A Note on using 'Forward Prediction' to Assess Precision and Bias of Genomic Predictions

C. Edel, S. Neuner, R. Emmerling and K.U. Götz
Bavarian State Research Center for Agriculture, 85586 Poing-Grub, Germany

Abstract

It has been argued repeatedly that cross-validation (CV) correlations should be used as a benchmark to assess matters of precision and bias of genomic predictions (GEBV) in practical applications. Irrespective of the fact that CV in this discussion is used in the very limited meaning of doing one single forward-prediction, we show by the use of standard formulae and simple simulation techniques that in traits underlying selection correlations derived from these techniques are considerably influenced by effects of selection on observed variances. As a consequence the squared correlations will be underestimations of the true precision of the estimates and the degree of underestimation depends on the number of selection steps, the selection intensity applied in any of them and the selection criteria applied in each step. The underestimation might be severe, so that without some reasonable assumptions about the selective conditions in a validation group no general conclusion about the precision and bias of genomic estimates can be drawn from a forward-prediction or comparable CV procedures. Additionally, we show that although correlation-measures are influenced by the effects of selection, linear combinations of estimates (e.g. differences between genomic predictions and daughter-based conventional estimates) are not much affected by the effects of selection. It is argued that the analysis of these differences could be a helpful extension to common validation tests for GEBV. We suppose that the aspects covered by this investigation will become increasingly important in the near future, when validation groups eventually will consist entirely of animals preselected on GEBV. However, in this case even the underlying conventional breeding value estimation will be influenced by the effects of selection.

Key words: genomic selection, cross-validation, forward-prediction, validation, selection effects

Material and Methods

All results were derived using assumptions based on multivariate normal distributions of true breeding values (TBV) and BLUP estimates hereof. In the most general form this joint distribution can be formulated as:

$$\begin{bmatrix} TBV \\ PA \\ GEBV \\ EBV \end{bmatrix} \sim N\{\boldsymbol{\mu}, \mathbf{V}\}$$

Following standard theory the variance of a BLUP estimate is the product of the reliability (R^2) and the assumed additive-genetic variance. Assigning an arbitrary mean of 100 and an additive-genetic variance of 144 gives:

$$\begin{bmatrix} TBV \\ PA \\ GEBV \\ EBV \end{bmatrix} \sim N\{\begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \end{bmatrix}, \begin{bmatrix} R^2_{TBV} & R^2_{PA} & R^2_{GEBV} & R^2_{EBV} \\ R^2_{PA} & R^2_{PA} & R^2_{PA} & R^2_{PA} \\ R^2_{GEBV} & R^2_{PA} & R^2_{GEBV} & R^2_{M1M2} \\ R^2_{EBV} & R^2_{PA} & R^2_{M1M2} & R^2_{EBV} \end{bmatrix} * 144\}$$

where TBV denotes the true breeding value, PA denotes the mean of estimated breeding values of parents or parent average, GEBV denotes the genomically enhanced breeding value and EBV a conventional BLUP estimate of the breeding value that already contains information on daughters. The PA is assumed to be a result from a breeding value estimation before daughter information on the animal was available and can thus be termed 'historical'.

Information available for the estimation of this PA is assumed to be a subset of the information available for the EBV. The GEBV is a combination of this 'historical' PA and the direct genomic value (DGV) that also does not contain information on daughters of the animal. In a standard two step procedure this combination is commonly the result of some sort of 'blending' (VanRaden, 2009; Harris and Johnson, 2010). As a consequence, the information available to estimate the 'historical' PA is also a complete subset of information in GEBV. Further

$$R^2_{M1M2} = R^2_{PA} + (R^2_{GEBV} - R^2_{PA})(R^2_{EBV} - R^2_{PA})/(1 - R^2_{PA})$$

by using a modification of a formulation originally developed by VanRaden (2009) in the context of 'blending'. This formulation can be shown to be approximately equivalent to formulations given by Harris and Johnson (1998) and Harris and Johnson (2010) in the context of the combination of two conventional estimates of the breeding value or the combination of DGV and conventional estimates, respectively. Both, the VanRaden approach and the approach of Harris and Johnson have been shown to give excellent approximations (Christensen and Lund, 2010; Ducrocq and Schneider, 2007). Both formulations can be interpreted without making reference to genomics, e.g. as a combination of two estimates of the breeding value including mendelian sampling, where the information on mendelian sampling originates from two independent sources (e.g. two independent groups of daughters) but estimates share common information on PA.

Studying effects of selection on means and (co-)variances of this joint distribution can either be done by simulation using sequences of pseudo random numbers from the multivariate distribution defined above or by explicit calculation of conditional means and (co-)variances. The first approach has been used to develop the figures used in this investigation, to prove the validity of explicit formulations where necessary and to assess scenarios with multiple steps of selection, where after the first selection step the assumption of multivariate normality does no longer hold. Simulations were done using the statistical software R (R Development Core Team, 2011). Explicit conditional means were calculated by standard formulae for multivariate distributions where the expectation of a set of variables A conditional on variable B is:

$$E(A|B) = \boldsymbol{\mu}_A + \mathbf{V}_{AB}\mathbf{V}_B^{-1}(B - \boldsymbol{\mu}_B)$$

Explicit derivations for conditional (co-) variances under selection were derived by the following formula given by Henderson (1975) where the joint covariance of a set of variables A conditional on a variable B on which selection has been applied is:

$$Var(A|B) = \begin{bmatrix} \mathbf{V}_A - \mathbf{V}_{AB}\mathbf{V}_0\mathbf{V}_{BA} & \mathbf{V}_{AB}\mathbf{V}_B^{-1}\mathbf{V}_{B_s} \\ \mathbf{V}_{B_s}\mathbf{V}_B^{-1}\mathbf{V}_{BA} & \mathbf{V}_{B_s} \end{bmatrix}$$

where

$$\mathbf{V}_0 = \mathbf{V}_B^{-1}(\mathbf{V}_B - \mathbf{V}_{B_s})\mathbf{V}_B^{-1}$$

and

$$\mathbf{V}_{B_s} = (1 - k) * \mathbf{V}_B$$

is the variance in the selected trait after selection. Assuming truncation selection on a trait that is normally distributed, the factor k depends on the intensity of selection i and the truncation point x resulting in

$$k = i(i - x)$$

(Falconer and Mackay, 1996).

