UNIVERSITY OF GEORGIA

ssGBLUP

Daniela Lourenco
UGA USA

Andres Legarra
INRA France

Ignacio Aguilar
INIA Uruguay

UGA TEAM, 08/2019

About prediction methods …

No Genotypes – Only Pedigree

BLUP $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$

No Pedigree – Only Genotypes

GBLUP $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{G}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$

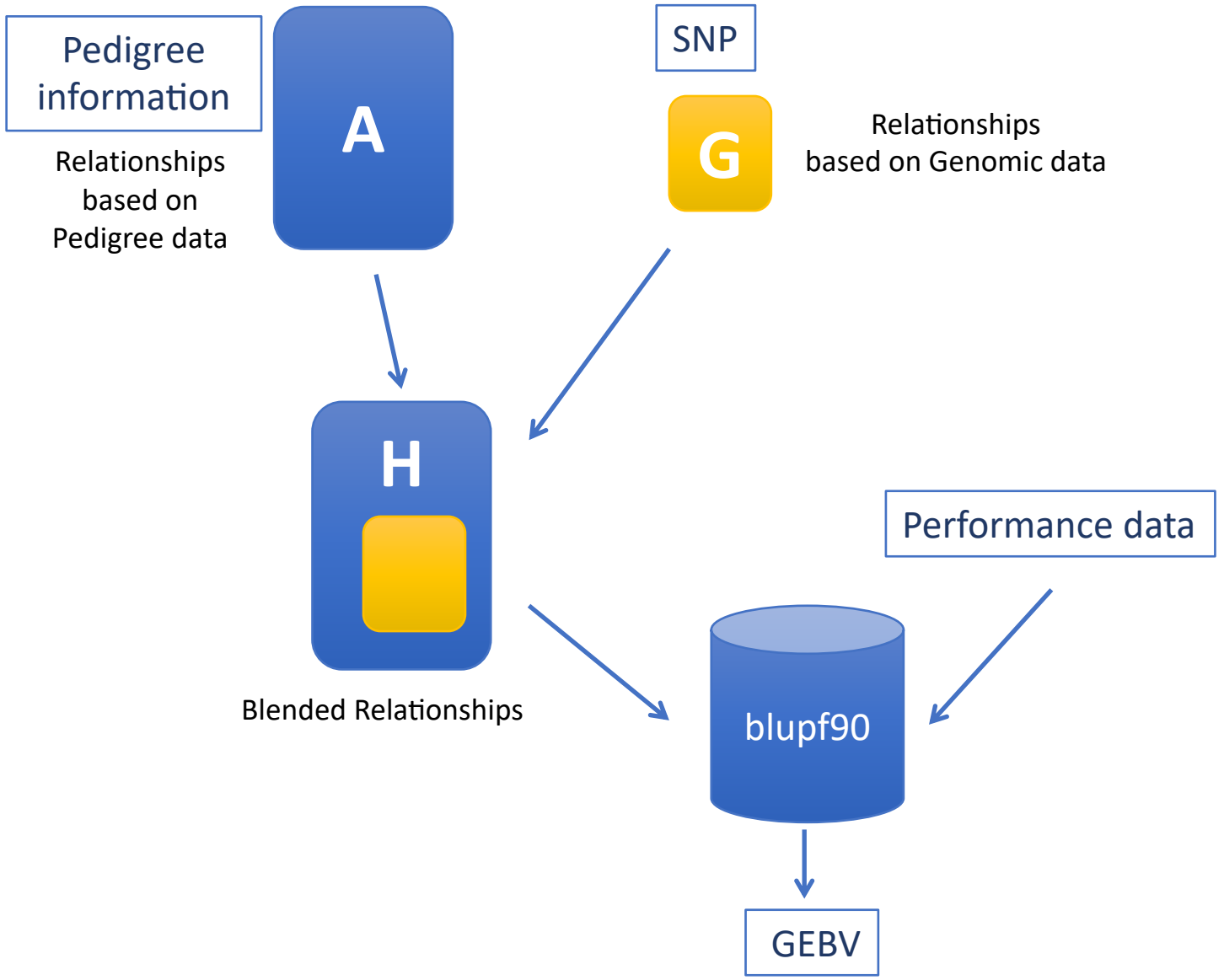
SNP-BLUP $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'M} \\ \mathbf{M'X} & \mathbf{M'M} + \mathbf{I}\lambda \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{M'y} \end{bmatrix}$

