Studies on inflation of GEBV in single-step GBLUP for type

Ignacy Misztal, Shogo Tsuruta, Yutaka Masuda, Daniela Lourenco, Heather Bradford

University of Georgia

A. Legarra, INRA, France

Tom Lawlor, Holstein Association

Paul VanRaden, USDA

Status of single step

- Large data problem solved
  - APY G inverse + indirect computation of A_{22}^{-1}

- Convergence problem solved
  - Inbreeding in A
  - UPG formulas for H matrix
  - Cutting unneeded pedigrees

- Slow time/round solved
  - Extensive parallel processing

- Inflation of GEBV – not quite solved
  - No problem with broilers, beef and pigs (after QC)

Initial effort to reduce inflation

$$H^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} + \lambda(G^{-1} - A_{22}^{-1}) \end{bmatrix}$$

λ	R^2 (%)	b1
1.0	41	0.76
0.9	41	0.81
0.8	41	0.84
0.7	40	0.88
0.6	40	0.90
0.5	39	0.92
0.3	35	0.91

Aguilar et al., 2010

Parameter omega

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

Trait	b1	
	ω = 1	ω = 0.7
Stature	0.84	0.95
Body depth	0.74	0.87
Foot angle	0.70	0.85
Udder depth	0.82	0.98
Teat length	0.78	0.97

Tsuruta et al., 2011

Why inflation and biases?

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

1-ungenotyped animals
2-genotyped animals

$$H = A + \begin{bmatrix} A_{12}A_{22}^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix} \begin{bmatrix} G - A_{22} \end{bmatrix} \begin{bmatrix} I & I \end{bmatrix} \begin{bmatrix} A_{22}^{-1}A_{21} & 0 \\ 0 & I \end{bmatrix}$$

Scales different $G - A_{22}$ Inflation/ deflation

Levels different $G - A_{22}$ Bias

Incomplete pedigree $G - A_{22}$???

Properties of G and A_{22}

- G - "infinite" pedigree
  - depends on gene frequencies, arbitrary scaling, genotyping accuracy & errors

- A_{22} - depends on pedigree completeness, depth, errors
  - Typical heterogeneous base population

- Adapt
  - a) G to A_{22}?
  - b) A_{22} to G?
  - c) Both?

Scaling G - gene frequencies

$$G \quad = \frac{(\mathbf{M} - 2\mathbf{P}\quad)(\mathbf{M} - 2\mathbf{P}\quad)'}{\sum_i 2p_i\, q_i}$$

M – gene content, P – gene frequencies

VanRaden (2008)

$$\left[(p_0 - q_0)^2 + 2\left(\frac{\sum p_j(1 - p_j)}{n}\right)\left(\frac{\alpha + \beta + 2}{\alpha + \beta}\right)\right]n$$

Different denominator (Gianola, 2009)

1. Use base population gene frequencies (Gengler, 2007; VanRaden, 2008; Christensen and Lund, 2010)
   - Hard to compute
   - Does not work if base population heterogeneous

Scaling G – fixed effects

2. Use a constant for phenotypes of genotyped animals

(Stranden and Christensen, 2010; Vitezica et al., 2011; Fernando et al., 2014)

$$G_j = \frac{(\mathbf{M} - 2\mathbf{P}_j)(\mathbf{M} - 2\mathbf{P}_j)'}{\sum_i 2p_i q_i} \qquad \text{Var}(\boldsymbol{u}_1) = \mathbf{G_1} , \text{Var}(\boldsymbol{u}_2) = \mathbf{G_2}$$

$$\boldsymbol{u}_1 = \boldsymbol{u}_2 + \boldsymbol{\mu} \qquad \text{Gene frequencies change the mean of EBV only}$$

$$\boldsymbol{y}_{genot} = \boldsymbol{hys} + \boldsymbol{u}_1 + \boldsymbol{e} \qquad \equiv \qquad \boldsymbol{y}_{genot} = \boldsymbol{hys} + \boldsymbol{\mu} + \boldsymbol{u}_2 + \boldsymbol{e}$$

