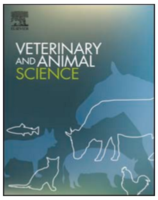
Contents lists available at ScienceDirect

Veterinary and Animal Science

journal homepage: www.elsevier.com/locate/vas



Short Communication: Reduced GBLUP equations to core animals in the algorithm for proven and young (APY)

Mohammad Ali Nilforooshan

Livestock Improvement Corporation, Private Bag 3016, Hamilton, 3204, New Zealand

ARTICLE INFO

Keywords:
Absorption
APY
Back-solve
BLUP
Core
GBLUP

ABSTRACT

The number of animal genotypes is rapidly increasing, and a major challenge for animal models is inverting the genomic relationship matrix (**G**). Matrix **G** has a limited dimensionality, and the algorithm for proven and young (APY) makes inverting a large **G** possible via the inverse of a block diagonal of **G** with a size equivalent to the dimensionality of **G**. APY divides genotyped animals into core and non-core groups, and breeding values of non-core animals are conditioned on the breeding values of core animals. Therefore, there is the possibility of opting out equations for non-core animals from the model. A methodology was presented for a reduced APY genomic BLUP (GBLUP) to equations for core animals. Using a small example dataset, the method was validated by the equality of the full and the reduced model analysis results. Absorption of fixed effect equations into random effect equations was successful in reducing the number of equations to solve and producing the same random effect solutions. Extending the method to APY single-step GBLUP (ssGBLUP) was not computationally justifiable. Other reduction techniques exist for ssGBLUP (regardless of APY or non-APY) that work by reducing the number of equations for non-genotyped animals. The number of equations can further be reduced by data pruning.

1. Introduction

The development of GBLUP (VanRaden, 2008) was a major methodological advancement toward the incorporation of genomic information in large-scale commercial genetic evaluations. Later, the development of single-step GBLUP (ssGBLUP) made possible the joint evaluation of genotyped and non-genotyped animals, and the best use of information from both groups (Aguilar et al., 2010; Christensen and Lund, 2010). The major computational challenge for both GBLUP and ssGBLUP is inverting the genomic relationship matrix (**G**), which has a cubic computational cost relative to the number of genotyped animals (Misztal et al., 2014). As the number of genotyped animals rapidly increases in livestock populations, so will the computational costs.

Misztal et al. (2014) discovered the limited dimensionality of **G** and that a limited number of genotypes can explain most (*e.g.*, 99%) of the variation in **G**. The number of those genotypes (core size) is a function of the effective population size, and the genome length (Pocrnic et al., 2016). Thus, the inverse of a block of **G** corresponding to those genotypes can explain the inverse of the whole **G**. Misztal et al. (2014) developed the algorithm for proven and young (APY), in which genotyped animals are split into core (c) and non-core (n) groups. The block of **G** for core animals ($\mathbf{G}_{cc}$) is directly inverted, and the other blocks of $\mathbf{G}^{-1}$ become a function of $\mathbf{G}_{cc}^{-1}$. This reduces the (cubic) computational cost of inverting **G** (to the total number of genotyped animals) to the (cubic) computational cost of inverting $\mathbf{G}_{cc}$ plus the linear computational cost for non-core animals in $\mathbf{G}_{nn}$ (Misztal et al., 2014). In this algorithm, relationships among non-core animals are explained by the coefficients provided by core animals, and the breeding values of non-core animals are conditioned on the breeding values of core animals and independent from each other. This leads to the diagonal $\mathbf{M}_{nn}^{-1}$ matrix replacing $\mathbf{G}^{nn}$ (the block of $\mathbf{G}^{-1}$ for non-core animals). The APY approximate of $\mathbf{G}^{-1}$ ($\mathbf{G}_{\text{APY}}^{-1}$) is:

$$\mathbf{G}^{-1} \approx \mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix} \mathbf{G}_{\text{APY}}^{cc} & \mathbf{G}_{\text{APY}}^{cn} \\ \mathbf{G}_{\text{APY}}^{nc} & \mathbf{G}_{\text{APY}}^{nn} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} + \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\mathbf{M}_{nn}^{-1}\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}\mathbf{M}_{nn}^{-1} \\ -\mathbf{M}_{nn}^{-1}\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathbf{M}_{nn}^{-1} \end{bmatrix}. \quad (1)$$

The diagonal elements of $\mathbf{M}_{nn}$ are:

$$m_{ii} = g_{ii} - \mathbf{g}_{ic}\mathbf{G}_{cc}^{-1}\mathbf{g}_{ci}, \quad (2)$$

Abbreviations: APY, Algorithm for Proven and Young; GBLUP, Genomic BLUP; ssGBLUP, single-step GBLUP.
E-mail address: mohammad.nilforooshan@lic.co.nz.

https://doi.org/10.1016/j.vas.2024.100334
Received 13 November 2023; Received in revised form 2 January 2024; Accepted 2 January 2024
Available online 4 January 2024

where g_{ii} is the ith diagonal element of $\mathbf{G}_{nn}$, and $\mathbf{g}_{ic}$ is the ith row of $\mathbf{G}_{nc}$.

