Deregression formula for single-step genomic BLUP

Y. Masuda[12], Z. Liu[3] and P. Sullivan[4]

[1] Rakuno Gakuen University, Ebetsu, Hokkaido, Japan
[2] University of Georgia, USA
[3] IT solutions for animal production (vit), Verden, Germany
[4] Lactanet, Sainte-Anne-de-Bellevue, Canada

Abstract

The objective of this study was to describe an algorithm to deregress GEBV in single-step genomic BLUP (ssGBLUP). The iterative method suggested by Jairath et al. (1998), which was for the pedigree-based EBV, was extended to support GEBV. The inverse of the unified relationship matrix ($\mathbf{H}^{-1}$), which is a function of the inverse of genomic relationship matrix ($\mathbf{G}^{-1}$) and the inverse of the additive relationship matrix ($\mathbf{A}^{-1}$), was considered in the deregression method. With more genotypes, $\mathbf{G}^{-1}$ can be replaced with a sparser inverse by the Algorithm of Proven and Young (APY) or a transformed matrix (single-step GTBLUP). The deregression algorithm consists of a series of matrix-vector multiplications, and the number of iterations is expected to be limited. Therefore, the total computing cost should be lower than PCG to solve the mixed model equations in the original single-step genomic evaluation. A validation study using real data is needed to confirm this method works as expected.

Key words: deregression, single-step genomic BLUP, genomic evaluation, algorithm

Introduction

The traditional, pedigree-based estimated breeding values (EBV) are "deregressed" to give pseudo-phenotypes for further genetic analyses. This technique is essential in MACE; EBV are deregressed for the across-country evaluation. As a result, each country receives the MACE EBV of foreign bulls on their own country scale; The MACE EBV are comparable with the EBV of domestic bulls.

The deregressed proofs are also useful for genomic evaluation, particularly "multi-step" methods in dairy cattle. Genomically enhanced breeding values (GEBV) are predicted using pseudo-phenotypes and genotypes of animals in a reference population, which typically consists of domestic and foreign sires. The pseudo-phenotype for domestic bulls can be daughter yield deviations (DYD) or deregressed proofs calculated based on EBV. The foreign bulls have deregressed MACE EBV. The entire system of multi-step approaches relies on unbiased EBV.

As more young bulls have been selected based on GEBV, domestic EBV tend to be biased downward (Masuda et al., 2018) because the pedigree-based BLUP cannot account for the genomic pre-selection of young bulls (Patry and Ducrocq 2011). Some scientists developed methods to adjust EBV to reduce the pre-selection bias (e.g., Mäntysaari and Strandén, 2010; Wiggans et al., 2012). However, the methods are ad hoc, and the bias may not be removed.

An essential solution to remove pre-selection bias is a "single-step" method combining all available phenotypes, pedigree, and genotypes. Two models have been tested in genomic prediction in dairy cattle: single-step genomic BLUP (ssGBLUP; Aguilar et al., 2010; Christensen and Lund, 2010) and single-step SNP BLUP (ssSNPBLUP; Liu et al., 2014). The single-step GBLUP is simpler in implementation and has better convergence behavior than ssSNPBLUP. However, a concern was the computational cost of the inverse of genomic relationship matrix $\mathbf{G}^{-1}$, which is integrated into the inverse of unified

relationship matrix ($\mathbf{H}^{-1}$) together with the inverse of additive relationship matrix ($\mathbf{A}^{-1}$). The issue has been solved using a sparse version of $\mathbf{G}^{-1}$ (Misztal, 2016) or a transformation of relationship matrices (Mäntysaari et al., 2017).

Some countries have already implemented ssGBLUP in the genetic evaluation of dairy cattle (https://interbull.org/ib/nationalgenoforms).

In the future, it may be possible for countries to contribute to the traditional MACE evaluation by providing deregressed GEBV to Interbull. For the traditional EBV, an iterative method (Jairath et al., 1998) has been used for deregression. An extension of this method to ssGBLUP is straightforward. Esa Mänysaari and Zengting Liu suggested the technique at a meeting of an Interbull working group. Masuda et al. (2021a) independently developed the method and applied it to simulated data. Nevertheless, the approach has not been formally described.

