Use of bivariate EBV-DGV model to combine genomic and conventional breeding value evaluations

E.A. Mäntysaari and I. Strandén*

Introduction

In all genomic evaluations implemented so far, the prediction model is first calibrated using reliably tested reference bulls. The solutions are used to derive direct genomic evaluations (DGV) for candidate animals. The candidate animals have relatives with information not accounted in the calibration data. VanRaden et al. (2009) proposed to combine the DGV and the national estimated breeding values (EBV) into genomically enhanced breeding value (GEBV) by selection index approach.

In conventional evaluations the first crop daughters are assumed to be progeny of unselected sons of the bull sire. This is violated, if the bull calves are selected based on DGVs. Patry et al. (2009) and Liu et al. (2009) have suggested methods to reduce the bias. An obvious solution is to back blend the genomic information to all published EBVs. Van Doormaal et al. (2009) implemented this in the Canadian genomic evaluation by modifying the parent averages of the descendants of the genotyped animals according to parental genomic information.

Ducrocq and Liu (2009) suggested pseudo bull evaluations where DGVs were included as records. Information in DGVs are first separated into pedigree relationship information, and the genomic information. The latter they call genomic effective daughter contribution (gEDC). Once the gEDC are available, they de-regress the DGVs to genomic daughter performances (gEDP). In the pseudo evaluations, the daughter performances of bulls (DYD, Daughter yield deviations) and corresponding daughter numbers (EDC) are accompanied by gEDP and gEDC, and the genomic information will become combined into information of all ungenotyped animals.

The method of Ducrocq and Liu (2009) is an approximation because the DGV observations will not follow the same genetic rules as the DYDs. This is most obvious if a genotyped bull sire has several genotyped sons. His own genotype has all the information, and nothing more can be derived from the sons. Ducrocq and Liu (2009) made an constraint that gEDC cannot be negative. This constraint can lead into undesired increase of reliability (R^2_{EBV}) of bulls that have genotyped offspring.

Discrepancy in information from gEDC and EDC is due to difference in reliability in original observations. Genomic information can have 100% reliability, but daughter performance has much lower accuracy. The DGV can be treated as an indicator trait with a high genetic correlation to the considered trait. With a simple selection index calculation, one can show that a single observation of an indicator trait like DGV, gives a reliability R^2_{EBV} equal to r^2_g, where r_g is the genetic correlation between the traits. When the pseudo national evaluations suggested by Ducrocq and Liu (2009) are calculated with a bivariate model no approximation is needed, and the accuracy obtained through pedigree will be exactly accounted.

The purpose of this study is to illustrate the bivariate DGV-EBV approach for combining the information from genomic evaluations and from national EBV runs.

*MTT Agrifood Research, Biotechnology and Food Research, 31600 Jokioinen Finland

Material and methods

Simple bivariate model for DYD and DGV for the bull *i* is:

$$\begin{bmatrix} DYD \\ DGV \end{bmatrix}_i = \begin{bmatrix} Z_{1i} & 0 \\ 0 & Z_{2i} \end{bmatrix} \begin{bmatrix} a_{DYD} \\ a_{DGV} \end{bmatrix}_i + \begin{bmatrix} e_{DYD} \\ e_{DGV} \end{bmatrix}_i$$

where a_{DGV} and a_{DYD} are the genomic evaluation and the true breeding value for corresponding traits, respectively. For the genotyped animals Z_{2i} is one, and for the ungenotyped animals zero. Similarly, for animals that have no daughter records $Z_{1i} = 0$. For simplicity, assume that the data are scaled so that the variances of the breeding values become unity:

$$Var \begin{bmatrix} a_{DYD} \\ a_{DGV} \end{bmatrix} = \begin{bmatrix} \sigma^2_{a,DYD} & \sigma_{a(DYD,DGV)} \\ \sigma_{a(DGV,DYD)} & \sigma^2_{a,DGV} \end{bmatrix} = \begin{bmatrix} 1 & r_a \\ r_a & 1 \end{bmatrix} = \mathbf{G} \qquad (1)$$

Note that since the DGV is estimated directly from the genome, we can assume it is estimated without error. However, having a zero residual variance can not be used in a standard Mixed Model solving program. Therefore, the residual variance is assigned to be a small number (0.01):

$$Var \begin{bmatrix} e_{DYD} \\ e_{DGV} \end{bmatrix} = \begin{bmatrix} \sigma^2_{e,Y}/EDC_Y & 0 \\ 0 & 0.01 \end{bmatrix} \qquad (2)$$

where $\sigma^2_{e,Y} = {}^{(4-h^2)}/_{h^2}$ is the residual variance of the trait after scaling. Note that (1) implies $Var[a_{DYD}|a_{DGV}] = (1 - r_a^2)$ which suggests the proper r_a to be R^2_{DGV}, i.e., the accuracy of a_{DGV} in predicting a_{DYD}.

