Genetics Selection Evolution

SHORT COMMUNICATION

Open Access

Computing strategies for multi-population genomic evaluation

Andrés Legarra[1*], David González-Diéguez[1,2] and Zulma G. Vitezica[1]

Abstract

Background: Multiple breed evaluation using genomic prediction includes the use of data from multiple populations, or from parental breeds and crosses, and is expected to lead to better genomic predictions. Increased complexity comes from the need to fit non-additive effects such as dominance and/or genotype-by-environment interactions. In these models, marker effects (and breeding values) are modelled as correlated between breeds, which leads to multiple trait formulations that are based either on markers [single nucleotide polymorphism best linear unbiased prediction (SNP-BLUP)] or on individuals [genomic(G)BLUP)]. As an alternative, we propose the use of generalized least squares (GLS) followed by backsolving of marker effects using selection index (SI) theory.

Results: All investigated options have advantages and inconveniences. The SNP-BLUP yields marker effects directly, which are useful for indirect prediction and for planned matings, but is very large in number of equations and is structured in dense and sparse blocks that do not allow for simple solving. GBLUP uses a multiple trait formulation and is very general, but results in many equations that are not used, which increase memory needs, and is also structured in dense and sparse blocks. An alternative formulation of GBLUP is more compact but requires tailored programming. The alternative of solving by GLS + SI is the least consuming, both in number of operations and in memory, and it uses only single dense blocks. However, it requires dedicated programming. Computational complexity problems are exacerbated when more than additive effects are fitted, e.g. dominance effects or genotype x environment interactions.

Conclusions: As multi-breed predictions become more frequent and non-additive effects are more often included, standard equations for genomic prediction based on Henderson's mixed model equations become less practical and may need to be replaced by more efficient (although less general) approaches such as the GLS + SI approach proposed here.

Background

For genomic prediction in livestock and crops, marker effects are often modelled as different but correlated effects across populations [1, 2]. This results in a multiple trait setting, in which each environment or population is modelled as a different trait. As individuals are present in a single environment, there is no covariance between other random effects (such as residual or permanent environmental effects). This leads to particular structures of the incidence matrices that make general computational strategies less efficient. In this work, we discuss some of these strategies, show that general strategies such as standard individual-based multiple trait genomic best linear unbiased prediction (GBLUP) leads to high computational redundancy, and we highlight that the old method of generalized least squares (GLS) followed by selection index (SI) [3, 4] is a competing strategy in terms of efficiency. The motivation for the comparison of these strategies was the need to obtain estimates of marker effects in a computationally efficient manner, for their use in planning assortative

*Correspondence: andres.legarra@inrae.fr
[1] INRAE, INP, UMR 1388 GenPhySE, 31326 Castanet-Tolosan, France
Full list of author information is available at the end of the article

matings in a two-way breeding scheme [5], using data from two purebred populations and a crossbred population, for a total of ~ 50K animals and genotypes at ~ 50K single nucleotide polymorphisms (SNPs). Here, we are concerned with medium-sized data sets, but with very complex models, where genetic evaluation can be done by exact (non-iterative) methods, i.e. by numerical inversion. This is a popular strategy for genomic evaluation in crops and in some populations of monogastric livestock.

Methods

For our argumentation, and following the example of [5], we assume additive (a) and dominant (d) SNP effects that are correlated across the different populations. In principle, the model can be extended to higher-order effects such as epistasis. The dominance effects have an a priori mean of 0 because genomic inbreeding is included in the model as a covariate [6]. For the example, we consider three populations but any number of populations can be accommodated. For each population i $(i = \{1, 2, 3\})$, we have a single trait vector of phenotypes, $\mathbf{y}_i$, with the following linear model:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_i\mathbf{a}_i + \mathbf{W}_i\mathbf{d}_i + \mathbf{e}_i,$$

where $\boldsymbol{\beta}_i$ contains the fixed effects specific to population i (e.g., a mean and an inbreeding depression covariate), $\mathbf{a}_i$ and $\mathbf{d}_i$ are additive and dominance SNP effects, respectively, specific to population i, with incidence matrices coded, e.g., as $\mathbf{Z}_i = \{-1, 0, 1\}$ and $\mathbf{W}_i = \{0, 1, 0\}$. Other codings are possible, such as in terms of breeding values and dominance deviations to achieve orthogonality [7]. The covariance structure, written using the Kronecker product ⊗, is:

