Proc. Assoc. Advmt. Anim. Breed. Genet. 22:93-96

APPROXIMATE GBLUP FOR EFFICIENT ROUTINE EVALUATIONS

T. P. Hancock[1,2]

[1] DEDJTR, Victoria, Australia
[2] DataGene Limited

SUMMARY

We present a computationally efficient approach to GBLUP which approximates inverse reference set matrix by optimally selecting the most informative animal cohort. The optimal animal cohort, named core reference animals, are identified through a Partial Incomplete Cholesky Decomposition (PICD) and selected such that the reconstruction error is at a specified user percentage. Our application of PICD on the Australian Holstein and Jersey reference sets shows that allowing a small error halves the effective size of reference set, resulting in significant gains in performance with only minor differences between exact and approximate breeding values and reliabilities ($r > 0.99$). Overall our results show that application of methods like PICD aimed at eliminating redundancy within large reference sets, significant performance gains can be made without sacrificing accuracy.

INTRODUCTION

Genomic evaluations are routinely used to evaluate the performance of dairy cattle world-wide. These genomic evaluations impose a significant and ever increasing computational burden on the evaluation organisations. This computational burden must be offset by the requirement to maintain a meaningful animal reference set to ensure that accurate and reliable predictions are made for the young animals entering the system. Up to now the focus has been on increasing the accuracy and reliability of genomic evaluations with projects such as GINFO (Pryce et al, in press) succeeding in increasing the overall reliability of the Australian genomic evaluations between 2 and 7 percent, by doubling the number of animals in the reference set. The cost of doubling the size of the reference set results in a dramatic increase in computational burden. GBLUP (Van Raden, 2008) like algorithms can be solved for breeding values using gradient techniques highly efficiently, however the reliability computation requires the explicit inverse of the genomic reference set matrix which scales at cubic complexity. With reference sets continuing to grow, and now including more than 35000 Australian dairy animals, more efficient solutions for genomic evaluations are required.

The accuracy and reliability of a genomic breeding value for a young, non-reference animal, is not based on the size of the reference set, but how related that animal is to the reference set. Additionally, the genomic relationship structure within the reference set animals are not related to the quality of their phenotypic information. Therefore simply adding animals to the reference set based on the quality of their phenotype alone will not ensure more reliable predictions into the future and is likely to make routine evaluations computationally infeasible.

In this paper we investigate the feasibility of a Partial Incomplete Cholseky Decomposition PICD (Foster et al, 2009) to identify a smaller cohort of reference set animals, named core reference set animals, which can be used to optimally represent the structure within full reference set. PICD has been shown in kernel regression literature to provide a robust approximate solution to a related model to GBLUP (Foster et al, 2009). In this paper we extend PICD for application to the GBLUP model by accounting for the diagonal weighting of all reference set animals to ensure that phenotype accuracy information is included in the evaluation of all animals. We show that application of PICD with a small degree of error can significantly reduce computational time without dramatically moving from the estimated breeding values or reliability from the full model.

MATERIALS AND METHODS

The equations for the GBLUP breeding values $\hat{a}$ and reliabilities *rel* are as follows (Van Raden, 2008),

$$\hat{\boldsymbol{a}} = \boldsymbol{G}_{cr}(\boldsymbol{G}_{rr} + \boldsymbol{R})^{-1}\boldsymbol{y} \qquad \text{and} \qquad \boldsymbol{rel} = \frac{\mathbf{diag}[\boldsymbol{G}_{cr}(\boldsymbol{G}_{rr} + \boldsymbol{R})^{-1}\boldsymbol{G}_{cr}^{T}]}{\mathbf{diag}[\boldsymbol{G}]} .$$

where $\boldsymbol{G}_{rr}$ is the genomic relationship matrix of the reference set animals, $\boldsymbol{R}$ is a diagonal matrix of observation weights and $\boldsymbol{G}_{cr}$ is the genomic covariance matrix of all animals with the reference set animals. The cost of a GBLUP model is in the evaluation of $(\boldsymbol{G}_{rr} + \boldsymbol{R})^{-1}$ where the number of required operations scales cubically, $O(r^3)$, as the number of reference set animals, *r*, increases.

Partial Incomplete Cholesky Decomposition (PICD) (Foster et al, 2009) is a variant of the Cholesky decomposition which employs both row pivoting and a diagonal error tolerance to create a rank-reduced decomposition. The purpose of PICD is to select from $\boldsymbol{G}_{rr}$ a reduced cohort of animals, called core reference animals, which are representative of the entire population. This cohort can then be used to reconstruct $\boldsymbol{G}_{rr}$ by,

$$\boldsymbol{G}_{rr} = \boldsymbol{L}^{T}\boldsymbol{L} \approx \boldsymbol{L}_{k}^{T}\boldsymbol{L}_{k} ,$$

where *k* is the set of core reference animals, $k < r$, and $\boldsymbol{L}_k$ is the Cholesky complement only including the currently selected *k* animals.

