The algorithm for proven and young (APY) from a different perspective

Mohammad Ali Nilforooshan
Livestock Improvement Corporation, Private Bag 3016, Hamilton 3240, New Zealand
mohammad.nilforooshan@lic.co.nz

Abstract

The inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$) is used in genomic BLUP (GBLUP) and the single-step GBLUP. The rapidly growing number of genotypes is a constraint for inverting $\mathbf{G}$. The APY algorithm efficiently resolves this issue. Matrix $\mathbf{G}$ has a limited dimensionality. Dividing individuals into core and non-core, $\mathbf{G}^{-1}$ is approximated via the inverse partition of $\mathbf{G}$ for core individuals. The quality of the approximation depends on the core size and composition. The APY algorithm conditions genomic breeding values of the non-core individuals to those of the core individuals, leading to a diagonal block of $\mathbf{G}^{-1}$ for non-core individuals ($\mathbf{M}_{nn}^{-1}$). Dividing observations into two groups (*e.g.*, core and non-core, genotyped and non-genotyped, *etc*), any symmetric matrix can be expressed in APY and APY-inverse expressions, equal to the matrix itself and its inverse, respectively. The change of $\mathbf{G}^{nn}$ to $\mathbf{M}_{nn}^{-1} = diag(\mathbf{G}^{nn})$ makes APY an approximate. This change is projected to the other blocks of $\mathbf{G}^{-1}$ as well. The application of APY is extendable to the inversion of any large symmetric matrix with a limited dimensionality at a lower computational cost. Furthermore, APY may improve the numerical condition of the matrix or the equation system.
Keywords: APY, conditional, diagonal, dimensionality, GBLUP, inverse, relationship matrix, single-step

1 Introduction

Genomic evaluations are mainly performed using the genomic relationship matrix $\mathbf{G}$ in the so-called method genomic BLUP (GBLUP, VanRaden (2008)) or random regression SNP marker models called SNP-BLUP (Koivula et al., 2012). The first predicts genomic breeding values of genotyped individuals, and the latter predicts marker effects (*i.e.*, allele substitution effects). Simultaneous genetic evaluation of genotyped and non-genotyped individuals for obtaining optimal and unbiased evaluations not limited to genotyped individuals, both methods were elevated to single-step GBLUP (ssGBLUP, Aguilar et al. (2010); Christensen and Lund (2010)), and single-step SNP-BLUP (ss-SNP-BLUP, Fernando et al. (2014)), also called the single-step marker effect model.

The number of genotyped individuals is rapidly growing, and the most expensive operation in GBLUP and ssGBLUP is inverting matrix $\mathbf{G}$. As the number of genotyped individuals reaches the number of markers, the numerical condition of $\mathbf{G}$ deteriorates. By the number of genotypes exceeding the number of markers, $\mathbf{G}$ becomes singular and non-invertible. Furthermore, the cost of inverting $\mathbf{G}$ and $\mathbf{A}_{22}$ (the block of $\mathbf{A}$ corresponding to genotyped individuals, where $\mathbf{A}$ is the pedigree-based additive genetic relationship matrix) required for ssGBLUP is cubic, and there is a bottleneck of direct inversion of a matrix of size about 150,000 (Fragomeni et al., 2015). Three solutions were proposed for this problem (Misztal et al., 2014; Fernando et al., 2016; Mäntysaari et al., 2017), one being the algorithm for proven and young (APY, Misztal et al. (2014)). This algorithm belongs to a group of methods called approximate kernel methods or Gaussian process approximations (Snelson and Ghahramani, 2007). APY forms a sparse representation of $\mathbf{G}^{-1}$ ($\mathbf{G}_{\text{APY}}^{-1}$), dividing genotyped individuals to core (c) and non-core (n) subsets. Direct inversion is only required for the block of $\mathbf{G}$ corresponding to core individuals ($\mathbf{G}_{cc}$). Consequently, the $O\left((c+n)^3\right)$ computational cost is reduced to $O(c^3) + O(n)$. In the APY algorithm, genomic breeding values of non-core individuals are conditioned on the genomic breeding values of core individuals. This algorithm is based on the assumption that the dimensionality of $\mathbf{G}$ is limited and that independent chromosome segments explain the rank of $\mathbf{G}$ (Misztal, 2016). As long as the number of core individuals is greater than the number of independent chromosome segments (Misztal et al., 2014), and the core subset covers the $\mathbf{G}$ spectrum (Bermann et al., 2022), it may not take all the genotyped individuals to explain the variation