The objective of this study was to describe an algorithm to deregress GEBV in ssGBLUP. Note that Liu and Masuda (2021) comprehensively discussed how the deregression should be applied to ssSNPBLUP. The points that they make are also valid to ssGBLUP. Therefore, this paper focuses more on the computational efficiency of deregression in ssGBLUP.

Methods

Deregression model

The deregression is based on the following single-trait model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{u} + \mathbf{Qt} + \mathbf{e}$$

where $\mathbf{y}$ is a vector of deregressed GEBV, μ is the overall mean, $\mathbf{u}$ is a vector of GEBV, $\mathbf{t}$ is a vector of unknown parent groups (UPG), also known as phantom parent groups, $\mathbf{e}$ is a vector of residuals, $\mathbf{1}$ is the vector of ones, and $\mathbf{Q}$ is a matrix relating UPG to deregressed values (Quaas, 1988). We assume the additive genetic variance (σ_u^2), the residual variance (σ_e^2), and the variance ratio ($\lambda = \sigma_e^2/\sigma_u^2$) are known.

We define three groups for animals; group 1 is for non-genotyped animals with the phenotype(s), group 2 is for all genotyped animals, regardless of whether they are phenotyped or not, and group 0 for ancestors of the animals in groups 1 and 2.

Let

$$\theta = \begin{bmatrix} \mathbf{u}_0^* \\ \mathbf{u}_1^* \\ \mathbf{u}_2^* \\ \mathbf{t} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_0 \\ \mathbf{u}_1 + \mathbf{Q}_1\mathbf{t} \\ \mathbf{u}_2 + \mathbf{Q}_2\mathbf{t} \\ \mathbf{t} \end{bmatrix},$$

where $\mathbf{Q}_1$ is $\mathbf{Q}$ for group 1, and $\mathbf{Q}_2$ for group 2. The inverse of additive relationship matrix for θ is

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} & \mathbf{A}^{02} & \mathbf{A}^{0t} \\ \mathbf{A}^{10} & \mathbf{A}^{11} & \mathbf{A}^{12} & \mathbf{A}^{1t} \\ \mathbf{A}^{20} & \mathbf{A}^{21} & \mathbf{A}^{22} & \mathbf{A}^{2t} \\ \mathbf{A}^{t0} & \mathbf{A}^{t1} & \mathbf{A}^{t2} & \mathbf{A}^{tt} \end{bmatrix}.$$

This matrix can be constructed using Henderson's algorithm (Henderson, 1976; Quaas, 1988). When the UPG effect is treated as random, i.e, $var(\mathbf{t}) = \mathbf{\Sigma}\sigma_u^2$, $\mathbf{A}^{tt}$ is replaced with $\mathbf{A}^{tt} + \mathbf{\Sigma}^{-1}$ (Masuda et al., 2021b). Applying QP-transformation (Misztal et al., 2013, or Tsuruta et al, 2019), the inverse of unified relationship matrix is available as

$$\mathbf{H}^* = \begin{bmatrix} \mathbf{H}^{00} & \mathbf{H}^{01} & \mathbf{H}^{02} & \mathbf{H}^{0t} \\ \mathbf{H}^{10} & \mathbf{H}^{11} & \mathbf{H}^{12} & \mathbf{H}^{1t} \\ \mathbf{H}^{20} & \mathbf{H}^{21} & \mathbf{H}^{22} & \mathbf{H}^{2t} \\ \mathbf{H}^{t0} & \mathbf{H}^{t1} & \mathbf{H}^{t2} & \mathbf{H}^{tt} \end{bmatrix}$$

$$= \mathbf{A}^* + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2 \\ \mathbf{0} & \mathbf{0} & -\mathbf{Q}_2'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}_2'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2 \end{bmatrix}.$$

Under the metafounder model (Legarra et al., 2015) or an alternative UPG model (Masuda et al., 2021b), the inverse of the unified relationship matrix has the following form:

$$\mathbf{H}^* = \mathbf{A}^{\#} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\mathbf{G}^{-1} - \mathbf{A}_{22}^{\#}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{A}^{\#}$ is the "inverse" of additive relationship matrix including metafounders' (or

UPG's) contributions, and $\mathbf{A}_{22}^{\#}$ is the inverse of the additive relationship matrix based on $\mathbf{A}^{\#}$ for genotyped animals.