The PICD algorithm identifies the core reference animal by performing single Cholesky updates to $\boldsymbol{L}_k$, animal-by-animal in a stage-wise and greedy fashion where the next animal to be added $\boldsymbol{L}_k$ is selected such that it maximally reduces the reconstruction error. The reconstruction error is a measurement of how well $\boldsymbol{L}_k^T\boldsymbol{L}_k$ predicts $\boldsymbol{G}_{rr}$. The addition of all *r* animals completely reconstructs the full Cholesky complement with no error. Therefore the reconstruction error can be measured as a percentage of complete reconstruction.

The algorithm requires as input the acceptable amount of error as a percentage, and from this will create a Cholskey complement, $\boldsymbol{L}_k$, of size (*N, k*) where *k* number of animals required to approximate the original matrix at that error percentage. The advantage of using this approach to others such as Singular Value Decomposition (SVD) is its ability to pick the specific animals required for the reconstruction, whereas SVD projects each animal onto every eigenvector. Therefore PICD is a means of selecting the most informative animals from the reference set.

PCID when used in the kernel regression setting reduces the cost complexity from order $O(r^3)$ to $O(kr^2)$ (Rasmussen and Williams, 2006). However, within the reference set of GBLUP there are also observation weightings defined. To allow for all reference set animals to have their observation weight applied we must derive a subset-of-regressors approximation of $(\boldsymbol{G}_{rr} + \boldsymbol{R})^{-1}$ using the Nystrom approximation of $\boldsymbol{G}_{rr}$ (Rasmussen and Williams, 2006). The Nystrom approximation of $\boldsymbol{G}_{rr}$ is the approximation of the $\boldsymbol{G}_{rr}$ using a subset of rows and can be expressed as,

$$\widehat{\boldsymbol{G}}_{rr} = \boldsymbol{G}_{rk}\boldsymbol{G}_{kk}^{-1}\boldsymbol{G}_{kr}$$

where the *k* animals are selected from the reference set using PICD. From this representation of $\widehat{\boldsymbol{G}}_{rr}$ we can apply the Woodbury matrix identity to gain an approximation of the whole system inclusive of the observation weights,

$$(\boldsymbol{G}_{rr} + \boldsymbol{R})^{-1} \approx \left(\widehat{\boldsymbol{G}}_{rr} + \boldsymbol{R}\right)^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{G}_{rk}(\boldsymbol{G}_{kk} + \boldsymbol{G}_{kr}\boldsymbol{R}^{-1}\boldsymbol{G}_{rk})^{-1}\boldsymbol{G}_{kr}\boldsymbol{R}^{-1}$$

where $\boldsymbol{G}_{rk}$ is the covariance between the all reference animals and the core reference animals. This approximation to GBLUP allows for a selection of core animals from the reference set, without losing any phenotypic information from the model. Once the solution to the approximate GBLUP is attained the pre and post multiplication by $\boldsymbol{G}_{cr}$ is still required to compute the breeding values and reliabilities respectively. If no error tolerance is specified the approximation will yield exactly the same results as solving the system directly. It is suggested that this be treated like a heritability analysis and run once annually, out of scope of an evaluation.

PICD is also similar in idea to the sparse inverse of $\boldsymbol{G}$ with the APY algorithm of (Misztal, 2014) however PICD is a reduced rank approximation where as APY is a sparse approximation. The main advantage of PICD over APY is reducing the size of the entire system required to be solved through the efficient use of the Woodbury matrix identity above. APY on the other hand approximates only $\boldsymbol{G}$ or $\boldsymbol{G^{-1}}$ which still requires the addition of observation weights, $\boldsymbol{R}$, and solution of the entire system to be computed.

MATERIALS AND METHODS

To evaluate our proposed PICD approximated GBLUP we perform a simple parameter sweep on the percent error for the PICD algorithm and evaluate three different metrics.

1. The computational elapsed time.
2. The number of animals in the core reference set.
3. The correlation between breeding values and reliabilities as compared to the exact solutions.

The PICD program was developed in-house and implemented in R using Rcpp and compiled using the Intel MKL library. The datasets under consideration are the 58961 non-duplicated Holstein bulls and cows as well as the 11768 non-duplicated Jersey bulls and cows from the December 2016 ABV ADHIS release. Of these animals 32481 Holstein and 8846 Jersey bulls and cows were found in the full Protein GEBV reference set. The parameter sweep is run between 0 and 50% allowable error in increments of 5%.

