# The Single Step: Genomic Evaluation for all

A. Legarra, I. Misztal, I. Aguilar

*INRA, UR 631 SAGA, BP 52627, F-31326 Castanet Tolosan, France*
*University of Georgia, Department of Animal and Dairy Science, Athens, 30602, USA*
*INIA, Las Brujas, 90200, Uruguay*

*andres.legarra@toulouse.inra.fr*

# Consider evolution of *genetic evaluation* methods

- we (animal breeders) like generality
- Animal populations (particularly ruminants) are complex:

| | |
|---|---|
| pedigree loops | overlapping generations, |
| culling | heterogeneous information |

- For this, popular methods consider all data simultaneously

| | |
|---|---|
| All relationships (**A** & $\mathbf{A}^{-1}$) | Environmental factors (BLUP) |
| All records (test-day, repeatability models)<br>I mean, *all* records<br>(decades of records, avoiding bias due to selection) | |
| All traits - missing data | Unknown parent groups |

# Consider evolution of *genetic evaluation* methods

- Computing was made simpler with more powerful computers, but
  - once a coherent & elegant framework is established, (almost) everything is feasible
  - smart people are much more important than brutal force

- Inversion of **A**
- Iteration on data
- Sparse matrices
- Approximate/iterative methods for reliabilities
- …

# Consider evolution of *genomic evaluation* methods

- Very fast use of powerful algorithms
  - Gauss-Seidel with Residual Update, PCG
  - Lasso / Elastic Net
  - EM

- Inclusion of pedigree & fixed effects
- records?

# Single Step as a missing data problem

- Methods for genomic evaluation lack of a general way of using traits recorded in relatives
  - If relatives do *not* have genotype of their own
- We can see genotype as a missing data problem (Christensen & Lund, 2010)
- « Genotype » :
  - at the SNPs
  - at multiallelic markers (haplotypes)
  - at the genes/QTLs themselves
- the following derivations are very general

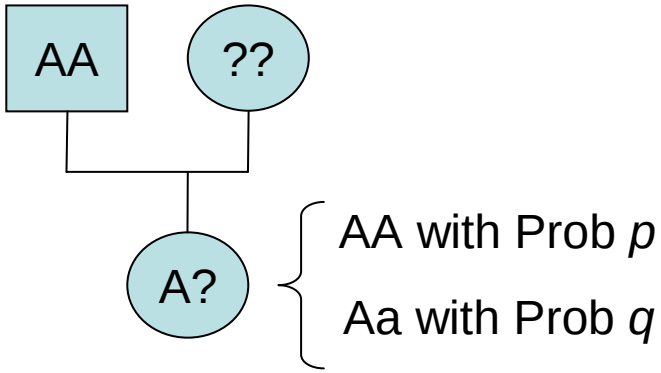
# Missing data

## Fill-in missing data: data augmentation

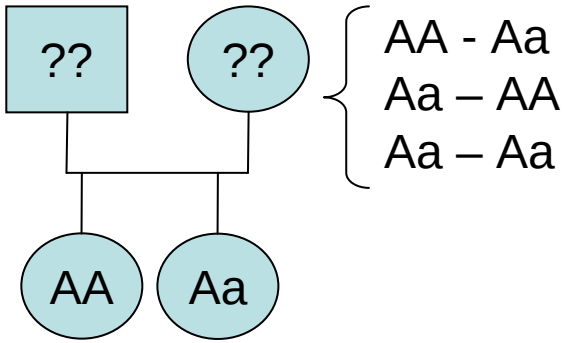
- *« data augmentation refers to a scheme of augmenting the observed data so as to make it more easy to analyze »* (Tanner & Wong, 1987)
  - Two flavors: EM and Bayesian (Posterior distributions)
- Augmenting = imputation
- In both flavors (EM and Bayesian), the *joint distribution* of the imputations needs to be considered
- Consider for instance a very far ancestor
  - Its predicted genotype will be the highest of ($p^2$,$2pq$,$q^2$)
  - But actually its distribution is « AA, Aa, aa » with *Pr* = ($p^2$,$2pq$,$q^2$)
    - Using a point estimator is a poor solution

