Studies on Inflation of GEBV in Single-Step GBLUP for Type

I. Misztal[1], H. L. Bradford[1], D. A. L. Lourenco[1], S. Tsuruta[1], Y. Masuda[1], A. Legarra[3] and T. J. Lawlor[4]

[1] Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA
[3] INRA, UMR1388 GenePhySE, Castanet Tolosan, 31326, France
[4] Holstein Association USA Inc., Brattleboro, VT 05302, USA

Abstract

The purpose of this study was to evaluate sources of inflation of GEBV in single-step GBLUP (ssGBLUP) evaluations. Tests involved 10 102 702 records of 18 type traits from 6 930 618 Holstein cows. A total of 576k animals with genotypes were used in the analyses and included 23 174 sires, 27 215 cows and 49 611 young animals. The genomic relationship matrix (**G**) was scaled for compatibility with the pedigree matrix for genotyped animals ($\mathbf{A}_{22}$). Genomic estimated breeding values (GEBV) using phenotypes up to 2010 or up to 2014 were calculated considering a) inbreeding in A_{22} but not in A, 2) inbreeding in both A and A_{22}, 3) as previously but considering nonzero inbreeding of phantom parents, and 4) as previously but reducing the additive variance by one half. Reliabilities (R^2) and regression factors (b_1) were derived based on DYD2014 and GEBV2010 of 1711 sires with at least 50 daughters in 2014 but no daughters in 2010. For cases 1 to 4, the reliabilities were 0.48, 0.49, 0.49, and 0.50, respectively. The average regressions were 0.75, 0.85, 0.90, and 0.96, respectively. The b_1 factors could be close to 1 by multiplying G^{-1} - A_{22}^{-1} by lambda and by multiplying A_{22}^{-1} by omega. The optimal omega was 0.7 for case 1) and 0.9 for case 2). Both parameters are shown to account for ignored inbreeding in A. Inflation of GEBV in ssGBLUP can be eliminated by considering inbreeding in A, correction factors, and by reducing the additive variance. A comprehensive theory extending unknown parent groups to metafounders may automatically eliminate inflation of GEBV for arbitrarily complex populations including multibreed.

Key words: inflation, inbreeding, single-step method

Introduction

Genomic evaluation in dairy is usually validated by forward prediction using a regression of daughter yield deviation (DYD) of bulls with daughters on the genomic prediction obtained at an earlier time. Desired evaluations should have high reliability (high R^2) and a minimum inflation (parameter b_1 close to 1.0). In the first validation of single-step GBLUP (ssGBLUP) for Holsteins (Aguilar *et al.,* 2010), the parameter b_1 was as low as 0.7, indicating 1/0.7 inflation of GEBV. In contrast, little or no inflation was found for broilers, pigs and beef. The inflation could be reduced by ad-hoc modifications at a slight reduction of reliability, using parameters λ (Aguilar *et al.*, 2010; Harris *et al.,* 2012) and ω (Tsuruta *et al.,* 2013), however the meaning of those parameters was unknown.

In general, ssGBLUP relies on compatibility of genomic and pedigree relationships, and the inflation is likely caused by a mismatch. The issue of compatibility is a complex one. The GRM indirectly incorporates "infinite" pedigree but depends on scaling, gene frequencies, quality control and the number of SNPs. The pedigree relationships depend on pedigree depth and completeness, and on pedigree accuracy. While scale differences are likely to result in inflation/deflation, the differences in levels likely result in biases. Matching both relationship may involve matching G only, A_{22} only or both. The purpose of this paper is to look at the factors of ssGBLUP in dairy that influence the inflation of GEBV.

Materials and Methods

Matrix H and compatibility

Legarra *et al.* (2009) presented a matrix that combines pedigree and genomic relationships:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

This matrix assumes compatibility between pedigree and genomic relationships. Different scales would cause inflation/deflation of GEBV, while different levels would mostly cause biases. Other potential sources of inflation/deflation are incomplete pedigrees and pedigree errors.

Scaling genomic relationships

The most popular GRM is given by VanRaden (2008):

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{2P})(\mathbf{M} - \mathbf{2P})'}{\sum_i 2p_i\, q_i}$$

Such a matrix is a direct byproduct of a SNP model.

This matrix can be scaled using several methods.

1. Scale G by using gene frequencies from the base population. While theoretically an optimal solution, in practice such gene frequencies are not known and need to be estimated at a large cost (Gengler *et al.,* 2007). Also, most populations have heterogenous base populations.
2. Scale G by using equal allele frequencies with a fixed effect in the model. Different gene frequencies add a constant to GEBV. Therefore, a constant (or group effect) can be added to the phenotypes of genotyped animals. Such scaling accounts mainly for biases and not for inflation. Note that this scaling has no meaning when phenotypes are only for ungenotyped animals (e.g., production traits when only bulls are genotyped).
3. Scale G for compatibility with A. The scaling could be either by regression (VanRaden, 2008) or by enforcing the equality of means of diagonal and off-diagonal elements of G with those of A_{22}. This method was linked to the Fst index.

