Advances in single-step genomic evaluations

Daniela Lourenco, M. Bermann, I. Aguilar, A. Legarra, I. Misztal

05/2026 - Interbull

UNIVERSITY OF GEORGIA
College of Agricultural & Environmental Sciences
Animal Breeding and Genetics Group

Single-step is strong now, but …

Reaction to single step

- Proof that it cannot work
- Found in conflict with NBCEC goals
- Denied request to collaborate with Clay Center
- Graduate students discouraged from coming to UGA

Adapted from Ignacy Misztal - UGA Symposium - April 8-9, 2025

The path was not very easy…but it was appealing

Single-step developments - UGA

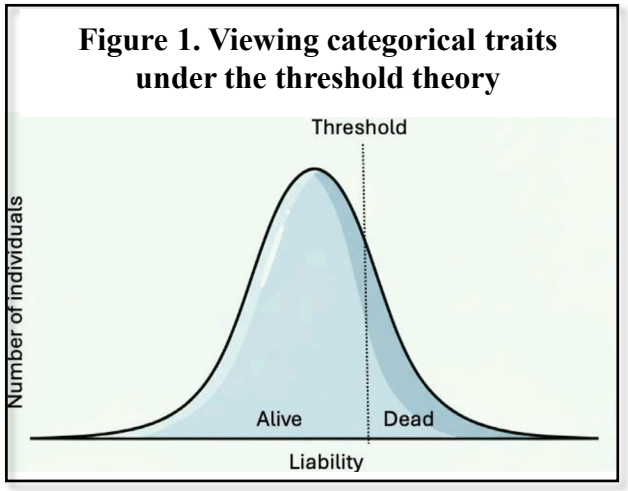
- Compatibility between **G** and $\mathbf{A}_{22}$
- UPG/Metafounders
- APY for large-scale evaluations
- Including MACE information
- Reliability methods
- Indirect predictions
- Reliability of Indirect predictions
- ssGWAS for any population size
- Multi-trait threshold models
- Methods for VCE in large populations
  - GPP
  - MCssGREML
- Efficient algorithms

Case study

- Multiple categorical traits related to health
- Some organizations use single-trait threshold models
- Other organizations use multi-trait linear models
- Why not using multi-trait threshold models?
- No theory for multi-trait threshold models under BLUP and ssGBLUP → Let's develop it!
  - Large-scale routine evaluations
- Gibbs Sampling: not suitable for routine evaluations and large data

Solving a 30-year-old issue

- No theory for multi-trait threshold models under BLUP and ssGBLUP
  - Large-scale evaluation of multiple categorical traits



- Solution: Maximum a posteriori (MAP)
  - Newton-Raphson
  - Expectation-Maximization

→ Optimization method: find the single point value $\hat{\theta}$ or $\hat{\boldsymbol{\theta}}$ that maximizes the posterior probability

Multi-trait threshold-linear models

- Model

$$\begin{bmatrix} \mathbf{l} \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 & 0 \\ 0 & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

- Joint log-likelihood

$$L(\mathbf{y}_1, \mathbf{y}_2, \boldsymbol{\theta}) = L(\mathbf{y}_1|\mathbf{y}_2, \boldsymbol{\theta}) + L(\mathbf{y}_2|\boldsymbol{\theta}) + L(\boldsymbol{\theta})$$

- Maximize $L(\mathbf{y}_1, \mathbf{y}_2, \boldsymbol{\theta})$ to obtain $\boldsymbol{\theta}$
  - Newton-Raphson

$$\boldsymbol{\theta}^{j+1} = \boldsymbol{\theta}^{j} - \mathbf{H}(j)^{-1}\nabla(j)$$

$$\left(\mathbf{W}'\tilde{\mathbf{R}}^{-1}\mathbf{W} + \mathbf{S}\right)\boldsymbol{\theta}^{j+1} = \mathbf{W}'\tilde{\mathbf{R}}^{-1}\tilde{\mathbf{y}}$$

calculate $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{R}}^{-1}$ at each iteration
Solve the MME until convergence