# (Joint) Uncertainty

- Consider a cow daughter of a genotyped bull




- Consider the parents of two genotyped bulls



# Imputation

- Long-range imputation, linkage-based imputation, peeling, etc
- These are the most exact forms of imputation and work well for 1 or 2 generations or if a subset of markers is genotyped, but…
- Most often *one* imputation is the result
- Very hard to come up with the distribution of the imputations
  - This is in principle feasible by sampling (but very long)

# Linear form of imputation

In the linear world everything is simpler

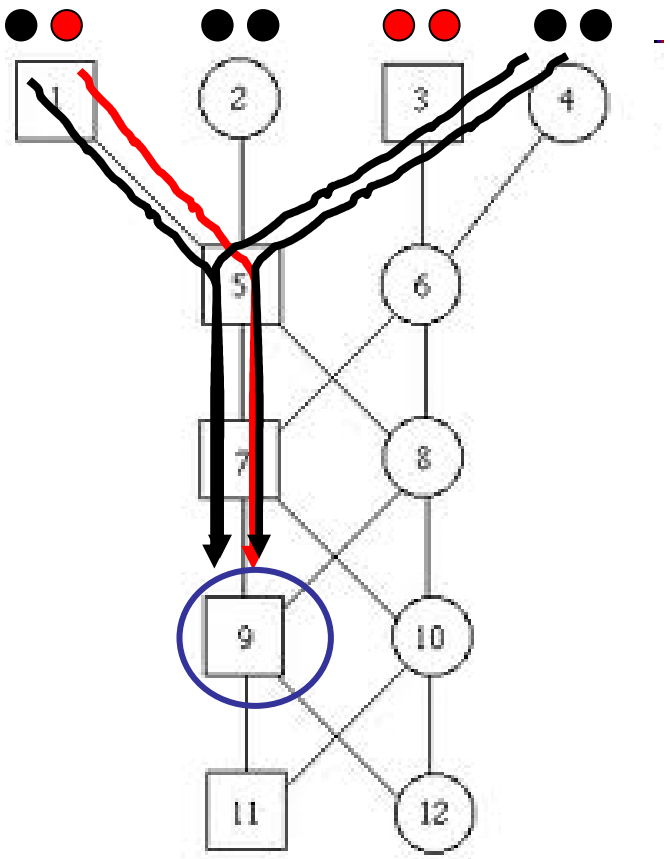
- Consider gene content at a locus

  *g = {0,1,2 for aa,aA,AA}*

- Consider two individuals *i* and *j*
- The basic identity is (Falconer; Cockerham, 1969):
  - $Cov(g_i,g_j) = Pr(IBD)2pq$
- Can we predict gene content of *j* from gene content of *i ?*

# Understanding covariance of gene content

- To each one of the 2M founder alleles we assign a tag *g* saying if the allele is A (g=1) or a (g=0) with probability *p* and *q=1-p*
- What is the covariance between $g_1$ and $g_9$ ?
- 9 *might* inherit alleles from 1
  - With probability Pr(IBD) between 1 and 9
- 9 *might* inherit alleles from 4
  - With probability Pr(IBD) between 4 and 9
- …and so on



# Linear form of imputation

- Therefore we can predict gene content of *j* from gene content of *i*
  - And its distribution (uncertainty)

$$\hat{\mathbf{g}}_j = E\left(\mathbf{g}_j \mid \mathbf{g}_i\right) = \mathbf{2}p + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{g}_i - \mathbf{2}p\right)$$

$$Var\left(\mathbf{g}_j \mid \mathbf{g}_i\right) = \left(\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)2pq$$

This is simple selection index machinery

- This is an approximation: linkage & mendelian rules (incompatibilities) are *not* used
  - But the same approximation is done working with pseudo-data (DYD's)
  - For individuals far away, the linear approximation is very good
- The same expression works for linear functions of gene contents (i.e. breeding values)
  - This is why Legarra et al. (2009) and Christensen & Lund (2010) arrive to the same expression

# Joint distributions

- Using these identities, and summing over all SNPs, we can derive a joint distribution of breeding values

The assumption of normality of the distributions implies no major genes… as in pedigree BLUP

# Joint distributions

- Using these identities, and summing over all SNPs, we can derive a joint distribution of breeding values

Unconditional distribution of genetic values of Genotyped individuals

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}\sigma_u^2) \text{ and}$$

The assumption of normality of the distributions implies no major genes… as in pedigree BLUP

