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 the single-step genomic BLUP, which incorporates genomic, pedigree, and phenotype information for simultaneous genetic evaluation of genotyped and non-genotyped individuals. The rapidly growing number of genotypes is a constraint for inverting a huge $\mathbf{G}$. The APY algorithm is an efficient method of solving 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, or genotyped and non-genotyped), 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 $\mathbf{M}_{nn}^{-1}$ makes APY an approximate. The application of APY is extendable to the inversion of any large symmetric matrix with a limited dimensionality at a lower computational cost. Possible applications are: computing the pedigree relationship matrix ($\mathbf{A}$) from the APY inverse of $\mathbf{A}^{-1}$, a diagonal block of $\mathbf{A}$ (same as the previous one, but avoiding unnecessary calculations), and the block of the block-diagonal preconditioner matrix corresponding to marker effects for iterative solving of marker effect model equations. Furthermore, APY may improve the matrix's numerical condition.
Keywords: APY, diagonal, dimensionality, GBLUP, single-step, relationship matrix

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 $O((c+n)^3)$ 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 $\mathbf{G}^{-1}_{\text{APY}}$ matrix is calculated as (Bermann et al., 2022):

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

where, $\mathbf{M}_{nn} = \mathbf{G}_{nn} - \mathbf{P}_{nc}\mathbf{G}_{cn}$, and $\mathbf{P}_{cn} = \mathbf{G}^{-1}_{cc}\mathbf{G}_{cn}$. In practice, $\text{diag}(\mathbf{M}_{nn})$ is used instead of $\mathbf{M}_{nn}$. Strandén et al. (2017) and Bermann et al. (2022) showed that:

$$\begin{aligned}\mathbf{G}_{\text{APY}} &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{P}_{nc} & \mathbf{I}\end{bmatrix}\begin{bmatrix}\mathbf{G}_{cc} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_{nn}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{P}_{nc} & \mathbf{I}\end{bmatrix}'\\ &= \begin{bmatrix}\mathbf{G}_{cc} & \mathbf{G}_{cn}\\ \mathbf{G}_{nc} & \mathbf{M}_{nn} + \mathbf{P}_{nc}\mathbf{G}_{cn}\end{bmatrix}\\ &= \mathbf{G} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_{nn} + \mathbf{P}_{nc}\mathbf{G}_{cn} - \mathbf{G}_{nn}\end{bmatrix}.\end{aligned} \tag{2}$$

The aim of this study is to provide new insights and possible applications for 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. A new way of understanding the properties of the APY inverse expression of G (*i.e.*, $\mathbf{G}^{-1}_{\text{APY}}$) is through understanding the hybrid pedigree-genomic relationship matrix ($\mathbf{H}$) used in ssGBLUP. Legarra et al. (2009) derived various forms of the same relationship matrix, including full pedigree and genomic information. Denoting genotyped and non-genotyped individuals as 2 and 1: $\mathbf{H} =$

$$\begin{bmatrix}\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}^{-1}_{22}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}^{-1}_{22}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}^{-1}_{22}\mathbf{G}\\ \mathbf{G}\mathbf{A}^{-1}_{22}\mathbf{A}_{21} & \mathbf{G}\end{bmatrix}, \tag{3}$$

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

$$\mathbf{A} + \begin{bmatrix}\mathbf{A}_{12}\mathbf{A}^{-1}_{22}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}^{-1}_{22}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}^{-1}_{22}(\mathbf{G} - \mathbf{A}_{22})\\ (\mathbf{G} - \mathbf{A}_{22})\mathbf{A}^{-1}_{22}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22}\end{bmatrix}. \tag{5}$$

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:

$$\mathbf{H} = \begin{bmatrix}(\mathbf{A}^{11})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \begin{bmatrix}-\mathbf{P}_{12}\\ \mathbf{0}\end{bmatrix}\mathbf{G}\begin{bmatrix}-\mathbf{P}_{12}\\ \mathbf{0}\end{bmatrix}', \tag{6}$$

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix}-\mathbf{P}_{12}\\ \mathbf{0}\end{bmatrix}(\mathbf{G} - \mathbf{A}_{22})\begin{bmatrix}-\mathbf{P}_{12}\\ \mathbf{0}\end{bmatrix}', \tag{7}$$

where, the projection matrix $\mathbf{P}_{12} = (\mathbf{A}^{11})^{-1}\mathbf{A}^{12} = -\mathbf{A}_{12}\mathbf{A}^{-1}_{22}$. 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):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22}^{-1} \end{bmatrix}. \tag{8}$$

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. 6 to Eq. 1. This shows that Eq. 6 is the APY inverse expression of $\mathbf{H}^{-1}$. Following Eq. 2, the APY expression of $\mathbf{H}^{-1}$ is:

$$\mathbf{H}_{\text{APY}}^{-1} = \mathbf{H}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^{22} + \mathbf{P}^{21}\mathbf{A}^{21} - \mathbf{H}^{22} \end{bmatrix}, \tag{9}$$

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 $\mathbf{H}$.