RESULTS AND DISCUSSION

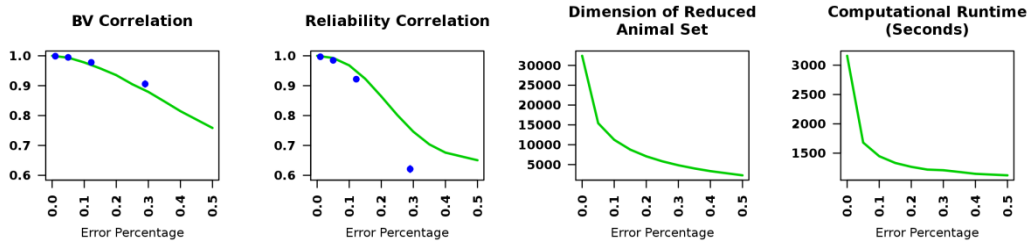

Figure 1. Holstein parameter screening results. Green line is the correlation between the exact solution and the PICD algorithm and the blue dots are the average correlation of 10 repeats of randomly selecting rows at four specified error tolerance

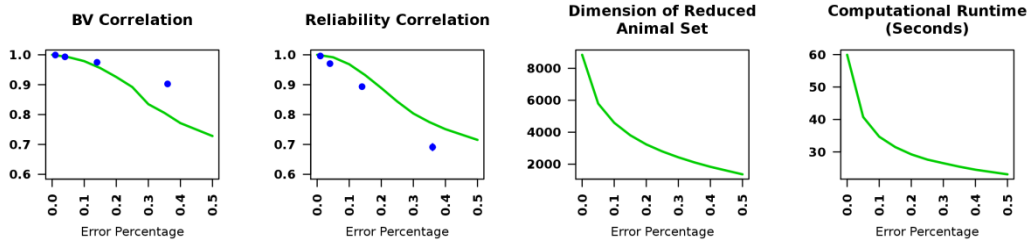

Figure 2. Jersey parameter screening results. Green line the correlation between the exact solution and PICD algorithm and the blue dots are the average correlation of 10 repeats of randomly selecting rows at four specified error tolerance

Figure 1 and 2 present the parameter sweep results for the Holstein and Jersey analyses respectively. The results include the computation of breeding values and reliability for all animals

in the analysis, including non-reference animals with no phenotype. From left to right, the first two plots are the correlation between approximate breeding values and reliabilities compared to exact GBLUP calculation, the dimension of the core reference animal set, *k*, and the run time.

Both Holstein and Jersey sets share the same profile, where at small amount of acceptable errors the approximate methods correlate very well ($r > 0.99$, % error = 0.05) with the exact solutions. The animals removed are predominantly bulls rather than cows. In of the 7754 cows and 1092 bulls in the Jersey reference set 2464 (32 %) cows and 591 (54 %) bulls were removed by PICD at 0.05 error tolerance. Of the 28228 cows and 4253 bulls in the Holstein reference set 13761 (49 %) cows and 3295 (78 %) bulls were removed PICD at 0.05 error tolerance. The removal of bulls from the reference is likely due to the selection of bulls results in stronger relationships between them, and therefore they produce more redundant set in terms of genotypic variation. The surprising result from these parameter sweeps by imposing only a small error the amount of animals in the core reference set is approximately.

The observed massive reduction in the reference set size is a result of the genomic redundancy within the reference set created by one-sided selection of animals. Reference set inclusion is based bulls having more than 10 daughters or cows in specific projects with phenotypic records, not on how related the animal is to the existing reference set. This approach is likely to select a reference set with a large number of highly related animals who collectively contribute very little to the performance of the overall evaluation. Algorithms like PICD are able to parse this redundant set and capture the key animals required to maintaining accuracy and reliability. The availability of such algorithms therefore encourages the continued collection of phenotypes and from the ever increasing pool of reference set animals timely evaluations are still possible.

At larger amounts of acceptable errors we observe that the PICD approximated reliabilities are significantly closer to the exact reliabilities than those computed from a random sample. However, the breeding values estimated by PICD are more poorly estimated, in particular within the Jersey analysis. This drop in performance is because PICD seeks to remove all redundancy within the genomic relationship matrix, without any knowledge of the phenotype. This style of selection may inadvertently remove animals with phenotypes that are highly informative for the trait under analysis because their relatives are already included in core reference set. This reduces the accuracy of breeding value estimation, but not reliability estimation, as the reliability is a function only of the relationship matrix (the target of PICD) not the phenotypic importance. This problem is well known and could potentially be overcome by selecting an animal subset using more complex objective functions which seek to balance the contributions from both left and right hand side GBLUP equations (Rasmussen and Williams, 2006).

In conclusion we have shown that it is possible to dramatically decrease the running time of genomic evaluations, without a significant impact on accuracy or reliability, by defining a smaller set of core reference animals. The implementation PICD with only small amount of error will reduce the computational burden on evaluation organisations allowing them to screen more animals, faster and more often.

REFERENCES

Foster L. et al, (2009) *Journal of Mach. Learn. Research*, **10**:857.
Misztal et al, (2014) *J Dairy Sci*, **97**:3943–3952.
Pryce J.E. (2016) *Interbull Bulletin*, **50**: in press.
Rasmussen C.E. and Willams C.K.I , (2006) 'Gaussian Processes for Machine Learning'. MIT Press.
VanRaden P.M. (2008) *J Dairy Sci*, **91**:4414.