Add a mean (or group effect) to model for genotyped animals
Works for means, not inflation
No effect for production if only bulls genotyped

Scaling G – compatibility with A

3. Scale G for compatibility with $\mathbf{A}_{22}$ (VanRaden, 2008; Chen et al., 2011; Vitezica et al., 2011)
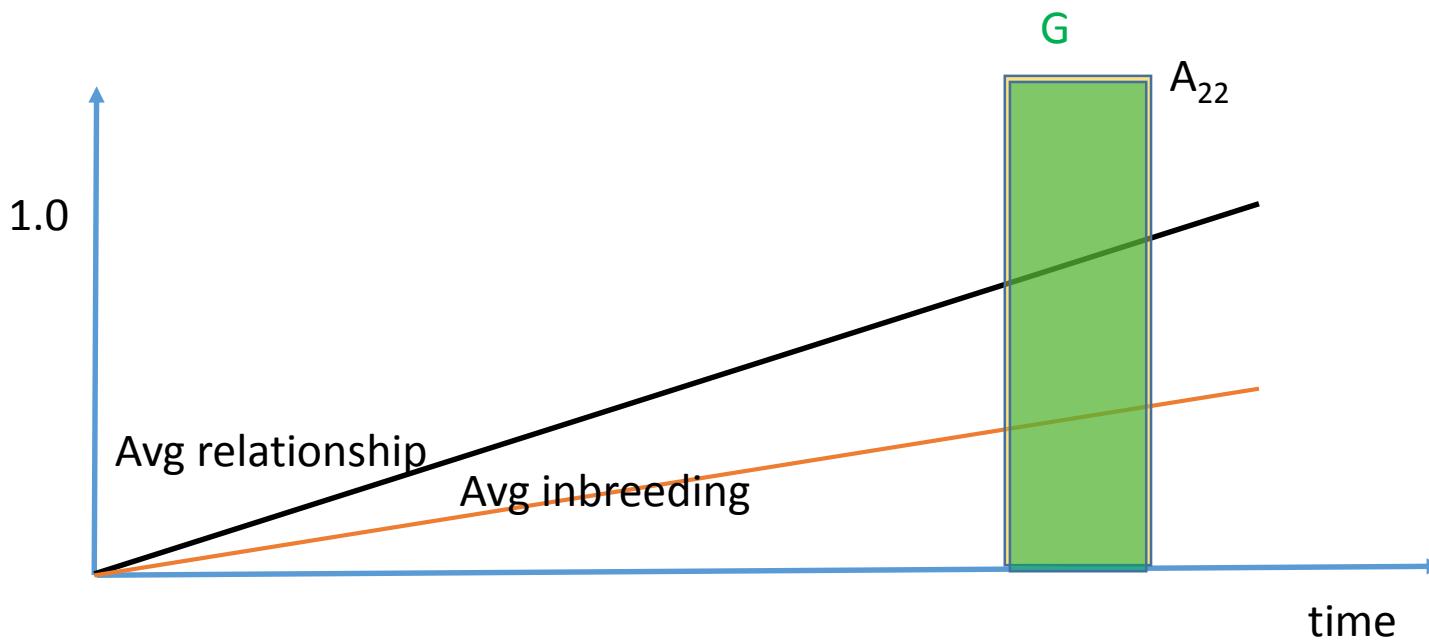

$$\mathbf{G} = \alpha + \left(1 - \frac{\alpha}{2}\right)\mathbf{G}_0, \qquad \alpha: \text{avg}(a_{22,ij}) = \text{avg}(g_{ij})$$

