Lab 2 Comparison between BLUP and Single-Step GBLUP

Prepared by D. Lourenco, A. Legarra, Z. Vitezica, and I. Aguilar

The data for this lab was simulated by D. Lourenco using QMSim (Sargolzaei & Schenkel, 2009). A single trait animal model was simulated assuming heritability of 0.4. All the genetic variance was explained by 500 QTL. Animals were genotyped for 45,000 SNP and the average LD was 0.18. The simulated additive genetic variance was 0.40 and the residual variance was 0.60. The simulated phenotype was generated using the following model:

Phenotype = sex_effect + animal_effect + residual

Description of files:

data3.txt:
1: animal ID
2: generation
3: sex
4: phenotype
5: true breeding value (TBV)

ped3.txt:
1: animal ID
2: sire ID
3: dam ID

snp3.2k:
1: animal ID
2: SNP genotype

mrkmap.txt:
1: SNP ID
2: Chromosome
3: position

1. Data and parameter files for the following exercises are in this link http://nce.ads.uga.edu/wiki/doku.php?id=course_materials_-_moscow_russia_2019

2. Look into the renumf90 parameter file (renum.par) and identify the keywords. Check if the keyword used to include genomic information is there. Check also if the keyword to compute inbreeding was included.

3. Run renumf90 program to renumber the data.

4. Check the renf90.par, renf90.dat, and renaddxx.ped. From the renaddxx.ped file, identify genotyped animals, and check in the wiki (http://nce.ads.uga.edu/wiki/doku.php?id=readme.renumf90) the content of each column. What is the content of **snp3.2k_XrefID**?

5. Run blupf90 without genomic information (this is the regular BLUP).

6. Run blupf90 using genomic information (this is ssGBLUP).

   The quality control of genomic data was performed inside blupf90.
   a) Which quality checks for both SNP and animals were done by default?
   b) Are there any duplicated genotypes?
   c) Why some SNPs were removed?

d) What is the correlation between **G** and $\mathbf{A}_{22}$?
e) Check averages of **G** and $\mathbf{A}_{22}$.

Clean genotype files can be saved after quality control. If you want to save those files, use the following option: OPTION saveCleanSNPs

7. Compare the solution files from exercises 6 and 7. If you want, you can compute correlation between true breeding values (column 5 of data3.txt) and EBV or GEBV for genotyped animals.

OPTIONAL EXERCISES

8. Do a validation on young selection candidates (individuals from 5th generation with genotypes and no phenotypes). Compare EBV and GEBV with true breeding value (TBV). Remember that correlation between (G)EBV and a benchmark (i.e., TBV) is a measure of prediction accuracy. What happened with prediction accuracy when genomic information was included? Check also intercept and regression coefficient from a regression of TBV on EBV and TBV on GEBV.

   Hint 1: remove the phenotypic information from the 5th generation and obtain solutions from a model with SNP information and with no SNP information.
   Hint 2: have renumf90 passing to the renumbered data a column containing generation number.

9. A very common validation method used in beef cattle and other species is the correlation between phenotypes adjusted to fixed effects and EBV or GEBV. This is called predictive ability or ability to predict future performance. Compute predictive ability for young genotyped animals in the 5th generation.

   Hint 1: the benchmark is now adjusted phenotypes obtained using the complete data and no genomic information. Run blupf90 with complete data and no genomic information. Run predictf90 in the same folder you ran blupf90. Before running, you should include the following option in the parameter file:
   OPTION include_effects X
   Where X is the number of the animal effect. If animal effect is effect number 2 in your model, X is 2. This means that phenotypes will be adjusted for all effects, but effect number 2. Adjusted phenotypes will be in a file called yhat_residual, with the following format:

   Animal_id, Y*, Yhat, residual
   where: Y* = Phenotype – fixed effects
   Yhat = EBV (or animal effect)
   Residual = Phenotype – EBV

   Hint 2: Correlate Y* with EBV and GEBV computed using reduced data.