Supplemental material

Equivalence between Standard Selection Index and BLUP

Consider a standard single-trait model of the form

$$\boldsymbol{y} = \boldsymbol{u} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} = (y_1, ..., y_n)'$, $\boldsymbol{u} = (u_1, ..., u_n)'$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ are vectors of phenotypes, genetic, and environmental effects, respectively. Here, for simplicity we assume that all these vectors have zero-mean.

In a standard G-BLUP model, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are assumed to be independent (i.e., $cov(\boldsymbol{u}, \boldsymbol{\varepsilon}') = \boldsymbol{0}$), both have null means (i.e., $\mathbb{E}(\boldsymbol{u}) = \mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$), and (co)variance matrices $var(\boldsymbol{u}) = \sigma^2_u \boldsymbol{G}$ and $var(\boldsymbol{\varepsilon}) = \sigma^2_\varepsilon \boldsymbol{I}$, respectively; here $\boldsymbol{G}$ is a relationship matrix that could be derived from a pedigree or from DNA sequences.

Consider now a partition of each of the data in into a training (trn) and a testing (tst) set. The objective is to predict the genetic values of the individuals in the testing set ($\boldsymbol{u}_{tst}$) using the phenotype data available from the training set ($\boldsymbol{y}_{trn}$). The (co)variance matrix of the vector of breeding values can be partitioned as follows

$$var\left(\begin{bmatrix}\boldsymbol{u}_{trn}\\ \boldsymbol{u}_{tst}\end{bmatrix}\right) = \sigma^2_u \begin{bmatrix}\boldsymbol{G}_{trn} & \boldsymbol{G}_{trn,tst}\\ \boldsymbol{G}'_{trn,tst} & \boldsymbol{G}_{tst}\end{bmatrix}$$

where $\boldsymbol{G}_{trn}$ and $\boldsymbol{G}_{tst}$ are the genetic relationship submatrices for the training and testing data points, respectively, and $\boldsymbol{G}_{trn,tst}$ is the genetic relationship submatrix between training and testing subjects. The Best Linear Predictor (BLP) of $\boldsymbol{u}_{tst}$ ($\hat{\boldsymbol{u}}_{tst}$) takes the form (e.g., Searle *et al.* 1992):

$$\begin{aligned}\mathbb{E}(\boldsymbol{u}_{tst}|\boldsymbol{y}_{trn}) &= \mathbb{E}(\boldsymbol{u}_{tst}) + cov(\boldsymbol{u}_{tst}, \boldsymbol{y}'_{trn}) \left[var(\boldsymbol{y}_{trn})\right]^{-1} (\boldsymbol{y}_{trn} - \mathbb{E}(\boldsymbol{y}_{trn}))\\ &= \boldsymbol{G}'_{trn,tst} \left(\boldsymbol{G}_{trn} + \lambda_0 \boldsymbol{I}\right)^{-1} \boldsymbol{y}_{trn}.\end{aligned}$$

Alternatively, one can write $\hat{\boldsymbol{u}}_{tst} = \boldsymbol{H} \cdot \boldsymbol{y}_{trn}$, where $\boldsymbol{H} = \boldsymbol{G}'_{trn,tst}(\boldsymbol{G}_{trn} + \lambda_0 \boldsymbol{I})^{-1}$ is a "Hat" matrix. Thus, the BLUP of the genetic value of the i^{th} testing individual is $\hat{\boldsymbol{u}}_{tst(i)} = \boldsymbol{H}'_i \boldsymbol{y}_{trn}$ where $\boldsymbol{H}'_i$ is the i^{th} row of $\boldsymbol{H}$, that is $\boldsymbol{H}'_i = \boldsymbol{G}'_i(\boldsymbol{G}_{trn} + \lambda_0 \boldsymbol{I})^{-1}$ which is equal to the weights of the standard selection index, $\hat{\beta}'_i = \boldsymbol{G}'_i(\boldsymbol{G}_{trn} + \lambda_0 \boldsymbol{I})^{-1}$ (see Equation 2 in the manuscript).

References

Searle S. R., G. Casella, and C. E. McCulloch, 1992 *Variance components.* John Wiley & Sons, Inc.