Computations

- Holstein type data up to 2014

- 18 traits

- 569k genotyped animals

- 1711 validation bulls with 50 daughters

- Single-step by blup90iod

- DYD2014=b_0+b_1GEBV2010

R^2 and b$_1$ without and with inbreeding in A
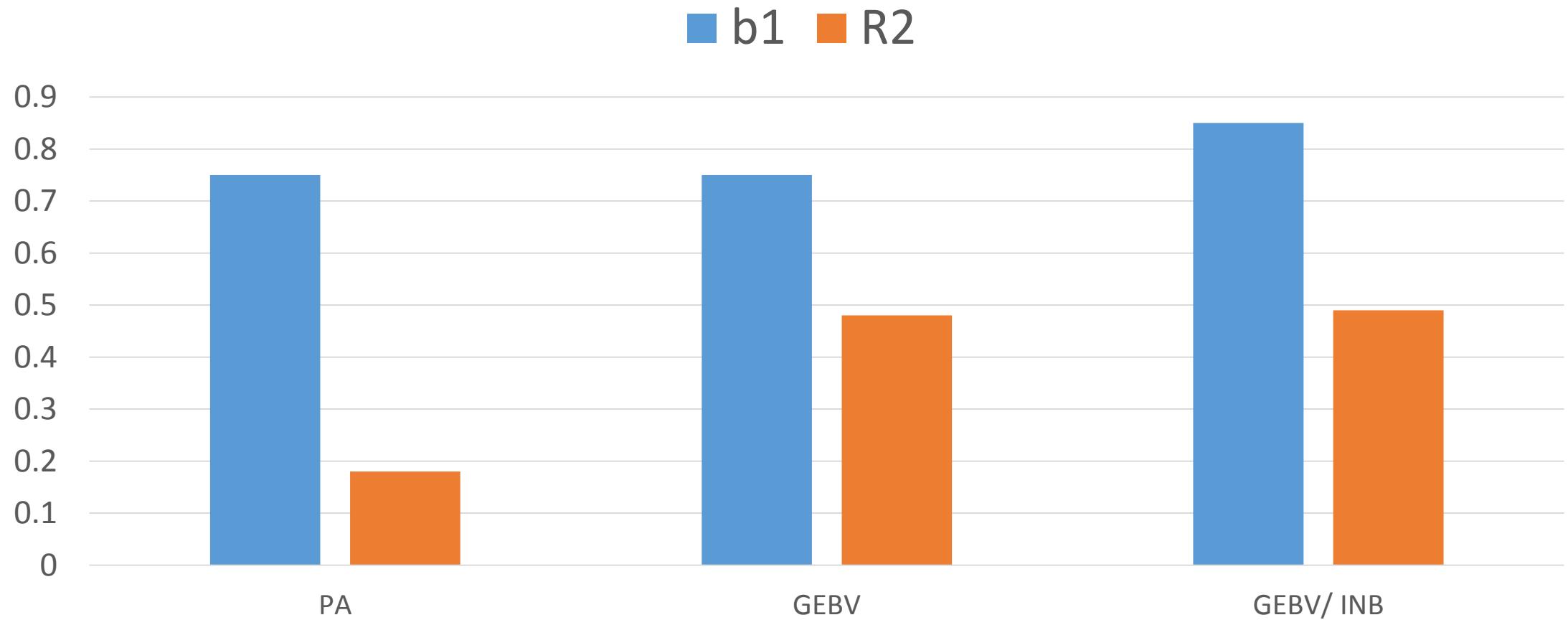
R^2 and b$_1$ with omegas

