UNIVERSITY OF
GEORGIA

Creating genomic relationship matrices with preGSf90

Daniela Lourenco
UGA USA

Ignacio Aguilar
INIA Uruguay

BLUPF90 TEAM, 02/2022

preGSf90

- Performs Quality Control of SNP information ✓

- Creates the genomic relationship matrix (**G**)
  - and relationships based on pedigree ($\mathbf{A}_{22}$)
  - Inverse of relationship matrices

BLUP-based models

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{A}^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

BLUP

Henderson, 1963

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{G}^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

GBLUP

Nejati-Javaremi et al., 1997
Fernando, 1998
VanRaden, 2008

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W}+\mathbf{H}^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{W'y} \end{bmatrix}$$

ssGBLUP

Misztal et al. (2009)
Legarra et al. (2009)
Aguilar et al. (2010)
Christensen & Lund (2010)

$$\mathbf{H}^{-1} = \left[\begin{array}{c:c} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \hdashline \mathbf{A}^{21} & \mathbf{A}^{22} \end{array}\right] + \left[\begin{array}{c:c} \mathbf{0} & \mathbf{0} \\ \hdashline \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{array}\right]$$

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

PreGSf90

- Created to construct the matrices using in ssGBLUP

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

$\mathbf{G}$ $\mathbf{G}^{-1}$

$\mathbf{A}_{22}$ $\mathbf{A}_{22}^{-1}$

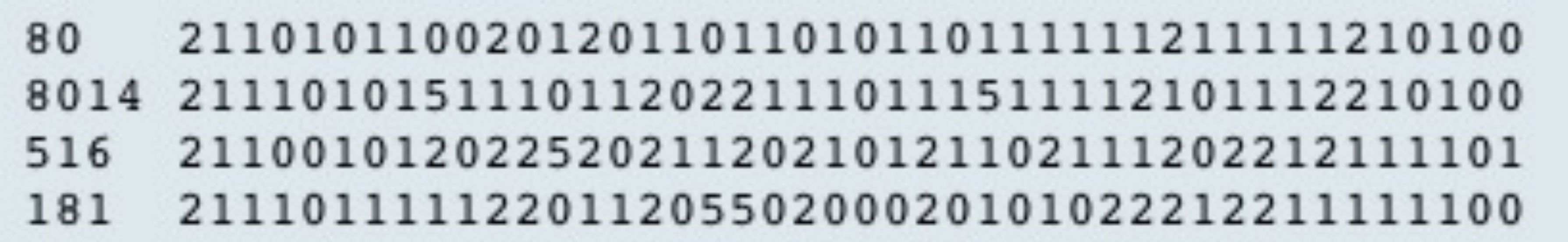
$$\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$$

Genomic Relationship Matrix - **G**

- $\mathbf{G} = \dfrac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)
  - Z = matrix for SNP marker
  - Dimension of Z = $n*i$
  - *n* animals
  - *i* markers

Genotype Codes
0 – Homozygous
1 – Heterozygous
2 – Homozygous
5 – No Call (Missing)

SNP file



```
80    21101011002012011011010110111111211111210100
8014  21110101511101120221110111511112101112210100
516   21100101202252021120210121102111202212111101
181   21110111112201120550200020101022212211111100
```

HOW TO: Creation of Genomic Matrix

- Read SNP marker information => **M** $\begin{bmatrix} 2 & 1 & 2 & .. \\ 0 & 1 & 0 & .. \\ .. & .. & .. & .. \end{bmatrix}.$
- Get 'means' to center
  - Calculate allele frequency from observed genotypes (p_i)
  - p_i= sum(SNPcode_i)/2n
- Centered matrix $\mathbf{Z} = \mathbf{M} - 2\mathbf{P}$
- $\mathbf{G} = \dfrac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)

Creation of Genomic matrix

- Issues
  - Large number of genotyped individuals
  - Large number of SNP markers
  - Matrix multiplication ~ cost $n^2 * i$

- Large amounts of data put in (cache) memory to do matrix multiplication for each pair of animals and indirect memory access (center)

PreGSf90

- Efficient methods
  - create the genomic relationship matrix and the relationship matrix based on pedigree
  - Invert the relationship matrices

- Computes statistics for the matrices
  - Means, Var, Min, Max
  - Correlations between diagonals
  - Correlations for off-diagonals
  - Correlations for the full matrices
  - Regression coefficients

OPTIONS – preGS90 parameter file

- PreGSF90
  - controled by adding OPTION commands to the parameter file

```
OPTION SNP_file marker.geno.clean
```

  - Reads:
    - `marker.geno.clean`
    - `marker.geno.clean.XrefID` (created by renumf90)

