Lazy computation of Residual Polygenic effects

***A. Legarra*[1]*, S. Savoia*[1]*, G. Jansen*[1]*, M. Bermann*[2] *and D. Lourenco*[2]**

[1] *Council on Dairy Cattle Breeding, Bowie, MD 20716, USA*

[2] *Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA*

Corresponding author: Andres Legarra: andres.legarra@uscdcb.com

Abstract

Single step methods use a blended relationship matrix that contains a fraction (typically 5%-20%) of pedigree-based relationships, called Residual Polygenic Effect (RPG). Indirect predictions of animals not included in the Mixed Model Equations (MME) are composed by a Direct Estimated Genomic Value (DGV), a sum of SNP readings times their solutions, and an RPG. Computation of RPG is not straightforward, and involves some complicated algebra and software, including separate relationships for genotyped and non-genotyped animals. We propose an alternative, equivalent computation that infers RPG for genotyped animals in the MME as RPG=GEBV-DGV, and then solves BLUP equations with RPG as "record" with heritability close to 1. The solution is the RPG for all animals in the MME, from which Parent Average can be used for Indirect Predictions. We show feasibility in practice with a US data set with millions of animals genotyped and in pedigree.

Key words: single step, residual polygenic effects

Introduction

Genomic evaluations typically include a fraction of pedigree-based relationships usually called Residual Polygenic effect (RPG). This fraction considers the relationships not well covered by markers, regress the genomic relationships towards pedigree ones, and generally prevents the evaluations from overdispersion (VanRaden, 2008; Liu et al., 2011). The resulting final GEBV $\hat{\mathbf{u}} = \hat{\mathbf{u}}^*$ can be separated into genomic-based and pedigree-based parts that we will call DGV (Direct Genomic Value). For genotyped animals g this is equal to the sum of SNP solutions $\hat{\mathbf{u}}_g^* = \mathbf{Z}\hat{\mathbf{a}}$, and the RPG part that is $\hat{\mathbf{d}}_g = \mathbf{A}_{g,g}\mathbf{G}^{-1}\hat{\mathbf{u}}_g$ (Legarra and Ducrocq, 2012, eq [8]).

RPGs are conceptually transmitted as an infinitesimal trait and therefore their covariances are modelled using pedigree-based relationships $\mathbf{A}$. The RPG of non-genotyped animals n as a function of genotyped ones can be obtained using some equivalent expressions (Vandenplas et al. 2023) which involve dedicated programming, among them $\hat{\mathbf{d}}_n = \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$ which solves RPG for non-genotyped animals.

New animals with genotypes (selection candidates) are typically evaluated, at least at first, based on solutions from the previous Single-step run. The DGV part is easily computed from the newly read genotype $\boldsymbol{z}$ as $\hat{u}^* = \boldsymbol{z}\hat{\boldsymbol{a}}$. The RPG part can be obtained as Parent Average of RPG from the ancestors, proceeding in pedigree order if needed from animals in the Single Step equations. This needs all solutions of RPG for non-genotyped animals, e.g. from $\hat{\mathbf{d}}_n = \mathbf{A}_{n,g}\mathbf{A}_{g,g}^{-1}\hat{\mathbf{d}}_g$.

The last equation can be a bit cumbersome to obtain, and Vandenplas et al. (2023) propose a few equivalent expressions, which need to be programmed. Here we propose an alternative shortcut that uses BLUP to obtain $\hat{\mathbf{d}}_n$ and therefore "regular" BLUP solvers can be used.

Materials and Methods

We arrange all animals (and Unknown Parent Groups or Metafounders) in the single step evaluation into non-genotyped and genotyped animals. Consider the following MME:

$$\begin{pmatrix}\mathbf{A}^{n,n}\lambda & \mathbf{A}^{n,g}\lambda \\ \mathbf{A}^{g,n}\lambda & \mathbf{A}^{g,g}\lambda+\mathbf{I}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{d}}_n \\ \hat{\mathbf{d}}_g^*\end{pmatrix}=\begin{pmatrix}0 \\ \hat{\mathbf{d}}_g\end{pmatrix} \quad [1]$$

For $\lambda = \frac{1-h^2}{h^2}$. When $h^2 \rightarrow 1$, $\lambda \rightarrow 0$ and $\hat{\mathbf{d}}_g^* \approx \hat{\mathbf{d}}_g$. From the top equation we obtain $\hat{\mathbf{d}}_g = -(\mathbf{A}^{n,n})^{-1}\mathbf{A}^{n,g}\hat{\mathbf{d}}_g = \mathbf{A}_{n,g}\mathbf{A}_{g,g}^{-1}\hat{\mathbf{d}}_g$. In other words, we obtain RPG solving BLUP equations.