Ad-hoc adjustments for inflation

The inverse of matrix H is:

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

Two types of adjustments were found to reduce the inflation: lambda (Aguilar *et al.*, 2010):

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \lambda\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \end{bmatrix},$$

where the use of lambda equal to 0.7 increased b_1 from 0.76 to 0.88, and omega:

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

where the use of omega equal to 0.7 increased b_1 by about 0.15.

Effects of inbreeding and correction factors

GEBV for young animals can be presented as:

$$GEBV = w_1PA + w_2DGV - w_3PI.$$

When inbreeding is ignored in A and two parents are known, the formula is:

$$GEBV = \frac{2}{2 + g^{ii} - a_{22}^{ii}}\, PA + \frac{g^{ii}}{2 + g^{ii} - a_{22}^{ii}}\, DGV - \frac{a_{22}^{ii}}{2 + g^{ii} - a_{22}^{ii}}\, PI.$$

When inbreeding is considered, the formula changes to:

$$GEBV = \frac{2/(1-Fi)}{2/(1-Fi) + g^{ii} - a_{22}^{ii}}\, PA + \frac{g^{ii}}{2/(1-Fi) + g^{ii} - a_{22}^{ii}}\, DGV - \frac{a_{22}^{ii}}{2/(1-Fi) + g^{ii} - a_{22}^{ii}}\, PI,$$

where

$$F_i = (F_{sire} + F_{dam}).$$

Inbreeding increases the denominator, resulting in smaller GEBV comprised of a larger fraction of PA. This was found by Lawlor *et al*. (2010).

When inbreeding is ignored in A but the lambda parameter is used, the equation becomes:

$$\text{GEBV} = \frac{2/\lambda}{2/\lambda + g^{ii} - a^{ii}_{22}} \text{PA} + \frac{g^{ii}}{2/\lambda + g^{ii} - a^{ii}_{22}} \text{DGV} - \frac{a^{ii}_{22}}{2/\lambda + g^{ii} - a^{ii}_{22}} \text{PI},$$

noting that:

$$2/\lambda \equiv 2/(1 - F_i).$$

In the absence of (or incomplete) inbreeding, the parameter lambda accounts for average inbreeding. If individual inbreeding of young animals varies, the use of average inbreeding causes some accuracy, as found experimentally.

With the omega parameter:

$$\text{GEBV} = \frac{2/\omega}{2/\omega + g^{ii}/\omega - a^{ii}_{22}} \text{PA} + \frac{g^{ii}/\omega}{2/\omega + g^{ii}/\omega - a^{ii}_{22}} \text{DGV} - \frac{a^{ii}_{22}}{2/\omega + g^{ii}/\omega - a^{ii}_{22}} \text{PI}.$$

The omega parameter less than 1 decreases the fraction of PP and increases the denominator.

Calculation of inbreeding coefficients

Calculating inbreeding from the data depends on the depth and completeness of pedigrees. Truncating pedigrees to a few generations eliminates effects of old missing pedigrees with a minimal effect on accuracy (Pocrnic *et al.*, 2017). Missing pedigrees can be accounted for by assuming nonzero inbreeding for the phantom parents of unknown parent groups (UPG; Westell *et al.*, 1988) by using the VanRaden (1992) algorithm. Lutaaya and Misztal (1999) found that such an algorithm only partially recovered missing inbreeding, however, their algorithm had an error (Aguilar and Misztal, 2008).

Data

Tests involved 10 102 702 records of 18 type traits of 6 930 618 U.S. Holsteins, with genotypes for 576k animals.

Computations

The data were analyzed using the blup90iod2 software with the following options:

1. Inbreeding in A_{22} but not in A (A)
2. Inbreeding for both A and A_{22} (A INB)
3. Inbreeding including UPG for both A and A_{22} (A UPG)
4. As above with 50% reduction of additive variance (A UPG 50%).

Validation was by the formula:

$$\text{DYD2014} = b_0 + b_1\text{GEBV2010}.$$

Results and Discussion

Table 1 shows average b_1, average R^2 and correlation between b_1 and heritability.

Option	b_1	R^2	Corr(b_1,h^2)
(PA-BLUP)	0.76	0.18	0.25
A	0.75	0.48	0.65
A INB	0.85	0.49	0.59
A UPG	0.90	0.49	0.45
A UPG 50%	0.96	0.50	0.13

As A becomes more complete, the inflation decreases, and R^2 increases incrementally. Also, the correlation between b_1 and h^2 lowers but is still high even with UPG inbreeding considered. This strong correlation may result from stronger selection for more heritable traits, which modifies genetic variances differentially depending on the selection pathway. Lowering the additive variance (and h^2) reduces the correlation to almost 0, improves the b_1 to a point where it is sufficiently close to 1.0, and marginally increases R^2. Rationale for decreasing the heritability is in Wiggans *et al.* (2012).

The actual inflation in ssGBLUP may be less because DYD can be biased by preselection. Masuda *et al.* (2017) found that the EBV trend is less than the GEBV trend starting for bulls born in 2010, with lager differences for younger bulls. Those potentially biased EBV were then used to calculate DYD. This study used only data up to 2014 due to inability to obtain genotypes past 2014.