$$Var\begin{pmatrix}\mathbf{a}_1\\ \mathbf{a}_2\\ \mathbf{a}_3\end{pmatrix} = \mathbf{G}_{0a}\otimes\mathbf{I} = \begin{pmatrix}\mathbf{I}\sigma^2_{a1} & \mathbf{I}\sigma_{a12} & \mathbf{I}\sigma_{a13}\\ \mathbf{I}\sigma_{a21} & \mathbf{I}\sigma^2_{a2} & \mathbf{I}\sigma_{a23}\\ \mathbf{I}\sigma_{a31} & \mathbf{I}\sigma_{a32} & \mathbf{I}\sigma^2_{a3}\end{pmatrix},$$

for $\mathbf{G}_{0a} = \begin{pmatrix}\sigma^2_{a1} & \sigma_{a12} & \sigma_{a13}\\ \sigma_{a21} & \sigma^2_{a2} & \sigma_{a23}\\ \sigma_{a31} & \sigma_{a32} & \sigma^2_{a3}\end{pmatrix}$, the covariance matrix of additive marker effects, with inverse $\mathbf{G}^{-1}_{0a} = \begin{pmatrix}g^{11}_{0a} & g^{12}_{0a} & g^{13}_{0a}\\ g^{21}_{0a} & g^{22}_{0a} & g^{23}_{0a}\\ g^{31}_{0a} & g^{32}_{0a} & g^{33}_{0a}\end{pmatrix}$

.

$$Var\begin{pmatrix}\mathbf{d}_1\\ \mathbf{d}_2\\ \mathbf{d}_3\end{pmatrix} = \mathbf{G}_{0d}\otimes\mathbf{I} = \begin{pmatrix}\mathbf{I}\sigma^2_{d1} & \mathbf{I}\sigma_{d12} & \mathbf{I}\sigma_{d13}\\ \mathbf{I}\sigma_{d21} & \mathbf{I}\sigma^2_{d2} & \mathbf{I}\sigma_{d23}\\ \mathbf{I}\sigma_{d31} & \mathbf{I}\sigma_{d32} & \mathbf{I}\sigma^2_{d3}\end{pmatrix},$$

for $\mathbf{G}_{0d} = \begin{pmatrix}\sigma^2_{d1} & \sigma_{d12} & \sigma_{d13}\\ \sigma_{d21} & \sigma^2_{d2} & \sigma_{d23}\\ \sigma_{d31} & \sigma_{d32} & \sigma^2_{d3}\end{pmatrix}$, with inverse

$$\mathbf{G}^{-1}_{0d} = \begin{pmatrix}g^{11}_{0d} & g^{12}_{0d} & g^{13}_{0d}\\ g^{21}_{0d} & g^{22}_{0d} & g^{23}_{0d}\\ g^{31}_{0d} & g^{32}_{0d} & g^{33}_{0d}\end{pmatrix}.$$

$$Var\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{e}_3\end{pmatrix} = \mathbf{R} = \begin{pmatrix}\mathbf{I}\sigma^2_{e1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma^2_{e2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_{e3}\end{pmatrix},$$

$$\begin{aligned}Var\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_3\end{pmatrix} &= \mathbf{V} = \begin{pmatrix}\mathbf{Z}_1\sigma^2_{a1}\mathbf{Z}_1^{'} & \mathbf{Z}_1\sigma_{a12}\mathbf{Z}_2^{'} & \mathbf{Z}_1\sigma_{a13}\mathbf{Z}_3^{'}\\ \mathbf{Z}_2\sigma_{a21}\mathbf{Z}_1^{'} & \mathbf{Z}_2\sigma^2_{a2}\mathbf{Z}_2^{'} & \mathbf{Z}_2\sigma_{a23}\mathbf{Z}_3^{'}\\ \mathbf{Z}_3\sigma_{a31}\mathbf{Z}_1^{'} & \mathbf{Z}_3\sigma_{a32}\mathbf{Z}_2^{'} & \mathbf{Z}_3\sigma^2_{a3}\mathbf{Z}_3^{'}\end{pmatrix}\\ &+ \begin{pmatrix}\mathbf{W}_1\sigma^2_{d1}\mathbf{W}_1^{'} & \mathbf{W}_1\sigma_{d12}\mathbf{W}_2^{'} & \mathbf{W}_1\sigma_{d13}\mathbf{W}_3^{'}\\ \mathbf{W}_2\sigma_{d21}\mathbf{W}_1^{'} & \mathbf{W}_2\sigma^2_{d2}\mathbf{W}_2^{'} & \mathbf{W}_2\sigma_{d23}\mathbf{W}_3^{'}\\ \mathbf{W}_3\sigma_{d31}\mathbf{W}_1^{'} & \mathbf{W}_3\sigma_{d32}\mathbf{W}_2^{'} & \mathbf{W}_3\sigma^2_{d3}\mathbf{W}_3^{'}\end{pmatrix}\\ &+ \begin{pmatrix}\mathbf{I}\sigma^2_{e1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\sigma^2_{e2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_{e3}\end{pmatrix}\\ &= \mathbf{G}_A + \mathbf{G}_D + \mathbf{R}\end{aligned}$$

Based on this notation, several equivalent estimators are derived in the following. We assume that variance components are known and that all individuals have a single record. Note that, by construction, we have a highly parameterized model, which in the GBLUP case has (many) more unknowns than records. I.e., each individual has an unknown additive and a dominance effect for each population but only one record in only one of the populations.