Though APY was developed in the ssGBLUP context, it targets $\mathbf{G}$, and it also applies to GBLUP (Bermann et al., 2022a; 2022b; Fernando et al., 2016). However, while the problem with inverting a large $\mathbf{G}$ has already been addressed by the APY algorithm, an equation system with the size equal to the total number of genotyped animals (corresponding to the additive genetic effect) needs to be solved. With iterative algorithms applicable to solving sparse equation systems, and access to powerful computational resources, genetic evaluation centres have no problem solving the equations. Nevertheless, that does not mean that computational costs cannot be reduced and become more affordable. Furthermore, many researchers and small organizations do not have access to advanced computational resources. The aim of this study is to introduce a methodology for reducing the number of GBLUP equations to core animals, based on the limited dimensionality of $\mathbf{G}$, to make solving large GBLUP equations feasible on more affordable computers.

2. Theory

In this study, the notations c and n were used for core and non-core animals, 1 and 2 for non-genotyped and genotyped animals, and p and q for parents and non-parents, respectively.

2.1. Reduced BLUP

Quaas and Pollak (1980) developed the reduced animal model, in which the breeding values of non-parents were conditioned on the breeding values of parents. As such, the number of equations for all animals in the pedigree was reduced to parents. The number of parents is considerably less than the number of non-parents in most populations, resulting in a considerable reduction in the number of equations. Consider a simple BLUP:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}, \tag{3}$$

where $\widehat{\mathbf{b}}$, $\widehat{\mathbf{a}}$, and $\mathbf{y}$ are the vectors of solutions for fixed effects, solutions for animals' random genetic effects, and phenotypes; $\mathbf{X}$ and $\mathbf{Z}$ are matrices relating phenotypes to fixed effects and animals, $\mathbf{A}$ is the pedigree-based additive genetic relationship matrix, $\lambda = \sigma^2_e / \sigma^2_a$, σ^2_e is the residual variance, and σ^2_a is the additive genetic variance. Reducing BLUP to the equations for parents:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{A}^{-1}_{pp}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \tag{4}$$

where $\mathbf{y}' = [\mathbf{y}'_p \quad \mathbf{y}'_q]$, $\mathbf{W}' = [\mathbf{Z}'_p \quad \mathbf{Z}^{*'}_q]$ replacing

$$\mathbf{Z}' = \begin{bmatrix} \mathbf{Z}'_p & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}'_q \end{bmatrix},$$

$\mathbf{Z}^*_q$ is a parent incidence matrix divided by 2,

$$\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{qq} \end{bmatrix},$$

$\mathbf{I}_p$ is an identity matrix with the size of $\mathbf{y}_p$, $\mathbf{R}^{qq} = (\mathbf{I}+\mathbf{D}_{qq}/\lambda)^{-1}$, and $\mathbf{D}_{qq}$ is a block of the diagonal matrix $\mathbf{D}$ in $\mathbf{A} = \mathbf{TDT}'$, here corresponding to phenotyped q rather than all q. Nilforooshan (2022) simplified the back-solving equation for the $\widehat{\mathbf{a}}_q$ solutions (Quaas and Pollak, 1980) to:

$$\begin{aligned} \widehat{\mathbf{a}}_q &= \mathbf{B}\big(\mathbf{y}_q - \mathbf{X}_q\widehat{\mathbf{b}} - (\widehat{\mathbf{a}}_s + \widehat{\mathbf{a}}_d)/2\big) + (\widehat{\mathbf{a}}_s + \widehat{\mathbf{a}}_d)/2, \\ \mathbf{B} &= (\mathbf{I}+\mathbf{D}^{qq}\lambda)^{-1} \end{aligned} \tag{5}$$

for phenotyped q, and $\widehat{\mathbf{a}}_q = (\widehat{\mathbf{a}}_s + \widehat{\mathbf{a}}_d)/2$ for non-phenotyped q, where s and d denote the sire and dam of q.