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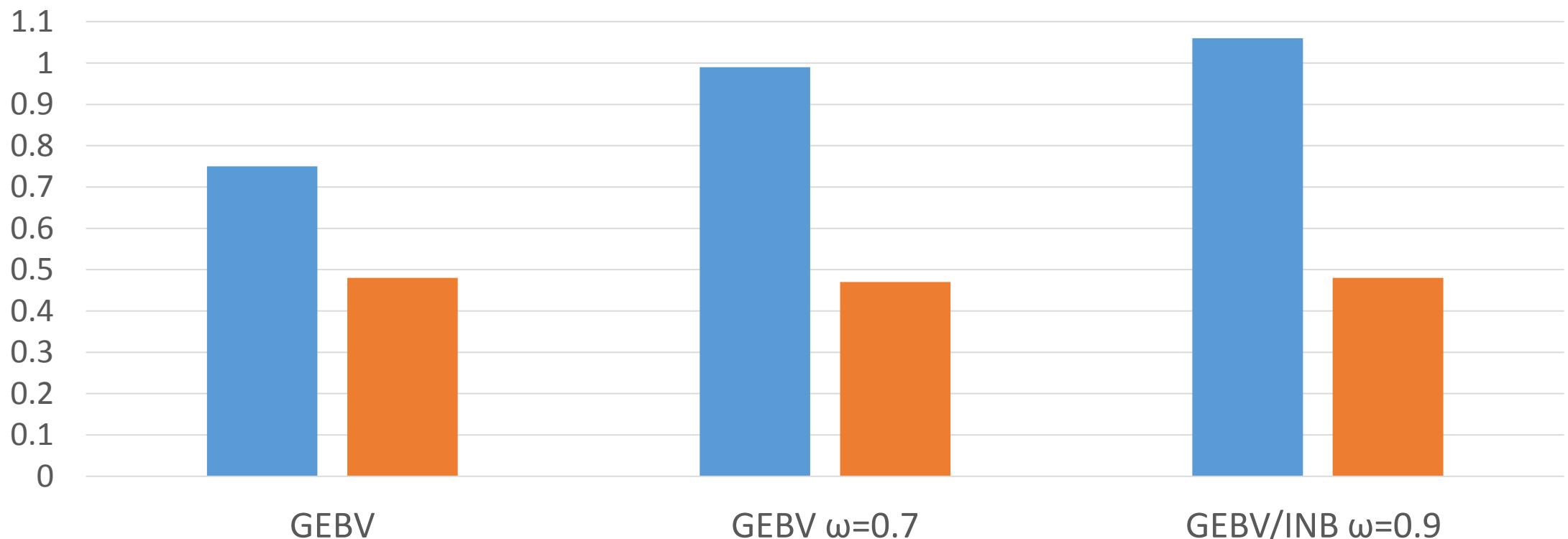

Inbreeding in A reduces optimal omega

GEBV Decomposition for Young Animals

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

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$
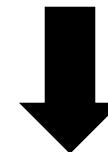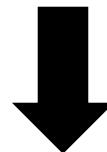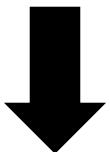
GEBV = w1 PA + w2 DGV - w3 PI

GEBV Decomposition for Young Animals

Two parents known -no inbreeding in A

$$\text{GEBV} = \frac{2}{2 + g^{ii} - a_{22}^{ii}} \text{ PA} + \frac{g^{ii}}{2 + g^{ii} - a_{22}^{ii}} \text{ DGV} - \frac{a_{22}^{ii}}{2 + g^{ii} - a_{22}^{ii}} \text{ PI}$$

Inbreeding F_i

$$\text{GEBV} = \frac{2/(1 - Fi)}{2/(1 - Fi) + g^{ii} - a_{22}^{ii}} \text{ PA} + \frac{g^{ii}}{2/(1 - Fi) + g^{ii} - a_{22}^{ii}} \text{ DGV} - \frac{a_{22}^{ii}}{2/(1 - Fi) + g^{ii} - a_{22}^{ii}} \text{ PI}$$

$$F_i = \frac{F_s}{2} + \frac{F_d}{2}$$

Under inbreeding:
Smaller weights of DGV and PI
Larger share of PA

No inbreeding in A if at most 3 generations

Meaning of λ parameter

$$H^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} + \lambda(G^{-1} - A_{22}^{-1}) \end{bmatrix}$$