# Joint distributions

- Using these identities, and summing over all SNPs, we can derive a joint distribution of breeding values

Unconditional distribution of genetic values of Genotyped individuals

$$p\left(\mathbf{u}_2\right) = N\left(\mathbf{0}, \mathbf{G}\sigma_u^2\right) \text{ and}$$

Conditional distribution of Non-Genotyped individuals

$$p\left(\mathbf{u}_1 \middle| \mathbf{u}_2\right) = N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11}\sigma_u^2 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\sigma_u^2\right)$$

The assumption of normality of the distributions implies no major genes… as in pedigree BLUP

# Joint distributions

- Using these identities, and summing over all SNPs, we can derive a joint distribution of breeding values

Unconditional distribution of genetic values of Genotyped individuals

$$p(\mathbf{u}_2) = N(\mathbf{0}, \mathbf{G}\sigma_u^2) \text{ and}$$

Conditional distribution of Non-Genotyped individuals

$$p(\mathbf{u}_1|\mathbf{u}_2) = N(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2, \mathbf{A}_{11}\sigma_u^2 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\sigma_u^2)$$

$$p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_2)\, p(\mathbf{u}_1|\mathbf{u}_2)$$

**Joint distribution**

The assumption of normality of the distributions implies no major genes… as in pedigree BLUP

# For BLUP: only covariances are needed

## ➔Model in one step (Single Step GBLUP)

Aguilar et al., 2010; Christensen & Lund, 2010

$$Var\begin{pmatrix}\mathbf{u}_1\\ \mathbf{u}_2\end{pmatrix}=\mathbf{H}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}=\begin{bmatrix}\overbrace{\mathbf{A}_{11}+\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\left(\mathbf{G}-\mathbf{A}_{22}\right)\mathbf{A}_{22}^{-1}\mathbf{A}_{21}}^{\textit{non genotyped}} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G}\\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \underbrace{\mathbf{G}}_{\textit{genotyped}}\end{bmatrix}$$

- Incredibly: $\mathbf{H}^{-1}$ is very simple:

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

# Single step GBLUP

**W**: incidence matrix of all animals on all data

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W} + \mathbf{H}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

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

**G**

**A**: pedigree relationship matrix

$\mathbf{A}_{22}$: pedigree matrix among genotyped individuals

This **G** is *any* matrix describing « genomic » covariances of breeding values;

it does not restrict to VanRaden's (2008) GBLUP

# Single Step Bayes?

- **G** can be (pre) computed by some method (BayesB, Bayesian Lasso, etc.) to be plugged in:
  - TABLUP (Zhang et al. 2010), HetVarGBLUP (Legarra et al. 2011)

- In principle, one can extend the Single Step to non-linear (Bayesian) models
- Monte Carlo SingleStep BayesB:

```
do i=1,niter
  sample missing genotypes from
  a=a+BayesB(all genotypes, all y)
enddo
a=a/niter
```

$$\begin{cases} \hat{\mathbf{g}}_1 = E(\mathbf{g}_1 \mid \mathbf{g}_2) = \mathbf{2}p + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{g}_2 \\ Var(\mathbf{g}_1 \mid \mathbf{g}_2) = \left(\mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)2pq \end{cases}$$

# Computing stuff

- Working with $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$, is a challenge. Because cost of inversion is cubic, this is tenable for < 100,000 genotypes
  - See Aguilar et al. 2011 for details
- However, most modern iteration on data methods (Jacobi, PCG) solve **Cx**=**b** by computing repeteadly **Cx**.
- We know how to do this (very) efficiently for

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W}+\mathbf{A}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Iteration on data

# Computing stuff

- Working with $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$, is a challenge. Because cost of inversion is cubic, this is tenable for < 100,000 genotypes
  - See Aguilar et al. 2011 for details
- However, most modern iteration on data methods (Jacobi, PCG) solve **Cx**=**b** by computing repeteadly **Cx**.
- We know how to do this (very) efficiently for

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W}+\mathbf{A}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Iteration on data

as **Gx**=**Z**(**D**(**Z**'**x**))

