(u^*\) for each animal; hence, it reduces the dimension and, thus, number of equations in the model as well as the NNZ elements in \(C^*_u\). Depending on the proportion of PC that can be “safely” omitted, this can result in substantial, additional savings in computational requirements. At convergence, solutions for the \(r\) elements of \(u^*\) are readily transformed into \(q\) predicted breeding values on the original scale.

An alternative choice for \(Q\) is the matrix of “factor loadings,” \(Q = E\Lambda^{-1/2}\). This gives \(\Sigma^* = I\) and avoids multiplication of \(H^{-1}\) by the diagonal elements of \(\Lambda\). Generally, \(E\) and, thus, \(E\Lambda^{-1/2}\) are dense, that is, contain \(q^2\) nonzero elements. However, the factor matrix can be rotated to lower triangular form, \(Q = E\Lambda^{-1/2}T\), with \(T\) an orthogonal matrix (i.e., \(TT' = I\)). This reduces the NNZ elements in \(Q\) to \((q + 1)/2\) and, consequently, the number of multiplications to set up the MME and the NNZ elements in \(C^*_u\). As above, a reduced rank model can be obtained by restricting \(Q\) to the first \(r < q\) columns of the (rotated) factor matrix (Meyer and Kirkpatrick, 2005).

The reparameterization proposed is reminiscent of the canonical transformation, for example, Thompson (1976). However, the latter requires that all individuals have all traits recorded with the same design matrices. The parameterization proposed does not involve these restrictions as it is applied at the effect level rather than to the data. Moreover, it is readily applied to models fitting additional random effects, for example, maternal genetic effects (see application below). Closely related is the so-called extended, factor-analytic model (Thompson et al., 2003), in which \(u\) is partitioned into \(r\) common and \(q\) specific factors, which have zero within-animal covariances. Although this model may result in proportionally even sparser coefficient matrices, it requires \(\Sigma\) to have an appropriate structure.

**APPLICATION**

We demonstrate the utility of the PC models by comparing standard multivariate (MV) and PC implementations for a subset of traits considered in genetic evaluation of meat sheep. Data consisted of 5.28 million
records for 11 traits recorded on 1.77 million animals in the terminal sire breeds evaluation in LAMBPLAN, the national genetic evaluation scheme for the Australian sheep industry (Brown et al., 2007). The first 5 traits represented the most commonly recorded traits in these breeds, namely birth, weaning, and postweaning weights and postweaning eye muscle area and fat depth. Numbers of records for these traits ranged from 888,365 to 1,386,636. In contrast, there were few records—between 5,786 and 8,154—for the other traits, comprising carcass measurements for muscle depth, fat depth, intramuscular fat percentage, shear force at Day 5, lean meat yield, and dressing percentage.

The model of analysis included contemporary groups as fixed effects and animals’ additive genetic effects and genetic groups (93 levels) as random effects for all traits. The latter were fitted “explicitly”—assigning proportions of membership for each animal—as augmenting the pedigree by phantom parents in single-step applications has been found to be problematic (Misztal et al., 2013). Genetic groups and additive genetic effects were fitted assuming the same covariance matrix with the same rank. In addition, dams’ permanent environmental effects (653,068 levels) were considered random effects for the 3 body weight traits. As in the routine LAMBPLAN evaluation, records for traits 1 to 5 were precorrected for the effects of birth-rearing type, age at measurement, and age of dam and body weight as a covariate for muscle and fat depth. For the carcass traits, birth-rearing type classes were fitted as fixed effects and age at slaughter and dam age were fitted as linear and quadratic covariables. For traits 6 to 9, the model also adjusted for HCW by including it as a linear covariable.

Including parents without records, there were 1,995,755 animals, of which 10,944 were genotyped for 48,599 SNP. To build $H^{-1}$, genomic relationships were computed following Yang et al. (2010). This yielded 66,455,483 nonzero elements in $H^{-1}$ (half-stored) compared with 6,584,393 elements in the corresponding pedigree-based matrix, $A^{-1}$.