Pedigree + Genotypes

ssGBLUP $\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{H}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$

ssBR $\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{M} & \mathbf{X}_n'\mathbf{Z}_n \\ \mathbf{M}'\mathbf{Z}'\mathbf{X} & \mathbf{M}'\mathbf{Z}'\mathbf{Z}\mathbf{M} + \mathbf{I}\frac{\sigma_e^2}{\sigma_\alpha^2} & \mathbf{M}_n'\mathbf{Z}_n'\mathbf{Z}_n \\ \mathbf{Z}_n'\mathbf{X}_n & \mathbf{Z}_n'\mathbf{Z}_n\mathbf{M}_n & \mathbf{Z}_n'\mathbf{Z}_n + \mathbf{A}^{nn}\frac{\sigma_e^2}{\sigma_g^2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\epsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{M}'\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}_n'\mathbf{y}_n \end{bmatrix}$

Single-Step Genetic Evaluation



H adjusts relationships for ungenotyped animals

Animal	Sire	Dam
1	0	0
2	0	0
3	1	1
4	2	2

Pedigree Relationship Matrix (**A**)

$$\begin{bmatrix} 1.0 & 0.0 & 0.5 & 0.5 \\ . & 1.0 & 0.5 & 0.5 \\ . & . & 1.0 & 0.5 \\ . & . & . & 1.0 \end{bmatrix}$$

Genomic Relationship Matrix (**G**) for animals 3 and 4

$$\begin{bmatrix} 1.0 & 0.52 \\ . & 1.0 \end{bmatrix}$$

Realized Relationship Matrix (**H**)

$$\begin{bmatrix} 1.004 & 0.0 & 0.507 & 0.507 \\ . & 1.004 & 0.507 & 0.507 \\ . & . & 1.0 & 0.52 \\ . & . & . & 1.0 \end{bmatrix}$$

Understanding the H matrix

- It is a projection of **G** matrix on the rest of individuals “so that” **G** matrix makes sense
  - e.g. parents of two animals related in **G** should be related in **A**

- It is a Bayesian updating of the pedigree matrix based on new information from genotypes

- Typically
  - $\mathbf{A}^{-1}$ in the millions but extremely sparse
  - **G** and $\mathbf{A}_{22}$ in the thousands
  - Leads to a very efficient method of genomic evaluation:
    - **Single Step GBLUP**

Some properties of H

- Semi-positive definite always
  - eigenvalues are always positive or zero
- Positive definite & invertible if **G** is invertible
- In practice, if **G** is too different (wrong pedigree or genotyping) from $\mathbf{A}_{22}$, this gives lots of numerical problems
- If everyone is genotyped, Single Step is GBLUP
- If no one is genotyped, Single Step is BLUP

H matrix from Legarra et al. (2009)

- Genomic evaluation would be simpler if all animals were genotyped (2)
- Genomic info can be extended to ungenotyped (1) animals
  - joint distribution of EBV for ungenotyped (u_1) and genotyped (u_2)

$$p(u_1, u_2) = p(u_2)p(u_1|u_2)$$

$$\mathbf{H} = \begin{pmatrix} var(u_1) & cov(u_1, u_2) \\ cov(u_2, u_1) & var(u_2) \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{pmatrix}$$

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22}) \\ (\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$

Legarra et al., 2009

Combining two sources of relationship

- **A**
  - Contains expected relationships
  - Is limited by the pedigree depth and completeness
  - Depends on accuracy of recording pedigrees

- **G**
  - Contains number of alleles shared between animals weighted by heterozygosity
  - No limitations regarding to number of past generations
  - Depends on allele frequency and quality of genomic data