2.2. Reduced APY GBLUP

The reduced APY GBLUP is based on the concept that the breeding values of non-core animals are conditioned on the breeding values of core animals. As such, there would be no need for non-core animals to remain in the analysis, and their solutions can be obtained via a back-solving procedure following solving the mixed model equations, including equations for core animals only. Considering a simple APY GBLUP:

$$\begin{bmatrix} \mathbf{X}'_2\mathbf{X}_2 & \mathbf{X}'_2\mathbf{Z}_2 \\ \mathbf{Z}'_2\mathbf{X}_2 & \mathbf{Z}'_2\mathbf{Z}_2+\mathbf{G}^{-1}_{\mathrm{APY}}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_2\mathbf{y}_2 \\ \mathbf{Z}'_2\mathbf{y}_2 \end{bmatrix}. \tag{6}$$

Reducing the above equation to core animals:

$$\begin{gathered} \begin{bmatrix} \mathbf{X}'_2\mathbf{R}^{22}\mathbf{X}_2 & \mathbf{X}'_2\mathbf{R}^{22}\mathbf{W}_2 \\ \mathbf{W}'_2\mathbf{R}^{22}\mathbf{X}_2 & \mathbf{W}'_2\mathbf{R}^{22}\mathbf{W}_2+(\mathbf{G}_{\mathrm{APY}})^{-1}_{cc}\lambda \end{bmatrix} \\ \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_c \end{bmatrix} = \begin{bmatrix} \mathbf{X}'_2\mathbf{R}^{22}\mathbf{y}_2 \\ \mathbf{W}'_2\mathbf{R}^{22}\mathbf{y}_2 \end{bmatrix}, \end{gathered} \tag{7}$$

where

$$\begin{aligned} (\mathbf{G}_{\mathrm{APY}})^{-1}_{cc} &= \mathbf{G}^{-1}_{cc}, \\ \mathbf{y}_2 &= \begin{bmatrix} \mathbf{y}_c \\ \mathbf{y}_n \end{bmatrix}, \\ \mathbf{W}_2 &= \begin{bmatrix} \mathbf{Z}_c \\ \mathbf{Z}^*_n \end{bmatrix}, \\ \mathbf{R}^{22} &= \begin{bmatrix} \mathbf{I}_c & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{nn} \end{bmatrix}, \end{aligned}$$

$\mathbf{I}_c$ is an identity matrix with the size of $\mathbf{y}_c$, and $\mathbf{R}^{nn} = (\mathbf{I}+\mathbf{D}_{nn}/\lambda)^{-1}$. Matrices $\mathbf{Z}^*_n$ and $\mathbf{D}_{nn}$ are defined in section ”Find $\mathbf{Z}^*_n$ and $\mathbf{D}_{nn}$”. Strandén et al. (2017) showed that:

$$\mathbf{G}_{\mathrm{APY}} = \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{M}_{nn}+\mathbf{G}_{nc}\mathbf{G}^{-1}_{cc}\mathbf{G}_{cn}-\mathbf{G}_{nn} \end{bmatrix}. \tag{8}$$

Thus, $(\mathbf{G}_{\mathrm{APY}})_{cc} = \mathbf{G}_{cc}$. Extending Eq. (5) from pedigree to genomic information, and limiting n to phenotyped n, back-solving for $\widehat{\mathbf{a}}_n$ solutions involves:

$$\begin{aligned} \widehat{\mathbf{a}}_n &= \mathbf{B}\big(\mathbf{y}_n - \mathbf{X}_n\widehat{\mathbf{b}} - \mathbf{G}_{nc}\mathbf{G}^{-1}_{cc}\widehat{\mathbf{a}}_c\big) + \mathbf{G}_{nc}\mathbf{G}^{-1}_{cc}\widehat{\mathbf{a}}_c \\ &= \mathbf{B}(\mathbf{y}_n - \mathbf{X}_n\widehat{\mathbf{b}}) + (\mathbf{I}-\mathbf{B})\mathbf{G}_{nc}\mathbf{G}^{-1}_{cc}\widehat{\mathbf{a}}_c, \\ \mathbf{B} &= (\mathbf{I}+\mathbf{D}^{nn}\lambda)^{-1}. \end{aligned} \tag{9}$$

Limiting n to non-phenotyped n, $\widehat{\mathbf{a}}_n = \mathbf{G}_{nc}\mathbf{G}^{-1}_{cc}\widehat{\mathbf{a}}_c$.