SNP-BLUP

This model is parameterized in terms of SNP effects across the different populations, resulting in the following structure of the left-hand side (LHS) of the mixed model equations (MME) (for clarity, only the random effects part is shown):

$$\begin{pmatrix} \cdots & & & & & \\ \mathbf{Z}_1^{'}\mathbf{Z}_1\sigma_{e1}^{-2}+\mathbf{I}g_{0a}^{11} & \mathbf{I}g_{0a}^{12} & \mathbf{I}g_{0a}^{13} & \mathbf{Z}_1^{'}\mathbf{W}_1\sigma_{e1}^{-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}g_{0a}^{21} & \mathbf{Z}_2^{'}\mathbf{Z}_2\sigma_{e2}^{-2}+\mathbf{I}g_{0a}^{22} & \mathbf{I}g_{0a}^{23} & \mathbf{0} & \mathbf{Z}_2^{'}\mathbf{W}_2\sigma_{e2}^{-2} & \mathbf{0} \\ \mathbf{I}g_{0a}^{31} & \mathbf{I}g_{0a}^{32} & \mathbf{Z}_3^{'}\mathbf{Z}_3\sigma_{e3}^{-2}+\mathbf{I}g_{0a}^{33} & \mathbf{0} & \mathbf{0} & \mathbf{Z}_3^{'}\mathbf{W}_3\sigma_{e3}^{-2} \\ \mathbf{W}_1^{'}\mathbf{Z}_1\sigma_{e1}^{-2} & \mathbf{0} & \mathbf{0} & \mathbf{W}_1^{'}\mathbf{W}_1\sigma_{e1}^{-2}+\mathbf{I}g_{0d}^{11} & \mathbf{I}g_{0d}^{12} & \mathbf{I}g_{0d}^{13} \\ \mathbf{0} & \mathbf{W}_2^{'}\mathbf{Z}_2\sigma_{e2}^{-2} & \mathbf{0} & \mathbf{I}g_{0d}^{21} & \mathbf{W}_2^{'}\mathbf{W}_2\sigma_{e2}^{-2}+\mathbf{I}g_{0d}^{22} & \mathbf{I}g_{0d}^{23} \\ \mathbf{0} & \mathbf{0} & \mathbf{W}_3^{'}\mathbf{Z}_3\sigma_{e3}^{-2} & \mathbf{I}g_{0d}^{31} & \mathbf{I}g_{0d}^{32} & \mathbf{W}_3^{'}\mathbf{W}_3\sigma_{e3}^{-2}+\mathbf{I}g_{0d}^{33} \end{pmatrix}$$

This LHS is composed of several dense blocks of the form $\mathbf{Z}_1^{'}\mathbf{Z}_1\sigma_{e1}^{-2}+\mathbf{I}g_{0a}^{11}$, several sparse blocks of the form $\mathbf{I}g_{0a}^{12}$, and several zero blocks. This makes the use of current "state of the art" sparse matrix software inefficient, because most gains due to sparsity are lost. As a result, it may be more efficient to work with full storage, i.e. dense matrices. However, the size of the equations is rather large, i.e. $6m$, where m is the number of SNPs. For instance, an analysis with 50 K SNP panels would involve MME of size 300 K and the inversion of the LHS would consume $O(6m)^3$ operations. The size of these equations are, however, invariant to the number of records n.

GBLUP using a standard multiple trait formulation

There are two possible implementations of GBLUP. The first considers three additive values and three dominance values per individual (one for each population). For each random effect, the MME use a single matrix that is arbitrarily scaled across all populations (the scale is arbitrary because there is no meaningful scaling factor that yields coherent "relationships" across the three populations). For instance, a possible matrix for the additive effect is:

$$\mathbf{G}_A^{*}=\begin{pmatrix} \mathbf{Z}_1\mathbf{Z}_1^{'} & \mathbf{Z}_1\mathbf{Z}_2^{'} & \mathbf{Z}_1\mathbf{Z}_3^{'} \\ \mathbf{Z}_2\mathbf{Z}_1^{'} & \mathbf{Z}_2\mathbf{Z}_2^{'} & \mathbf{Z}_2\mathbf{Z}_3^{'} \\ \mathbf{Z}_3\mathbf{Z}_1^{'} & \mathbf{Z}_3\mathbf{Z}_2^{'} & \mathbf{Z}_3\mathbf{Z}_3^{'} \end{pmatrix}$$

with $\left(\mathbf{G}_A^{*}\right)^{-1}=\begin{pmatrix} \mathbf{G}_A^{*(11)} & \mathbf{G}_A^{*(12)} & \mathbf{G}_A^{*(13)} \\ \mathbf{G}_A^{*(21)} & \mathbf{G}_A^{*(22)} & \mathbf{G}_A^{*(23)} \\ \mathbf{G}_A^{*(31)} & \mathbf{G}_A^{*(32)} & \mathbf{G}_A^{*(33)} \end{pmatrix}$.