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

Considering Eq. 1 and 2, as long as no change is made to $\mathbf{M}_{nn}$, the APY and the APY inverse expressions of $\mathbf{G}$ are equal to $\mathbf{G}$ and $\mathbf{G}^{-1}$, respectively. Matrix $\mathbf{G}_{\text{APY}}^{-1}$ becomes an approximate $\mathbf{G}^{-1}$ when $\mathbf{M}_{nn}$ is changed to a diagonal matrix with diagonal elements:

$$m_{ii} = g_{ii} - \mathbf{g}_{ic}\mathbf{G}_{cc}^{-1}\mathbf{g}_{ci}, \tag{10}$$

representing genomic Mendelian sampling (Misztal et al., 2014). Using Eq. 10, calculations can be paralleled across all genotyped individuals. Compared with $\mathbf{G}^{nn}$:

$$\begin{aligned} \mathbf{G}_{\text{APY}}^{nn} &= \left(\text{diag}(\mathbf{M}_{nn})\right)^{-1} \\ &= \left(\text{diag}(\mathbf{G}_{nn} - \mathbf{G}_{nc}\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn})\right)^{-1} \\ &= \left(\text{diag}\left((\mathbf{G}^{nn})^{-1}\right)\right)^{-1}. \end{aligned} \tag{11}$$

The change of $\mathbf{M}_{nn}$ to $\text{diag}(\mathbf{M}_{nn})$ is propagated to the other blocks of $\mathbf{G}_{\text{APY}}^{-1}$ via the projection matrix $\mathbf{P}_{cn}$ (Eq. 1). No change is made to $\mathbf{G}_{\text{APY}}$ other than to the off-diagonal elements of $\mathbf{G}_{nn}$ (Strandén et al., 2017). Following Eq. 2, $\mathbf{M}_{nn} + \mathbf{P}_{nc}\mathbf{G}_{cn} - \mathbf{G}_{nn} = \mathbf{0}$. Thus, replacing $\mathbf{M}_{nn}$ with $\text{diag}(\mathbf{M}_{nn})$ replaces $\text{offdiag}(\mathbf{G}_{nn})$ with $\text{offdiag}(\mathbf{P}_{nc}\mathbf{G}_{cn})$. Therefore, it can be articulated that genomic relationships among non-core individuals become a function of $\mathbf{G}_{cc}$ and $\mathbf{G}_{cn}$. The efficiency of the APY algorithm depends on how well $\text{offdiag}(\mathbf{P}_{nc}\mathbf{G}_{cn})$ replaces $\text{offdiag}(\mathbf{G}_{nn})$.

2.3 Other applications

The application of the APY algorithm is not limited to $\mathbf{G}^{-1}$, nor to ssGBLUP and GBLUP. 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. (2014). They suggested the APY algorithm for the $\mathbf{A}_{22}$ inversion, which is required in ssGBLUP (Eq. 8). They derived an equivalent formula for the APY approximation of $\mathbf{A}_{22}^{-1}$:

$$\mathbf{A}_{22}^{-1} \approx \left(\mathbf{A}_{22}^{-1}\right)_{\text{APY}} = \begin{bmatrix} \mathbf{A}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{A}_{cc}^{-1}\mathbf{A}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{A}_{cc}^{-1}\mathbf{A}_{cn} \\ \mathbf{I} \end{bmatrix}'. \tag{12}$$

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. The $\mathbf{a}_{ci}$ vectors (rows of $\mathbf{A}_{cn}$) can be efficiently computed using the Colleau algorithm (Colleau, 2002), which can be done in parallel for many vectors at a time. The a_{ii} elements ($\text{diag}(\mathbf{A}_{nn})$) are easy to compute applying the fast and efficient algorithms available for computing inbreeding coefficients (Tier, 1990; Meuwissen and Luo, 1992; Sargolzaei and Iwaisaki, 2005; Sargolzaei et al., 2005). 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),

even for large populations. The computational cost of inverting $\mathbf{A}^{-1}$ to $\mathbf{A}$ can be reduced by obtaining an APY inversion of $\mathbf{A}^{-1}$:

$$\mathbf{A}_{\text{APY}} = \begin{bmatrix} (\mathbf{A}^{cc})^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -(\mathbf{A}^{cc})^{-1}\mathbf{A}^{cn} \\ \mathbf{I} \end{bmatrix} (\mathbf{M}^{nn})^{-1} \begin{bmatrix} -(\mathbf{A}^{cc})^{-1}\mathbf{A}^{cn} \\ \mathbf{I} \end{bmatrix}', \tag{13}$$

where $\mathbf{M}^{nn}$ is a diagonal matrix with diagonal elements $m^{ii} = a^{ii} - \mathbf{a}^{ic}(\mathbf{A}^{cc})^{-1}\mathbf{a}^{ic}$. Matrix $\mathbf{A}^{nn}$ is sparse. Thus, compared to $\mathbf{G}^{nn}$, there are considerably fewer non-zero off-diagonal elements set to 0. On the other hand, the choices of core size and core composition are likely to be more important. In the APY algorithm, relationships among non-core individuals are conditioned on the information from core individuals. In $\mathbf{A}$, the number of relatives that can explain the relationships between a non-core individual with other non-core individuals is limited. Thus, the choice of core individuals becomes more difficult. Contrarily, in $\mathbf{G}$ all individuals share information via many markers, regardless of whether they are relatives.