2.2.1. Find $\mathbf{Z}^*_n$ and $\mathbf{D}_{nn}$

In the BLUP context, Quaas and Pollak (1980) described $\mathbf{Z}^*_q$ as a matrix with rows and columns corresponding to non-parents and parents, respectively, in which parentage incidences are coded as 0.5 and the other elements are 0. They also defined $\mathbf{D}_{qq}$ as a block of the diagonal matrix $\mathbf{D}$ (in $\mathbf{A} = \mathbf{TDT}'$, where $\mathbf{T}$ is a lower triangular matrix). Their reduced model made use of the sparsity of $\mathbf{A}^{-1}$, and that the only non-zero off-diagonal elements are those between parents and progeny and between mates (Henderson, 1976). Consequently, they conditioned solutions for non-parents on the solutions of their parents, and reduced the model to parents only. Phenotypes from non-parents contain information and needs to be used, this was achieved via $\mathbf{Z}^*_q$. The phenotype information from non-parents is transferred via their parents through the 0.5 coefficients in $\mathbf{Z}^*_q$. The $\mathbf{A}^{-1}$ structure provides the same pattern of information flow. According to Henderson (1976), $\mathbf{A}^{-1} = (\mathbf{T}^{-1})^{'}\mathbf{D}^{-1}\mathbf{T}^{-1}$, where $\mathbf{T}^{-1} = \mathbf{I} - \mathbf{P}$, and $\mathbf{P}$ is a lower triangle parentage incidence matrix divided by 2. Quaas and Pollak (1980) showed the $\mathbf{A}^{-1}$ structure as (with some notation changes compared to the original article):

$$\begin{aligned}\mathbf{A}^{-1} &= \begin{bmatrix} \mathbf{A}_{pp}^{-1} + \mathbf{P}_{pq}\mathbf{D}^{qq}\mathbf{P}_{pq}^{'} & -\mathbf{P}_{pq}\mathbf{D}^{qq} \\ -\mathbf{D}^{qq}\mathbf{P}_{pq}^{'} & \mathbf{D}^{qq} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_{pp}^{-1} + \mathbf{P}_{pq}\mathbf{D}_{qq}^{-1}\mathbf{P}_{pq}^{'} & -\mathbf{P}_{pq}\mathbf{D}_{qq}^{-1} \\ -\mathbf{D}_{qq}^{-1}\mathbf{P}_{pq}^{'} & \mathbf{D}_{qq}^{-1} \end{bmatrix}.\end{aligned} \tag{10}$$

Matrices $-\mathbf{P}_{pq}^{'}$ and $\mathbf{D}^{qq}$ can be associated with blocks of $\mathbf{T}^{-1}$ and $\mathbf{D}^{-1}$, respectively. Note that $\mathbf{Z}_q^*$ is equivalent to rows of $\mathbf{P}_{pq}^{'}$ corresponding to phenotyped q. Following Eq. (1), $\mathbf{A}_{\mathrm{APY}}^{-1}$ is created for parents and non-parents being the core and non-core groups.

$$\mathbf{A}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{A}_{pp}^{-1} + \mathbf{A}_{pp}^{-1}\mathbf{A}_{pq}\mathbf{M}_{qq}^{-1}\mathbf{A}_{qp}\mathbf{A}_{pp}^{-1} & -\mathbf{A}_{pp}^{-1}\mathbf{A}_{pq}\mathbf{M}_{qq}^{-1} \\ -\mathbf{M}_{qq}^{-1}\mathbf{A}_{qp}\mathbf{A}_{pp}^{-1} & \mathbf{M}_{qq}^{-1} \end{bmatrix}, \tag{11}$$

where $\mathbf{M}_{qq}$ is a diagonal matrix equal to $\mathbf{D}_{qq}$ (see the *proof*), diagonal elements $m_{ii} = \alpha_{ii} - \boldsymbol{\alpha}_{ip}\mathbf{A}_{pp}^{-1}\boldsymbol{\alpha}_{pi}$, α_{ii} is the ith diagonal element of $\mathbf{A}_{nn}$, and $\boldsymbol{\alpha}_{ic}$ is the ith row of $\mathbf{A}_{nc}$.

$$\begin{aligned} \mathit{Proof}:\ \mathbf{M}_{qq}^{-1} &= \mathbf{D}_{qq}^{-1} \\ m_{ii} &= \alpha_{ii} - \boldsymbol{\alpha}_{ip}\mathbf{A}_{pp}^{-1}\boldsymbol{\alpha}_{pi} \\ &\Rightarrow \mathbf{M}_{qq}^{-1} = \mathrm{diag}\left(\left(\mathbf{A}_{qq} - \mathbf{A}_{qp}\mathbf{A}_{pp}^{-1}\mathbf{A}_{pq}\right)^{-1}\right) \\ &= \mathrm{diag}(\mathbf{A}^{qq}) = \mathbf{A}^{qq} = \mathbf{D}^{qq} = \mathbf{D}_{qq}^{-1}. \end{aligned}$$