Inflation and other species

No inflation has been found in broilers, where pedigrees are complete and pedigrees included only 3 generations; adding more generations did not increase accuracy. Little or no inflation has been found in commercial evaluation of Angus in the US, where pedigrees are quite complete and selection is relatively weak. Inflation for Holstein in the US could be due to a substantial fraction of missing parentage and very strong selection for production.

Crossbred data and metafounders

Ensuring compatibility between A and G in the purebred with heterogeneous base populations or the multibreed context is more complex. The completeness of pedigree will still be an issue, but re-ranking due to scale may be automatically eliminated if a model includes breed effects. See Christensen *et al.* (2014) for a specialized case of a terminal cross. Out of many choices in pigs, ignoring breed differences was the best choice in Lourenco *et al.* (2016).

Automatic scaling for arbitrarily complex populations with missing pedigree may be possible using metafounders (Legarra *et al.*, 2015). As G accounts for all past pedigrees, this method proposes to adjust A to G as follows:

1. Create G using equal gene frequencies and some base scale.
2. Create as many UPG as necessary and call them metafounders.
3. Based on G, create a covariance matrix between metafounders.
4. Construct A and A_{22} using the metafounders and the covariance matrix.

In tests, ssGBLUP using the metafounders was superior for a crossbred prediction (Xiang *et al.*, 2017). In a simulated single population, ssGBLUP with metafounders was superior to regular ssGBLUP with inbreeding ignored but similar with inbreeding considered (Garcia-Baccino *et al.*, 2017).

Conclusions

Evaluations by ssGBLUP are inflated when the pedigree is long but incomplete and inbreeding in A is ignored. Inflation can be reduced by a combination of pedigree truncation, incorporation of inbreeding in A in addition to inbreeding for unknown parents, and by reducing the heritability.

Acknowledgements

This research was mainly supported by grants from Holstein Association USA (Brattleboro, VT) and by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 from the US Department of Agriculture's National Institute of Food and Agriculture (Washington, DC).

References

Aguilar, I. & Misztal, I. 2008. Recursive algorithm for inbreeding coefficients assuming non-zero inbreeding of unknown parents. *J. Dairy Sci. 91,* 1669-1672.

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2010. 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., Madsen, P., Nielson, B. & Su, G. 2014. Genomic evaluation of both purebred and crossbred performances. *Genet. Sel. Evel. 46:23.*

Garcia-Baccino, C.A., Legarra, A., Christensen, O.F., Misztal, I., Pocrnic, I., Vitezica, Z.G. & Cantet, R.J.C. 2017. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. *Genet. Sel. Evol. 49:34.*

Gengler, N., Mayeres, P. & Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. *Anim. 1,* 21-28.

Harris, B.L., Winkelman, A.M. & Johnson, D.L. 2012. Large-scale single-step genomic evaluation for milk production traits. *Interbull Bulletin 46,* 20-24.

Lawlor, T.J., Misztal, I., Tsuruta, S., Aguilar, I. & Legarra, A. 2010. Decomposition and interpretation of genomic breeding values from a unified one-step national evaluation. *Proc. 9th WCGALP,* Leipzip, Germany.

Legarra, A., Aguilar, I. & Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci. 92,* 4656-4663.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I. & Misztal, I. 2015. Ancestral relationships using metafounders: Finite ancestral populations and across population relationships. *Genet. 200,* 455-468.

Lourenco, D.A.L., Tsuruta, S., Fragomeni, B.O., Chen, C.Y., Herring, W.O. & Misztal, I. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. *J. Anim. Sci. 94,* 909-919.

Lutaaya, E. & Misztal, I. 1999. Inbreeding in populations with incomplete pedigrees. *J. Anim. Breed. Genet. 116:6,* 475-480.

Masuda, Y., VanRaden, P.M., Misztal, I. & Lawlor, T.J. 2017. Differing genetic trend estimates from traditional and genomic evaluations for genotyped animals as evidence of pre-selection bias in U.S. Holsteins. *Submitted.*

Pocrnic, I., Lourenco, D.A.L., Bradford, H.L., Chen, C.Y. & Misztal, I. 2017. Technical note: Impact of pedigree depth on convergence of single-step genomic BLUP in a purebred swine population. *J. Anim. Sci. 95,* 3391-3395.

Tsuruta, S., Misztal, I. & Lawlor, T. 2013. Short communication: Genomic evaluations of final score for US Holsteins benefit from the inclusion of genotypes on cows. *J. Dairy Sci. 96,* 3332-3335.

VanRaden, P. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *J. Dairy Sci. 75:11,* 3136-3144.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91,* 4414-4423.

Westell, R.A., Quaas, R.L. & Van Vleck, L.D. 1988. Genetic Groups in an Animal Model. *J. Dairy Sci. 71,* 1310-1318.

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

Xiang, T., Christensen, O.F. & Legarra, A. 2017. Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. *J. Anim. Sci. 95,* 1472-1480.