**COMPUTING ENVIRONMENT AND STRATEGY**

Mixed model equations were solved iteratively, using double precision computations in a preconditioned conjugate gradient algorithm (PCG; e.g., Tsuruta et al., 2001) with partial Cholesky decomposition preconditioner, as implemented in the single-step module of our mixed model package WOMBAT (Meyer, 2007). Nonzero elements of the coefficient matrix (half-stored) were held in core using a combination of sparse matrix and dense storage. Equations for animals’ additive genetic effects were grouped by availability of genomic information. For nongenotyped animals, these were ordered by traits within animal, yielding small, dense blocks of size $q \times q$ in the coefficient matrix $C_q$ in the standard MV parameterization. For genotyped animals, equations were ordered by animal within trait. This order resulted in an advantageous structure of the coefficient matrix for the PC models. Let $N$ denote the number of genotyped animals. Dense diagonal blocks for genotyped animals were stored in 2-dimensional arrays, with additional vectors for diagonal elements. This required a single matrix of size $qN \times qN$ for MV analyses. For the PC models, however, ordering equations by animal within trait yielded $r$ dense diagonal blocks in the coefficient matrix of size $N \times N$, whereas the “between-trait” (or, more accurately, the between-PC) off-diagonal blocks were sparse. Storing only these diagonal blocks as dense matrices, storage requirements were thus reduced from $q^2N^2$ to $rN^2$ elements. Diagonal blocks for genetic groups were held in a single dense block for both MV and PC models, as the transformation to PC scale resulted in almost dense, between-trait off-diagonal blocks. Similarly, the diagonal block for the small number of equations representing additional fixed effects and covariables fitted for the traits 6 to 11 were also treated as a single dense block. All remaining nonzero coefficients were held in compressed sparse row format.

Cholesky factors of dense blocks to form the preconditioning matrix were obtained using LAPACK routine DPOTRF (Anderson et al., 1999). Each PCG iterate then required solutions to corresponding subsystems of equations, which were computed using LAPACK routine DPOTRS. These routines access only one triangle of a symmetric matrix and the user is able to select which part. This allowed storage of the dense blocks of the coefficient matrix and their Cholesky factors in the same array, swapping diagonals as required. Computations for each PCG iterate also required the product of the (complete) coefficient matrix and a vector. This was formed using routines DSYMV from the Basic Linear Algebra Subprograms (BLAS) library (Blackford et al., 2002) for the dense parts of the MME and the sparse matrix equivalent, MKL_DCSRSYMV, for the sparse parts. Furthermore, the vector “inner” products needed for each PCG iterate were evaluated using the BLAS function DDOT.

Computations were carried out under Linux on a machine with 256 GB of RAM and 16 Intel Xeon CPU E5-2630 cores (Intel Corporation, Santa Clara, CA), rated at 2.4 GHz with a cache size of 20 MB. Basic Linear Algebra Subprograms (BLAS) library (Blackford et al., 2002) and LAPACK (Anderson et al., 1999) routines used were loaded from the Intel Math Kernel Library 11.1 (Intel Corporation), comparing single and multithreaded versions.
RESULTS AND DISCUSSION

Characteristics of the MME and computational requirements to solve them for the standard MV parameterization (MV11) and PC models fitting \( r = 11 \), \( \ldots \), 7 components (denoted as PC\( r \)) are summarized in Table 1. With equal numbers of equations, equivalent models MV11 and PC11 differed substantially in the NNZ elements in the coefficient matrix, with approximately corresponding ratios for memory requirements (values given pertain to the solution step) and computing time needed. With few records for traits 6 to 11, there was a high proportion of additional elements in \( C^* \) from the data, that is, due to the higher NNZ coefficients in \( Z^* \) than in \( Z \). In spite of this and the fact that only a small proportion of animals (0.55%) had genotypes, the PC model proved highly advantageous. Clearly, this advantage will rapidly increase as the amount of genomic information available rises.

Reducing the number of PC further decreased the total number of equations and computational requirements with decreasing \( r \), and with fewer effects to be estimated, solutions converged in fewer iterations. Eigenvalues of the genetic covariance matrix ranged from 10.2 to 0.015, with the last 3 eigenvalues together explaining 1.1% of the total genetic variation. Hence, we expect to be able to reduce \( r \) by at least 1 or even 2 or 3 with little effect on the accuracy of genetic evaluation. It should be emphasized, though, that such choice must be carefully made, inspecting the impact for individual traits and selected groups of target animals. Even for cases where little dimension reduction is considered advisable, the parameterization to PC remains very effective in reducing computational requirements.