Comparing Eq. (10) and (11), it can be interpreted that Quaas and Pollak (1980) had been arrived at an APY inverse of $\mathbf{A}$ with animals divided into parents and non-parents. Conditioning non-parents' breeding values on parents' breeding values, they reduced the model to parents and back-solved breeding values of non-parents as a linear function to the breeding values of parents (Eq. (5)). Also, $\mathbf{D}$ and $\mathbf{P}_{pq}^{'}$ in Eq. (10) are $\mathbf{M}$ and $\mathbf{A}_{qp}\mathbf{A}_{pp}^{-1}$ in Eq. (11). Matrix $\mathbf{M}$ in both Eqs. (1) and (11) represents individuals' Mendelian Sampling variances in different ways (observed (Eq. (1)) *vs.* expected (Eq. (11))). Due to the properties of $\mathbf{A}^{-1}$ and that the relationships of progeny are fully conditional on parents in $\mathbf{A}$, the $\mathbf{A}_{\mathrm{APY}}^{-1}$ with parents and non-parents as core and non-core is an exact $\mathbf{A}^{-1}$ (unlike $\mathbf{G}_{\mathrm{APY}}^{-1}$, which is an approximate $\mathbf{G}^{-1}$).

Putting the above evidences together, $\mathbf{Z}_n^*$ and $\mathbf{D}_{nn}$ (Eq. (7)) are equal to the rows of $\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}$ for phenotyped n (*i.e.*, rows of $\mathbf{G}_{nc}$ for phenotyped n multiplied by $\mathbf{G}_{cc}^{-1}$) and the block of $\mathbf{M}_{nn}$ corresponding to phenotyped n, respectively. Both $\mathbf{Z}_n^*$ and $\mathbf{D}_{nn}$ are available from $\mathbf{G}_{\mathrm{APY}}^{-1}$.

2.2.2. *Caveats*

1. Though computationally simple, back-solving is an additional step for the reduced APY GBLUP compared to the full APY GBLUP.
2. Both $\mathbf{W}_2^{'}\mathbf{R}^{22}\mathbf{W}_2 + (\mathbf{G}_{\mathrm{APY}})_{cc}^{-1}\lambda$ (Eq. (7)) and $\mathbf{Z}_c^{'}\mathbf{Z}_c + \mathbf{G}_{\mathrm{APY}}^{cc}$ (the diagonal block of $\mathbf{Z}_2^{'}\mathbf{Z}_2 + \mathbf{G}_{\mathrm{APY}}^{-1}$ (Eq. (6)) corresponding to core animals) are dense. However, because $\mathbf{W}_2^{'}\mathbf{R}^{22}\mathbf{W}_2$ is dense, its summation cost is greater than for $\mathbf{Z}_c^{'}\mathbf{Z}_c$, but $\mathbf{W}_2^{'}\mathbf{R}^{22}\mathbf{W}_2$ is a relatively small matrix.
3. Though, $\mathbf{X}_2^{'}\mathbf{Z}_2$ is not sparse, $\mathbf{X}_2^{'}\mathbf{R}^{22}\mathbf{W}_2$ is dense. However, $\mathbf{X}_2^{'}\mathbf{R}^{22}\mathbf{W}_2$ (Eq. (7)) is smaller than $\mathbf{X}_2^{'}\mathbf{Z}_2 = [\mathbf{X}_c^{'}\mathbf{Z}_c \quad \mathbf{X}_n^{'}\mathbf{Z}_n]$ (Eq. (6)) and has a size equal to the size of $\mathbf{X}_c^{'}\mathbf{Z}_c$.

2.2.3. *Absorption of fixed effect equations*

Usually, fixed effect solutions are of little interest. Therefore, fixed effect equations can be absorbed into random effect equations. If solutions for some fixed effects are of interest, those fixed effects can be exempt from absorption. Garrick et al. (2019) showed that the order of a complete set of mixed model equations can be reduced from the number of fixed effects (f) plus the number of random effects to the number of random effects, by the absorption of fixed effect equations into the random effect equations. For example, such absorption transforms Eq. (3) into:

$$(\mathbf{Z}^{'}\mathbf{Z} + \mathbf{A}^{-1}\lambda - \mathbf{Z}^{'}\mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Z})\widehat{\mathbf{a}}_c = \mathbf{Z}^{'}\mathbf{y} - \mathbf{Z}^{'}\mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y}. \tag{12}$$

3. Materials and methods

An example dataset (dataset Nilforooshan, 2022) was used to illustrate the model setup for the full APY GBLUP, the reduced APY GBLUP, the absorbption of fixed effect equations for both the full APY GBLUP and the reduced APY GBLUP, and testing the hypotheses of the equivalence of solutions with those from the full APY GBLUP. The example dataset contained pedigree and genotypes on eight animals. The R package pedSimulate (Nilforooshan, 2023) was used to simulate genotypes on twenty genetic markers based on the pedigree. Four animals had phenotype data, four animals were considered as core, and one animal was shared between the core and the phenotyped sets. The fixed effects were the overall mean and the sex. The ratio of the residual to the additive genetic variance was considered as 1.5. The code (in R programming language) to reproduce the simulated data, process and analyze the data, and check the results of the experiments are available in a public data repository (dataset Nilforooshan, 2022).