Note that the cost of inversion of $\mathbf{G}_A^{*}$ is $O\left(n^3\right)$. Note also that, in order to describe covariances between individuals, $\mathbf{G}_A^{*}$ must be multiplied by the SNP effect variance (e.g. σ_{a1}^2), rather than by the population variance, which is why we prefer not to refer to $\mathbf{G}_A^{*}$ as containing "relationships". A similar matrix $\mathbf{G}_D^{*}$ is defined for dominance effects. The two matrices $\mathbf{G}_A^{*}$ and $\mathbf{G}_D^{*}$ are used in a multiple trait formulation with missing records for each trait, using the Kronecker factorization. This results in the following structure of LHS, as for multiple trait analysis (for clarity, only a portion of the structure is shown, for additive effects for two of the three populations):

$$\begin{pmatrix} \cdots & & & & & \\ \mathbf{I}\sigma_{e1}^{-2}+\mathbf{G}_A^{*(11)}g_{0a}^{11} & \mathbf{G}_A^{*(12)}g_{0a}^{11} & \mathbf{G}_A^{*(13)}g_{0a}^{11} & \mathbf{G}_A^{*(11)}g_{0a}^{12} & \mathbf{G}_A^{*(12)}g_{0a}^{12} & \mathbf{G}_A^{*(13)}g_{0a}^{12} \\ \mathbf{G}_A^{*(21)}g_{0a}^{11} & \mathbf{G}_A^{*(22)}g_{0a}^{11} & \mathbf{G}_A^{*(23)}g_{0a}^{11} & \mathbf{G}_A^{*(21)}g_{0a}^{12} & \mathbf{G}_A^{*(22)}g_{0a}^{12} & \mathbf{G}_A^{*(23)}g_{0a}^{12} \\ \mathbf{G}_A^{*(31)}g_{0a}^{11} & \mathbf{G}_A^{*(32)}g_{0a}^{11} & \mathbf{G}_A^{*(33)}g_{0a}^{11} & \mathbf{G}_A^{*(31)}g_{0a}^{12} & \mathbf{G}_A^{*(32)}g_{0a}^{12} & \mathbf{G}_A^{*(33)}g_{0a}^{12} \\ \mathbf{G}_A^{*(11)}g_{0a}^{21} & \mathbf{G}_A^{*(12)}g_{0a}^{21} & \mathbf{G}_A^{*(13)}g_{0a}^{21} & \mathbf{G}_A^{*(11)}g_{0a}^{22} & \mathbf{G}_A^{*(12)}g_{0a}^{22} & \mathbf{G}_A^{*(13)}g_{0a}^{22} \\ \mathbf{G}_A^{*(21)}g_{0a}^{21} & \mathbf{G}_A^{*(22)}g_{0a}^{21} & \mathbf{G}_A^{*(23)}g_{0a}^{21} & \mathbf{G}_A^{*(12)}g_{0a}^{21} & \mathbf{I}\sigma_{e2}^{-2}+\mathbf{G}_A^{*(22)}g_{0a}^{22} & \mathbf{G}_A^{*(23)}g_{0a}^{22} \\ \mathbf{G}_A^{*(31)}g_{0a}^{21} & \mathbf{G}_A^{*(32)}g_{0a}^{21} & \mathbf{G}_A^{*(33)}g_{0a}^{21} & \mathbf{G}_A^{*(13)}g_{0a}^{21} & \mathbf{G}_A^{*(23)}g_{0a}^{22} & \mathbf{G}_A^{*(33)}g_{0a}^{22} \\ & & & & & \cdots \end{pmatrix}.$$

A similar pattern results for the dominance effects. There are also corresponding cross-products of incidence matrices in the additive x dominance blocks, e.g. $\mathbf{I}\sigma_{e1}^{-2}$. Thus, the LHS are composed of a dense formulation where each row has two times three blocks of size n, leading to $6n$ equations (and inversion cost of $O(6n)^3$). To our knowledge, this is the formulation used by standard BLUP/REML software programs (blupf90, Wombat, ASREML) with elements stored in memory. Note that these MME require $\mathbf{G}_{\boldsymbol{A}}$ and $\mathbf{G}_{\boldsymbol{D}}$ to be full rank, which is often not the case (because of the presence of clones, or as a result of "centering" the $\mathbf{Z}$ and $\mathbf{W}$

matrices, or because $n > m$). Matrices $\mathbf{G}_A$ and $\mathbf{G}_D$ may, therefore, require some "blending".

