Labs 3 and 4 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 + true_breeding_value + 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. Copy the full folder into your directory
   `cp -r /home/labs/lab3_4 .`

2. Modify an existent renumf90 parameter file (or create a new one), according to the data file, to fit the following model:

$$y = sex + animal + e$$

3. Run renumf90 program to renumber the data. Do not forget to add the keyword to include genomic information. Add also the keyword to compute inbreeding based on the pedigree.

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 preGSf90 including the option to save clean SNP file after quality control. Check the output. Which quality checks for both SNP and animals were done by default? Are there any duplicated genotypes? Why some SNPs were removed? What is the correlation between **G** and $\mathbf{A}_{22}$? Check averages of **G** and $\mathbf{A}_{22}$.

6. Working with the clean SNP files, add an option to create PCA plots and run preGSf90 again. As your SNP file is already clean, do not forget to include an option to skip quality control. Based on the PCA plot, what can you conclude about the population structure?

7. Run blupf90 without SNP information. Now run blupf90 using genomic information (this is ssGBLUP) and compare cpu time and solutions.
   Hint: use the following command to provide computing time and to save outputs to a log file:
   ```
   time echo renf90.par | blupf90 | tee blup1.log
   ```

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.
   Hint 3: if generation column is number 4, the new data can be created using the AWK Unix tool:
   ```
   awk '$4<5' renf90.dat > renf90.dat.reduced
   ```

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.

OPTIONAL EXERCISE

10. Running GBLUP in blupf90: This software suite was created specifically for ssGBLUP, however, the programs can also be used to run GBLUP if small tricks are applied. Remember that in GBLUP pedigree information is not used and only genotyped animals can have phenotypes. Here are the tricks:
    a) create a dummy pedigree file (animal ID, 0 for sire, and 0 for dam)
    b) create a file with phenotypes only for genotyped animals
    c) run `renumf90` setting the PED_DEPTH to 1

    d) run `blupf90`, but first check the options you need to include in the parameter file in order to run a pure GBLUP:
    http://nce.ads.uga.edu/wiki/doku.php?id=how_to_run_pure_gblup

    Check the output of blupf90 and the solution file.

11. Let's assume you are working on a ssGBLUP project and your objective is to test different models using the same data. You can run preGSf90 with clean data once and this program will automatically save a file called GimA22i ($\mathbf{G^{-1}} - \mathbf{A_{22}^{-1}}$). Every time you change your model, you can just read the GimA22i file avoiding the creation of this matrix every time. This can save some computing resources. Alternatively, you can save $\mathbf{G}$, $\mathbf{A_{22}}$, $\mathbf{G^{-1}}$, $\mathbf{A_{22}^{-1}}$ in binary (not user-friendly) or text format (user-friendly). Check the documentation for preGSf90 and explore the options to save the matrices
    http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90