About the matrices

- Inverse of **H** is used in MME

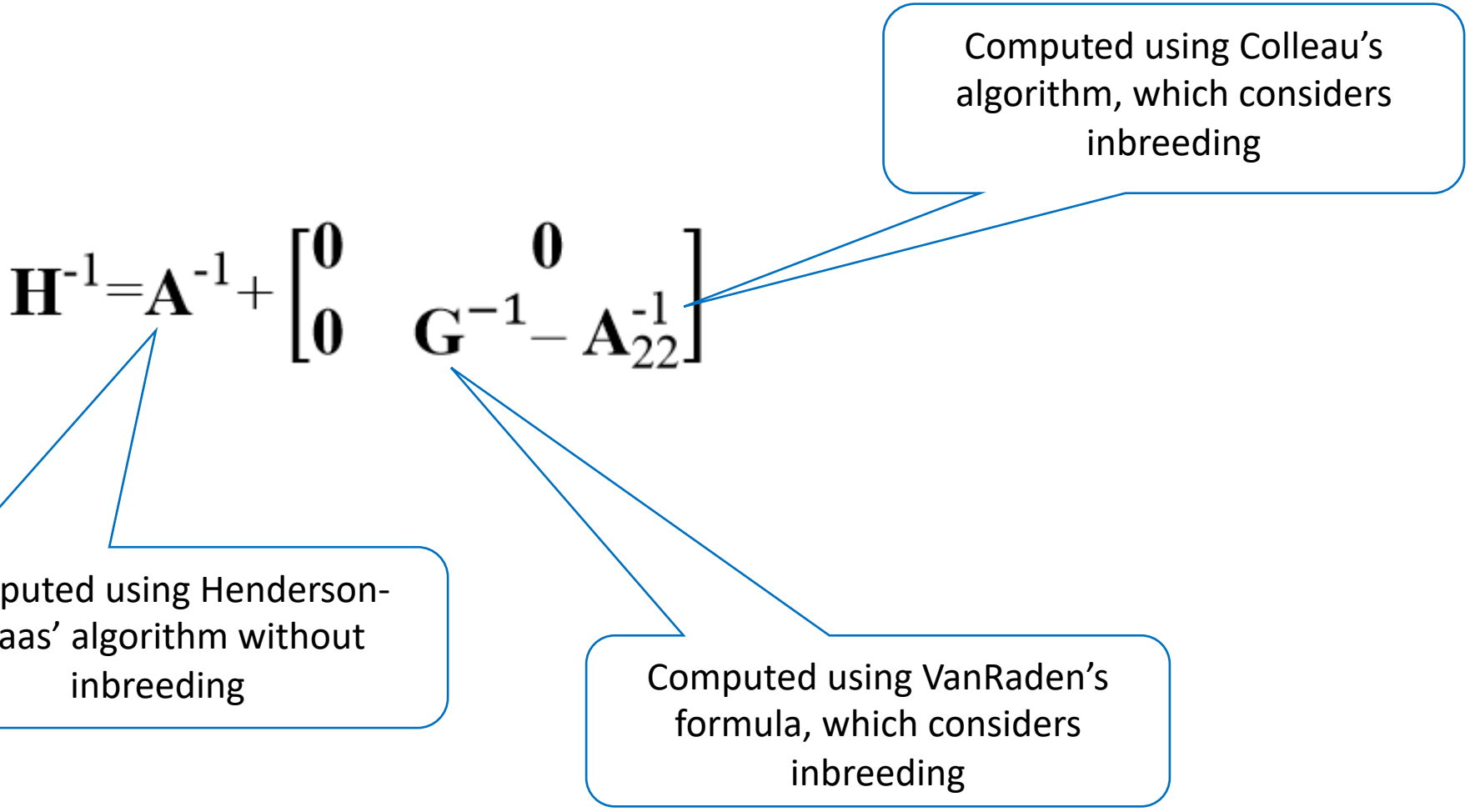
$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

…and avoiding 'double counting'

Inverse of the regular pedigree relationship matrix

Correcting for genomic relationships

Computing all matrices before 2016



Initial tests with ssGBLUP

- Tsuruta et al. (2011)
  - US Holsteins final score
  - 8.9M phenotypes | 7.9M pedigree | 17.3k genotypes (6.9k validation)
  - Inflated GEBV for young bulls (validation)
  - Solution: to reduce $\mathbf{A}_{22}^{-1}$

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

- τ and ω to reduce inflation in GEBV

Experience with simulated data

- Pocrnic et al. (2016)
- Each of the 10 generations: 5 males mated 12.5k females
- 138k pedigree | 75k genotyped animals
- Average inbreeding in generation 10 = 0.21
- **No convergence after 5000 iterations**

- Ideal simulated population
- No missing pedigree
- All recent generations were in the pedigree file
- **Convergence obtained with ω = 0.70**

Computing all matrices after 2016

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

Computed using Colleau's algorithm, which considers inbreeding

Computed using Henderson-Quaas' algorithm **with inbreeding**

Computed using VanRaden's formula, which considers inbreeding

To prepare data for ssGBLUP with inbreeding in A^{-1}

- renumf90

```
EFFECT
1 1 1 cross alpha
RANDOM
animal
OPTIONAL
mat mpe
FILE
aaaped.dat
FILE_POS
1 2 3 4 5
SNP_FILE
allsnp.dat_clean
PED_DEPTH
4
INBREEDING
pedigree
(CO)VARIANCES
28.904825
```

UNIVERSITY OF GEORGIA

Compatibility between G and A_{22}

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

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

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

$\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}$ Inflation/deflation???