in $\mathbf{G}$. Therefore, the variation in $\mathbf{G}$ can be explained by the core subset, and genomic breeding values of the non-core individuals are expressed as a linear function of those from the core individuals (Bermann et al., 2022). As such, the accuracy of the APY algorithm depends on the core size and composition.

The $\mathbf{G}_{\text{APY}}^{-1}$ matrix is calculated as (Bermann et al., 2022):

$$\mathbf{G}_{\text{APY}}^{-1} = \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{P}_{cn} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}' = \begin{bmatrix} \mathbf{G}_{cc}^{-1} + \mathbf{P}_{cn}\mathbf{M}_{nn}^{-1}\mathbf{P}_{nc} & -\mathbf{P}_{cn}\mathbf{M}_{nn}^{-1} \\ -\mathbf{M}_{nn}^{-1}\mathbf{P}_{nc} & \mathbf{M}_{nn}^{-1} \end{bmatrix}, \tag{1}$$

where, $\mathbf{P}_{cn} = \mathbf{G}_{cc}^{-1}\mathbf{G}_{cn}$, $\mathbf{M}_{nn}$ is a diagonal matrix with diagonal elements:

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

g_{ii} is a diagonal element of $\mathbf{G}_{nn}$, and $\mathbf{g}_{ic}$ is a row vector of $\mathbf{G}_{nc}$. Strandén et al. Strandén et al. (2017) and Bermann et al. Bermann et al. (2022) showed that:

$$\begin{aligned} \mathbf{G}_{\text{APY}} &= \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{P}_{nc} & \mathbf{I} \end{bmatrix}' \\ &= \begin{bmatrix} \mathbf{G}_{cc} & \mathbf{G}_{cn} \\ \mathbf{G}_{nc} & \mathbf{M}_{nn} + \mathbf{P}_{nc}\mathbf{G}_{cn} \end{bmatrix} \\ &= \mathbf{G} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_{nn} + \mathbf{P}_{nc}\mathbf{G}_{cn} - \mathbf{G}_{nn} \end{bmatrix}. \end{aligned} \tag{3}$$

The aim of this study is to provide new insights and understanding about the APY algorithm.

2 Theory and Discussion

2.1 The APY and APY-inverse expressions

In this subsection, it is shown that any covariance or inverse covariance (generally any symmetric) matrix has expressions, here called APY and APY-inverse expressions. To understand the properties of the APY and APY-inverse expressions, we get help from the hybrid pedigree-genomic relationship matrix ($\mathbf{H}$) used in ssGBLUP. Legarra et al. (Legarra et al., 2009) derived various forms of the same relationship matrix, including the full pedigree and genomic information. Denoting genotyped and non-genotyped individuals as 2 and 1: $\mathbf{H} =$

$$\begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix}, \tag{4}$$

$$\begin{bmatrix} (\mathbf{A}^{11})^{-1} + (\mathbf{A}^{11})^{-1}\mathbf{A}^{12}\mathbf{G}\mathbf{A}^{21}(\mathbf{A}^{11})^{-1} & -(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}\mathbf{G} \\ -\mathbf{G}\mathbf{A}^{21}(\mathbf{A}^{11})^{-1} & \mathbf{G} \end{bmatrix}, \tag{5}$$

$$\mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22}) \\ (\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}. \tag{6}$$