GBLUP using a compact formulation

The second option for the formulation of GBLUP, which requires much less memory, directly uses the inverses of the covariance matrices $\mathbf{G}_A$ and $\mathbf{G}_D$ (again, assuming that they are, or have been made, invertible), rather than the Kronecker factorization of covariances, e.g.:

$$\mathbf{G}_A^{-1} = \begin{pmatrix} \mathbf{Z}_1\sigma^2_{a1}\mathbf{Z}_1' & \mathbf{Z}_1\sigma_{a12}\mathbf{Z}_2' & \mathbf{Z}_1\sigma_{a13}\mathbf{Z}_3' \\ \mathbf{Z}_2\sigma_{a21}\mathbf{Z}_1' & \mathbf{Z}_2\sigma^2_{a2}\mathbf{Z}_2' & \mathbf{Z}_2\sigma_{a23}\mathbf{Z}_3' \\ \mathbf{Z}_3\sigma_{a31}\mathbf{Z}_1' & \mathbf{Z}_3\sigma_{a32}\mathbf{Z}_2' & \mathbf{Z}_3\sigma^2_{a3}\mathbf{Z}_3' \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{G}_A^{(11)} & \mathbf{G}_A^{(12)} & \mathbf{G}_A^{(13)} \\ \mathbf{G}_A^{(21)} & \mathbf{G}_A^{(22)} & \mathbf{G}_A^{(23)} \\ \mathbf{G}_A^{(31)} & \mathbf{G}_A^{(32)} & \mathbf{G}_A^{(33)} \end{pmatrix},$$

and then it uses the following LHS structure:

$$\begin{pmatrix} \cdots & & & & & & \\ & \mathbf{I}\sigma_{e1}^{-2} + \mathbf{G}_A^{(11)} & \mathbf{G}_A^{(12)} & \mathbf{G}_A^{(13)} & \mathbf{I}\sigma_{e1}^{-2} & \mathbf{0} & \mathbf{0} \\ & \mathbf{G}_A^{(21)} & \mathbf{I}\sigma_{e2}^{-2} + \mathbf{G}_A^{(22)} & \mathbf{G}_A^{(23)} & \mathbf{0} & \mathbf{I}\sigma_{e2}^{-2} & \mathbf{0} \\ & \mathbf{G}_A^{(31)} & \mathbf{G}_A^{(32)} & \mathbf{I}\sigma_{e3}^{-2} + \mathbf{G}_A^{(33)} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_{e3}^{-2} \\ & \mathbf{I}\sigma_{e1}^{-2} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_{e1}^{-2} + \mathbf{G}_D^{(11)} & \mathbf{G}_D^{(12)} & \mathbf{G}_D^{(13)} \\ & \mathbf{0} & \mathbf{I}\sigma_{e2}^{-2} & \mathbf{0} & \mathbf{G}_D^{(21)} & \mathbf{I}\sigma_{e2}^{-2} + \mathbf{G}_D^{(22)} & \mathbf{G}_D^{(23)} \\ & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_{e3}^{-2} & \mathbf{G}_D^{(31)} & \mathbf{G}_D^{(32)} & \mathbf{I}\sigma_{e3}^{-2} + \mathbf{G}_D^{(33)} \end{pmatrix},$$

which is of size $2n$, but is still considerably dense.

None of the above alternatives (SNP-BLUP or either GBLUP) are very satisfying since they all involve large matrices of size $> n$ or $> m$, and only SNP-BLUP directly leads to estimates of SNP effects, which are required to predict newly genotyped animals (without re-running the evaluation) or to arrange assortative matings.