When **G** is large, the direct inversion is infeasible. Two solutions have been suggested. Misztal et al. (2014) developed the Algorithm for Proven and Young (APY), which results in a sparse inverse ($\boldsymbol{G}_{APY}^{-1}$) accounting for all the additive variation in **G**. Mäntysaari et al. (2017) suggested single-step GTBLUP, which transforms **G** inexpensively to compute the inverse ($\boldsymbol{G}_{w}^{-1}$). Both the methods are expected to be capable of millions of genotyped animals and facilitate solving the mixed model equations in a practical time. In deregression, $\mathbf{G}^{-1}$ can be replaced with $\boldsymbol{G}_{APY}^{-1}$ or $\boldsymbol{G}_{w}^{-1}$.

This model considers all genotyped animals used in the original genetic evaluation because $\mathbf{G}^{-1}$ accounts for precise relationships among genotyped animals. Liu and Masuda (2021) emphasized that the deregression method should be based on the animal model. Therefore, **y** includes cows with phenotypes(s) only. The deregressed GEBV for bulls can be calculated as in Liu and Masuda (2021).

Algorithm

The mixed model equations are

$$\begin{bmatrix} \mathbf{1}'\mathbf{W}_1\mathbf{1}+\mathbf{1}'\mathbf{W}_2\mathbf{1} & \mathbf{0} & \mathbf{1}'\mathbf{W}_1 & \mathbf{1}'\mathbf{W}_2 & \mathbf{0} \\ \mathbf{0} & \lambda\mathbf{H}^{00} & \lambda\mathbf{H}^{01} & \lambda\mathbf{H}^{02} & \lambda\mathbf{H}^{0t} \\ \mathbf{W}_1\mathbf{1} & \lambda\mathbf{H}^{10} & \mathbf{W}_1+\lambda\mathbf{H}^{11} & \lambda\mathbf{H}^{12} & \lambda\mathbf{H}^{1t} \\ \mathbf{W}_2\mathbf{1} & \lambda\mathbf{H}^{20} & \lambda\mathbf{H}^{21} & \mathbf{W}_2+\lambda\mathbf{H}^{22} & \lambda\mathbf{H}^{2t} \\ \mathbf{0} & \lambda\mathbf{H}^{t0} & \lambda\mathbf{H}^{t1} & \lambda\mathbf{H}^{t2} & \lambda\mathbf{H}^{tt} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\mathbf{u}_0^*} \\ \widehat{\mathbf{u}_1^*} \\ \widehat{\mathbf{u}_2^*} \\ \hat{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{W}_1\mathbf{y}_1+\mathbf{1}'\mathbf{W}_2\mathbf{y}_2 \\ \mathbf{0} \\ \mathbf{W}_1\mathbf{y}_1 \\ \mathbf{W}_2\mathbf{y}_2 \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{y}_1$ is **y** for non-genotyped animals, and $\mathbf{y}_2$ for genotyped animals. The weight matrices are $\mathbf{W}_1 = \text{diag}\{n_{i1}\}$ and $\mathbf{W}_2 = \text{diag}\{n_{i2}\}$, where n_{i1} (n_{i2}) is effective record contribution (ERC) of i-th non-genotyped (genotyped) animal. If an animal has no phenotype(s), the weight is set to 0.

The deregressed GEBV for cows ($\mathbf{y}_1$ and $\mathbf{y}_2$) can be available using an iterative method. Following Jairath et al. (1998), a possible algorithm to obtain **y** can be found as follows.