It worth mentioning that replacing $\mathbf{G}$ with $\mathbf{A}_{22}$ in any of these equations turns $\mathbf{H}$ to $\mathbf{A}$. Similarly, replacing $\mathbf{G}$ with $\mathbf{G}_{nn}$ and $\mathbf{A}$ with $\mathbf{G}$ turns $\mathbf{H}$ to $\mathbf{G}$. The above equations can be simplified to:

$$\mathbf{H} = \begin{bmatrix} (\mathbf{A}^{11})^{-1} + \mathbf{P}_{12}\mathbf{G}\mathbf{P}_{21} & -\mathbf{P}_{12}\mathbf{G} \\ -\mathbf{G}\mathbf{P}_{21} & \mathbf{G} \end{bmatrix} = \begin{bmatrix} (\mathbf{A}^{11})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{P}_{12} \\ \mathbf{0} \end{bmatrix} \mathbf{G} \begin{bmatrix} -\mathbf{P}_{12} \\ \mathbf{0} \end{bmatrix}', \tag{7}$$

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} -\mathbf{P}_{12} \\ \mathbf{0} \end{bmatrix} (\mathbf{G} - \mathbf{A}_{22}) \begin{bmatrix} -\mathbf{P}_{12} \\ \mathbf{0} \end{bmatrix}', \tag{8}$$

where, the projection matrix $\mathbf{P}_{12} = (\mathbf{A}^{11})^{-1}\mathbf{A}^{12} = -\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$. A nice property of $\mathbf{H}$ is that its inverse can be derived directly with no need to form and invert $\mathbf{H}$ (Aguilar et al., 2010; Christensen and Lund, 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22}^{-1} \end{bmatrix}. \tag{9}$$

Matrix $\mathbf{H}^{-1}$ replaces $\mathbf{A}^{-1}$ in BLUP for ssGBLUP. Replacing $\mathbf{G}$ with $\mathbf{M}_{nn}^{-1}$, $\mathbf{A}^{-1}$ with $\mathbf{G}$, and notations 1 and 2 with c and n, respectively, turns Eq. 7 to Eq. 1. This shows that Eq. 7 is the APY-inverse expression of $\mathbf{H}^{-1}$. Following Eq. 3, the APY expression of $\mathbf{H}^{-1}$ equals:

$$\mathbf{H}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{22} + \mathbf{P}^{21}\mathbf{A}^{21} - \mathbf{H}^{22} \end{bmatrix}, \tag{10}$$

where $\mathbf{M}^{22} = \mathbf{H}^{22} - \mathbf{P}^{21}\mathbf{A}^{12}$, $\mathbf{P}^{12} = (\mathbf{A}^{11})^{-1}\mathbf{A}^{12}$, and $\mathbf{H}^{22} = \mathbf{A}^{22} + \mathbf{G} - \mathbf{A}_{22}^{-1}$. Similarly, there are APY and APY-inverse expressions for $\mathbf{H}$. The APY and APY-inverse expressions of $\mathbf{G}$ are obtained by replacing $\mathbf{M}_{nn}^{-1}$ with $\mathbf{G}^{nn}$ in Eq. 3 and Eq. 1, respectively. Algebraically, it is possible to prove that the APY and APY-inverse expressions of $\mathbf{G}$ are equal to $\mathbf{G}$ and $\mathbf{G}^{-1}$, respectively.

2.2 Understanding the differences between $\mathbf{G}^{-1}$ and $\mathbf{G}_{\mathbf{APY}}^{-1}$

As shown in the previous subsection, replacing $\mathbf{G}^{nn}$ with $\mathbf{M}_{nn}^{-1}$ turns $\mathbf{G}^{-1}$ to $\mathbf{G}_{\text{APY}}^{-1}$. Note that, though, $\mathbf{G}_{\text{APY}}^{-1}$ is an approximate $\mathbf{G}^{-1}$, the APY-inverse expression of $\mathbf{G}$ and the APY expression of $\mathbf{G}^{-1}$ are the exact $\mathbf{G}^{-1}$. In this subsection, the differences between $\mathbf{G}^{nn}$ and $\mathbf{M}_{nn}^{-1}$, and between $\mathbf{G}^{-1}$ and $\mathbf{G}_{\text{APY}}^{-1}$ are explained.