$$\text{GEBV} = \frac{2/\lambda}{2/\lambda + g^{ii} - a_{22}^{ii}} \text{PA} + \frac{g^{ii}}{2/\lambda + g^{ii} - a_{22}^{ii}} \text{DGV} - \frac{a_{22}^{ii}}{2/\lambda + g^{ii} - a_{22}^{ii}} \text{PI}$$

$$2/\lambda \equiv 2/(1 - F_i)$$ λ compensates for lack of inbreeding in A

Meaning of ω parameter

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

$$\text{GEBV} = \frac{2/\omega}{2/\omega + g^{ii}/\omega - a_{22}^{ii}} \text{PA} + \frac{g^{ii}/\omega}{2/\omega + g^{ii}/\omega - a_{22}^{ii}} \text{DGV} - \frac{a_{22}^{ii}}{2/\omega + g^{ii}/\omega - a_{22}^{ii}} \text{PI}$$

Denominator larger
Fraction of PI down

Coefficient behind A_{22}^{-1} more important
(Misztal et al., 2010)

Inbreeding and missing pedigree

- Inbreeding = f(pedigree depth)

- Options if incomplete pedigree:
    - Truncate
    - Nonzero inbreeding for unknown parents
        - Prediction by A – UPGs (VanRaden, 1992)
        - Prediction by G – metafounders (Legarra et al., 2015)