Use of individual reliabilities of DGV. Unlike the equations in VanRaden et al. (2009) and Ducrocq and Liu (2009) the (1) requires a same accuracy R^2_{DGV} of DGV for every animal. This can be relaxed by a modification inspired by random regression models. The design matrix of breeding values is modified so that the transformed breeding values have a simpler (co)variance structure. Let us replace Z_i by the Cholesky transformation of $\mathbf{G}_i = \mathbf{L}_i\mathbf{L}_i^t$:

$$\mathbf{Z}^*_i = \mathbf{L}_i = \begin{bmatrix} 1 & 0 \\ r_a & \sqrt{1 - r_a^2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \sqrt{r^2_{DGV,i}} & \sqrt{1 - r^2_{DGV,i}} \end{bmatrix} \qquad (3)$$

When $\mathbf{L}_i$ is a Cholesky decomposition of $\mathbf{G}_i$, (i.e. (1) but with the $R^2_{DGV,i}$ of the bull *i*), then the variance $var[\mathbf{L}_i\mathbf{a}_i] = \mathbf{L}_i var[\mathbf{a}_i]\mathbf{L}_i^t = \mathbf{I}$. So after the transformation the new breeding value coefficients are independent, and do not depend on $R^2_{DGV,i}$ of the individual bull. The original breeding values can be obtained by multiplying the solutions from transformed model by matrix $\mathbf{L}_i$. This, however, is unnecessary because the combined index GEBV $\hat{a}_{DYD}$ is not affected by the transformation.

Discounting the DGV information in bulls used as reference group. In the bivariate model information from DGV is used to improve accuracy of $\hat{a}_{DYD}$. If the DYD of a bull is already used in estimation of genomic prediction equations, the information will become double counted. This can be avoided by decreasing the weight of DYD information in (2) as $EDC^*_i = EDC_i - df_i$. The discount factor, $df_i = {}^{R^2_{DGV,i\in ref}}/_{(1-R^2_{DGV,i\in ref})}$, can be derived by absorbing the a_{DGV} equation into a_{DYD} in conceptual mixed model equations. Simpler alternative would be to decrease the $R^2_{DGV,i}$

of the bulls in reference bull group. This would have the desired effect on R^2_{DYD} of the bull itself, but would have opposite effect on the relatives of the reference bulls.

Example data. Ducrocq and Liu (2009) data was used to illustrate the method (Table 1). The pedigree has 14 animals of which one (animal 2) had no genomic information. Animals 5, 6 and 8-14 had the same sire. All genotyped animals were assumed to have $R^2_{DGV} = 0.4$. The example was modified to resemble a more realistic scenario. In the first modification, labeled B, animals 4, 5 and 6 were assumed to have $R^2_{DGV} = 0.6$. In scheme C, the animals 1,4,5 and 6 were assumed to also have 50 daughters each. Heritability of the trait was 0.36. In the final scheme (D) the progeny tested bulls 1, 4 and 5 were assumed to be included in a DGV calibration reference group.

Table 1: Pedigree and data availability of example animals [a]

Animal(s)	Sire	Dam	B R^2_{DGV}/EDC	C R^2_{DGV} /EDC	D[b] R^2_{DGV}/EDC
1	-	-	0.4 / -	0.4 / 50	0.4* / 50
2	1	-	-/-	-/-	-/-
3	-	-	0.4 / -	0.4 / -	0.4 / -
4	-	2	0.7 / -	0.6 / 50	0.6* / 50
5	4	-	0.7 / -	0.6 / 50	0.6* / 50
6	4	-	0.7 / -	0.6 / 50	0.6 / 50
7	4	3	0.4 / -	0.4 / -	0.4 / -
8-14	4	-	0.4 / -	0.4 / -	0.4 / -

