A Modification of VanRaden's Index for the Blending of Genomic Breeding Values

C. Edel[*], R. Emmerling[*], K.-U. Götz[*]

Introduction

The term 'blending' is commonly used to describe a form of combination of outcomes of a genomic evaluation (direct genomic values, *dgv*) with information from standard genetic routine evaluations, i.e. the parent averages (*pa*) in the majority of relevant cases. The combined estimate is usually termed genomically enhanced breeding value (*gebv*). For this combination several authors have suggested an index-like procedure (de Roos et al. (2009); VanRaden (2008)) that has advantages over more complex methods (i.e. Ducrocq and Liu (2009)) due to the ease of use under practical conditions. However, more 'holistic' methods have been already formulated (Aguilar et al. (2009)). A special approach to our knowledge is that of Kachman (Kachman, personal communication) who uses multivariate variance component estimation techniques to estimate the (co-)variance structure of the information involved in calculating the *gebv*. Pursuing a similar intention, we propose an index of a general form, additionally introducing a scaling factor c. We use cross-validation techniques (*cv*) to derive trait-specific optimal values for c using the correlation between daughter yield deviations (*dyd*) and the *gebv* of the animals in the validation sets as the criterion guiding optimization.

Material and Methods

The Index Similar to the index proposed by VanRaden (2008) (VRI) our approach considers three sources of information about the true breeding value of an animal i (a_i):

- dgv_i the direct genomic value,
- ebv_i a polygenic breeding value without inclusion of genomic information estimated for the animal by using only the group of genotyped animals (the 'calibration' or 'learn'-set) and their phenotypic information (animal itself excluded)
- pa_i the polygenic breeding value estimate from the routine evaluation (usually the parent average).

All three sources of information are estimated breeding values, where

$$cov(a, \hat{a}) \;=\; Var(\hat{a}) \;=\; R^2 * \sigma_a^2.$$

R^2 denotes for the reliability of the estimation (squared accuracy) and σ_a^2 the true additive-genetic variance. Assuming these variables to be distributed multivariate normal

[*]Bavarian State Research Center for Agriculture, Institute of Animal Breeding, Poing/Grub, Germany

$$\left[\begin{array}{c|c} a & \\ \hline dgv & \\ ebv & \\ pa & \end{array}\right] \sim N\{\mathbf{0}, \mathbf{V}\} \sim N\{\mathbf{0}, \left[\begin{array}{c|c} V_{gg} & V_{gy} \\ \hline V_{yg} & V_{yy} \end{array}\right]\}$$

where

$$\mathbf{V} = \left[\begin{array}{c|ccc} \sigma_a^2 & cov(a,dgv) & cov(a,ebv) & cov(a,pa) \\ \hline & \sigma_{dgv}^2 & cov(dgv,ebv) & cov(dgv,pa) \\ & & \sigma_{ebv}^2 & cov(ebv,pa) \\ & symm. & & \sigma_{pa}^2 \end{array}\right]$$

$$= \left[\begin{array}{c|ccc} \sigma_a^2 & R_{dgv}^2\sigma_a^2 & R_{ebv}^2\sigma_a^2 & R_{pa}^2\sigma_a^2 \\ \hline & R_{dgv}^2\sigma_a^2 & cov(dgv,ebv) & cov(dgv,pa) \\ & & R_{ebv}^2\sigma_a^2 & cov(ebv,pa) \\ & symm. & & R_{pa}^2\sigma_a^2 \end{array}\right]$$

The covariance of estimated breeding values for a simple additive-genetic model might be formulated as

$$cov(\hat{a}_1, \hat{a}_2) = \boldsymbol{b}_1\boldsymbol{b}_2 cov(\boldsymbol{y}_1, \boldsymbol{y}_2) = R_1^2 R_2^2\, cov((a_1 + e_1), (a_2 + e_2))$$

treating the regression coefficients as constants. The problem reduces to the question of the degree of dependency in the residuals, resulting in

$$cov(\hat{a}_1, \hat{a}_2) = R_1^2 R_2^2\, \sigma_y^2 \quad \text{or} \quad cov(\hat{a}_1, \hat{a}_2) = R_1^2 R_2^2\, \sigma_a^2$$

in case of completely dependent or for completely independent residuals, respectively. In a similar situation Weigel et al. (1998) introduced a scaling factor *c* as a measure for the lack of independence. Following this nomenclature one might formulate

$$cov(\hat{a}_1, \hat{a}_2) = R_1^2 R_2^2\, (\sigma_a^2 + c * \sigma_e^2)$$

where c has the advantage to be bounded between 0 and 1. Division of the whole term by σ_a^2 results in

$$= R_1^2 R_2^2\, (1 + c * \lambda) \quad \text{with} \quad \lambda = \sigma_e^2/\sigma_a^2$$

Filling in the terms developed above, the following form of the matrix $\mathbf{V}_{yy}$ results:

$$\mathbf{V}_{yy} = \begin{bmatrix} R_{DGV}^2 & R_{DGV}^2 R_{EBV}^2\, (1 + c_1 * \lambda) & R_{DGV}^2 R_{PA}^2\, (1 + c_2 * \lambda) \\ & R_{EBV}^2 & R_{EBV}^2 \\ symm. & & R_{PA}^2 \end{bmatrix}$$

This part of the index is different to the formulation of VRI, where $\mathbf{V}_{yy}$ is:

$$\mathbf{V}_{yy} = \begin{bmatrix} R_{DGV}^2 & R_{EBV}^2 & R_{EBV}^2 + (R_{DGV}^2 - R_{EBV}^2)(R_{PA}^2 - R_{EBV}^2)/(1 - R_{EBV}^2) \\ & R_{EBV}^2 & R_{EBV}^2 \\ symm. & & R_{PA}^2 \end{bmatrix}$$

Optimal weights for the aggregation of the three sources of information of animal i result from solving the individual index equation of the animal

$$b_i \quad = \quad \mathbf{V}_{gy}\mathbf{V}_{yy}^{-1}$$

$c1$ and $c2$ are interpreted as population parameters and optimal values for these parameters were derived by using a sufficient number of cv-samples and a grid-search algorithm. However, preliminary studies showed that all in cases c1$\sim$c2 (results not shown), so the simplifying assumption c1=c2 was used and a linear optimization between 0 and 1 was applied.