    - Pedigree file
    - Map file (optional)

Genomic Matrix default options

- $\mathbf{G}_0 = \dfrac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}$ (VanRaden, 2008)
- With:
  - **Z** centered using current allele frequencies
    - Current genotyped animals

Genomic Matrix Options

- OPTION whichfreq *x*
  - 0: read from file *freqdata* or other specified name (needs OPTION FreqFile)
  - 1: 0.5
  - 2: current calculated from genotypes (default)

- OPTION FreqFile *file*
  - Reads allele frequencies from a file

Genomic Matrix default options

- **Blending** - to avoid singulatiry problems

  $$\mathbf{G} = 0.95*\mathbf{G}_0 + 0.05*\mathbf{A}_{22}$$

  – OPTION AlphaBeta 0.95 0.05   #(default)

  – Beta may vary from 0.2 to 0.01

Genomic Matrix default options

- **Tuning**
  - Adjust **G** to have mean of diagonals and off-diagonals equal to $\mathbf{A}_{22}$
  - OPTION tunedG 2 #(default) Chen et al. (2011)

- Base of GBLUP is *genotyped* animals
- Base of pedigree is *founders of the pedigree*
- For SSGBLUP modelled as a mean for genotyped animals
  - $p(\boldsymbol{u}_2) = N(\mathbf{1}\mu, \mathbf{G})$
  - Integrate μ : $\mathbf{G}^* = 11'\lambda + (1 - {}^{\lambda}/_{2})\mathbf{G}$
  - μ = (Genomic base) – (Pedigree base)
  - Vitezica et al. 2011

Options for matching **G** to **A**$_{22}$

- OPTION tunedG *x*
  - 0: no adjustment
  - 1: mean(diag(G))=1, mean(offdiag(G))=0
  - 2: mean(diag(G))=mean(diag(A$_{22}$)),
    mean(offdiag(G))=mean(offdiag(A$_{22}$)) (default)
  - 3: mean(G)=mean(A$_{22}$)
  - 4: Use Fst adjustment. Powell et al. (2010) & Vitezica et al. (2011)

$$\lambda = \frac{1}{n^2}\left(\sum_{i}\sum_{j}\mathbf{A}_{22_{ij}} - \sum_{i}\sum_{j}\mathbf{G}_{ij}\right) \qquad \mathbf{G}^* = 11'\lambda + (1 - {}^{\lambda}/_{2})\mathbf{G}$$

Storing and Reading Matrices

- preGSf90 saves $\mathbf{G^{-1}} - \mathbf{A_{22}^{-1}}$ by default (file: GimA22i)

To save ‘raw’ genomic matrix:

- OPTION saveG [all]
  - If the optional *all* is present all intermediate **G** matrices will be saved!!!

To save $\mathbf{G}^{-1}$

- OPTION saveGInverse
  - Only the final **G**, after blending, scaling, etc. is inverted !!!