- Expectation-Maximization
  - E-step

$$Q\left(\boldsymbol{\theta}\left|\boldsymbol{\theta}^{j}\right.\right) = L(\tilde{\boldsymbol{l}}|\mathbf{y}_2, \boldsymbol{\theta}) - \frac{1}{2}\Sigma_{i=1}^{n}\frac{1}{\sigma_i^2}Var(l_i) + L(\mathbf{y}_2|\boldsymbol{\theta}) + L(\boldsymbol{\theta})$$

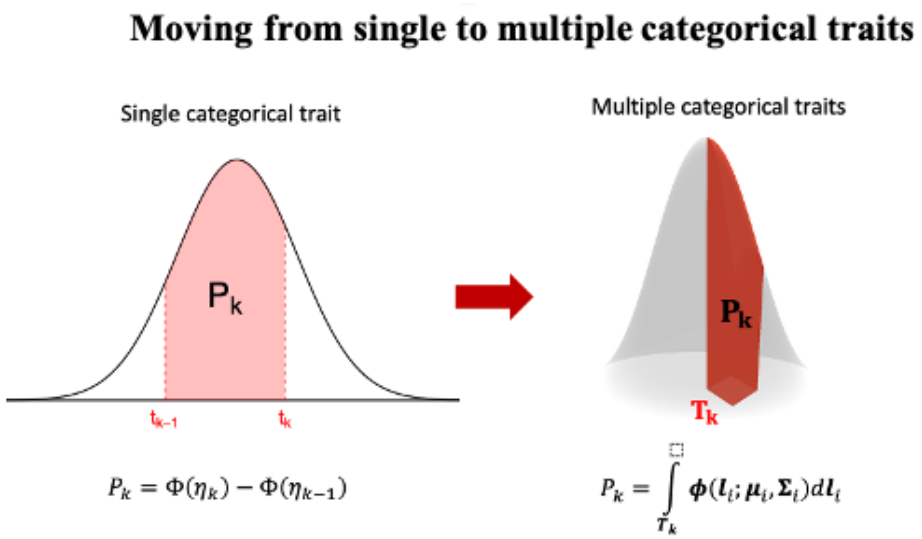
  - M-step

$$\text{argmax}_{\boldsymbol{\theta}}\left[Q\left(\boldsymbol{\theta}\left|\boldsymbol{\theta}^{j}\right.\right)\right] = \text{argmax}_{\boldsymbol{\theta}}\left[L(\tilde{\boldsymbol{l}}|\mathbf{y}_2, \boldsymbol{\theta}) + L(\mathbf{y}_2|\boldsymbol{\theta}) + L(\boldsymbol{\theta})\right]$$

At each iteration, calculate $\tilde{\boldsymbol{l}}_i$ for each animal with records
Solve a regular MM using $(\tilde{\boldsymbol{l}}, \mathbf{y}_2)$ as phenotypes

- Difference from linear models: Phenotypes and residual covariances are modified every iteration of NR or EM

Multi-trait threshold-linear models



- Consider all categorical traits jointly, so the joint probability:

$$P_k = P(\boldsymbol{t}_{k-1} < \boldsymbol{l}_i \leq \boldsymbol{t}_k) = \int_{\boldsymbol{T_k}} \boldsymbol{\phi}(\boldsymbol{l}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) d\boldsymbol{l}_i$$

