Assessment of Cross-validation Strategies for Genomic Prediction in Cattle

M. Erbe, E.C.G. Pimentel, A.R. Sharifi and H. Simianer[*]

Introduction

The basic idea of cross-validation procedures is to divide a data set into a reference and a validation set, to omit any kind of information of the validation set and to predict this information, e.g. phenotypes, with a model trained exclusively in the reference set. The accuracy of prediction can be used to evaluate the underlying model and to compare alternative models. In the field of genomic selection, cross-validation can be used for assessing the accuracy of genomic breeding values (GEBVs) predicted with a specific model (e.g. Blonk et al., 2010; Luan et al., 2009) and for comparing the quality of different approaches used for estimation of GEBVs (Lund et al., 2009; Goddard and Hayes, 2007). However, the way of subdividing the data set is known to influence the results obtained by cross-validation (Luan et al., 2009; Lee et al., 2008). Therefore, we studied the changes in results when using different numbers of animals for the reference and the validation set and tried to find optimal subdivision strategies for the different objectives of cross-validation.

Material and Methods

Data. We used a sample of 2,294 Holstein bulls, which were genotyped with the Illumina 50K SNP chip. SNPs with a minor allele frequency lower than 5%, with missing position or a call rate lower than 95% were excluded. After filtering, there were 39,557 SNPs remaining for further analyses. Missing genotypes at these SNP positions were imputed using fastPHASE (Scheet and Stephens, 2006). All bulls had pedigree information and breeding values for somatic cell score with an accuracy > 0.87, which were used as quasi-phenotypes for the following analyses.

Methods to predict the GEBVs. We used two best linear unbiased prediction (BLUP) models for the estimation and prediction of the GEBVs. The first model included a random genomic and a random polygenic effect (Model A), while the second one included a random genomic component only (Model B). For model A, we fitted

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{Wg} + \mathbf{e}$$

where *y* is a vector of the phenotypes (breeding values for somatic cell score) for all bulls in the reference set, μ is the overall mean, **Z** is the incidence matrix for the random polygenic effect, *u* is a vector containing a random polygenic effect for each individual, **W** is the incidence matrix for the random genomic effect, *g* is a vector containing a random genomic effect for each animal and *e* is a vector of random residual terms. *u* is assumed to follow $N(0, \mathbf{A}\sigma_u^2)$ where **A** is the pedigree based relationship matrix. *g* is distributed

[*] Georg-August-University, Animal Breeding and Genetics Group, 37075 Goettingen, Germany

$g \sim N(0, \mathbf{G}\sigma_g^2)$ where **G** is a marker based relationship matrix, which was built according to VanRaden (2008) based on all available SNPs (n=39,557). In model A, the total breeding value was the sum of the polygenic and the genomic breeding value.

In model B, the polygenic component was omitted, hence the model was

$$\mathbf{y} = \mu + \mathbf{Wg} + \mathbf{e}.$$

Here, information of only one eighth of all available SNPs (n=4,945) was used to build **G**. The prediction accuracy of model A is expected to exceed the one of model B. As can be seen in the model design, we did not estimate an effect for each single SNP, but used the genomic relationship matrix to model a genomic effect for each individual. It was thus possible to estimate variance components in each step in each replicate by using ASReml (Gilmour et al., 2009). With the corresponding variance components, effects were estimated and GEBVs for the bulls in the validation set were predicted.

Cross-validation procedure. The whole data set was divided into a reference and a validation set. Phenotypes of the animals in the validation set were assumed to be unknown. First, a random sample of 100 animals was drawn for the validation set while the remaining 2,194 bulls built the reference set. In the next step, the size of the validation set was increased by 100 by moving 100 randomly chosen individuals from the reference to the validation set. This was done stepwise until 1,500 bulls were in the validation set. For each step, GEBVs were predicted for the bulls in the validation set with the information from the animals in the reference set. The whole procedure was repeated 60 times.