To save $\mathbf{A}_{22}$ and inverse

- OPTION saveA22 and OPTION saveA22Inverse

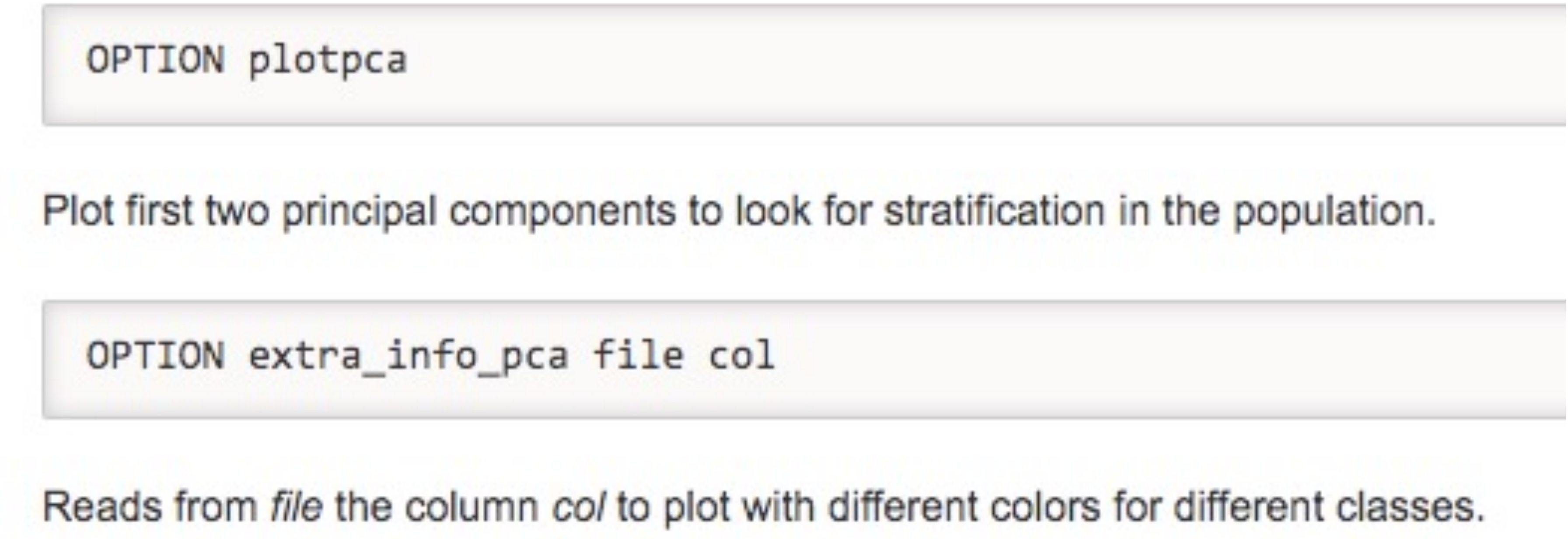
Storing and Reading Matrices

- OPTION saveG [all] , OPTION saveGInverse, …
  - Saves in binary format
  - “Dumped” format to save space and time
  - To save as row, column, value:
    - OPTION no_full_binary
    - Still binary, but can be easily read and converted to text

Storing with Original IDs

- Some matrices could be stored in text files with the original IDs extracted from *renaddxx.ped* created by the RENUMF90 program (col #10)
- For example:
  - OPTION saveGOrig
  - OPTION saveDiagGOrig
  - OPTION saveHinvOrig
- Values
  - origID_i, origID_j, val