R^2 and b$_1$ with inbreeding for phantom parents (UPG)
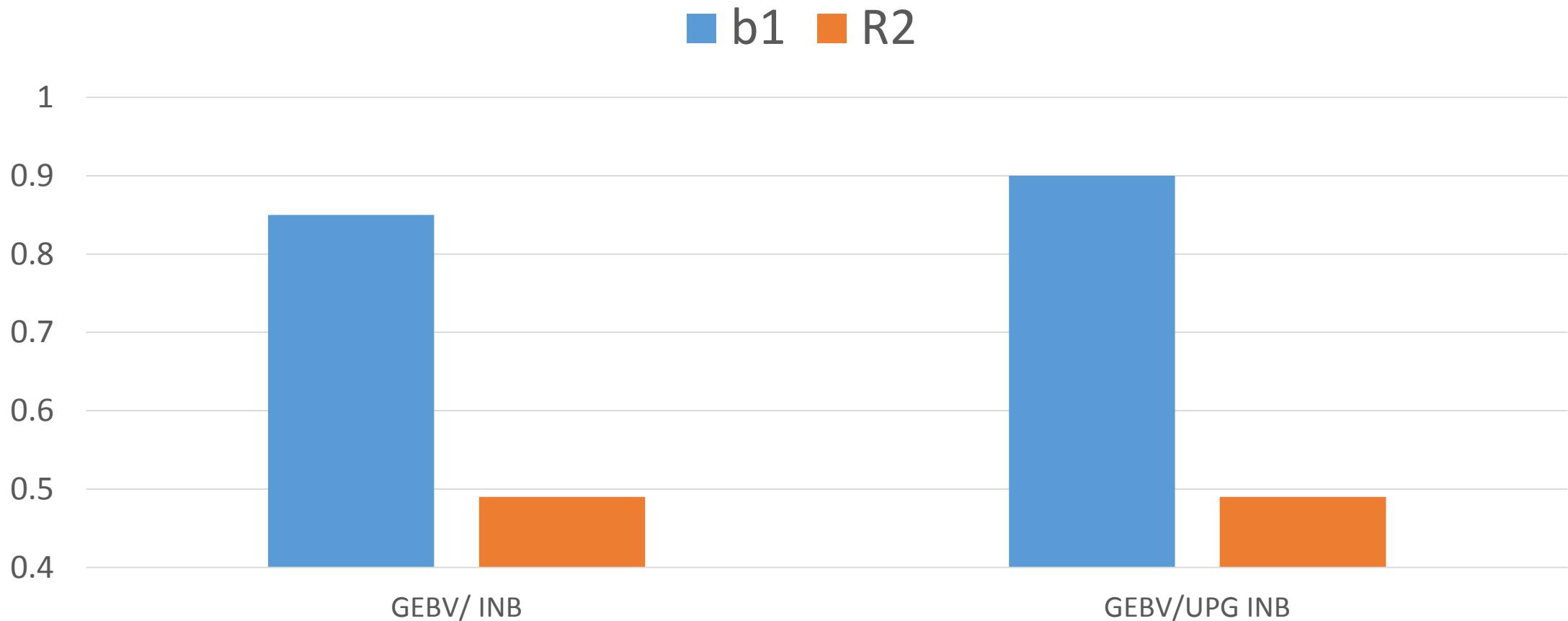
Anything else to raise b_1 above 0.90? Why high corr(b_1, h^2)
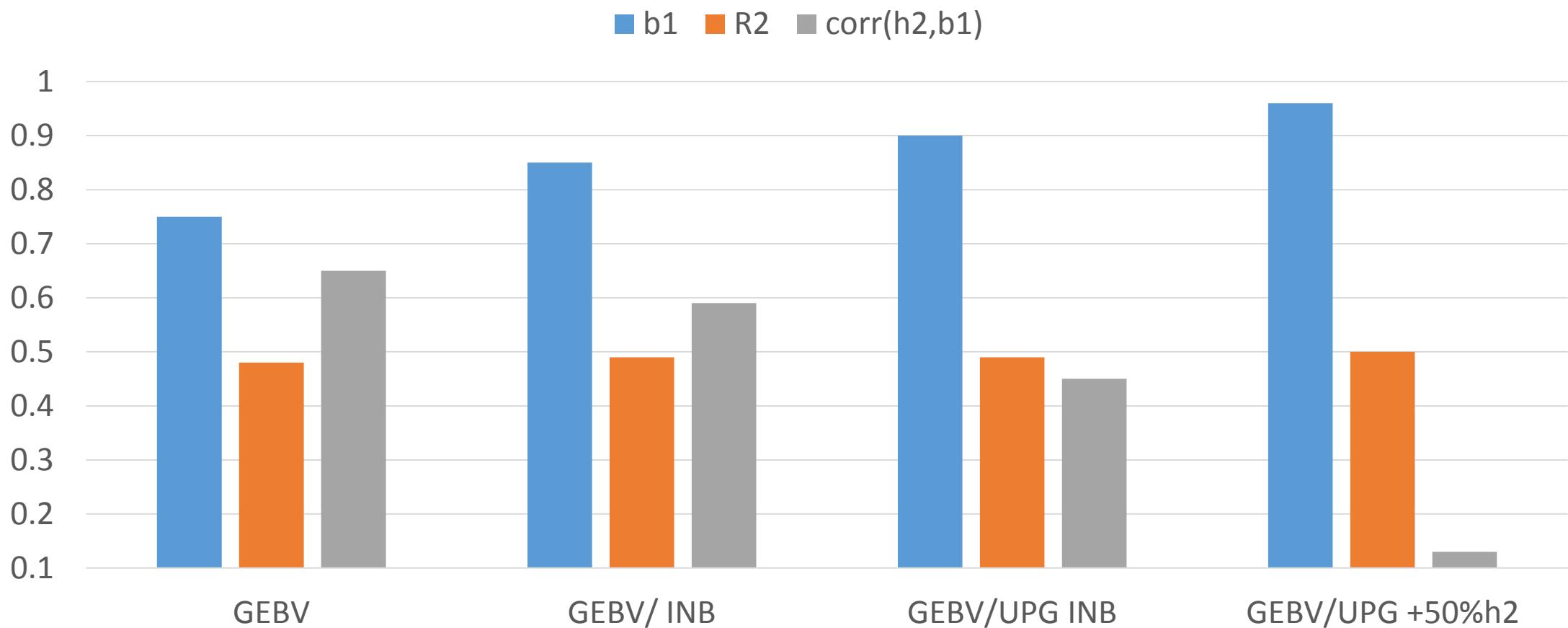
Reduction of h^2 -- Wiggans et al., 2011
Optimal reduction related to intensity of selection (Lawlor, 2017)

Reducing heritability for genotyped animals only?

