Optimal core definitions for the APY model with an example in large-scale pig dataset

Optimal core definitions for the APY model with an example in large-scale pig dataset

I. Pocrnic[1*], F. Lindgren[2], D. Tolhurst[1], W.O. Herring[3] and G. Gorjanc[1]

[1]The University of Edinburgh, The Roslin Institute, EH25 9RG, Edinburgh, United Kingdom;
[2]The University of Edinburgh, School of Mathematics, EH9 3FD, Edinburgh, United Kingdom;
[3]Genus PIC, TN 37075, Hendersonville, USA;
[*]ivan.pocrnic@roslin.ed.ac.uk

Abstract
Genetic evaluation databases accumulated vast amounts of genomic information that are skyrocketing computational needs for standard genomic evaluation models due to their cubic computational complexity. Several scalable approaches have been proposed, such as the Algorithm for Proven and Young (APY), where genotyped animals are usually randomly partitioned into core and noncore subsets to obtain sparse approximation of the inverse of genomic relationship matrix. Here we show a deterministic optimisation of the core set using a sequential sampling approach. We test this optimisation on a large pig dataset. Our results confirm that the APY is robust and that the size of the core set is critical. Our results further show that the stability of APY can be enhanced with an optimal spread of core animals across a given domain of genotyped animals. We discuss possibilities of an alternative optimisation based on the Nearest Neighbours Gaussian Process that also results in a sparse inverse.

Introduction
In the era of genomics, computational burden is inevitable with the standard genomic evaluation models, predominately due to the cubic cost of inverting the dense genomic relationship matrix. Several scalable genomic approaches have been proposed, such as reparametrizing the models with marker effects, inducing sparsity in the inverse, or dimensionality reduction. The Algorithm for Proven and Young (APY) is one of the approaches, where genotyped animals are usually randomly partitioned into core and noncore subsets. The core animals then form the dense part of the inverse, while the non-core animals are conditionally independent given the core animals and form the diagonal part of the inverse. While the APY provides a good approximation of the full standard model, random partitioning can make results unstable, possibly affecting accuracy or even reranking animals (Misztal *et al*., 2020). In this contribution we (i) show application of optimal core construction using conditional sequential sampling algorithm on a large pig dataset and (ii) discuss alternative core construction using the Nearest Neighbours Gaussian Process (NNGP).

Materials & Methods
Phenotypic data on 42,868 animals collected from 1999 to 2021, for a moderately heritable trait measured on purebred (33,544), crossbred (114) and backcross (9,210) pigs were provided by PIC (a Genus company, Hendersonville, TN, USA). Genomic information was available for 49,788 pigs, including all animals with phenotypes, and consisted of 42,707 single-nucleotide polymorphism markers. Validation subset included 478 phenotyped and genotyped youngest animals born in 2021 (their phenotypes were removed from the analysis). We used GBLUP in the following model $\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Wc} + \mathbf{e}$, where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{b}$ is a vector of contemporary group fixed effects, $\mathbf{a}$ is a vector of random animal effects with $\mathbf{a} \sim N(\mathbf{0}, \mathbf{G}\sigma_a^2)$, $\mathbf{c}$ is a vector of random litter effects with $\mathbf{c} \sim N(\mathbf{0}, \mathbf{I}\sigma_c^2)$, $\mathbf{e}$ is a vector of random residuals with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{W}$ are corresponding design matrices, $\mathbf{G}$ is the genomic relationship matrix with added $\mathbf{I}0.01$ to enable inversion, and σ_a^2, σ_c^2, and σ_e^2 are respectively known

additive, litter and residual variances. The model was run using BLUP90IOD software (Tsuruta *et al*., 2001), either with the full inverse of **G** or with the APY inverse of **G** as defined in Misztal *et al*. (2014). For the APY we applied two core selection strategies; (i) strategy where the core animals were randomly sampled among all the genotyped animals and (ii) optimised core selection strategy where we iteratively selected core animals with the largest variance conditionally to previously selected core animals (Pocrnic *et al*., 2021). For the random strategy, sampling was replicated five times to manifest potential instability of genetic evaluations and we present the mean ± 95% CI. The number of core animals was the same in both strategies and was based on the number of eigenvalues that explained from 10 to 99 % of the variation in **G** (Pocrnic *et al*., 2016). We compared strategies based on the correlations between genomic estimated breeding values (GEBV) obtained with the full inverse and GEBV obtained by each APY strategy. Furthermore, we assessed the predictive ability for the validation animals as correlations between their GEBV and phenotypes adjusted for the fixed effects in the model. To gain insight into the behaviour or the algorithms, we have visualised population structure with a non-linear dimension reduction technique called Uniform Manifold Approximation and Projection (UMAP; McInnes *et al*., 2020).