Genomic Matrix - Population structure



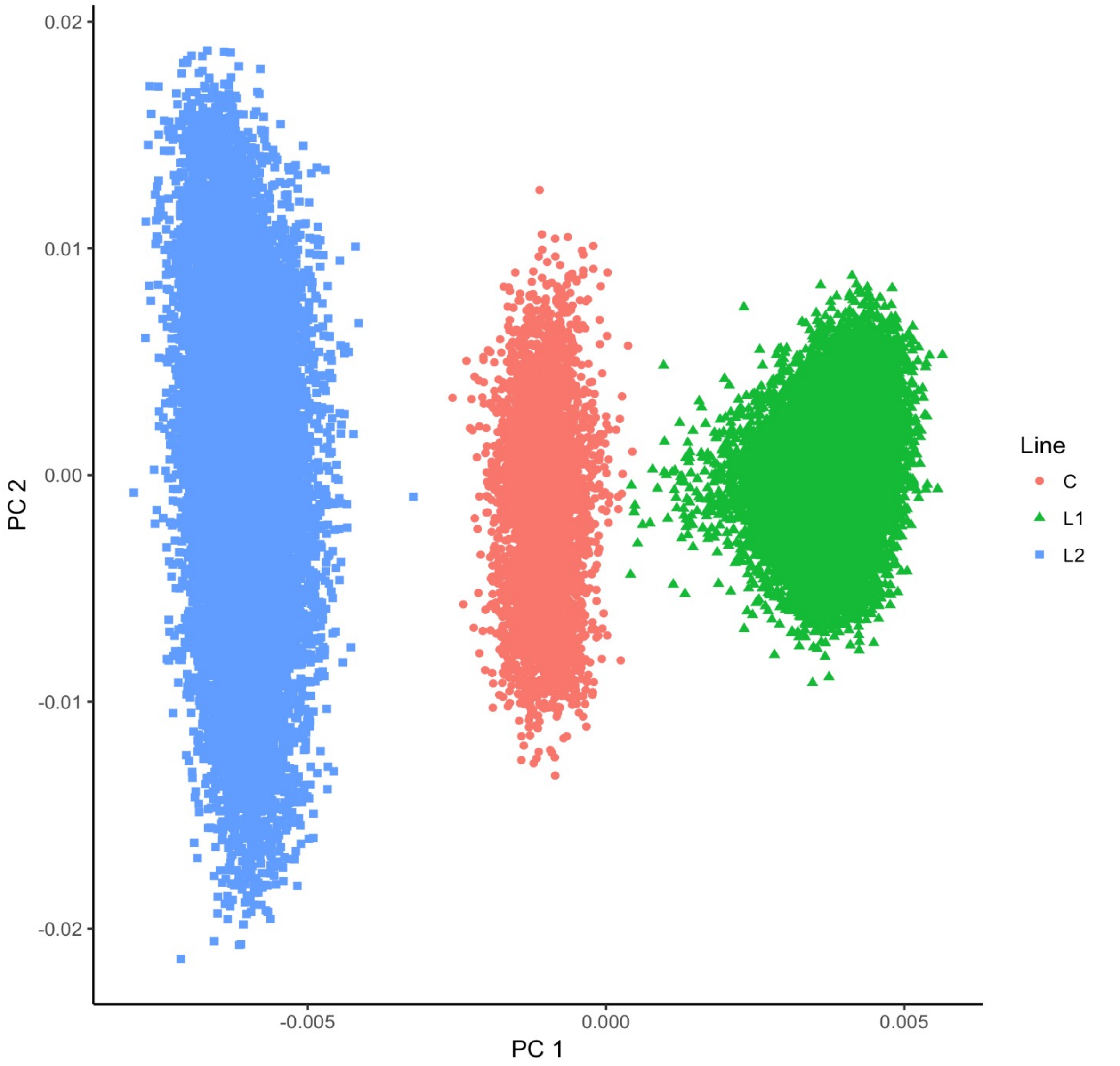
```
OPTION plotpca
```

Plot first two principal components to look for stratification in the population.

```
OPTION extra_info_pca file col
```

Reads from *file* the column *col* to plot with different colors for different classes.

Genomic Matrix - Population structure



Tricks to setup **G** for GBLUP

- Tricks are needed because preGSf90 is set up for ssGBLUP

1) Use a dummy pedigree

```
1 0 0
2 0 0
…
```

2) Use PED_DEPTH 1 in renumf90

3) Change blending parameters

- OPTION AlphaBeta 1.00 0.00 → G = 1.00***G** + 0.00***I**
- OPTION AlphaBeta 0.99 0.01 → G = 0.99***G** + 0.01***I**

4) No adjustment for compatibility with $\mathbf{A}_{22}$

- OPTION tunedG 0

Tricks to setup **G** for GBLUP

- Yet another way to run GBLUP in BLUPF90

1) In renum.par, remove any information about the pedigree. Example:

```
FILE
pedigree.txt
FILE_POS
1 2 3 0 0
PED_DEPTH
3
```

3) Change blending parameters

- `OPTION AlphaBeta 1.00 0.00` → G = 1.00***G** + 0.00***I**
- `OPTION AlphaBeta 0.99 0.01` → G = 0.99***G** + 0.01***I**

4) No adjustment for compatibility with $\mathbf{A}_{22}$

- `OPTION tunedG 0`

PreGSf90 inside BLUPF90 ??

- Almost all programs from BLUPF90 support the creation of genomic relationship matrices
- OPTION SNP_file xxxx

- Why preGSF90 ?
  - Same genomic relationship matrix for several models, traits, etc.
  - Just do it once and store GimA22i

Use in application programs

- Use renumf90 for renumbering and creation of XrefID and files

```
SNP_FILE
marker.geno
```

- Run preGSf90 with quality control, saving clean files

- Option 1:

  run blupf90 with clean files

- Option 2:

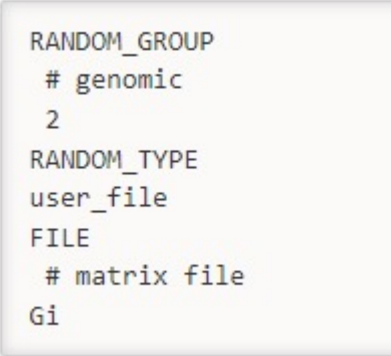
  run preGSf90 with clean files (program saves **GimA22i**)

  run blupf90 with option to read **GimA22i** from the file

Reading external matrices

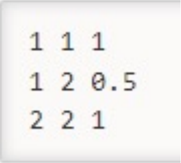
- BLUPF90 programs accept external matrices created outside
- http://nce.ads.uga.edu/wiki/doku.php?id=user_defined_files_for_covariances_of_random_effects
- File should be row, column, value in plain text format (lower OR upper triangular)

renf90.par



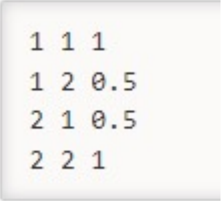
```
RANDOM_GROUP
 # genomic
 2
RANDOM_TYPE
user_file
FILE
 # matrix file
Gi
```

Valid format



```
1 1 1
1 2 0.5
2 2 1
```

Non-valid format



```
1 1 1
1 2 0.5
2 1 0.5
2 2 1
```

- user_file: if providing the inverse of the covariance structure

- user_file_inv: if the program has to invert the covariance structure