MVN evaluated at $\boldsymbol{l}_i$

Upper thresholds

Lower thresholds

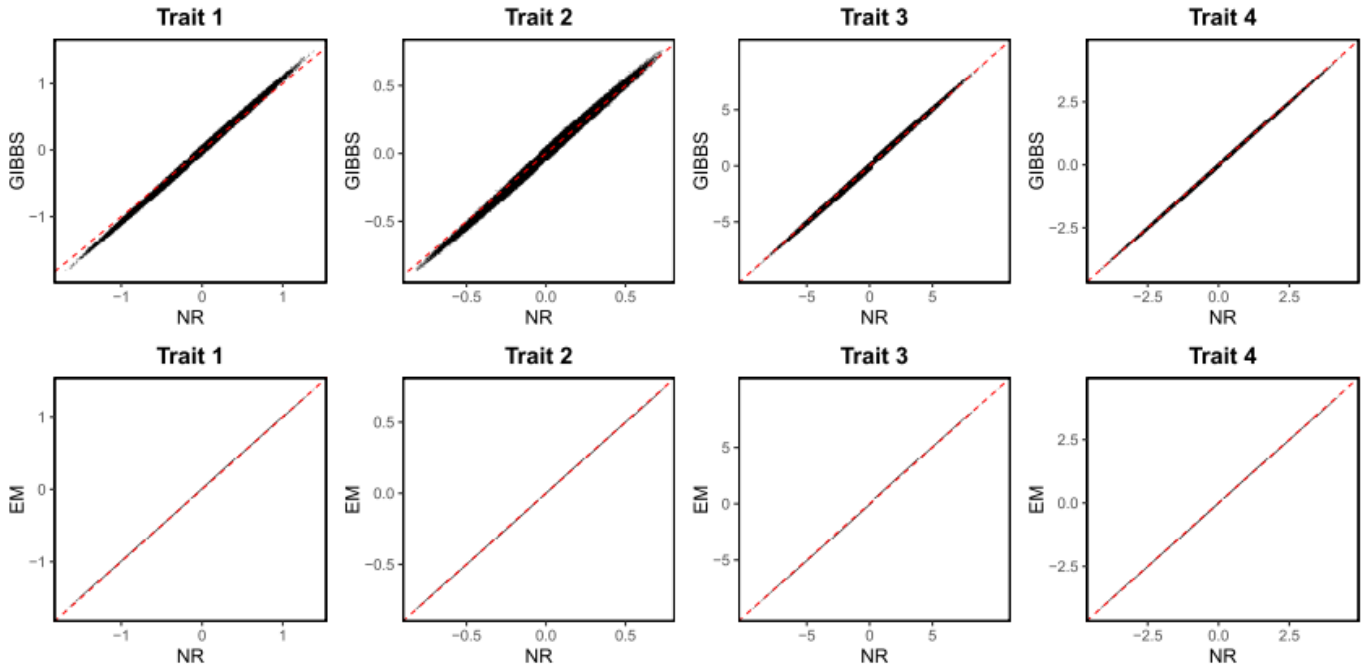
$\boldsymbol{l}_i$: vector of unobserved liabilities for the categorical traits for the i^{th} animal

- Example:
  - Two categorical traits (with 2 and 3 categories)
  - Each animal has 6 possible joint observations (6 possible probabilities)
  - $P_{2,3}$ is the probability of recording the second and third categories for the first and second traits, in the same animal

Multi-trait threshold-linear models - Simulation

- Solution: Maximum a posteriori (MAP)
  - Newton-Raphson
  - Expectation-Maximization

2 categories 3 categories Continuous



GENETICS, 2026, 233(1), iyag086
https://doi.org/10.1093/genetics/iyag086
Advance Access Publication Date: 26 March 2026
Investigation

OXFORD GENETICS

Methods for joint genetic prediction of multiple ordinal categorical and continuous traits

Matias Bermann [1,*] Andres Legarra [1,2] Ignacy Misztal [1] Daniela Lourenco [1]

- 1.8 M animals
- 30 min in the paper
- 4 min now

- CE for CDCB (Andres + Simone)
- 23M records
- 17.7M animals in ped
- 965k genotyped (39k core)
- 29 hours

Reliability of GEBV in threshold models

- Reliability of GEBV from multi-trait threshold models

Alvarez-Munera et al.

