HSSGBLUP: a single-step SNP-BLUP genomic evaluation software adapted to large livestock populations.

Thierry Tribout, Vincent Ducrocq, D Boichard

HSSGBLUP: a Single-Step SNP BLUP genomic evaluation software adapted to large livestock populations

Tribout T.[1], Ducrocq V.[1], Boichard D.[1] [1]*Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France*

Context

In France, current dairy and beef cattle genomic evaluations are based on multi-step approaches.

Preselection of genotyped animals generates biased estimated breeding values and genetic trends.

Single Step GBLUP evaluations are being implemented to solve this issue, with the development of a software fitting the French bovine evaluations requirements:

- Large populations (up to 20 million animals)
- Hundreds of thousands of informative genotyped animals
- Genetic Groups
- Multiple traits evaluations, possibly with maternal genetic effects and heterogeneous variances
- Inclusion of effects of QTL or causal variants
- Inclusion of foreign phenotypic information for international populations (Holstein, BSW)
- …

INRAE develops a software covering these features: *HSSGBLUP*. The main strategies adopted are presented here.

Current status & Perspectives

The software is completed. Optimizations (computing times) are in progress.

- All new evaluations have already been implemented with *HSSGBLUP* (e.g. see poster « Toward a genomic evaluation of cheese-making traits including candidate SNP in Montbéliarde cows » #110 by Sanchez et al).
- All current French bovine polygenic and multi-step genomic evaluations will be progressively replaced by Single Step SNP BLUP evaluations before april 2022 (dairy populations) and april 2023 (beef populations).

The model considered is the **Hybrid Single Step model** proposed by **Fernando et al (2016)**:

$$\begin{bmatrix} y_n \\ y_g \end{bmatrix} = \underbrace{\begin{bmatrix} X_n \\ X_g \end{bmatrix}\beta}_{\text{Environmental component}} + \underbrace{\begin{bmatrix} Z_n J_n \\ Z_g J_g \end{bmatrix}\mu_g + \begin{bmatrix} C_n \\ 0 \end{bmatrix}\varphi + \begin{bmatrix} Z_n & 0 \\ 0 & Z_g M_g \end{bmatrix}\begin{bmatrix} u_n \\ \alpha \end{bmatrix}}_{\text{Breeding value}} + e$$

$\square_n$ = non genotyped animals

$\square_g$ = genotyped animals

To ensure consistency of pedigree and genomic relationships:

μ_g = mean of unselected base animals (**Hsu et al, 2017**)

$J_n = -A_{ng}A_{gg}^{-1}1$, computed as in **Tribout et al (2019)**

$J_g = -\mathbf{1}$

Genetic Groups component

φ = vector of genetic groups effects

C_n = contributions of the Genetic Groups to non-genotyped animals

Genetic component

u_n = breeding value of non-genotyped animals

α = vector of markers effects

M_g = genotypes at markers of genotyped animals

Mixed Model Equations *(example for a single trait, without maternal genetic effect):*

$$\begin{bmatrix} X'X & X_g'Z_gM_g & X_n'Z_n \\ M_g'Z_g'X_g & M_g'Z_g'Z_gM_g + M_n'A^{nn}M_n\frac{\sigma_e^2}{\sigma_g^2} + I\frac{\sigma_e^2}{\sigma_{\alpha_i}^2} & M_g'A^{gn}\frac{\sigma_e^2}{\sigma_g^2} \\ Z_n'X_n & A^{ng}M_g\frac{\sigma_e^2}{\sigma_g^2} & Z_n'Z_n + A^{nn}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix}\begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \\ \hat{u}_n \end{bmatrix} = \begin{bmatrix} X'y \\ M_g'Z_g'y_g \\ Z_n'y_n \end{bmatrix} * \exp(-0{,}5\,\hat{\gamma}_i)$$

Inverse of pedigree relationship matrix $A^{-1} = \begin{bmatrix} A^{nn} & A^{ng} \\ A^{gn} & A^{gg} \end{bmatrix}$

M_n=(imputed) genotypes at markers of non-genotyped animals

σ_e^2 =residual variance

σ_g^2 = genetic variance

$\sigma_{\alpha_i}^2$=genetic variance associated to the i[th] SNP, QTL, causal variant

$M_n'A^{nn}M_n$ computed as $M_g'A^{gn}(A^{nn})^{-1}A^{ng}M_g$ (**Taskinen et al, 2017**), using an efficient algorithm proposed by **Vandenplas et al (2018)**

Inclusion of QTL or causal mutations:

$\sigma_{\alpha_i}^2$ is a function of the proportion of genetic variance explained by the i[th] SNP, QTL, causal variant

Heterogeneous variances:

Here, $\sigma_{e_i}^2 = \exp(\gamma_i)\,\sigma_e^2$ is the residual variance in the i[th] level of heterogeneity

$\hat{\gamma}_i$ are iteratively estimated on the data, as described in **Meuwissen et al (1996)**

The genomic relationship matrix is neither built nor inverted → the model is well adapted for populations with hundreds of thousands of genotyped animals

Programming strategies

- Coded in Fortran 90
- Solver: Preconditionned Conjugate Gradient
- Iteration on data
- Use of sparse matrices

Memory-saving strategies, making the software suitable for very large populations

- Portions of code are parallelized (openMP)
- Use of intel MKL-Pardiso library, optimized for parallelized computations on sparse matrices

References

- Fernando L.R., Cheng H., Golden B.L. Garrick D.J., Genet. Sel. Evol (2016) 48:96
- Hsu W. L., Garrick D.J., Fernando R.L., G3 (Bethesda) (2017) 7(8):2685-2694
- Meuwissen T.H.E., De Jong G., Engel B., J. Dairy Sci. (1996), 79:310-316
- Taskinen M., Mantysaari E.A., Strandén I, Genet. Sel. Evol. (2017) 49:36
- Tribout T., Boichard D., Ducrocq V., Vandenplas J., 70th EAAP, Ghent (2019)
- Vandenplas J., Eding H., Calus M.P.L., 69th EAAP, Dubrovnik (2018)

Centre de Recherche de Jouy-en-Josas
78350 Jouy-en-Josas
thierry.tribout@inrae.fr