Results

The size of the core set in the APY ranged from 8 to 8,348 animals and corresponded to the number of eigenvalues that explained from 10 to 99% of the variation in **G**. Figure 1 shows that correlations between GEBV obtained with the full and APY inverse of **G** were almost a linear function of the percentage of explained variation. Both optimised and random core selection methods performed equally well when the number of core animals was equal to the number of eigenvalues explaining more than 98% of variation in **G**. The difference between the methods was larger in the validation population (Figure 1B), than in the entire genotyped population (Figure 1A). There was more uncertainty with random core sampling when the number of core animals was low, especially in the validation population (Figure 1B).

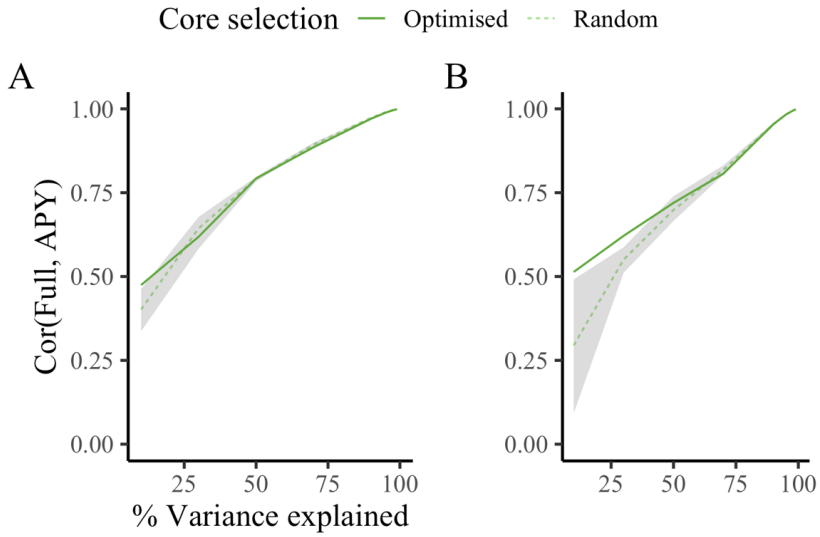

Figure 1. Correlation between GEBV obtained with the full and APY inverse of the genomic relationship matrix for all genotyped (A) and validation animals (B). APY is based on the optimised or random (mean ± 95% CI) core selection, with the number of core animals defined as the number of eigenvalues that explained certain percentage of variance.

As expected from previous studies (Pocrnic *et al*., 2016), the predictive ability was largely determined by the number of core animals (Figure 2), but with no difference between the APY and full inverse when the number of core animals was equal to the number of eigenvalues explaining more than 98% of variation in **G**. Also, differences between 90 to 99% were minimal. In this sense, the optimised core selection achieved the same satisfactory predictive ability as the random selection when the number of core animals was large, but also achieved notable improvements when the number of core animals was smaller. This improvement can be demonstrated by comparison with a large confidence interval in the random sampling strategy for a small number of core animals. This interval reduces as the core gets larger. The apparently higher predictive ability of random core selection with the larger core subsets (>95%) was due to rounding and was always less than 0.005, compared to optimised core selection.

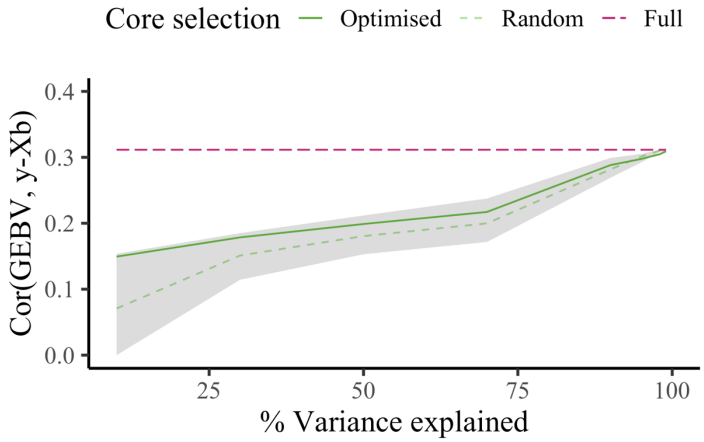

Figure 2. Predictive ability for validation animals. Predictive ability is the correlation between GEBV obtained with the full or APY inverse of the genomic relationship matrix and phenotypes adjusted for fixed effects. APY is based on the optimised or random (mean ± 95% CI) core selection, with the number of core animals defined as the number of eigenvalues that explained certain percentage of variance.