GLS and selection index formulation

An alternative is to use GLS followed by SI theory, as shown by Henderson [3, 4]. First, we estimate the fixed effects by GLS:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

after which SNP effects are estimated through covariances as:

$$\widehat{\mathbf{a}} = \mathbf{C}_a'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right),$$

$$\widehat{\mathbf{d}} = \mathbf{C}_{\mathrm{d}}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right).$$

where $\mathbf{C}_a = Cov(\mathbf{y}, \mathbf{a}')$ and $\mathbf{C}_d = Cov\left(\mathbf{y}, \mathbf{d}'\right)$. Constructing $\mathbf{V}$ is actually easy, because it is just a sum of matrices. Provided $\mathbf{V}$ is invertible, and using a single inversion, we can solve for $\boldsymbol{\beta}$ first and then backsolve for $\mathbf{a}$ and $\mathbf{d}$. In the multi-population case, this may be computationally simpler than the SNP-BLUP and GBLUP strategies because $\mathbf{V}$ is smaller (of size n) than several of the $\mathbf{G}$ matrices or LHS matrices in the SNP-BLUP and GBLUP formulations. Moreover, $\mathbf{V}$ is invertible by construction, while $\mathbf{G}_A$ and $\mathbf{G}_D$ may not be full rank and may need some "blending". Several authors [8–10] have already pointed out that the GLS formulation is computationally more compact when MME are non-sparse and requires, in principle, fewer computations.

Therefore, to estimate SNP effects for multi-population evaluation, we propose the following algorithm based on GLS and SI. Assume that there are n_1, n_2, and n_3 records for each population:

1. Read data, build $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$.
2. Create empty $\mathbf{V}$ of the right size ($n = n_1 + n_2 + n_3$).
3. Add residual variances to $\mathbf{V}$.
4. Add contributions from populations 1, 2, 3 and a, d, e to $\mathbf{V}$.
5. Invert $\mathbf{V}$ (with associated cost $O(n^3)$).
6. Solve for fixed effects: $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.
7. Solve for random effects using covariances as described in the following. This can be done in steps ($\mathbf{Z}_1$, then $\mathbf{Z}_2$, etc.), specifically:

$$\widehat{\mathbf{a}} = \mathbf{C}_{\mathrm{a}}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\mathbf{d}} = \mathbf{C}_{\mathrm{d}}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right).$$

To compute $\widehat{\mathbf{a}}$ we need $\mathbf{C}_{\mathrm{a}}^{'} = \mathrm{Cov}\left(\mathbf{a}, \mathbf{y}^{'}\right)$ the covariance of $\mathbf{a}$ with $\mathbf{y}$, which is:

$$\mathbf{C}_{a}^{'} = \begin{pmatrix} \mathbf{I}\sigma_{a1}^{2} & \mathbf{I}\sigma_{a12} & \mathbf{I}\sigma_{a13} \\ \mathbf{I}\sigma_{a21} & \mathbf{I}\sigma_{a2}^{2} & \mathbf{I}\sigma_{a23} \\ \mathbf{I}\sigma_{a31} & \mathbf{I}\sigma_{a32} & \mathbf{I}\sigma_{a3}^{2} \end{pmatrix} \begin{pmatrix} \mathbf{Z}_{1}^{'} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{2}^{'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_{3}^{'} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{1}^{'}\sigma_{a1}^{2} & \mathbf{Z}_{2}^{'}\sigma_{a12} & \mathbf{Z}_{3}^{'}\sigma_{a13} \\ \mathbf{Z}_{1}^{'}\sigma_{a21} & \mathbf{Z}_{2}^{'}\sigma_{a2}^{2} & \mathbf{Z}_{3}^{'}\sigma_{a23} \\ \mathbf{Z}_{1}^{'}\sigma_{a31} & \mathbf{Z}_{2}^{'}\sigma_{a32} & \mathbf{Z}_{3}^{'}\sigma_{a3}^{2} \end{pmatrix}.$$

The algorithm to compute $\widehat{\mathbf{a}}$ then is:

7a. Build $\mathbf{y}^{\mathrm{c}} = \mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$.

7b. Then $\mathbf{v} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{1}^{'} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{2}^{'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_{3}^{'} \end{pmatrix} \begin{pmatrix} \mathbf{y}_1^{\mathrm{c}} \\ \mathbf{y}_2^{\mathrm{c}} \\ \mathbf{y}_3^{\mathrm{c}} \end{pmatrix} = \begin{pmatrix} \mathbf{Z}_{1}^{'}\mathbf{y}_1^{\mathrm{c}} \\ \mathbf{Z}_{2}^{'}\mathbf{y}_2^{\mathrm{c}} \\ \mathbf{Z}_{3}^{'}\mathbf{y}_3^{\mathrm{c}} \end{pmatrix}.$