Criteria for comparison. Pearson's correlation coefficient between the realized and the predicted phenotypes for the animals in the validation set was calculated for each step in each replicate. In case the model did not converge during the process of variance component estimation, the correlation for this step in the particular replicate was considered to be NA.

The correlation between realized and predicted phenotypes was also used for testing whether the models differed significantly from each other. Therefore, the correlation coefficients were transformed so that they follow approximately a normal distribution and the difference between the correlation coefficient was tested against being zero (Sachs and Hedderich, 2009). The test of significantly different correlations was applied in each step in each replicate and the obtained p-values were averaged over the replicates.

Results and discussion

Figure 1 shows a Box-Whisker-Plot of the correlations between realized and predicted phenotypes for the animals in the validation set. As expected, Model A was found to be more accurate than Model B. With both methods, the highest correlations could be found when the number of animals in the validation set (n_v) was small. With n_v=100, the median of the correlations obtained was 0.689 and 0.627 with Model A and B, respectively. The median was almost constant with n_v ranging between 100 and 600 and then decreased continuously with both methods. The median of the correlation was 0.577 and 0.536 for n_v=1500 with Models A and B, respectively. Due to the design of cross-validation the smaller the validation set, the larger is the reference set. A larger reference set will lead to a more accurate estimation of the variance components and thus to a better estimation of the effects.

Therefore, also a better prediction of the phenotypes for the animals in the validation set is possible and higher correlations are expected with small validation sets.

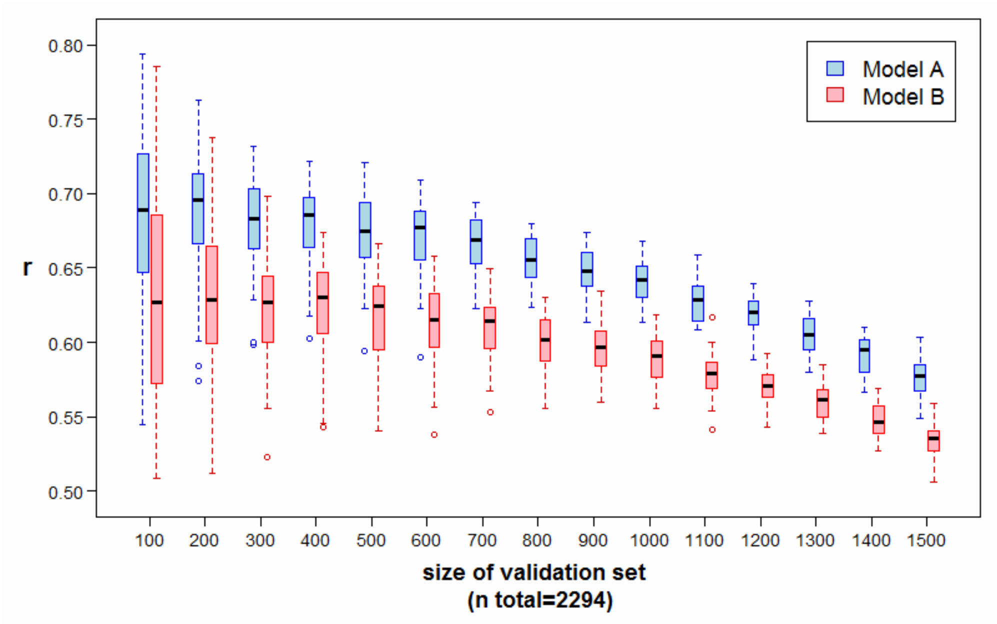

Figure 1: Box-Whisker-Plot of the correlations between realized and predicted phenotypes for the animals in the validation set.