Given Eq. 2, let's reconstruct $\mathbf{M}_{nn}$ as if $\mathbf{M}_{nn}^{-1}$ had non-zero off-diagonal elements to calculate.

$$\begin{aligned} \mathbf{M}_{nn} &= \mathbf{G}_{nn} - \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} = (\mathbf{G}^{nn})^{-1} \\ \Rightarrow \mathbf{M}_{nn}^{-1} &= \mathbf{G}^{nn} \\ \Rightarrow diag(\mathbf{M}_{nn}^{-1}) &= diag(\mathbf{G}^{nn}). \end{aligned}$$

Thus, as long as the off-diagonal elements of $\mathbf{M}_{nn}^{-1}$ are not set to zero, $\mathbf{M}_{nn}^{-1} = \mathbf{G}^{nn}$. Eq. 2 is a computationally optimised way of calculating the reciprocals of the diagonal elements of $\mathbf{G}^{nn}$ (*i.e.*, avoiding the calculations for off-diagonal elements, and the parallel computing possibility). Ignoring the off-diagonal elements of $\mathbf{G}^{nn}$ requires other block of $\mathbf{G}^{-1}$ compensate the missing information (making genomic relationships among non-core animals conditional to core animals), and this change should be projected to the other blocks of $\mathbf{G}^{-1}$ via the projection matrix $\mathbf{P}_{cn}$ (Eq. 1). Consequently, $\mathbf{G}^{nn}$ in the other blocks of $\mathbf{G}^{-1}$ are changed to $\mathbf{M}_{nn}^{-1} = diag(\mathbf{G}^{nn})$ to turn $\mathbf{G}^{-1}$ to $\mathbf{G}_{\text{APY}}^{-1}$ (Eq. 1).

In comparison with $\mathbf{G}$, no change is made to $\mathbf{G}_{\text{APY}}$ other than to $\mathbf{G}_{nn}$ (Eq. 3). It can be articulated that the efficiency of the APY algorithm depends on how well $\mathbf{M}_{nn} + \mathbf{P}_{nc}\mathbf{G}_{cn}$ represents $\mathbf{G}_{nn}$.

2.3 Other applications

The application of the APY algorithm is not limited to $\mathbf{G}^{-1}$, nor to GBLUP and ssGBLUP. This algorithm can be applied to approximate the inverse of any large symmetric matrix, where the rank of the matrix is smaller than its dimension. Representing any such matrix with $\mathbf{G}$, only $\mathbf{G}_{cc}$ needs to be inverted. Besides reduced matrix inversion cost, there are sparsity-related reduced computational costs.

The first and the only time the APY algorithm was suggested for inverting a matrix other than $\mathbf{G}$ was by Misztal et al. Misztal et al. (2014). They suggested the APY algorithm for the $\mathbf{A}_{22}$ inversion, which is required in ssGBLUP (Eq. 9). They derived an equivalent formula for the APY approximation of $\mathbf{A}_{22}^{-1}$:

$$\mathbf{A}_{22}^{-1} \approx \left(\mathbf{A}_{22}^{-1}\right)_{\text{APY}} = \begin{bmatrix} \mathbf{A}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{A}_{cc}^{-1}\mathbf{A}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{A}_{cc}^{-1}\mathbf{A}_{cn} \\ \mathbf{I} \end{bmatrix}'. \tag{11}$$

Here, the diagonal elements of $\mathbf{M}_{nn}$ equal $m_{ii} = a_{ii} - \mathbf{a}_{ic}\mathbf{A}_{cc}^{-1}\mathbf{a}_{ci}$, where i is a non-core genotyped individual. However, computing $\mathbf{A}_{22}^{-1}$ via the APY algorithm is a problem in a loop, which means to obtain the inverse of a block of $\mathbf{A}$ (*i.e.*, $\mathbf{A}_{22}^{-1}$), the inverse of its sub-block ($\mathbf{A}_{cc}^{-1}$) is required. There are two other well-established methods for the calculation of $\mathbf{A}_{22}^{-1}$ (Colleau, 2002; Faux and Gengler, 2013).