If rather than $\mathbf{A}$, a diagonal block of it ($\mathbf{A}_{cc}$) is needed, some of the calculations in Eq. 13 become redundant, and $\mathbf{A}_{cc}$ can be calculated as:

$$\begin{aligned} (\mathbf{A}_{cc})_{\text{APY}} &= (\mathbf{A}^{cc})^{-1} + (\mathbf{A}^{cc})^{-1}\mathbf{A}^{cn}(\mathbf{M}^{nn})^{-1}\mathbf{A}^{nc}(\mathbf{A}^{cc})^{-1} \\ &= (\mathbf{A}^{cc})^{-1}(\mathbf{A}^{cc} + \mathbf{A}^{cn}(\mathbf{M}^{nn})^{-1}\mathbf{A}^{nc})(\mathbf{A}^{cc})^{-1}. \end{aligned} \tag{14}$$

Calculating $(\mathbf{A}_{cc})_{\text{APY}}$, there is no choice of the core size and composition, as the choice of individuals for which the relationship coefficients to be approximated is already made. The APY approximation of $\mathbf{A}_{cc}$ might be influenced by $\mathbf{A}^{nn}$ changed to the diagonal $(\mathbf{M}^{nn})^{-1}$. Should APY approximations need improvement, the researcher might consider adding a chosen group of non-core individuals to the core subset. An application for $\mathbf{A}_{cc}$ is to calculate $\mathbf{A}_{22}$ for blending with $\mathbf{G}$ to improve the numerical condition of $\mathbf{G}$, and to introduce residual polygenic variance not captured by the markers.

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. 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.

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 (Nilforooshan and Lee, 2019). The problem of nonidentical results for random cores and the same data can be addressed by saving the identification of the core individuals. 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 $\mathbf{G}$. Though with a sufficiently large core size, the gain from an optimal core subset would be marginal (Nilforooshan and Lee, 2019), if screening for the optimal core subset is computationally affordable, it would be proffered over a random core subset.

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 (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.

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.

Conclusions

This study aimed to open new insights and understanding about the APY algorithm and to introduce new possible applications to this algorithm. Starting from the $\mathbf{H}$ matrix formula, it was shown that every covariance or inverse covariance matrix could be shown as a combination of its two diagonal blocks (diagonal blocks for genotyped and non-genotyped individuals in $\mathbf{H}$). The projection matrix makes the combination (information flow) between the two diagonal blocks. Furthermore, it was shown that any covariance or inverse covariance matrix has APY and APY inverse expressions equal to the matrix itself and its inverse, respectively. The difference arises when a diagonal block of the APY inverse (corresponding to non-core individuals) changes to a specific diagonal matrix. That 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 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 $\mathbf{G}^{-1}$. Possible new applications for APY are: computing $\mathbf{A}$, a diagonal block of $\mathbf{A}$, and the block of the block-diagonal preconditioner matrix corresponding to marker effects 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

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(2):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(1):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(1):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(1):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(1):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(1):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(6):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(2):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. In *Proceedings of the 12th World Congress on Genetics Applied to Livestock Production: 3-8 July 2022; Roterdam.*

Henderson, C. R. (1975). Rapid method for computing the inverse of a relationship matrix. *J. Dairy Sci.*, 58(11):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(7):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(9):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(11):4728–4737. https://doi.org/10.2527/jas2017.1912.

Meuwissen, T. H. E. and Luo, Z. (1992). Computing inbreeding coefficients in large populations. *Genet. Sel. Evol.*, 24:305–313. https://doi.org/10.1186/1297-9686-24-4-305.

Misztal, I. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*, 202(2):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(6):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(3):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(1):573–581. https://doi.org/10.1534/genetics.116.187013.

Sargolzaei, M. and Iwaisaki, H. (2005). Comparison of four direct algorithms for computing inbreeding coefficients. *Anim. Sci. J.*, 76:401–406. https://doi.org/10.1111/j.1740-0929.2005.00282.x.

Sargolzaei, M., Iwaisaki, H., and Colleau, J. J. (2005). A fast algorithm for computing inbreeding coefficients in large populations. *J. Anim. Breed. Genet.*, 122:325–331. https://doi.org/10.1111/j.1439-0388.2005.00538.x.

Snelson, E. and Ghahramani, Z. (2007). Local and global sparse gaussian process approximations. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 524–531, San Juan, Puerto Rico. PMLR. https://proceedings.mlr.press/v2/snelson07a.html (accessed on 24 November 2022).

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(3):264–274. https://doi.org/https://doi.org/10.1111/jbg.12257.

Tier, B. (1990). Computing inbreeding coefficients quickly. *Genet. Sel. Evol.*, 22:419–430. https://doi.org/10.1186/1297-9686-22-4-419.

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(1):51. https://doi.org/10.1186/s12711-018-0429-3.

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

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.