Combining Genomic and Conventional Data in the Dutch National Evaluation

W.M. Stoop, H. Eding, M.L. van Pelt an G. de Jong
CRV, P.O. Box 454, 6800 AL Arnhem, The Netherlands
E-mail: Marianne.Stoop@crv4all.com

Abstract

In 2010, the Netherlands introduced Genomically Enhanced Breeding Values (GEBV) in their national evaluation. These GEBV were based on a post-processing step. For further optimization, a very basic method is proposed to incorporate the DGV based on genomic data directly into the conventional breeding value estimation: mass-selection.

The DGV are transformed to pseudo-observations which are analyzed as correlated trait (to the trait of interest) with a mass-selection model. The correlation between the trait of interest and this pseudo-trait is 1, and the heritability of the pseudo-trait is set to the predictive value of the DGV (estimated from cross-validation). In this way the pseudo-trait EBVaffects the EBV of the trait of interest of genotyped individuals, which in turn affects the EBV of relatives, effectively dropping the genomic information down the pedigrees of the trait of interest.

Results for a testrun on Udder health data are shown as application of this method.

1. Introduction

In 2010, the Netherlands introduced Genomically Enhanced Breeding Values (GEBV) in their national evaluation. These GEBV were based on a post-processing step (Van Raden *et al.*, 2009), where Direct Genomic Values (DGV) derived from deregressed proofs and SNP information, and conventional EBV were integrated. An obvious drawback of this method is that genomic information of an individual does not influence the breeding value of relatives.

Several methods have been proposed in literature to directly incorporate genomic information in the conventional breeding value estimation (e.g. Misztal *et al.*, 2009; Forni *et al.*, 2011, and Ducrocq and Liu, 2009). Especially the method where DGV are transformed into equivalent daughter performances (pseudo-records) is appealing, because 1) it corrects for pre-selection of young bulls (bulls with low DGV are no longer used as test bull, thus bull dams only get their best sons tested, adding the DGV avoids bull dams to be severely over-estimated due to selected testing of their best sons), 2) it allows genomic information to influence the (G)EBV of relatives (thus increasing the reliability), 3) the genomic information can quite easily be incorporated in existing BLUP software and is computationally feasible, 4) pseudo-records allow for the inclusion of DGV, rather than genotypes, and is therefore compatible with the current structure of reference populations and ownership of data.

The genomic part of the DGV (i.e. the markereffect) is a more or less simple summation of SNP effects and once you know the SNP the animal carries, the genotype of relatives does not add any information. Hence the genomic part of the DGV is independent of an animals pedigree or offspring information. This implies that pseudo-records derived from DGV must not be analyzed in the usual multi-trait animal model incorporating the numerator relationship matrix. In stead, pseudo-records derived from DGV should be analyzed using a mass-selection model.

Thus, pseudo-records for a certain trait of interest will be analyzed as a correlated trait – the pseudo-trait – to the original trait of interest.

This paper describes a method to combine genomic and conventional data, using a mass-selection model for the pseudo-records.

2. Material and Methods

2.1 Deriving pseudo-records

In mass-selection, the breeding value of an animal for a certain trait is a function of the deviation of the observation from the population mean and the heritability of this trait.

$$u_{ij} = h_j^2 \cdot \left(y_{ij} - \bar{y}_j\right)$$

where u_{ij} is the breeding value for the *i*-th animal for trait *j*, y_{ij} is the observation and $\bar{y}_j$ is the population mean.

If the breeding value - in this case the DGV - is known, the pseudo-observation that would result in that DGV is easily derived from the above:

$$y_{ij} = DGV_{ij} / h_j^2$$

In a mass-selection model, a single observation thus derived will result in a breeding value for the pseudo-trait equal to DGV_{ij} with reliability h_j^2.

2.2 Parameters

When the genotypes of all animals are determined with the same DNA-chip, and based on the same reference population, the genomic part of the DGV (i.e. markereffect) will have equal predictive reliability (γ^2). Setting the heritability of the pseudo-trait to the predictive reliability γ^2 of the DGV-markereffect, one pseudo-record per genotyped animal is sufficient to obtain a breeding value for the pseudo-trait equal to the DGV with a reliability γ^2.