4. Results

The solutions of the full APY GBLUP were identical to those for the reduced APY GBLUP and its following back-solving (for non-core animals), providing evidence on the equivalence of the full and the reduced APY GBLUP. This hold true when equations corresponding to fixed effects were absorbed into those corresponding to random effects (dataset Nilforooshan, 2022).

5. Discussion

Inverting $\mathbf{G}$ is a bottleneck for large-scale genomic animal models. Discovering the limited dimensionality of $\mathbf{G}$ and the development of the APY algorithm made inverting a large $\mathbf{G}$ feasible via a sparse representation of $\mathbf{G}^{-1}$. In the APY algorithm, genotyped animals are divided into core and non-core groups, and the breeding values of non-core animals are conditioned on the breeding values of core animals. This allows for a reduced model, where solving the equations for non-core animals is no longer needed. This study presented a reduced APY GBLUP followed by a back-solving procedure, producing solutions equivalent to a full APY GBLUP, where equations are reduced to those corresponding to core animals. Though this was the first study attempting to reduce APY GBLUP equations to core animals, reducing GBLUP equations has been proposed previously. With an emphasis on solving the problem with a singular $\mathbf{G}$, Fernando et al. (2016) studied APY GBLUP and developed other alternatives. Their strategies (III and IV) involved modifying the genotype matrix rather than the genomic relationship matrix, as the latter is the case for APY. One strategy (III) involved Gaussian elimination and pivoting to transform the genotype matrix to row echelon form. The resulting matrix is upper diagonal and contains independent rows. The other strategy (IV) involved orthonormalization of the rows of the genotype matrix. Aside from the orthogonalization cost of the genotype matrix, in dealing with residual polygenic effects, APY is more convenient as it can be directly applied to the blended $\mathbf{G}$ and $\mathbf{A}_{22}$ (*i.e.*, $k\mathbf{G} + (1-k)\mathbf{A}_{22}$). The methods that modify the genotype matrix cannot accommodate $\mathbf{A}_{22}$, but a block of it corresponding to orthogonal genotypes.

For a GBLUP with no random effects other than the direct additive genetic and residual effects, the size of the mixed model equations may get small enough (*e.g.*, 15K effective dimensionality of **G**) to even make direct solving of the equation system possible on a low-end computer, regardless of the number of genotypes. Iteration-on-data techniques such as Preconditioned Conjugate Gradient, though convenient, are more useful when solving large and sparse equation systems. Direct solving might be beneficial for small equation systems. If the equation system is small enough to make direct solving computationally feasible, the advantage of directly solving equations is obtaining exact breeding values and reliabilities rather than approximates. Reliabilities are a function of the diagonal elements of the inverse of the coefficient matrix of the mixed model equation corresponding to breeding values (prediction error variances, Henderson, 1984). Bermann et al. (2022b) introduced an efficient algorithm for approximating reliabilities, exploiting the sparse structure of $\mathbf{G}_{\text{APY}}^{-1}$ in APY GBLUP and APY ssGBLUP.

By dividing genotyped animals into core and non-core groups, the full APY GBLUP (Eq. (6)) can be written as:

$$\begin{bmatrix} \mathbf{X}_2^{'}\mathbf{X}_2 & \mathbf{X}_c^{'}\mathbf{Z}_c & \mathbf{X}_n^{'}\mathbf{Z}_n \\ \mathbf{Z}_c^{'}\mathbf{X}_c & \mathbf{Z}_c^{'}\mathbf{Z}_c + \mathbf{G}_{\text{APY}}^{cc}\lambda & \mathbf{G}_{\text{APY}}^{cn}\lambda \\ \mathbf{Z}_n^{'}\mathbf{X}_n & \mathbf{G}_{\text{APY}}^{nc}\lambda & \mathbf{Z}_n^{'}\mathbf{Z}_n + \mathbf{G}_{\text{APY}}^{nn}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}}_c \\ \widehat{\mathbf{a}}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_2^{'}\mathbf{y}_2 \\ \mathbf{Z}_c^{'}\mathbf{y}_c \\ \mathbf{Z}_n^{'}\mathbf{y}_n \end{bmatrix},$$

with the $\mathbf{Z}_n^{'}\mathbf{Z}_n + \mathbf{G}_{\text{APY}}^{nn}\lambda$ matrix being sparse, and the number of equations equal to $f + c + n$. The reduced APY GBLUP (Eq. (7)) reduces the number of equations to $f + c$, and further to c by the absorption of fixed effect equations into the random effect equations. Usually, f is hundreds, c is thousands, and n is millions in large livestock populations.