- We also know how to compute (very) efficiently **Gx** and $\mathbf{A}_{22}\mathbf{x}$ but *not* $\mathbf{G}^{-1}\mathbf{x}$ or $\mathbf{A}_{22}^{-1}\mathbf{x}$

*Two possible solutions follow*:

by Colleau's (2002) algorithm

# Extended MME

- Or the unsymmetric equations

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{W}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{W}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_1 + \mathbf{A}^{11}\sigma_u^{-2} & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_2 + \mathbf{A}^{12}\sigma_u^{-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_2\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_1 + \mathbf{A}^{21}\sigma_u^{-2} & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_2 + \mathbf{A}^{22}\sigma_u^{-2} & \mathbf{I}\sigma_u^{-2} & -\mathbf{I}\sigma_u^{-2} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_u^{-2} & \mathbf{A}_{22}\sigma_u^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}\sigma_u^{-2} & \mathbf{0} & -\mathbf{G}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hat{\boldsymbol{\varphi}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

For a total number of operations *O(n)+O(mp)*

# 1- Extended MME

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W}+\mathbf{H}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}$$

- Is equivalent to

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{W}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{W}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_1+\mathbf{A}^{11}\sigma_u^{-2} & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_2+\mathbf{A}^{12}\sigma_u^{-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_2\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_1+\mathbf{A}^{21}\sigma_u^{-2} & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_2+\mathbf{A}^{22}\sigma_u^{-2} & \mathbf{I}\sigma_u^{-2} & -\mathbf{I}\sigma_u^{-2} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_u^{-2} & \mathbf{A}_{22}\sigma_u^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}\sigma_u^{-2} & \mathbf{0} & -\mathbf{G}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hat{\boldsymbol{\varphi}} \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

For a total number of operations *O(n)+O(mp)*

as in regular BLUP

as in any genomic evaluation

# Extended MME

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W}+\mathbf{H}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

- Has the same solution as

$$\left[\begin{array}{ccc|cc} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{W}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{W}_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_1\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_1+\mathbf{A}^{11}\sigma_u^{-2} & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_2+\mathbf{A}^{12}\sigma_u^{-2} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_2\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_1+\mathbf{A}^{21}\sigma_u^{-2} & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_2+\mathbf{A}^{22}\sigma_u^{-2} & \mathbf{I}\sigma_u^{-2} & -\mathbf{I}\sigma_u^{-2} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_u^{-2} & \mathbf{A}_{22}\sigma_u^{-2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I}\sigma_u^{-2} & \mathbf{0} & -\mathbf{G}\sigma_u^{-2} \end{array}\right] \left[\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hline \hat{\boldsymbol{\varphi}} \\ \hat{\boldsymbol{\gamma}} \end{array}\right] = \left[\begin{array}{c} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y} \\ \hline \mathbf{0} \\ \mathbf{0} \end{array}\right]$$

Regular BLUP

Genomic stuff

Separate the two blocks of equations

# 2- Ducrocq's (& Legarra) iterative system (Interbull meeting)

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}\mathbf{R}^{-1}\mathbf{W}+\mathbf{H}^{-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad \mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1} \end{bmatrix}$$

- Can be solved iterating on

regular MME

RHS correction for genomic information

$$1: \quad \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{W}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{W}_2 \\ \mathbf{W}_1\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_1+\mathbf{A}^{11}\sigma_u^{-2} & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_2+\mathbf{A}^{12}\sigma_u^{-2} \\ \mathbf{W}_2\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_1+\mathbf{A}^{21}\sigma_u^{-2} & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_2+\mathbf{A}^{22}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y}+\sigma_u^{-2}\left(\hat{\boldsymbol{\gamma}}-\hat{\boldsymbol{\varphi}}\right) \end{bmatrix}$$

$2: \quad \mathbf{G}\hat{\boldsymbol{\gamma}} = -\hat{\mathbf{u}}_2$

Deviations due to genomics

$3: \quad \mathbf{A}_{22}\hat{\boldsymbol{\varphi}} = -\hat{\mathbf{u}}_2$

Avoid double counting of relationships

For a total number of operations *O(n)+O(mp)*

as in regular BLUP

as in any genomic evaluation

Similar schemes can iterate over pedigree BLUP and SNP effects

# Compatibility of **G** and **A**

- **G** and **A** need to be on the same scale (same base population, same genetic variance)
  - Large deviations of HW (e.g. in crossbreds) make theory inadequate
  - Solution: build **A** and **G** according to a crossbred theory (Lo et al., 1993; Harris & Johnson 2010)
  - More work needs to be done