Contrarily, one may apply the APY algorithm for inverting $\mathbf{A}^{-1}$ to $\mathbf{A}$. Though calculating $\mathbf{A}$ is computationally expensive, calculation of $\mathbf{A}^{-1}$ is computationally fast and efficient (Henderson, 1975), and it is sparse. The computational cost of inverting $\mathbf{A}^{-1}$ to $\mathbf{A}$ can be reduced by obtaining an APY inversion of $\mathbf{A}^{-1}$:

$$\mathbf{A}_{\text{APY}} = \begin{bmatrix} (\mathbf{A}^{cc})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -(\mathbf{A}^{cc})^{-1}\mathbf{A}^{cn} \\ \mathbf{I} \end{bmatrix} (\mathbf{M}^{nn})^{-1} \begin{bmatrix} -(\mathbf{A}^{cc})^{-1}\mathbf{A}^{cn} \\ \mathbf{I} \end{bmatrix}', \tag{12}$$

where $\mathbf{M}^{nn}$ is a diagonal matrix with diagonal elements $m^{ii} = a^{ii} - \mathbf{a}^{ic}(\mathbf{A}^{cc})^{-1}\mathbf{a}^{ic}$. If rather than $\mathbf{A}$, a diagonal block of it ($\mathbf{A}_{cc}$) is needed, some of the calculations in Eq. 12 become redundant, and $(\mathbf{A}_{cc})_{\text{APY}}$ can be calculated as:

$$\begin{aligned}(\mathbf{A}_{cc})_{\text{APY}} &= (\mathbf{A}^{cc})^{-1} + (\mathbf{A}^{cc})^{-1}\mathbf{A}^{cn}(\mathbf{M}^{nn})^{-1}\mathbf{A}^{nc}(\mathbf{A}^{cc})^{-1} \\ &= (\mathbf{A}^{cc})^{-1}(\mathbf{A}^{cc} + \mathbf{A}^{cn}(\mathbf{M}^{nn})^{-1}\mathbf{A}^{nc})(\mathbf{A}^{cc})^{-1}. \end{aligned} \tag{13}$$

The above formula can be used to calculate an approximate $\mathbf{A}_{22}$ to blend with $\mathbf{G}$ (*i.e.*, $k\mathbf{G} + (1 - k)\mathbf{A}_{22}$; $0 < k < 1$) for improving the numerical condition of $\mathbf{G}$, and to introduce residual polygenic variance not captured by the markers. However, the APY approximations of $\mathbf{A}$ and $\mathbf{A}_{cc}$ are expectedly inaccurate. For accurate approximations, non-core animals need to have their parents, progeny, and mates in the core group. The non-zero off-diagonal elements of $\mathbf{A}^{-1}$ are among parent-progeny and mates (Henderson, 1975). That considerably limits the choice and the number of animals in the non-core group. On the other hand, for $\mathbf{G}^{-1}_{\text{APY}}$, all animals in $\mathbf{G}_{cc}$ contribute information to each animal in $\mathbf{G}_{nn}$.

The APY algorithm helped overcome the limitations of inverting $\mathbf{G}$. On the contrary, this constraint does not exist for marker effect models (*i.e.*, SNP-BLUP and ss-SNP-BLUP) because a marker × marker matrix is used instead of $\mathbf{G}^{-1}$, which does not need to be inverted. This advantage comes at the price of dense matrix multiplications, and convergence complexities (Vandenplas et al., 2018; Bermann et al., 2022).

Unlike $\mathbf{G}$, the size of that matrix remains constant over time unless the genotyping platform changes, and the old genotypes are imputed to a genotyping platform with a higher marker density. In fact, GBLUP and SNP-BLUP are equivalent models (Bermann et al., 2022). Conversion formulas between these two models are presented in the Appendix.