And finally $\begin{pmatrix} \widehat{\mathbf{a}}_1 \\ \widehat{\mathbf{a}}_2 \\ \widehat{\boldsymbol{a}}_3 \end{pmatrix} = \begin{pmatrix} \mathbf{I}\sigma_{a1}^{2} & \mathbf{I}\sigma_{a12} & \mathbf{I}\sigma_{a13} \\ \mathbf{I}\sigma_{a21} & \mathbf{I}\sigma_{a2}^{2} & \mathbf{I}\sigma_{a23} \\ \mathbf{I}\sigma_{a31} & \mathbf{I}\sigma_{a32} & \mathbf{I}\sigma_{a3}^{2} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{pmatrix}.$

$$= \begin{pmatrix} \mathbf{v}_1\sigma_{a1}^{2} + \mathbf{v}_2\sigma_{a12} + \mathbf{v}_3\sigma_{a13} \\ \mathbf{v}_1\sigma_{a21} + \mathbf{v}_2\sigma_{a2}^{2} + \mathbf{v}_3\sigma_{a23} \\ \mathbf{v}_1\sigma_{a31} + \mathbf{v}_2\sigma_{a32} + \mathbf{v}_3\sigma_{a3} \end{pmatrix}$$

We then proceed similarly for $\mathbf{d}$.

Useful by-products of the inversion-based (as opposed to iterative) computation of BLUP are reliabilities random effect predictions, which are usually obtained from prediction error variances. Prediction error variances of estimates of SNP effects can also be used in genome-wide association studies to assess significance of SNP effects. For the particular case of GLS + SI, individual reliabilities can be obtained from the inverses $\mathbf{V}^{-1}$ and $\left(\mathbf{X}^{'}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$ and from the covariances of $\mathbf{y}$ with the different random effects [4]. Efficient algorithms for REML based on the GLS formulation also exist [10, 11].

Iterative methods for genetic evaluation

In contrast to the exact inversion methods described above, most genetic evaluation software used for livestock use iterative methods that do not invert matrices or even set up MME explicitly, e.g. [12]. Iterative methods have two inconveniences compared to exact inversion methods: (1) convergence may be slow and is a priori unpredictable, and (2) other information such as reliabilities from the inverses of the MME is lost. For the types of multi-population models considered here, convergence of iterative methods is not always good, because there are many more effects than records and, for a given number of records, the condition number of the MME worsens with each extra effect.

We are not aware of iterative methods that use the GLS + SI formulation. However, in principle, it is possible to solve $\left(\mathbf{X}^{'}\mathbf{V}^{-1}\mathbf{X}\right)\widehat{\boldsymbol{\beta}} = \mathbf{X}^{'}\mathbf{V}^{-1}\mathbf{y}$ without inversion or even storage of matrices, by first solving $\mathbf{V}\boldsymbol{\Theta} = \mathbf{y}$ (so $\boldsymbol{\Theta} = \mathbf{V}^{-1}\mathbf{y}$) and $\mathbf{VM} = \mathbf{X}$ (so $\mathbf{M} = \mathbf{V}^{-1}\mathbf{X}$), and then $\left(\mathbf{X}^{'}\mathbf{M}\right)\widehat{\boldsymbol{\beta}} = \mathbf{X}^{'}\boldsymbol{\Theta}$. To estimate SNP effects using SI, it is useful to note that $\mathbf{y}^{c} = \boldsymbol{\Theta} - \mathbf{M}\widehat{\boldsymbol{\beta}}$.

Discussion

Henderson's formulation of MME allowed the use of linear methods for genetic evaluation as opposed to, say, likelihoods in pedigrees [13]. The two key discoveries of the MME and of the fast (and sparse) construction of the inverse of the pedigree-based numerator relationship matrix led to computational efficiency, not only for estimation of breeding values, but also for estimation of variance components, in which most algorithms use predictions of random effects and elements from the inverse of the MME.

However, sparsity of the MME is only partly retained when using dense marker genotypes, as these invariably lead to dense cross-products, either as incidence matrices ($\mathbf{Z}'\mathbf{Z}$) or as covariance matrices ($\mathbf{ZZ}^{'}$). In addition, the latter ($\mathbf{ZZ}^{'}$) need to be inverted for their inclusion in MME. Within the framework of the use of linear models for genetic evaluation, the use of any computing strategy, including GBLUP, SNP-BLUP, and GLS + SI, is now mostly a matter of convenience for the user (availability of software) and for the programmer (general and/or easy formulations are preferred to complex ones). Computationally, the efficiency of the approach depends on the number of records and on the number of markers. In our particular problem of multi-breed prediction, generally, SNP-BLUP is computationally easier when $n > m$ (more records than markers), GLS + SI is easier when $m > n$ (more markers than records), and GBLUP is easier when $m > n$ and the model is notoriously complex to fit (e.g. random regressions on time or temperature, correlated animal effects, etc.).

SNP-BLUP models are interesting because they yield estimates of marker effects. These allow so-called "indirect" predictions (e.g. for young genotyped animals with no own record) to be calculated as the sum of SNP effect estimates weighted by gene content at SNPs. Dominance (or higher order interaction) effect estimates allow matings that maximize performance to be optimized. GBLUP or GLS + SI also allow estimates of marker effects to be obtained using covariances as explained before.