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

Misztal et al. (2010)

τ	ω	R^2	Regression
1.0	1.0	0.41	0.75
1.5	**0.9**	**0.42**	**0.87**
1.5	0.6	0.41	0.96
1.5	0.4	0.40	1.00
1.0	0.4	0.39	0.97
0.6	0.4	0.39	0.94

Optimal τ=1.5

66% genetic variance for genotyped animals

Other factors influencing inflation

- Pedigree errors

- Foreign animals

- UPG definitions

- Chromosomes XY and 0

- Imputation

- Real inflation lower if DYD biased down

What if multibreed?

- Fixed effect for gene frequencies – does not eliminate inflation
- Possibly ignore, use avg gene frequencies (Simeone et al., 2011; Lourenco et al., 2016)
- Possibly, UPG sufficient (Swan et al., 2012)
- Match A to G using gene frequencies – metafounders (Legarra et al., 2015)

- How many breeds can share SNP60k without loss of accuracy?

Concept of Metafounders

We need to adjust the UPG theory to match **A** to **G** instead of viceversa

In other words, we can infer the relationships across breeds from markers

Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation

Ole F Christensen

Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships

Andres Legarra,*[1] Ole F. Christensen,[†] Zulma G. Vitezica,[‡] Ignacio Aguilar,[§] and Ignacy Misztal**

Genetic evaluation for three-way crossbreeding

Ole F. Christensen[1*], Andres Legarra[2], Mogens S. Lund[1] and Guosheng Su[1]

Metafounders are related to F_{st} fixation indices and reduce bias in single-step genomic evaluations

Carolina A. Garcia-Baccino[1,2], Andres Legarra[3*], Ole F. Christensen[4], Ignacy Misztal[5], Ivan Pocrnic[5], Zulma G. Vitezica[3] and Rodolfo J. C. Cantet[1,2]

Legarra, AGBU 2017

Algorithm for Metafounders

- Construct **G** using equal gene frequencies

- Call UPG metafounders

- Make metafounder effects random and calculate their covariances based on **G**

- Construct **A** and **A**$_{22}$ using these covariances

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

Parameters of H⁻¹ in blupf90

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

Controls additive variance

Controls additive variance, blending
Default value 0.95

Blending for numerical stability
Beta or gamma: 0.01-0.05
Gamma better If causative SNPs

Controls bias

Mainly controls
inflation due to incomplete
pedigree

Pretty good choice

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

If
  * Truncated pedigree
  * Inbreeding in A
  * Inbreeding for UPG
  * G slightly blended and scaled

Why inflation in dairy but not in other species?

- Dairy
  - Strong selection
  - Missing pedigrees particularly for grade cows
- Broilers (Cobb)
  - 3 generations of pedigree and phenotypes – no parents' inbreeding
- Beef (Angus)
  - Data from breeding operations – nearly complete pedigrees
  - Less selection
- Pigs
  - Complete pedigrees – after truncation

Summary

- Incomplete relationships lead to inflation

  - Minimal impact on reliability

- Reducing heritability eliminates bias and increases reliability

- Possibly automatic multibreed scaling with "metafounders"