The mixed model equations (MME) of the marker effect models do not require direct matrix inversion (Fernando et al., 2014). Indirect inversion of $\mathbf{A}$ is needed, which is easy to obtain (Henderson, 1975). However, due to convergence difficulties, a specialised preconditioned conjugate gradient (PCG) solver with a block-diagonal Jacobi preconditioner matrix is applied, which is extended from single-trait to multi-trait analyses (Harris et al., 2022). As such, a marker × marker diagonal block of the MME (here called $\mathbf{Q}$) is inverted, which is expanded by the number of traits in the model. The APY algorithm is a good candidate for this scenario, where the markers are divided into core and non-core. Only the block corresponding to core markers ($\mathbf{Q}_{cc}$) is inverted. Similar rules applied to $\mathbf{G}^{-1}_{\text{APY}}$ are applied to this scenario, with the difference that the role of markers and genotyped individuals are switched. Due to collinearity in the marker × individual genotype matrix, this matrix is not of full rank. The main source of collinearity is the markers with low minor allele frequency. Also, it would probably not take all the genotyped individuals to explain marker effects. Therefore, $\mathbf{Q}$ has a limited dimensionality, and the off-diagonal elements of $\mathbf{Q}_{nn}$ (in the preconditioner matrix, not in the MME) are conditioned on $\mathbf{Q}_{cc}$ and $\mathbf{Q}_{cn}$. The $\mathbf{Q}^{-1}_{\text{APY}}$ is a preconditioner matrix with the preconditioning properties similar to those of $\mathbf{Q}^{-1}$. Though the number of PCG iterations might differ, the cost of storing $\mathbf{Q}^{-1}_{\text{APY}}$ in the memory is cheaper, and each PCG iteration is expected to be faster due to the sparsity of $\mathbf{Q}^{-1}_{\text{APY}}$.

The core size and composition define the APY accuracy. Core size, which its optimum is a function of the effective population size (Pocrnic et al., 2016), is the most important. As long as there is room to increase the core size to span over 98% of the eigenvalue spectra of $\mathbf{G}$, a random set of core individuals is shown to perform well because it gives good coverage over generations and breeds in the population. With a sufficiently large core size, the gain from an optimal core subset is marginal (Nilforooshan and Lee, 2019). There is ongoing research on finding the optimal core subset, and it is an important topic for admix populations and when the core size is constrained. When the core size is limited, an optimum core composition can harvest a larger variation of $\mathbf{G}$, and provide more accurate estimated of breeding values (Abdollahi-Arpanahi et al., 2022; Pocrnic et al., 2022).

It is unknown what proportion of random markers would cover over 98% of the eigenvalue spectra of $\mathbf{Q}$. Similar to the concept of effective population size defining the optimum number of core individuals for $\mathbf{G}^{-1}_{\text{APY}}$ might be the concept of effective marker size defining the optimum number of core markers for $\mathbf{Q}^{-1}_{\text{APY}}$. Such markers are likely segregating in the coding regions, with effects as independent and orthogonal as possible to other markers; a concept similar to independent chromosome segments equal to $2N_eL/log(4N_eL)$ (Goddard, 2009), where N_e is the effective population size, and L is the length of chromosome in Morgans. Therefore, $\mathbf{G}$ and $\mathbf{Q}$ might have similar dimensionality, and the required core size might be the same for both. Possibly, choosing markers corresponding to the highest diagonal elements of $\mathbf{Q}$ is better than a random set of core markers. This is because those markers cover a larger variation in $\mathbf{Q}$ (*i.e.*, $trace(\mathbf{Q}) = \sum eigenvalue(\mathbf{Q})$). This would favour choosing markers with lower

minor allele frequency. An optimised core subset may reduce the need for a larger core size (*i.e.*, the same variation in $\mathbf{Q}$ captured by a smaller set of markers). Future research is needed on this topic.

The APY accuracy is usually measured by the correlation between genomic breeding values obtained via $\mathbf{G}^{-1}$ and $\mathbf{G}^{-1}_{\text{APY}}$. However, it might be okay to have a correlation coefficient slightly less than 1. A small variation of $\mathbf{G}$ might be due to collinearity and noise-related and good to get discharged. The APY algorithm may help reduce the collinearity and noise in $\mathbf{G}$. Nilforooshan and Lee Nilforooshan and Lee (2019) showed that APY reduced the very large $\max(\text{diag}(\mathbf{G}^{-1}))$, which is a sign of reduced collinearity and improved condition of $\mathbf{G}$. Validation of genomic breeding values is a good complementary.