Compared to SNP-BLUP and GBLUP, an advantage of GLS + SI is the ability to fit increasingly complex models without increasing the dimensions of the GLS. Examples are genotype-by-genotype and genotype-by-environment interactions [14, 15], which are of interest, respectively, for planned matings and breeding for target environmental conditions.

The focus of this note was medium-sized populations of up to ~ 100K individuals and/or genotyped markers. In these populations, it is possible to fit rather complex models while maintaining the favorable features of exact methods, including shorter computing time, computed reliabilities, maximum likelihood algorithms, and even genome-wide associations studies. For very large data sets, iteration on data [12, 16–19] are good options, as they do not require cross-products to be explicitly computed (or only partially) or inversion of the MME.

Conclusions

We have shown that for multi-breed prediction, Henderson's MME (either in terms of marker effects or of individual animal effects—SNP-BLUP and GBLUP, respectively) do not necessarily lead to the most computationally efficient approach, although it is a very flexible one. If most individuals are genotyped, then other more parsimonious alternatives could be considered in addition to GBLUP or SNP-BLUP. The use of GLS combined with SI is one of these.

Acknowledgements

ZGV and AL greatly enjoyed visiting and being visited at different times by Rohan Fernando, from whom they learnt much. They also wish to recall that Rohan is definitely a sweet-tooth person, and we all crave for Dulce de Leche and Alfajores. We wish you the best for the future, Rohan.

Authors' contributions

AL conceived the algorithms, who were verified by ZGV, and wrote the R program and a custom Fortran program. DGD raised the computational problem and tested and debugged the software. AL wrote the first version of the manuscript that was completed by DGD and ZGV. All authors read and approved the final manuscript.

Funding

Much of this work was done during the PhD of DGD. DGD was funded by France Génétique Porc, RAGT and INRA (contract 15000416, project 00000837). Work was supported by INRA SELGEN metaprogram (project OptiMaGics). The project was partly supported by the Toulouse Midi-Pyrénées Bioinformatics platform.

Availability of data and materials

A program that implements SNP-BLUP and GLS followed by selection index is available at https://genoweb.toulouse.inra.fr/~alegarra/test.r.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] INRAE, INP, UMR 1388 GenPhySE, 31326 Castanet-Tolosan, France. [2] International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, Mexico.

Received: 22 October 2021 Accepted: 25 January 2022
Published online: 04 February 2022

References

1. Karoui S, Carabaño MJ, Díaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol. 2012;44:39.
2. Wientjes YCJ, Bijma P, Vandenplas J, Calus MPL. Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. Genetics. 2017;207:503–15.
3. Henderson CR. Selection index and expected genetic advance. In: Statistical Genetics and Plant Breeding. Washington: National Research Council Publication; 1963. p. 141–63.
4. Henderson CR. Applications of linear models in animal breeding. Guelph: University of Guelph; 1984.
5. González-Diéguez D, Tusell L, Bouquet A, Legarra A, Vitezica ZG. Purebred and crossbred genomic evaluation and mate allocation strategies to exploit dominance in pig crossbreeding schemes. G3 (Bethesda). 2020;10:2829–41.
6. Xiang T, Christensen OF, Vitezica ZG, Legarra A. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet Sel Evol. 2016;48:92.
7. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. Genetics. 2017;206:1297–307.
8. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23.
9. Strandén I, Garrick DJ. Technical note: derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 2009;92:2971–5.
10. Lee SH, van der Werf JHJ. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. Genet Sel Evol. 2006;38:25–43.
11. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28:2540–2.
12. Legarra A, Misztal I. Technical note: Computing strategies in genome-wide selection. J Dairy Sci. 2008;91:360–6.
13. Fernando R, Stricker C, Elston R. An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. Theor Appl Genet. 1993;87:89–93.
14. Vitezica ZG, Legarra A, Toro MA, Varona L. Orthogonal estimates of variances for additive, dominance and epistatic effects in populations. Genetics. 2017;206:1297–307.
15. Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet. 2014;127:595–607.
16. Strandén I, Lidauer M. Solving large mixed linear models using preconditioned conjugate gradient iteration. J Dairy Sci. 1999;82:2779–87.
17. Tsuruta S, Misztal I, Strandén I. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. J Anim Sci. 2001;79:1166–72.

18. Matilainen K, Mäntysaari EA, Lidauer MH, Strandén I, Thompson R. Employing a Monte Carlo algorithm in newton-type methods for restricted maximum likelihood estimation of genetic parameters. PLoS One. 2013;8:e80821.
19. Reverter A, Golden BL, Bourdon RM, Brinks JS. Method R variance components procedure: application on the simple breeding value model. J Anim Sci. 1994;72:2247–53.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.