- Tuning parameters in ssGBLUP useful – now understood

DYD Regression coefficient as function of τ and ω
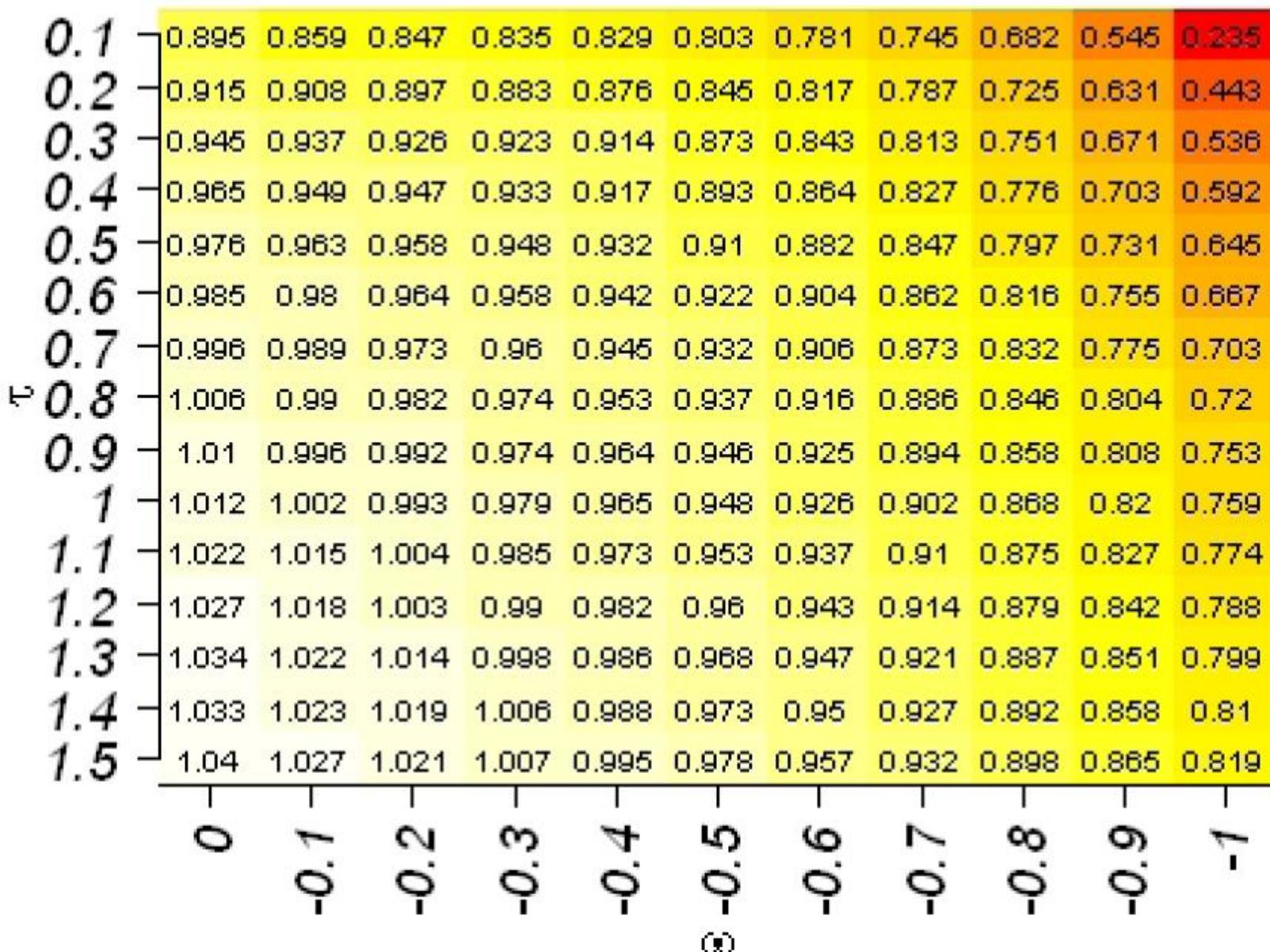
$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \tau\mathbf{G}^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix}.$$

τ	ω	R^2	Regression
1.0	1.0	0.41	0.75
1.5	0.9	0.42	0.87
1.5	0.6	0.41	0.96
1.5	0.4	0.40	1.00
1.0	0.4	0.39	0.97
0.6	0.4	0.39	0.94

Optimal G if divided by 1.5