The predictive power of the DGV-markereffect is usually assessed through cross-validation (De Roos *et al.*, 2009). In this procedure a certain cohort of sires, with sufficient offspring for reliable conventional breeding values (EBV), is genotyped and have their DGV predicted, including pedigree, but ignoring the data on offspring, and compared to a run without genomic data, so with pedigree only. The reliability of a breeding value is equal to the square of the correlation between estimated breeding value and true breeding value r_{CV}. Thus the predictive reliability of the DGV (γ^2) is the difference of the squared correlations between the DGV and the pedigree only, corrected for the average reliability γ_{EBV}^2 of the cohort of sires, and thus:

$$h^2 = \gamma^2 = (r^2_{DGV} - r^2_{Pedigree}) / \gamma_{EBV}^2$$

If multiple DNA-chips have been used, with different predictive value (e.g. a 500k chip versus a 6k chip), the heritability could be set to the lowest predictive reliability. Observation records must then be weighted according to which chip was used to arrive at an animals DGV with corresponding reliability (Mäntysaari and Strandén, 2010). Alternatively, a system where animals have repeated records allows for differences in reliability of the DGV. Allowing for repeated records (without a permanent environment effect), the reliability of a breeding value is:

$$r_{ij}^2 = \frac{N_i \cdot h_j^2}{1 + \left(N_i - 1\right) \cdot h_j^2}$$

Re-arranging the above, the number of repeated records needed to obtain a breeding value with a given reliability, for a trait with a heritability equal to the lowest predictive reliability γ^2:

$$N = \frac{r_{ij}^2\left(1 - \lambda_j^2\right)}{\gamma_j^2\left(1 - r_{ij}^2\right)} = \frac{EDC\left(r_{ij}^2\right)}{EDC\left(\gamma_j^2\right)}$$

Note that the number of pseudo-records *N* is always≥ 1, because h² is set equal to the lowest predictive reliability γ^2.

Both DGV and EBV are estimates on the same trait. In other words, both are estimates of the cumulative effect of the same QTL involved in the trait. If both DGV and EBV are estimated without error, the expectation is that DGV = EBV. Thus the genetic correlation between trait and pseudo-trait is 1.

It follows that the genetic correlation between a pseudo-trait, modelled on a trait of interest, and a third trait is equal to the correlation between the trait of interest and the third trait. And since the pseudo-trait and trait

of interest are modelled with the same genetic variance, the covariances for the pseudo-trait are also equal to that of the trait of interest and third trait(s). However, this implies that the covariances for the pseudo-trait are linear combinations of other covariances in the model-matrices. This causes singularity.

To lift the singularity, the covariances of the pseudo-trait with all other traits may be multiplied by a factor ß. To ensure full utilisation of the information of DGV in the EBV estimates, the heritability of the pseudo-traits should then be multiplied by $ß^{-2}$. This raises the apparent heritability of the pseudo-trait, but compensates for the loss of correlation caused by the multiplication with ß in the covariances. Note that when ß is chosen equal to the square root of the heritability (i.e. equal to γ), this model becomes identical to the model proposed by Mäntysaari and Strandén (2010), with h^2 = 1 and genetic correlation between trait and pseudo-trait = γ.

In the mixed model equations all 'conventional' traits are analyzed using the full relationship matrix A^{-1}. For pseudo-traits and covariances where at least one of the traits is a pseudo-trait, the relationship matrix is replaced by an identity matrix I (contrary to Mäntysaari and Strandén, 2010).

Table 1. Correlations in BV and differences in Reliability between a breeding value estimation where pseudo-records were either analyzed single trait or multitrait, for 3 groups of bulls: those without own pseudo-record, those with pseudo-record and daughter information, and those with pseudo-record, but without daughter information.[1,2]

		No PSR	PSR, with daughters	PSR, no daughters
UDH_index	Corr_BV	0.997	0.979	0.839
	Diff_Rel	0.0	+1.0	+13.2
PSR_UDH	Corr_BV	-0.005	0.977	0.962
	Diff_Rel	+16.5	+21.5	+0.9

[1] diff_BV is the absolute difference between the relative breeding values ~N (100, 4).
[2] diff_rel is the absolute difference between the reliabilities (reliability scale 0-100).