# Compatibility of **G** and **A**

- More generally: allelic frequencies *(p)* in the base population are unknown
  - This is not serious if there is no selection or data files are large (dairy)
  - In presence of (old) selection, deviations of both genetic base and genetic variance will exist (Chen et al., 2011; Vitezica et al. 2011; this congress)
- Correction through Wright's Fst (Powell et al; 2010):
  - matches « new » and « old » populations
  - considers both change of base *and* reduction in variance

$$\mathbf{G}^{*} = \left(1 - \frac{\alpha}{2}\right)\mathbf{G} + \mathbf{1}\mathbf{1}'\alpha$$

$$\alpha = mean\left(\mathbf{A}_{22} - \mathbf{G}\right)$$

# Bias & inflation

- Genomic predictions in dairy seem to be inflated (biased) (e.g. Aguilar et al. 2011)
  - The problem exists also for pedigree-based BLUP
    - even in simulations (Vitezica et al., 2011)
  - Seems to be alleviated (to some extent) by playing with weights of **G** and $\mathbf{A}_{22}$
  - Too odd to be luck…
- Is there anything wrong with basic theory?
  - Certainly unrelated base populations are a fallacy
  - …

# Why Single Step

- Generality
- DYD's are difficult…
  - for maternal traits,
  - species with some phenotypes recorded on candidates (beef, swine)
  - small progeny numbers (sheep)
  - weighting DYD for complex traits (i.e. RR models) is difficult (multivariate equivalent of *edc*'s)
- Consider Ducrocq's (& Legarra) iterative system

# Two-step *vs.* Single Step

Pedigree-based BLUP

$$1: \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}_1'\mathbf{R}^{-1}\mathbf{W}_1 & \mathbf{X}_2'\mathbf{R}^{-1}\mathbf{W}_2 \\ \mathbf{W}_1\mathbf{R}^{-1}\mathbf{X}_1 & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_1 + \mathbf{A}^{11}\sigma_u^{-2} & \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{W}_2 + \mathbf{A}^{12}\sigma_u^{-2} \\ \mathbf{W}_2\mathbf{R}^{-1}\mathbf{X}_2 & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_1 + \mathbf{A}^{21}\sigma_u^{-2} & \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{W}_2 + \mathbf{A}^{22}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}_2'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} + \sigma_u^{-2}\left(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\varphi}}\right)$$

$$2: \mathbf{G}\hat{\boldsymbol{\gamma}} = -\hat{\mathbf{u}}_2$$

~DYD's

$$3: \mathbf{A}_{22}\hat{\boldsymbol{\varphi}} = -\hat{\mathbf{u}}_2$$

correction for double counting

No need for weights because MME and the iterative process take care of them

The Single Step can be seen as an iterated « DYD + genomic evaluation » system

# Why single step

- Patry & Ducrocq (2011a) showed that bias will plague national evaluations if selection is based on genomic proofs
  - No way of including this in pedigree-BLUP except using pseudo-data in the RHS (Patry & Ducrocq 2011b)
  - which is what the Single Step does in an exact manner

- GWAS/estimation of SNP effects can still be done: easy jump between Single Step and SNP effects (Strandén and Garrick, 2009)

$$\hat{\mathbf{a}} = \mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}}$$

SNP effects

EBV's

# Take-home message

- Single Step is simpler than it seems
  - Computationally feasible
- Slightly more complex than national pedigree-BLUP

- Compatibility problems solved
- When *not* to use it?
  - If everybody is genotyped (and with no selective genotyping !)
  - If somebody comes with a « super-peeling like » algorithm:
    - using long-range phasing,
    - Mendelian coherence,
    - imputing all individuals in a pedigree *and*
    - considering uncertainty in the « data augmentation » procedure

INRA

# Acknowledgements


- ANR project Amasgen, Apisgene
- GENOMIA funding:

  www.poctefa.eu

  – V Ducrocq
  – P VanRaden
  – D Johnson
  – MA Toro
  – ZG Vitezica
- Toulouse bioinformatics platform (bioinfo.genotoul.fr)