Methods Tested

- **VRI** Van Raden's Index described above. Reliabilities were derived from the prediction error variances (*pev*) of the G-BLUP or the corresponding polygenic system. Reliabilities of *pa* were approximated by using $R^2_{PA} = 0.25(R^2_s + R^2_d)$. The accuracy of *gebv* (RHOGEBV) given in table 1 is the theoretical accuracy of the index.
- **ITZ** The Index described above using the simplifying assumption that c1=c2. Trait-specific c were derived using 12 cv-samples and a grid-search approach. For all traits the c maximizing the correlation between the *gebv* and the *dyd* of the animals in the validation set is consecutively considered to be the optimal value. The correlations shown in table 1 are derived at average optimal c over all cv-samples. The accuracy of *gebv* given in table 1 is the theoretical accuracy of the index.
- **Simple** *dgv* and *pa* are aggregated using weights that are reciprocals of their respective *pev* ($pev^{-1}_{dgv}/\sum pev^{-1}$ und $pev^{-1}_{ebv}/\sum pev^{-1}$). This approach does not account for a nonzero covariance between these two sources of information. The accuracy of the *gebv* given in table 1 is a simple combination of the reliabilities of *dgv* and *pa* respectively by summing performance equivalents (e.g. Edel et al. (2009)).
- **CRV** A 'QTL-contribution' (*cont*) is calculated by $cont_i = dgv_i - ebv_i$. The value of *cont* is then added to *pa* without weighting. This is basically the approach of de Roos et al. (2009). However, instead of using a sire-maternal grandsire *pa* we use a standard sire-dam *pa* in our case. The accuracy of *gebv* given in table 1 is calculated following de Roos et al. (2009).

Material The comparison of methods was based on 2444 genotyped Fleckvieh bulls. Genotyping was done using the Illumina 54k SNP-chip. Birthyears of genotyped bulls ranged from 1979 to 2003 with a focus on birthyears 1995 to 2002. Phenotypic information used for the analysis was taken from the November 2009 routine evaluation for milk production traits. Daughter yield deviations and weights were calculated following Edel et al. (2009). Direct genomic values were derived using a standard G-BLUP approach (VanRaden, 2008). For each of the 12 cross-validation samples, validation sets of size 194 were sampled from birth years >2000. All animals in the validation sets had breeding values based on daughter information in the latest routine evaluation (R^2 >0.9 for the trait milk yield). The remaining 2250 genotypes were used as the calibration-set. For the calculation of *gebv* with the methods described above, historical parent averages and reliabilities were taken from the August 2004 routine evaluation.

Results and Discussion

The results of the comparison are summarized in table 1. Method ITZ performs as well as VRI

Table1: Result of the comparison of methods. GEBVtoDYD and GEBVtoEBV are correlations to the *dyd* and *ebv* from the latest routine evaluation. Estimates are averages over all *cv*-samples. RHOGEBV is an accuracy theoretically derived (see text). MY, FY and PY denote for the traits milk yield, fat yield and protein yield respectively.

		GEBVtoDYD	GEBVtoEBV	RHOGEBV
MY	VRI	0.616	0.659	0.677
	ITZ ($c = 0.37$)	0.611	0.658	0.665
	Simple	0.603	0.645	0.784
	CRV	0.602	0.644	0.833
FY	VRI	0.558	0.603	0.676
	ITZ ($c = 0.29$)	0.572	0.617	0.675
	Simple	0.533	0.579	0.768
	CRV	0.528	0.575	0.805
PY	VRI	0.548	0.605	0.675
	ITZ ($c = 0.29$)	0.568	0.628	0.678
	Simple	0.525	0.581	0.752
	CRV	0.514	0.576	0.776

with MY and shows an increasing advantage over the other methods with FY and PY. Methods ITZ and VRI additionally show the smallest deviations between realized accuracies and the accuracies theoretically derived. With ITZ, finding an optimal value of c didn't cause any problems. For each of 12 *cv*-samples a unique maximum was identified inside the parameter space defined by the approach. The advantage found for ITZ is considered large enough to justify the additional calculations based on *cv*-samples.

Acknowledgements Part of the SNP-genotypes of bulls used in this study were gratefully contributed by the association of Austrian Fleckvieh breeders (AGÖF).

References

Aguilar, I., Misztal, I., Johnson, D., Legarra, S., Tsuruta, S., and Lawlor, T. (2009). *J. Dairy Sci.*, 93:743–752.

de Roos, A., Schrooten, C., Mullaart, E., van der Beek, S., de Jong, G., and Voskamp, W. (2009). *Interbull Bulletin*, 39:47–50.

Ducrocq, V. and Liu, Z. (2009). *Interbull Bulletin*, 40:172–177.

Edel, C., Emmerling, R., and Götz, K.-U. (2009). *Interbull Bulletin*, 40:178–183.

VanRaden, P. (2008). *J. Dairy Sci.*, 91:4414–4423.

Weigel, K., Lawlor, T., VanRaden, P., and Wiggans, G. (1998). *J. Dairy Sci.*, 81:2040–2044.