Another benefit of APY is reducing the computational cost of blending $\mathbf{G}$ with $\mathbf{A}_{22}$ (Hollifield et al., 2022). APY does not require storing $\mathbf{G}_{nn}$, but $diag(\mathbf{G}_{nn})$. Therefore, rather than calculating the whole $\mathbf{A}_{nn}$, calculation of $diag(\mathbf{A}_{nn}) = 1 + \mathbf{F}_n$ would be enough, where $\mathbf{F}_n$ is the vector of inbreeding coefficients for n. The inbreeding coefficients, however, need to be calculated before calculating any row of $\mathbf{A}_{22}$ (Colleau, 2002), including the rows for c. The average of the diagonal and off-diagonal elements of $\mathbf{A}_{22}$ might be needed for tuning $\mathbf{G}$. However, it can be assumed that $\mu(offdiag(\mathbf{A}_{nn})) = 0$ (*i.e.*, $\mu(offdiag(\mathbf{A}_{cc})) + \mu(\mathbf{A}_{cn}) = \mu(offdiag(\mathbf{A}_{22}))$).

3 Conclusions

This study aimed to open new insights and understanding about the APY algorithm. It was shown that any symmetric matrix can be expressed as a combination of its two diagonal blocks (for core and non-core animals, genotyped and non-genotyped animals, *etc*). The projection matrix makes the combination (information flow) between the two diagonal blocks. As such, any symmetric matrix has (the so-called) APY and APY-inverse expressions equal to the matrix itself and its inverse, respectively. The difference arises when the off-diagonal elements of a diagonal block of the APY-inverse expression (corresponding to the non-core set) are set to 0. This change is projected to the rest of the inverse matrix via the projection matrix. That diagonal matrix sets non-core individuals independent from each other conditional to the coefficients provided by the core individuals. The APY algorithm can also be understood as an approximate absorption of the off-diagonal elements of a diagonal block into the rest of the matrix.

The APY algorithm is based on the concept of the limited matrix dimensionality. A genomic relationship matrix has a limited dimensionality equivalent to the number of independent chromosome segments, which allows a reduction in the dimensionality of $\mathbf{G}$. Therefore, it would take the inverse of a diagonal block of $\mathbf{G}$ to invert $\mathbf{G}$. An APY-inverse of $\mathbf{G}$ with a sufficient core size and proper core composition produces genomic breeding values analogous to those using the exact $\mathbf{G}^{-1}$. A possible new application for APY is computing the block corresponding to marker effects, of the block-diagonal preconditioner matrix for iterative solving of marker effect model equations. The application of APY is not limited to obtaining the best sparse approximates of $\mathbf{G}^{-1}$, and new applications may emerge in the future.

References

Abdollahi-Arpanahi, R., Lourenco, D., and Misztal, I. 2022. A comprehensive study on size and definition of the core group in the proven and young algorithm for single-step GBLUP. *Genet. Sel. Evol.*, **54**: 34. https://doi.org/10.1186/s12711-022-00726-6.

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

Bermann, M., Lourenco, D., Forneris, N.S., Legarra, A., and 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. *Genet. Sel. Evol.*, **54**: 52. https://doi.org/10.1186/s12711-022-00741-7.

Christensen, O.F., and Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.*, **42**: 2. https://doi.org/10.1186/1297-9686-42-2.

Colleau, J.J. 2002. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.*, **34**: 409. https://doi.org/10.1186/1297-9686-34-4-409.

Faux, P., and Gengler, N. 2013. Inversion of a part of the numerator relationship matrix using pedigree information. *Genet. Sel. Evol.*, **45**: 45. https://doi.org/10.1186/1297-9686-45-45.

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

Fernando, R.L., Dekkers, J.C.M., and Garrick, D.J. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet. Sel. Evol.*, **46**: 50. https://doi.org/10.1186/1297-9686-46-50.

Fragomeni, B.O., Lourenco, D.A.L., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., Lawlor, T.J., and Misztal, I. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.*, **98**: 4090–4094. https://doi.org/10.3168/jds.2014-9125.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, **136**: 245–257. https://doi.org/10.1007/s10709-008-9308-0.

Harris, B.L., Sherlock, R.G., and Nilforooshan, M.A. 2022. Large-scale multiple-trait single-step marker model implementation. *Proc. 12th World Cong. Genet. Appl. Livest. Prod., Rotterdam, the Netherlands*. pp. 1360–1363. https://doi.org/10.3920/978-90-8686-940-4_324 (accessed on 2 February 2023).

Henderson, C.R. 1975. Rapid method for computing the inverse of a relationship matrix. *J. Dairy Sci.*, **58**: 1727–1730. https://doi.org/10.3168/jds.S0022-0302(75)84776-X.