2.3 Reliability of GEBV

Because the genetic correlation between pseudo-trait and the trait-of-interest is 1, the full reliability of the breeding value for the pseudo-trait is expressed in the reliability of the (G)EBV of the trait of interest. In terms of expected daughter contributions (EDC):

$$EDC_{GEBV} = EDC_{TOI} + EDC_{PSEUDO}$$

If you rewrite this to the derivation of the reliability, it follows that

$$rel_{GEBV} = \frac{EDC_{GEBV}}{\alpha + EDC_{GEBV}} \quad \wedge \quad \alpha = \frac{4 - h^2_{GEBV}}{h^2_{GEBV}}$$

Then the reliability of the GEBV incorporating conventional and pseudo-trait data is:

$$rel_{GEBV} = \frac{r_{TOI} + rel^2_{pseudo} - 2r_{TOI}\, rel^2_{pseudo}}{1 - r_{TOI}\, rel^2_{pseudo}}$$

Which agrees with results from the "Information source method" by Harris and Johnson (1998). Moreover, the reliability of the GEBV estimated in with this method is equal to the reliability of GEBV using the method of post-process integration by Van Raden *et al.* (2009).

3. Application

A first testrun using Udder Health data from the April 2011 evaluation was performed. In Table 1 a comparison is made between a testrun where no correlations between pseudo-record (PSR_UDH) and conventional trait (UDH_index) was used (single trait setting),

and a testrun where a full correlation matrix was used (multi-trait setting). Data of three groups of bulls was analyzed: those without own pseudo-record (no PSR), those with pseudo-record and daughter information (PSR, with daughters), and those with pseudo-record, but without daughter information (PSR, no daughters).

For bulls without pseudo-record, differences in the UDH_index (the trait of interest) between the runs are neglectable. This is expected, as the majority of bulls have no genotyped relatives and thus the pseudo-records do not affect their EBV for the UDH_index. However, if we analyze a subset of these bulls, the non-genotyped Holstein bulls with genotyped relatives, the reliability of the UDH_index increases on average 1 percent (data not shown). Note that the reliability for the pseudo-UDH trait increases significantly (+16.5) in the multitrait setting, because only in the multitrait setting information from the UDH_index flows to PSR_UDH.

Bulls with pseudo-record and daughter information saw some reranking in (G)EBV (correlation is 0.979) and increase slightly in reliability for their UDH_index, which means the pseudo-records still adds some info to the trait of interest. Note again that the daughter information adds information to the pseudo-record (diff_rel +21.5) in the multitrait setting.

For bulls with pseudo-records but no daughter information the effect on the trait of interest is most defined: their reliability for the UDH_index increases 11-13 percent. They also show more changes in the (G)EBV, decreasing the correlation between the two runs to 0.839 for the UDH_index. Note that for these young bulls the differences in reliability for the pseudo-trait is small: there is no daughter information to affect the pseudo-trait.

4. Conclusions

In this study we included DGV (transformed to pseudo-records) in the conventional breeding value estimation. These DGV pseudo-records were included as a correlated trait, with a h^2 equal to the predictive reliability estimated from the cross-validation (of the lest predictive micro-array), and had a correlation with the original trait of 1. Pseudo-record traits were analyzed with a mass-selection model, where no numerator relationship matrix is included.

First results look promising and validate expectations from theory.

5. References

De Roos, A.P.W., Schrooten, C., Mullart, E., van der Beek, S., de Jong, G. & Voskamp, W. 2009. Genomic Selection at CRV, *Interbulletin 39*, 47-50.

Ducrocq V. & Liu, Z. 2009. 'Combining genomic and classical information in national BLUP evaluations'. *Interbull Bulletin 40*, 172-177.

Forni, S., Aguilar I. & Misztal I. 2011. 'Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information' *Genetics Selection Evolution 43:1.*

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

Mäntysaari E.A. & Strandén I. 2010. 'Use of bivariate EBV-DGV model to combine genomic and conventional breeding value evaluations'. *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany.

Misztal, I., Aguilar, I., Johnson, D., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2009. 'Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information', *Interbull Bulletin 40*, 240-244.

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.