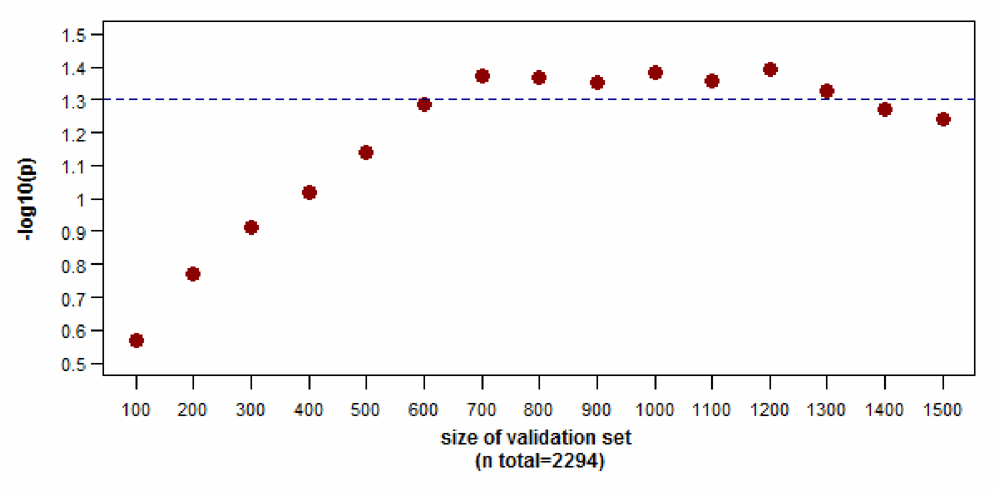

Figure 2: -log(p)-values (averaged over the replicates) of the test for a difference in correlation coefficients of Models A and B. The dashed line symbolizes the significance threshold on a 5% error level.

However, variation over the replicates was also highest in the case of very small validation sets. For example, with n_v=100, the results varied between 0.545 and 0.794 with Model A and ranged from 0.509 to 0.786 with Model B. Lee et al. (2008) described similar tendencies concerning the accuracy and the variation of prediction of phenotypes.

Since a higher number of values can be used for calculating the correlation coefficient when the number of animals in the validation set is higher, the correlation coefficient is estimated better with larger values of n_v. Therefore, even if the absolute distance between the models regarding the correlation coefficients seems to be similar with all sizes of the validation set, a significance differentiation between Models A and B was only possible with n_v ranging between 700 and 1,300 (Figure 2).

Conclusion

The optimal subdivision of a data set depends on the objective of cross-validation. The highest correlations can be obtained with the size of the reference set being large and therefore the validation set being small. A small validation set, however, also leads to a high variation in the obtained correlations and results depend strongly on the sample chosen for the validation set. A five-fold subdivision, using 20 per cent of the data as validation set, seems to be a good compromise. Larger validation sets provide more accurate estimation of correlation coefficients. Hence, if the aim is to differentiate significantly between two models, larger validation sets are recommended.

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr “Synbreed – Synergistic plant and animal breeding” and the project FUGATO-plus GenoTrack. We thank the Vereinigte Informationssyteme Tierhaltung v.W. (VIT), Verden, for providing data.

References

Blonk, R.J.W., Komen, H., Kamstra, A. *et al.* (2010). *Genetics*, 184: 213-219.

Gilmour, A.R., Gogel, B.J., Cullis, B.R. *et al.* (2009). ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Goddard, M.E. and Hayes, B.J. (2007). *J. Anim. Breed. Genet.*, 124: 323-330.

Lee, S.H., van der Werf, J.H.J., Hayes, B.J. *et al.* (2008). *PLoS Genetics*, 4(10): e1000231.

Luan, T., Woolliams, J.A., Lien, S. *et al.* (2009). *Genetics*, 183: 1119-1126.

Lund, M.S., Sahana, G., de Koning, D.-J. *et al.* (2009). *BMC Proceedings*, 3(Suppl 1): S1.

Sachs, L. and Hedderich, J. (2009) *Angewandte Statistik*. Springer, Dodrecht, Heidelberg, London, New York.

Scheet, P. and Stephens, M. (2006). *Am. J. Hum. Genet.*, 78: 629-644.

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