EXERCISES_GMATRIX

Andrés Legarra,
UR631 SAGA, INRA, 31326 Castanet Tolosan, France
andres.legarra at toulouse.inra.fr

We will use mice data from Legarra et al. (2008). Copy `/home/ads-guest3/andres/mice_cours` to your directory. This data is also used in "Exercises_optional".

A. Compute genomic relationships with compute_G

The objective of this exercise is to build genomic relationship matrices. I will use first home-made programs. The data set corresponds to

Go to `mice_cours/progs` and look at the program `compute_G.f90` . This program computes relationships following VanRaden's 2008 first ($\mathbf{G} = \mathbf{ZZ}'/2\sum_{all\ SNPs} p_i(1-p_i)$) or second ($\mathbf{G} = \frac{1}{nsnp}\sum \frac{\mathbf{z}_i\mathbf{z}_i'}{p_i q_i}$) **G** (which was used by Yang et al. in human genetics). Actually, to make **G** positive definite, it computes $\mathbf{G} = 0.95\,\mathbf{ZZ}'/2\sum_{all\ SNPs} p_i(1-p_i) + 0.05\mathbf{I}$.

Go to `../data` and run it:

```
$ ../progs/compute_G
 genofile?
mice_genotypes.txt
 out file for G :
 mice_genotypes.txt.G

 out file for G-1 :
 mice_genotypes.txt.Gi

 which G?
  1 - VanRaden firstG = ZZ'/sum(2pq)
  2 - VanRaden secondG = Yang et al. = mean (Z_i Z_i' /(2p_i
q_i))
1
 write out G ? (T:F)
T
 write out G inverse? (T:F)
F
 Column position in file for the first marker:           12
 Format to read SNP file:
 (i10,1x,10946i1)

 Number of SNPs :       10946
 nanim=         1884 nsnp=        10946
...
```

```
average freq  0.512276949805317      var(freq)
7.708975460608908E-002
 X re-setup
 G computed, time   108.8205
 G-1 computed, time   7.727821
 Rank G:         1884
```

It creates a file, `mice_genotypes.txt.G` y. Take a look. Let's see the aspect of diagonal and off-diagonal; if SNP's are in H-W equilibrium, they should average to 1 and 0 respectively. Take a look:

```
awk '$1==$2' mice_genotypes.txt.G | less
```

looks like this is around 1 but there are values higher and lower than 1. The minimum is 0.82; the maximum is 1.29. We can use awk to compute the mean (you can use R or whatever).

```
awk '$1==$2' mice_genotypes.txt.G | awk 'BEGIN{r=0}; {r=r+$5};
END{print r/NR}'
```

which is 1.029 so this population has an average inbreeding of 0.02. The base population was actually composed of 8 inbred lines, whose descendants mated during 50 generations. Now, off-diagonals:

```
awk '$1!=$2' mice_genotypes.txt.G | less
```

for off-diagonals. Off diagonal have a minimum of -0.28 and a maximum of 1.17. This maximum possibly correspond to a couple of twins (or clones) that do exist in the population.

Activities:

1) can you find those twins?
2) Plot a histogram of relationships. How does it look?
3) Take the code. Right after computation of allelic frequencies (line 117), put `freq(i)=0.5` to fix them in fact to 0.5. Compile and run the exercise again. How do diagonals and off-diagonals look like?

B. Computing and using genomic relationships with pregsf90

This software allows computing of genomic relationships in an efficient manner, and they can be used for GBLUP or Single Step. We'll see three programs in this session:

`preGSf90` : computes **G** and prepares matrix $\mathbf{H}^{-1}$ for SingleStep.
`blupf90`: does BLUP using (if it does exist) genomic information
`remlf90`: estimation of genetic parameters
They are in directory `mice_data/bin`

1. Cálculo de la matriz de parentesco genómica y H:

```
$ ../bin/preGSf90
 name of parameter file?
mice.par
```

File `mice.par` is a "standard" BLUPF90 parameter file, with one extra option:

```
OPTION SNP_file mice_genotypes.txt
```

This option makes pregsf90 reading a genotype file and an "equivalences" file `mice_genotypes.txt_XrefID.` Here, we have the OPTIONS that make G closest to "pure" $\mathbf{G} = \mathbf{ZZ}'/2\sum_{all\,SNPs} p_i\left(1-p_i\right)$.

The printout on screen is very informative, do read it !!

We are going to write unto a file matrix **G** and its inverse, $\mathbf{G}^{-1}$. We need to add the following OPTIONS:

```
OPTION saveAscii
OPTION saveG
OPTION saveGInverse
```

Run it and take a look at **G**.
There are files with 2sum(pq) (`sum2pq`) , frequencies (`freqdata.out`) and a file with $\mathbf{G}^{-1}$-$\mathbf{A}_{22}^{-1}$ (`GimA22i`) which is used by all applications.

Take a look at **G** so created. Compute means of diagonals and off-diagonals. Are they identical to above?
Take a look at $\mathbf{G}^{-1}$ (file `Gi`).
NOTE: you will see that **G** is indexed now by row, column numbers instead of the original id's in the genotype file (but if you use `renumf90` for recoding first there is no problem). These old id's can be retrieved using one of the OPTIONs.

2. Genetic evaluations using GBLUP

Although GBLUP can be seen as a SingleStep where all animals in data have genotype, we will use another way to fit a GBLUP here. The f90 series of programs has a utility to include external files with covariance structures; see the wiki
http://nce.ads.uga.edu/wiki/doku.php?id=user_defined_files_for_covariances_of_random_effects
for an explanation.

So you can use parameter file `mice_blup.par` to test GBLUP. You can check mice_blup.par to use pedigree relationships instead, with file `pedigri.dat`.

This file makes a genetic evaluation for Body Weight. You can try to plot EBV's of pedigree BLUP against GBLUP.

3. GREML.

This is basically the same but using `remlf90` or `airemlf90` instead of `blupf90`.

What you can do first is to estimate variance components for Body Weight using the files above. What do you get? Are parameter estimates the same?

Now analyse the 4th column (trait: Body Length) of the data file instead, both for pedigree or genomics.

What do you get? Are parameter estimates the same for pedigree or genomics?