Koivula, M., Strandén, I., Su, G., and Mäntysaari, E.A. 2012. Different methods to calculate genomic predictions—Comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *J. Dairy Sci.*, **95**: 4065–4073. https://doi.org/10.3168/jds.2011-4874.

Legarra, A., Aguilar, I., and Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.*, **92**: 4656–4663. https://doi.org/10.3168/jds.2009-2061.

Mäntysaari, E.A., Evans, R.D., and Strandén, I. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *J. Anim. Sci.*, **95**: 4728–4737. https://doi.org/10.2527/jas2017.1912.

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*, **202**: 401–409. https://doi.org/10.1534/genetics.115.182089.

Misztal, I., Legarra, A., and Aguilar, I. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci.*, **97**: 3943–3952. https://doi.org/10.3168/jds.2013-7752.

Nilforooshan, M.A., and Lee, M. 2019. The quality of the algorithm for proven and young with various sets of core animals in a multi-breed sheep population. *J. Anim. Sci.*, **97**: 1090–1110. https://doi.org/10.1093/jas/skz010.

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

Pocrnic, I., Lindgren, F., Tolhurst, D., William, O.H., and Gorjanc, G. 2022. Optimisation of the core subset for the APY approximation of genomic relationships. *Genet. Sel. Evol.*, **54**: 76. https://doi.org/10.1186/s12711-022-00767-x.

Snelson, E., and Ghahramani, Z. 2007. Local and global sparse gaussian process approximations. *Proc. 11th Int. Conf. AI Stat., San Juan, Puerto Rico*, **2**: 524–531. https://proceedings.mlr.press/v2/snelson07a.html (accessed on 2 February 2023).

Strandén, I., Matilainen, K., Aamand, G.P., and Mäntysaari, E.A. 2017. Solving efficiently large single-step genomic best linear unbiased prediction models. *J. Anim. Breed. Genet.*, **134**: 264–274. https://doi.org/https://doi.org/10.1111/jbg.12257.

Vandenplas, J., Eding, H., Calus, M.P.L., and Vuik, C. 2018. Deflated preconditioned conjugate gradient method for solving single-step blup models efficiently. *Genet. Sel. Evol.*, **50**: 51. https://doi.org/10.1186/s12711-018-0429-3.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**: 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Hollifield, M.K., Bermann, M., Lourenco, D., and Misztal, I. 2022. Impact of blending the genomic relationship matrix with different levels of pedigree relationships or the identity matrix on genetic evaluations. *JDS Comms.*, **3**: 343–347. https://doi.org/10.3168/jdsc.2022-0229.

Appendix

Considering the MME for GBLUP:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix},$$

and $\mathbf{G} = \mathbf{W}\mathbf{W}'$, conversion of GBLUP to SNP-BLUP MME follows:

$$\begin{aligned} &\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix}' \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix}' \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \\ \Rightarrow &\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{W} \\ \mathbf{W}'\mathbf{Z}'\mathbf{X} & \mathbf{W}'\mathbf{Z}'\mathbf{Z}\mathbf{W}+\mathbf{W}'\mathbf{G}^{-1}\mathbf{W}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{Z}'\mathbf{y} \end{bmatrix} \\ \Rightarrow &\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{V} \\ \mathbf{V}'\mathbf{X} & \mathbf{V}'\mathbf{V}+\mathbf{I}\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{V}'\mathbf{y} \end{bmatrix}. \end{aligned}$$

On the other hand, the conversion of SNP-BLUP to GBLUP is as follows:

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{V} \\ \mathbf{V}'\mathbf{X} & \mathbf{V}'\mathbf{V}+\mathbf{I}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^{-1} \end{bmatrix}' \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{V}'\mathbf{y} \end{bmatrix},$$

where $\hat{\mathbf{b}}$, $\hat{\mathbf{u}}$ and $\hat{\mathbf{a}}$ are the vectors of solutions for fixed effects, individuals' additive genetic merit and marker effects, $\alpha = \sigma_e^2/\sigma_m^2$, σ_e^2 is the residual variance, and σ_m^2 is the additive genetic variance captured by markers.