Convergence of this iterative system is slow, as it will be shown later, because there are many more non-genotyped than genotyped animals. Thus, a second, approximated model and associated MME are:

$$\begin{pmatrix}\mathbf{1}'\mathbf{1} & \mathbf{0} & \mathbf{1}' \\ \mathbf{0} & \mathbf{A}^{n,n}\lambda & \mathbf{A}^{n,g}\lambda \\ \mathbf{1} & \mathbf{A}^{g,n}\lambda & \mathbf{A}^{g,g}\lambda+I\end{pmatrix}\begin{pmatrix}\hat{\mu} \\ \hat{\mathbf{d}}_n \\ \hat{\mathbf{d}}_g^*\end{pmatrix} = \begin{pmatrix}\mathbf{1}'\hat{\mathbf{d}}_g \\ \mathbf{0} \\ \hat{\mathbf{d}}_g\end{pmatrix} \quad [2]$$

upon solution of this system, we should add back $\hat{\mu}$ to $\hat{d}_g^*$ and $\hat{d}_n$.

After running a single step evaluation with US data, at CDCB we tested both [1] and [2] with the "yield" group of traits (milk, fat and protein yields) with ~50M animals in pedigree, ~2.5M animals genotyped. Therefore there are 2.5M RPG "records" (obtained as $\hat{\mathbf{d}}_g = \hat{\mathbf{u}}_g - \mathbf{Z}\hat{\mathbf{a}}$) and 48M animals with "no records", for a total of 150M equations. We used PCG iteration in blup90iod3 from the BLUPF90 suite (Lourenco et al., 2022), with 8 threads. We tried several convergence criteria from 10^{-6} to 10^{-14}, and we considered 10^{-14} as pseudo-true. Then we computed the correlation with the pseudo-true solutions.

In addition, we verified number of rounds needed to achieve PCG convergence of 10^{-12} for all traits evaluated by CDCB except calving ease and stillbirth.

Results & Discussion

Results for yield traits are presented in Tables 1 and 2. Time per round is 20 seconds so 1000 iterations take ~5h, and this is a post-processing step after the single step run that need not be run again. For model [1], the convergence with the pseudo-true solution is quite fast, with a correlation of almost 1 at a PCG convergence of 10^{-10}.

On the other hand, model [2] is of faster apparent convergence but of actual slower one – good correlations with the pseudo-true solution appear at PCG convergence levels of 10^{-14} and smaller. So the total computing time is about the same, and the convergence is more misleading. Model [2] is not recommended.

Table 3 shows number of rounds to reach PCG convergence of 10^{-12}. Trait groups that take longest time (over 1,000 iterations) are those with large number of animals and lowest proportion of animals genotyped. For instance, "Fertility" has 50M animals in pedigree, 2.2M of them genotyped; "Health" has ~10M animals in pedigree, 1.2M of them genotyped; and "Residual Feed Intake" has 60K animals in pedigree, 9K of them genotyped.

Table 1: correlation of the solution for RPGs using [1] with the pseudo-true solution, yield traits.

PCG convergence	correlation with pseudo-true	iteration
10E-06	0.916	113
10E-07	0.958	244
10E-08	0.978	381
10E-09	0.995	655
10E-10	0.9995	898
10E-11	0.999988	1177
10E-12	0.999999	1385
10E-13	1	1549
10E-14	1	1808

Table 2: correlation of the solution for RPGs using [2] with the pseudo-true solution, yield traits

PCG convergence	correlation with pseudo-true	iteration
10E-10	0.872	31
10E-11	0.895	46
10E-12	0.914	79
10E-13	0.950	213
10E-14	0.995	568
10E-15	0.9992	712
10E-16	0.9996	804

Table 3: rounds to reach PCG convergence of 10E-12, all trait groups

Trait group	iteration
Fertility	1476
Gestation length	384
Heifer livability	460
Health	397
Livability	1751
Productive life	1537
Residual feed intake	104
Somatic cell score	1040
Yield	1385

Conclusions

In absence of dedicated software, the RPG solutions for animals in the single step equations can be computed using BLUP with $h^2 \approx 1$. Computing time is a few hours and it depends on the ratio genotyped animals/non-genotyped animals. This BLUP gives a simple and competitive solution to backsolve RPGs for all animals considered in the evaluation. Those RPGs can be used later for candidates to selection through “indirect predictions”.

References

Legarra, A., and Ducrocq, V. 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *Journal of dairy Science*, 95(8), 4629-4645.

Liu, Z., Seefried, F. R., Reinhardt, F., Rensing, S., Thaller, G., & Reents, R. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genetics Selection Evolution*, 43(1), 19.

Lourenco, D., Tsuruta, S., Aguilar, I., Masuda, Y., Bermann, M., Legarra, A., et al. 2022 Recent updates in the BLUPF90 software suite. In: Proceedings of 12th World Congress on Genetics Applied to Livestock Production 3–8 July 2022; Rotterdam.

Vandenplas, J., Ten Napel, J., Darbaghshahi, S. N., Evans, R., Calus, M. P., Veerkamp, R., Cromie, A., Mantysaari, E. and Strandén, I. 2023. Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genetics Selection Evolution,* 55(1), 37.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11), 4414-4423.