Theoretically, reduced APY GBLUP is extendible to reduced APY ssGBLUP, but it is computationally unjustified (Appendix), because the conditionality in ssGBLUP is not limited to genomic information but also pedigree information. There are two other possibilities for reducing ssGBLUP equations (Nilforooshan, 2022; Nilforooshan and Garrick, 2021), both reduce the number of equations for non-genotyped animals. Currently, the only computationally justifiable way of reducing the number of equations for genotyped animals in ssGBLUP is through data pruning. Such data pruning involves iteratively discarding non-phenotyped non-parents (parents with all progeny discarded change to non-parents) except animals with a non-genotyped parent or a non-genotyped mate.

Aiming to reduce the computational cost of large-scale ssGBLUP evaluations, Tsuruta et al. (2021) discarded genotyped animals with no phenotype and no progeny. The indirect genomic predictions obtained for the discarded genotyped animals, though not identical to the corresponding genomic predictions from the full model, were accurate.

Appendix

Reduced APY ssGBLUP

A simple APY ssGBLUP is written as

$$\begin{bmatrix} \mathbf{X}^{'}\mathbf{X} & \mathbf{X}^{'}\mathbf{Z} \\ \mathbf{Z}^{'}\mathbf{X} & \mathbf{Z}^{'}\mathbf{Z} + \mathbf{H}_{\text{APY}}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{b}} \\ \widehat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{'}\mathbf{y} \\ \mathbf{Z}^{'}\mathbf{y} \end{bmatrix},$$

where

6. Conclusions

The method presented in this study is "equivalent" to the full APY GBLUP in terms of producing the same solutions and "reduced" in terms of the number of equations. It made use of the conditional properties of APY (non-core animals to core animals) to discard equations for non-core animals from APY GBLUP. Then, breeding values of non-core animals were obtained from a simple back-solving procedure. Absorption of fixed effect equations into the random effect equations further reduced the number of equations to only the number of core animals. With that limited number of equations, any computer able to invert $\mathbf{G}_{cc}$ is capable of direct solving reduced APY GBLUP equations. The advantages of direct solving the equations are obtaining exact breeding values and reliabilities rather than approximates.

Data statement

The dataset used in this study were made publicly available. The data and the code can be found at the repository: https://doi.org/10.6084/m9.figshare.20539650.v3.

Ethical statement

Hereby, I confirm that,

1. the study does not involve any animal or human subject.

2. The research is original. It has not been published, nor is under consideration by any other journal.

3. There has been no use of AI at any stage of the study and writing of the manuscript.

CRediT authorship contribution statement

Mohammad Ali Nilforooshan: Writing – review & editing, Writing – original draft, Software, Investigation, Formal analysis, Data curation.

Declaration of competing interest

The author declares the following financial interests/personal relationships which may be considered as potential competing interests:

Mohammad Ali Nilforooshan reports financial support was provided by New Zealand Ministry for Primary Industries. Mohammad Ali Nilforooshan reports a relationship with Livestock Improvement Corporation that includes: employment.

Acknowledgment

This work was financially supported by the NZ Ministry for Primary Industries, SFF Futures Programme: Resilient Dairy–Innovative breeding for a sustainable dairy future [grant number PGP06-17006].

$$\mathbf{H}_{\mathrm{APY}}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathrm{APY}}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Dividing genotyped animals into core and non-core groups:

$$\mathbf{H}_{\mathrm{APY}}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{\mathrm{APY}}^{cc} - \mathbf{A}_{22}^{cc} & \mathbf{G}_{\mathrm{APY}}^{cn} - \mathbf{A}_{22}^{cn} \\ \mathbf{0} & \mathbf{G}_{\mathrm{APY}}^{nc} - \mathbf{A}_{22}^{nc} & \mathbf{M}_{nn}^{-1} - \mathbf{A}_{22}^{nn} \end{bmatrix}.$$

Matrix $\mathbf{M}_{nn}^{-1} - \mathbf{A}_{22}^{nn}$ is not diagonal. Due to the pedigree information conditionalities, the above $\mathbf{H}_{\mathrm{APY}}^{-1}$ cannot be reduced. Alternatively, APY can be applied to $\mathbf{H}$ rather than $\mathbf{G}$. *i.e.*,

$$\mathbf{H}_{\mathrm{APY}}^{-1} = \begin{bmatrix} \mathbf{H}^{11} & \mathbf{H}^{1c} & \mathbf{H}^{1n} \\ \mathbf{H}^{cn} & \mathbf{H}^{cc} & \mathbf{H}^{cn} \\ \mathbf{H}^{n1} & \mathbf{H}^{nc} & \mathbf{M}_{nn}^{-1} \end{bmatrix}.$$