Results & Discussion

Results for four different selection scenarios including selection on PA, selection on GEBV, selection on EBV and a combination of selection on PA and EBV are summarized in table 1. All results were derived by assuming reliabilities of 0.38, 0.65 and 0.89 for PA, GEBV and EBV. These reliabilities were used to construct the covariance matrix presented in material and methods. This covariance structure was assumed to characterize the true distribution of TBV and BLUPs before selection was applied. Selection on EBV was included to represent cases of so called 'selective genotyping'. Selection on GEBV was included as a proof of concept only, because in this case we would expect an additional biasing effect on underlying conventional EBV (Patry, 2011), something that is not covered by this investigation. In all cases the squared correlations, estimated after selection was applied, were considerably lower than before selection. However, the prediction error standard deviation (peSD) did not change, so that no direct conclusion can be drawn about the precision of the estimate from squared correlations estimated from a forward prediction. These "re-estimated" squared correlations are underestimates of the true precision in all cases, as can be shown by calculating the resulting standard deviations of GEBV-TBV and GEBV-EBV (peSD and peSD proxy) under the assumption that the re-estimated correlations are true (values in parentheses). With selective genotyping inspection of the peSD and the peSD proxy in both cases indicates, that these deviations are in fact even lower than expected in the case of no selection. However, in contrast to the effect on squared correlations, the effect of selection on these deviations is comparably weak. Corresponding values might be easily derived from any validation and - together with an inspection of mean-differences expected under selection (see table 2) - might be useful to assess whether a GEBV is a biased estimate or not.

Conclusions

Although it might be intuitive to assume that squared correlations derived from forward prediction should relate to precision of GEBV, we have demonstrated that this is not the case for traits under selection. It is true that these values provide information with respect to achievable selection response. However, farmers comparing different types of estimates of breeding values are in the majority interested in how much a future, more reliable estimate will deviate from the estimate they used to take their selection decision. This deviation is what we described here as the 'peSD proxy' and we have shown that this estimate is not directly related to the re-estimated squared-correlation. It might be argued, that using squared correlations as precision estimates should at least be 'conservative'. However, this should be done with some care and might lead to incorrect conclusions when it comes to validations of GEBV, e.g. in the case of the Interbull GEBV test. The method presented here can be adopted to calculate expected intercepts and slopes for regressions of TBV on GEBV or EBV on GEBV, respectively. This provides further insight into how selection influences these criteria (Edel et al., in preparation).

References

Christensen, O.F. & Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. *Gen. Sel. Evol., 42,* 2.

Ducrocq, V. & Schneider, M. del P. 2007. Generalization of the information source method to compute reliabilities in test day models. *Interbull Bulletin, 37,* 82-87.

Falconer, D.S. & Mackay, T. 1996. *Introduction to quantitative genetics.* Fourth Edition. Prentice Hall. ISBN 0582-24302-5.

Harris, B.L. & Johnson, D.L. 1998. Approximate reliabilities of genetic evaluations under an animal model. *J. Dairy Sci, 81,* 2723-2728.

Harris, B.L. & Johnson, D.L. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci, 93,* 1243-1252.

Henderson, C.R. 1975. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics 31,* 423-447.

Patry, C. 2011. Impacts of genomic selection on classical genetic evaluations. *Doctoral thesis, AgroParisTech 2011.*

R Development Core Team 2011. *R: A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci, 92,* 16-24.

Table 1. Effects of selection on squared correlations (ρ^2) and standard deviations of TBV-GEBV (peSD) for four different selection scenarios. Selection on EBV is included to represent cases of selective genotyping. 'peSD proxy' denotes the standard deviation of GEBV-EBV. The values in parentheses are peSD expected under the (wrong) assumption that *a posteriori* estimates of squared correlations directly relate to the precision of the GEBV.

selection applied to	**proportion selected (%)**	$\rho^2_{TBV\ GEBV}$	$\rho^2_{EBV\ GEBV}$	**peSD true**	**peSD proxy**
PA	100	0.65	0.63	7.10	6.95
	25	0.51	0.45	7.10 (8.42)	6.95 (8.88)
GEBV	100	0.65	0.63	7.10	6.95
	25	0.31	0.29	7.10 (9.97)	6.93 (10.12)
EBV	100	0.65	0.63	7.10	6.95
	75	0.51	0.47	6.65 (8.43)	6.49 (8.71)
PA/EBV	100	0.65	0.63	7.10	6.95
	25/75	0.40	0.33	6.40 (9.34)	6.24 (9.86)

Table 2. Effects of selection on averages.

selection applied to	**proportion selected (%)**	**Ø TBV**	**Ø PA**	**Ø GEBV**	**Ø EBV**
PA	100	100.0	100.0	100.0	100.0
	25	109.4	109.4	109.4	109.4
GEBV	100	100.0	100.0	100.0	100.0
	25	112.3	107.2	112.3	111.4
EBV	100	100.0	100.0	100.0	100.0
	75	104.8	102.0	103.2	104.8
PA/EBV	100	100.0	100.0	100.0	100.0
	25/75	113.3	109.9	111.4	113.3