Blending and compatibility

- These are two different things
- Many people do not understand this
- “compatibility” tries to put **G** and **A** in the same scale
- “blending” : $\mathbf{G} = \mathbf{G}*0.95 + \mathbf{A}_{22}*0.05$
  - used to have an invertible **G**
  - assigns part of the genetic variance to pedigree – not markers

Options for Blending and compatibility

- **Blending**
- OPTION AlphaBeta alpha beta
  - G = alpha*G + beta*A_{22}

- Compatibility
- OPTION tunedG
  - 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)

$$\rho = \frac{1}{n^2}\left(\sum_i \sum_j \mathbf{A}_{22\,i,j} - \sum_i \sum_j \mathbf{G}_{i,j}\right) \qquad \mathbf{G}^* = (1 - \rho/2)\,\mathbf{G} + \mathbf{1}\mathbf{1}'\,\rho$$

Forcing G to be similar to A_{22}

- Vitezica et al. (2011) and Christensen et al. (2012) provided an unbiased method that forces the same genetic base across **G** and $\mathbf{A_{22}}$:

$$\boldsymbol{G}^* = a + b\boldsymbol{G}$$

  - a accounts for old relationships among non-genotyped ancestors
  - b accounts for reduction in the genetic variance

$$a + b\,\overline{\boldsymbol{G}} = \overline{\boldsymbol{A}}_{22}$$

$$a + b\,\overline{diag(\boldsymbol{G})} = \overline{\left(diag(\boldsymbol{A_{22}})\right)}$$

Forcing G to be similar to A_{22}

Recipe (default in blupf90)

- Compute $\boldsymbol{G}$ with current allele frequencies
- Compute $\boldsymbol{A}_{22}$
- Solve equations $a + b\,\overline{\boldsymbol{G}} = \overline{\boldsymbol{A}}_{22}$, $a + b\,\overline{diag(\boldsymbol{G})} = \overline{\left(diag(\boldsymbol{A_{22}})\right)}$
- Get new $\boldsymbol{G}^{*} = a + b\boldsymbol{G}$
- Build final $\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} + \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G^{*-1}} - \boldsymbol{A_{22}^{-1}} \end{pmatrix}$

Does actually G resemble A_{22}?

- If pedigree is good and genotyping is good, yes it does!

- Usually
  - $Cor\left(A_{22_{ij}}, G_{ij}\right) \approx 0.8$
  - If $Cor\left(A_{22_{ij}}, G_{ij}\right) > 0.95$ genomic is not so informative
  - If $Cor\left(A_{22_{ij}}, G_{ij}\right) < 0.5$ mislabeling of samples or heterogeneous population
  - $Cor\left(F_{pedigree_i}, F_{genomic_i}\right) \approx 0.5$
- Useful for quality control

Main scaling parameters in ssGBLUP

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

$$\mathbf{G}=(\alpha\mathbf{G}+\beta\mathbf{A}_{22})$$

- **Blending**
- makes **G** positive-definite
- α = from 0.95 to 0.80
- Improves convergence

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \tau(\alpha\mathbf{G}+\beta\mathbf{A}_{22})^{-1}-\omega\mathbf{A}_{22}^{-1}\end{bmatrix}$$

- **Scaling**
- τ and ω
- Used for compatibility between $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$
- Reduces inflation

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \tau((\alpha\mathbf{G}+\beta\mathbf{A}_{22})+\mathbf{11}'a)^{-1}-\omega\mathbf{A}_{22}^{-1}\end{bmatrix}$$

- **Tuning**
- tunedG
- Accounts for selection in **A**
- Improves accuracy and reduces bias

$$a=\frac{1}{n^2}\left(\sum_i\sum_j\mathbf{A}_{22}-\sum_i\sum_j\mathbf{G}\right)$$

Should τ and ω be used in ssGBLUP evaluations?

- Need for scaling parameters depend on compatibility among matrices
  - Pedigree, genomic, pedigree for genotyped animals

- Most of the compatibility problems are caused by
  - Ignoring inbreeding in $\mathbf{A}^{-1}$
  - Missing pedigrees for genotyped animals
  - Wrong definition of UPG
  - Ignoring inbreeding for UPG

Validation of genomic models