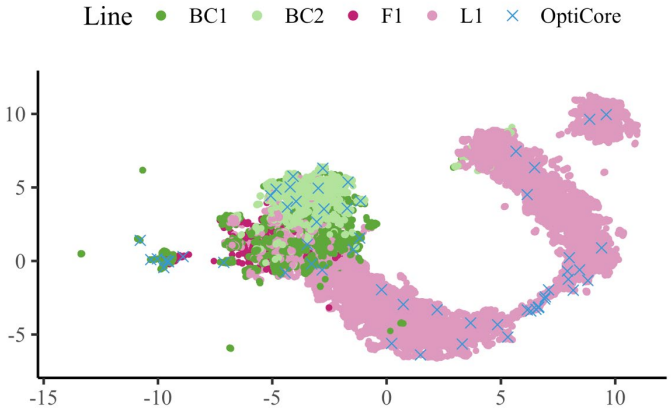

Figure 3. Uniform Manifold Approximation and Projection for purebred (L1), crossbred (F1), and backcross (BC1, BC2) pigs, with first 60 optimally selected core animals.

The UMAP demonstrates several clusters in the population. For example, Figure 3 shows the UMAP for the entire genotyped population, in which purebred animals (L1) cluster into three groups and the majority of crossbreed animals (F1, BC1, and BC2) are closely positioned to one of these clusters. Optimised core set spreads correspondingly.

Discussion

We have demonstrated that the APY is robust, largely dependent on the size of the core set and can be further enhanced by an optimal spread of core animals. While the optimised core selection strategy can introduce stability in the predictive ability (accuracy), especially when the core size is less than optimal, it also introduces additional computing cost compared to random sampling strategy. The optimised strategy builds upon sequential sampling algorithm and results in the spread of core animals far away from each other in a covariance sense. From that viewpoint, the additional computing cost could be justified in several cases, specifically when the genotyped population consists of multiple breeds, lines, and/or crossbred animals. In that sense, the optimised core selection strategy can be further extended towards building a unique core set for each animal based on the NNGP methodology, which recently gained popularity for its ability to deal with large-scale applications in spatial statistics (Datta *et al*., 2016). While the strategy we presented is depended on application through the APY, the extension into NNGP leads to a sparse inverse through a sparse Cholesky factor (e.g., Faux *et al*. 2012; Finley *et al*., 2019) and as such could substitute the APY-based inverse matrix when convenient. If n is the total number of genotyped animals, the computational complexity of the full inverse is $O(n^3)$ that is reduced to $O(m^3+2m^2(n-m))$ in the case of the APY or $O(nm^3)$ in the case of the NNGP, where m is either the number of the core animals in APY or the nearest neighbours in NNGP, and assuming $m << n$. On top of the computational complexity for the inversion itself, optimised strategy for APY and NNGP introduce additional computing cost of finding optimal core animals or nearest neighbours. Our future research will investigate faster alternatives for finding optimal set of core animals and nearest neighbours as well as assessing the practical benefit of proposed strategies within complex population structures. Furthermore, we will consider situations in which there is a benefit of using a set of nearest neighbours to form a sparse NNGP inverse relationship matrix rather than being fitted as a core in APY.

References

Datta A., Banerjee S., Finley A.O., and Gelfand A.E. (2016) J. Am. Stat. Assoc. 111:800-812. https://doi.org/10.1080/01621459.2015.1044091

Faux P., Gengler N., and Misztal I. (2012) J. Dairy Sci. 95:6093-6102. https://doi.org/10.3168/jds.2011-5249

Finley A.O., Datta A., Cook B.D., Morton D.C., Andersen H.E. *et al.* (2019) 28:401-414. https://doi.org/10.1080/10618600.2018.1537924

McInnes L., Healy J., and Melville J. (2020) arXiv preprint. https://arxiv.org/abs/1802.03426v3

Misztal I., Legarra A., and Aguilar I. (2014) J. Dairy Sci. 97:3943-3952. https://doi.org/10.3168/jds.2013-7752

Misztal I., Tsuruta S., Pocrnic I., and Lourenco D.A.L. (2020) J. Anim. Sci. 98:1-8. https://doi.org/10.1093/jas/skaa374

Pocrnic I., Lourenco D.A.L., Masuda Y., and Misztal I. (2016) Genet. Sel. Evol. 48:82. https://doi.org/10.1186/s12711-016-0261-6

Pocrnic I., Lindgren F., and Gorjanc G. (2021) Proc. of EAAP, Davos, Switzerland. https://doi.org/10.3920/978-90-8686-918-3

Tsuruta S., Misztal I., and Stranden I. (2001) J. Anim. Sci. 79:1166-1172. https://doi.org/10.2527/2001.7951166x