1. Initialize $\hat{\mu} = 0$, $\widehat{\mathbf{u}_0^*} = \mathbf{0}$, $\widehat{\mathbf{u}_1^*} = \mathbf{0}$, and $\widehat{\mathbf{u}_2^*} = \mathbf{0}$.
2. Compute $\widehat{\mathbf{u}_1^*} = \mathbf{u}_1 - \mathbf{1}\hat{\mu}$ and $\widehat{\mathbf{u}_2^*} = \mathbf{u}_2 - \mathbf{1}\hat{\mu}$, where $\mathbf{u}_1$ ($\mathbf{u}_2$) is a vector of GEBV for non-genotyped (genotyped) animals.
3. Solve
$$\begin{bmatrix} \mathbf{H}^{00} & \mathbf{H}^{0t} \\ \mathbf{H}^{t0} & \mathbf{H}^{tt} \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{u}_0^*} \\ \hat{\mathbf{t}} \end{bmatrix} = \begin{bmatrix} -\mathbf{H}^{01}\widehat{\mathbf{u}_1^*} - \mathbf{H}^{02}\widehat{\mathbf{u}_2^*} \\ -\mathbf{H}^{t1}\widehat{\mathbf{u}_1^*} - \mathbf{H}^{t2}\widehat{\mathbf{u}_2^*} \end{bmatrix}.$$
4. Compute
$$\begin{aligned} \mathbf{z}_1 = \mathbf{W}_1\mathbf{y}_1 &= \mathbf{W}_1\mathbf{1}\hat{\mu} + \lambda\mathbf{H}^{10}\widehat{\mathbf{u}_0^*} \\ &\quad + (\mathbf{W}_1 + \lambda\mathbf{H}^{11})\widehat{\mathbf{u}_1^*} \\ &\quad + \lambda\mathbf{H}^{12}\widehat{\mathbf{u}_2^*} + \lambda\mathbf{H}^{1t}\hat{\mathbf{t}} \\ \mathbf{z}_2 = \mathbf{W}_2\mathbf{y}_2 &= \mathbf{W}_2\mathbf{1}\hat{\mu} + \lambda\mathbf{H}^{20}\widehat{\mathbf{u}_0^*} \\ &\quad + \lambda\mathbf{H}^{21}\widehat{\mathbf{u}_1^*} + (\mathbf{W}_2 \\ &\quad + \lambda\mathbf{H}^{22})\widehat{\mathbf{u}_2^*} + \lambda\mathbf{H}^{2t}\hat{\mathbf{t}} \end{aligned}$$
and
$m = \mathbf{1}'\mathbf{W}_1\mathbf{y}_1 + \mathbf{1}'\mathbf{W}_2\mathbf{y}_2 = \mathbf{1}'\mathbf{z}_1 + \mathbf{1}'\mathbf{z}_2$.
5. Compute $\hat{\mu} = m/(\mathbf{1}'\mathbf{W}_1\mathbf{1} + \mathbf{1}'\mathbf{W}_2\mathbf{1})$.
6. Go to Step 2 until convergence is reached.
7. Compute
$$\mathbf{y}_1 = \mathbf{W}_1^{-1}\mathbf{W}_1\mathbf{y}_1 = \mathbf{W}_1^{-1}\mathbf{z}_1$$
$$\mathbf{y}_2 = \mathbf{W}_2^{-1}\mathbf{W}_2\mathbf{y}_2 = \mathbf{W}_2^{-1}\mathbf{z}_2.$$

Deregressed GEBV for bulls are available through the method by Liu and Masuda (2021).

Discussion

The algorithm consists of a series of matrix-vector multiplications. The number of iterations is expected to be limited to get convergence. Although step 3 needs to solve a system of sparse equations, it can be solved by preconditioned conjugate gradient (PCG), which is inexpensive. Therefore, the total computing cost should be lower than the original single-step evaluation.

The deregression method shown here is flexible to support any models defined in the ssGBLUP framework. The user can replace $\mathbf{G}^{-1}$ with $\boldsymbol{G}_{APY}^{-1}$ or $\boldsymbol{G}_{w}^{-1}$. A different assumption in UPG can be applicable.

The deregressed proofs should be reversible (Mäntysaari et al., 2011) in the sense that the recalculated evaluations based on deregressed proofs are identical to the original sire evaluations. To ensure reversibility, Liu and Masuda (2021) suggested a deregression formula should include all animals with phenotype(s), all genotyped animals, and the same pedigree animals used in the original single-step evaluation.