Using multithreaded libraries reduced elapsed time, substantially so for the standard MV model. For this analysis, evaluation of the preconditioning matrix involved the Cholesky decomposition of a matrix of size \( qN = 120,384 \), which represented most of the time needed for the setup steps, and it was in this decomposition routine where all 16 processors were used. Corresponding times for the PC analyses and improvements were much smaller—the operation count for the Cholesky factorization cubically increases with the size of the matrix \( n^3/6 + n^2/2 + n/3 \) multiplications for a matrix of size \( n \). Hence, decomposing a matrix of size \( N \) up to \( qN \) times was substantially less demanding than the same operation for a matrix of size \( qN \). Again, multithreaded computing will become more effective as the number of genotyped animals increases.

Calculations shown considered a relatively small example where the MME could be held in core. In

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Unit</th>
<th>MV11</th>
<th>PC11</th>
<th>PC10</th>
<th>PC9</th>
<th>PC8</th>
<th>PC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics of the mixed model equations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEQ (^3)</td>
<td></td>
<td>24.16</td>
<td>24.16</td>
<td>22.17</td>
<td>20.17</td>
<td>18.17</td>
<td>16.18</td>
</tr>
<tr>
<td>NNZ (^4)</td>
<td>Sparse</td>
<td>Data</td>
<td>234.5</td>
<td>1,301.8</td>
<td>1,157.2</td>
<td>964.7</td>
<td>790.5</td>
</tr>
<tr>
<td></td>
<td>Sparse</td>
<td>Data + random</td>
<td>891.5</td>
<td>1,352.2</td>
<td>1,203.0</td>
<td>1,005.7</td>
<td>827.1</td>
</tr>
<tr>
<td></td>
<td>Dense</td>
<td>Genotyped</td>
<td>7,246.2</td>
<td>658.8</td>
<td>598.9</td>
<td>539.0</td>
<td>479.1</td>
</tr>
<tr>
<td></td>
<td>Dense</td>
<td>Groups</td>
<td>0.26</td>
<td>0.40</td>
<td>0.33</td>
<td>0.27</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td>8,162.1</td>
<td>2,035.4</td>
<td>1,824.3</td>
<td>1,565.3</td>
<td>1,324.5</td>
</tr>
</tbody>
</table>

**Computational requirements**

| Memory | GB | 118.0 | 28.3 | 25.6 | 22.3 | 19.3 | 16.4 |
| Time Setup | min | 438 | 25 | 21 | 19 | 18 | 17 |
| Time Total | min | 1,981 | 526 | 303 | 248 | 205 | 149 |
| Iterates | 3,471 | 4,001 | 3,690 | 3,455 | 3,070 | 2,514 |

| Memory Time Setup | GB | 118.0 | 28.3 | 25.6 | 22.3 | 19.3 | 16.4 |
| Time Total | min | 1,981 | 526 | 303 | 248 | 205 | 149 |
| Iterates | 3,471 | 4,001 | 3,690 | 3,455 | 3,070 | 2,514 |

1 MV = multivariate; MV11 = standard MV parameterization.
2 PC = principal components; PC\( r \) = principal component model comprising \( r \) components.
3 NEQ = number of equation in the mixed model (in millions).
4 NNZ = number of nonzero; the NNZ elements in half-stored coefficient matrix (in millions).
5 Computing times using a single processor.
6 Number of iterates required in iterative solution scheme.
7 Computing times using multithreaded library routines and up to 16 processors.
practice, this is unlikely to be feasible and an “iteration on data” type strategy will need to be used instead (Schaeffer and Kennedy, 1986; Misztal and Gianola, 1987; Tier and Graser, 1991). However, the NNZ elements in the coefficient matrix will be similarly important in such schemes. Details will depend heavily on the particular model and scheme chosen. Indeed, in some current implementations for MV genetic evaluation, all animals are treated as if they have records for all traits assigning “dummy” observations where necessary (e.g., Brown et al., 2001), so that the denser nature of $Z^*$ will have much less impact on the NNZ elements in $C^*$ than in our example. As shown, in-core storage of dense blocks in the coefficient matrix together with the use of efficient, multithreaded library routines is highly advantageous and should, at least to some extent, also contribute to the efficiency of iteration on data schemes. Obviously, as the number of genotyped animals increases, we may only be able to resort to partitioned matrix operations. Computational efficiency of genomic evaluation is an active area of research. Other equivalent (or almost equivalent) models have been proposed and other strategies to increase sparseness of the MME are available that will offer alternative or complementary computational savings.

**Implications**

The models described provide an addition to our toolbox, which will help in making multitrait, single-step genomic evaluation in livestock improvement schemes computationally feasible.

**LITERATURE CITED**


