Validation Strategies for Single Step GBLUP

Peter von Rohr

2022-02-23

Disclaimer

Possible validation strategies are developed and documented

Background

[Legarra and Reverter, 2018] emphasize the importance of cross-validation to better estimate the population-level accuracy of GEBV. The work of [Legarra and Reverter, 2018] build on a procedure that use a series of statistics which describe the change of predictions from "old" to "recent" evaluations. Estimates of "population" accuracy which is the correlation between TBV and EBV across animals in a population are proposed. Population accuracy is relevant to compare the predictive ability of models and to maximize or to predict genetic progress. The study of [Legarra and Reverter, 2018] does not propose methods to estimate individual accuracies, which are a measure of the risk when choosing a particular animal for breeding.

Strategy

Based on the section "Statistics to test the quality of evaluation methods in brief" of [Legarra and Reverter, 2018], the following strategy applied in the paper is extracted. This strategy is used to define a procedure for our own validation procedure.

Successive evaluations with "partial" and "whole" data ($\hat{u}_p$ and $\hat{u}_w$, respectively) are considered. Those are based on "old" (p) and "recent + old" (w) phenotypic data. It is important to note that $\hat{u}_p$ and $\hat{u}_w$ are vectors of the same length and may contain the predicted breeding values of a set of "focal" animals (e.g. young candidates for selection) or of the entire dataset (i.e. all the animals in the pedigree).

Most statistics that are used for validation can be computed based on $\hat{u}_p$ and $\hat{u}_w$. Further parameters are the genetic variance ($\sigma^2_{u,\infty}$) which is corrected for the Bulmer effect and average values of the inbreeding coefficients and the numerator relationship matrix.

Procedure

The main statistics that are required for the validation of breeding values are based on the following components.

- $\hat{u}_p$
- $\hat{u}_w$
- $\sigma^2_{u,\infty}$
- $\bar{F}$ and $\bar{f}$ which are computed based on numerator relationship matrix

The required input for computing a first set of validation statistics are

- $\hat{u}_p$
- $\hat{u}_w$

Both of these vectors have the same length and hence define the set of focal animals. *

References

Andres Legarra and Antonio Reverter. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genetics Selection Evolution*, 50(1):1–18, nov 2018. ISSN 12979686. DOI: 10.1186/s12711-018-0426-6. URL `https://link.springer.com/articles/10.1186/s12711-018-0426-6https://link.springer.com/article/10.1186/s12711-018-0426-6`.