Masuda et al. (2021a) tested a prototype of the current method on simulated data with 6900 genotypes. However, they did not use the same pedigree and genomic information as ssGBLUP evaluation. Also, their method incorrectly treated contributions from genotyped daughters. Although they did not check the reversibility, the resulting deregressed proofs would likely have been biased because of the missing information. Regarding the cost of iterative deregression with $\mathbf{H}^{-1}$, each round required nearly the same computing time as a round of PCG (preconditioned conjugate gradient) in ssGBLUP. Less than 20 rounds were needed to get convergence for deregression, compared to more than 500 rounds in ssGBLUP.

The current method is useful for an evaluation center computing the ssGBLUP evaluations, because pre-calculated $\mathbf{G}^{-1}$ and the other parts of $\mathbf{H}^{*}$ are immediately available. In other words, there is no advantage to use this method for populations where $\mathbf{H}^{*}$ has not been calculated. Liu and Masuda (2021) suggested a deregression approach derived from ssSNPBLUP. Their method is, in theory, applicable to GEBV calculated from any single-step model. However, neither method has been tested with real data. Further research is needed to verify reversibility and computational feasibility for both methods.

Conclusions

This study suggested an algorithm to deregress GEBV from ssGBLUP evaluation. The method is computationally feasible. It supports any options for modelling missing parents (UPG and MF) and the inverse of genomic relationship matrix ($\boldsymbol{G}_{APY}^{-1}$ or $\boldsymbol{G}_{w}^{-1}$). Further study is required to verify this approach using real data.

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. 93:743-752.

Christensen, O.F. and Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Henderson, C.R., 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics. 32:69-83.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B. and Kolstad, B. 1998. Genetic evaluation for herd life in Canada. J. Dairy Sci. 81:550-562.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I. and Misztal, I. 2015. Ancestral

relationships using metafounders: finite ancestral populations and across population relationships. Genetics. 200:455-468.

Liu, Z., Goddard, M.E., Reinhardt, F. and Reents, R. 2014. A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97:5833-5850.

Liu, Z. and Masuda, Y. 2021. A deregression method for single-step genomic model using all genotype data. Interbull Bull. 56.

Mäntysaari, E.A. and Strandén, I. 2010. Use of bivariate EBV-DGV model to combine genomic and conventional breeding value evaluations. In Proceedings of the 9th World Congress on Genetics Applied to Livestock Production; August 1 (Vol. 6).

Mäntysaari, E.A., Koivula, M., Strandén, I., Pösö, J. and Aamand, G.P. 2011. Estimation of GEBVs using deregressed individual cow breeding values. Interbull Bull. 44:19-24.

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.

Masuda, Y., VanRaden, P.M., Misztal, I. and Lawlor, T.J. 2018. Differing genetic trend estimates from traditional and genomic evaluations of genotyped animals as evidence of preselection bias in US Holsteins. J. Dairy Sci. 101:5194-5206.

Masuda, Y., Liu, Z., and Sullivan, P. 2021a. Preliminary results on de-regressed proof in single-step GBLUP. Presented at Virtual Interbull Meeting. https://interbull.org/ib/programme_virtual_2021

Masuda, Y., Tsuruta, S., Bermann, M., Bradford, H.L. and Misztal, I. 2021b. Comparison of models for missing pedigree in single-step genomic prediction. J. Anim. Sci. 99:skab019.

Misztal, I., Vitezica, Z.G., Legarra, A., Aguilar, I. and Swan, A.A. 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Breed. Genet. 130:252-258.

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.

Misztal, I., 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*, *202*(2), pp.401-409

Patry, C. and Ducrocq, V. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. J. Dairy Sci. 94:1011-1020.

Quaas, R. L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71:91–98.

Tsuruta, S., Lourenco, D.A.L., Masuda, Y., Misztal, I. and Lawlor, T.J., 2019. Controlling bias in genomic breeding values for young genotyped bulls. J. Dairy Sci. 102:9956-9970.

Wiggans, G.R., VanRaden, P.M. and Cooper, T.A. 2012. Adjustment of all cow evaluations for yield traits to be comparable with bull evaluations. J. Dairy Sci. 95:3444-3447.