[a] Base pedigree and DGV reliability as in Ducrocq and Liu (2009) [b] bulls with * are included in DGV calibration

Results and discussion

Ducrocq and Liu (2009) noted that their method tends to overestimate R^2_{EBV} of animals, like number 2, having many genotyped offspring . For the ungenotyped animal 2, their method gave too high (0.197) reliability while the bivariate method gave 0.160. When some animals have higher R^2_{DGV} for genomic evaluations, reliability of their relatives is also affected (scheme B). However, this effect is small compared to rapid increase in R^2_{EBV} when the bulls have own daughters (scheme C). As expected this affects reliability of their relatives much more than small change in R^2_{DGV}. Note that, here, the R^2_{EBV} based on 50 daughters would be 0.83. As the effect of DGV on reliability of progeny tested bulls was not large, the discounting of the information from the bulls in reference set had only minor effect (scheme D). For example, R^2_{EBV} for animal 1 dropped from 0.852 to 0.835.

The bivariate combining method is easier to automate than the selection index equations (de Roos et al. (2009); Ducrocq and Liu (2009); Van Doormaal et al. (2009)). The only requirement is a multivariate BLUP program that allows random regression models and weights for observations. A BLUP program will give optimal weights to DGVs and EBVs, and if the program can estimate the accuracies of estimates, also the R^2_{GEGV} are obtained.

Methods that directly combine the genomic information and phenotypic records have been presented (Aguilar et al. (2010); Christensen and Lund (2010)). When the genomic evaluations are solved directly, the solutions are directly GEBVs, and the genomic information will be transferred to all animals in pedigree. For some traits, however, the 2-stage genomic evaluations might be preferred,

since the simultaneous modeling of genomic and phenotypic data are based on simple BLUP models which assume equal variance for all the SNP effects.

Table 2: Estimated reliability of 14 example animals by method[a]

Animal(s)	D&L[b]	A	B	C	D
1	0.400	0.400	0.405	0.852	0.835
2	0.197	0.160	0.213	0.355	0.348
3	0.400	0.400	0.400	0.400	0.400
4	0.490	0.400	0.640	0.888	0.863
5	0.400	0.400	0.610	0.875	0.845
6	0.400	0.400	0.610	0.875	0.874
7	0.400	0.400	0.441	0.509	0.501
8-14	0.400	0.400	0.441	0.509	0.501

[a]In A all genotyped animals have $R^2_{DGV} = 0.4$; in B animals 4,5 and 6 have $R^2_{DGV} = 0.6$, in C animals 1,4,5,6 have 50 daughters, and in D the animals 1,4 and 5 were used to derive the genomic model. [b]Reliability estimated in Ducrocq and Liu (2009) for data with all 14 the animals, except 2, genotyped with a reliability 0.40.

Conclusion

The bivariate method was easy to implement and based on small example, results were logical. The strength of the method is the automatic operation over all relatives and across generations. The bivariate combining and blending equations were extended to cases where some bulls have been included in the reference group used for solving the genomic equations. This, however, had less effect than expected.

References

Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., and Lawlor, T. (2010). *J. Dairy Sci.*, 93(2):743–752.

Christensen, O. and Lund, M. S. (2010). *Genet. Sel. Evol.*, 42(2).

de Roos, A., Schrooten, C., Mullaart, E., van der Beek, S., de Jong, G., and Voskamp, W. (2009). In *Proc. Interbull Workshop, Bull.*, 39, pages 47–50.

Ducrocq, V. and Liu, Z. (2009). In *Proc. Interbull Meeting, Bull.*, 40, pages 172–177.

Liu, Z., Seefried, F., Reinhardt, F., and Reents, R. (2009). In *Proc. Interbull Meeting, Bull.*, 40, pages 184–188.

Patry, C., Ducrocq, V., and Patry, C. (2009). In *Proc. Interbull Meeting, Bull.*, 40, pages 167–171.

Van Doormaal, B., Kistemaker, G., Sullivan, P., Sargolzaei, M., and Schenkel, F. (2009). In *Proc. Interbull Meeting, Bull.*, 40, pages 214–218.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., and Schenkel, F. S. (2009). *J. Dairy Sci.*, 92(1):16–24.