Grouping 1 and c into d, $m_{ii} = h_{ii} - \mathbf{h}_{id}\mathbf{H}_{dd}^{-1}\mathbf{h}_{di}$, where $h_{ii} = g_{ii}$, $\mathbf{h}_{id}$ is the ith row of $\mathbf{H}_{nd} = [\mathbf{H}_{n1} \quad \mathbf{H}_{nc}]$, $\mathbf{H}_{nc} = \mathbf{G}_{nc}$, and

$$\mathbf{H}_{dd}^{-1} = \mathbf{A}_{dd}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{cc}^{-1} - \mathbf{A}_{cc}^{-1} \end{bmatrix}.$$

However, calculating $\mathbf{H}_{n1}$ involves $\mathbf{GA}_{22}^{-1}\mathbf{A}_{21}$ and is not computationally justifiable.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science, 93*(2), 743–752. https://doi.org/10.3168/jds.2009-2730

Bermann, M., Lourenco, D., Forneris, N. S., Legarra, A., & Misztal, I. (2022). On the equivalence between marker effect models and breeding value models and direct genomic values with the Algorithm for Proven and Young. *Genetics Selection Evolution, 54*(1), 52. https://doi.org/10.1186/s12711-022-00741-7

Bermann, M., Lourenco, D., & Misztal, I. (2022). Efficient approximation of reliabilities for single-step genomic best linear unbiased predictor models with the Algorithm for Proven and Young. *Journal of Animal Science, 100*(1), skab353. https://doi.org/10.1093/jas/skab353

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution, 42*(1), 2. https://doi.org/10.1186/1297-9686-42-2

dataset Nilforooshan, M. A. (2022). Code & Data – Genomic evaluations reduced to equations for core animals in the algorithm for proven and young (APY). *Journal Contribution*. https://doi.org/10.6084/m9.figshare.20539650.v3

Fernando, R. L., Cheng, H., & Garrick, D. J. (2016). An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. *Genetics Selection Evolution, 48*(1), 80. https://doi.org/10.1186/s12711-016-0260-7

Garrick, D. J., Golden, B. L., & Garrick, D. P. (2019). Alternative implementations of preconditioned conjugate gradient algorithms for solving mixed model equations. *Proceedings of the AAABG 23rd conference* (pp. 250–253). Armidale, New South Wales, Australia.Accessed 25 December 2023, http://www.aaabg.org/aaabghome/AAABG23papers/61Garrick23250.pdf

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics, 32*(1), 69–83. https://doi.org/10.2307/2529339

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. Guelph: University of Guelph.

Misztal, I., Legarra, A., & Aguilar, I. (2014). Using recursion to compute the inverse of the genomic relationship matrix. *Journal of Dairy Science, 97*(6), 3943–3952. https://doi.org/10.3168/jds.2013-7752

Nilforooshan, M. A. (2022). pedSimulate – An R package for simulating pedigree, genetic merit, phenotype, and genotype data. *Revista Brasileira de Zootecnia, 51*, e20210131. https://doi.org/10.37496/rbz5120210131

Nilforooshan, M. A. (2023). Technical note: Extension of the reduced animal model to single-step methods. *Journal of Animal Science, 101*, skac272. https://doi.org/10.1093/jas/skac272

Nilforooshan, M. A., & Garrick, D. (2021). Reduced animal models fitting only equations for phenotyped animals. *Frontiers in Genetics, 12*, 637626. https://doi.org/10.3389/fgene.2021.637626

Pocrnic, I., Lourenco, D. A. L., Masuda, Y., Legarra, A., & Misztal, I. (2016). The dimensionality of genomic information and its effect on genomic prediction. *Genetics, 203*(1), 573–581. https://doi.org/10.1534/genetics.116.187013

Quaas, R. L., & Pollak, E. J. (1980). Mixed model methodology for farm and ranch beef cattle testing programs. *Journal of Animal Science, 51*, 1277–1287. https://doi.org/10.2527/jas1981.5161277x

Strandén, I., Matilainen, K., Aamand, G. P., & Mäntysaari, E. A. (2017). Solving efficiently large single-step genomic best linear unbiased prediction models. *Journal of Animal Breeding and Genetics, 134*(3), 264–274. https://doi.org/10.1111/jbg.12257

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Lawlor, T. J., & Misztal, I. (2021). Reducing computational cost of large-scale genomic evaluation by using indirect genomic prediction. *JDS Communications, 2*(6), 356–360. https://doi.org/10.3168/jdsc.2021-0097

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science, 91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980