New method

$$(W'\tilde{R}^{-1}W + S)\theta^{j+1} = W'\tilde{R}^{-1}\tilde{y}$$ → Mixed model equations

$$(W'\tilde{R}^{-1}W + S)^{-1}$$ → PEC and PEV for random effects

New method

Effective contributions

1. Value of observations for an animal

$$D_i = Z_i'(R_i^{-1} - R_i^{-1}(S_i^{-1})R_i^{-1})Z_i$$

2. Value of observations on descendants

$$E_i = \frac{1}{3}G_0^{-1} - \frac{4}{9}G_0^{-1}\left(D_i + \sum_{l=1}^{p_i} E_l + \frac{4}{3}G_0^{-1}\right)^{-1} G_0^{-1}$$

3. Value of observations on ancestors

$$E_j^* = \frac{1}{3}G_0^{-1} - \frac{4}{9}G_0^{-1}\left(-E_i + F_j + \frac{4}{3}G_0^{-1}\right)^{-1} G_0^{-1}$$

1. Final information block

$$F_i = D_i + \sum_{l=1}^{p_i} E_l + \sum_{j=1}^{t_i} E_j^*$$

2. Prediction error variance (PEV)

$$T_i = (F_i + G_0^{-1})^{-1}$$

3. Approximated reliability

$$\rho_{iu}^2 = 1 - \frac{PEV_{iu}}{Var_u}$$

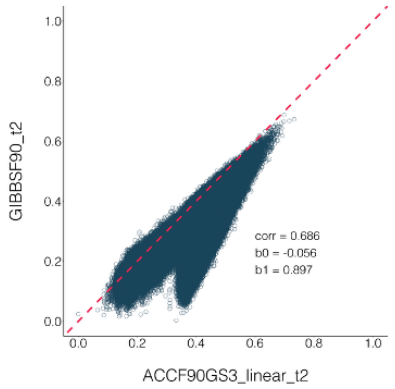
Lin. $Var(e) = R$ Thr. $\tilde{R}$

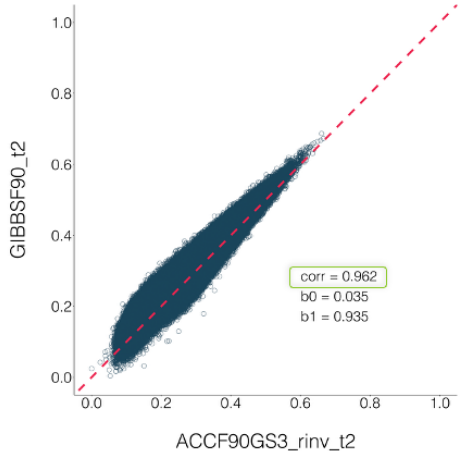
Reliability of GEBV in threshold models

- Reliability of GEBV from multi-trait threshold-linear models

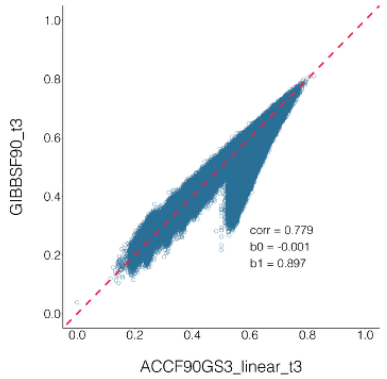
Alvarez-Munera et al.

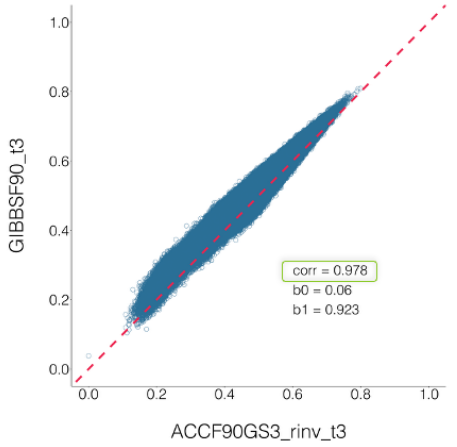
Trait 2: Categorical





Trait 3: Continuous





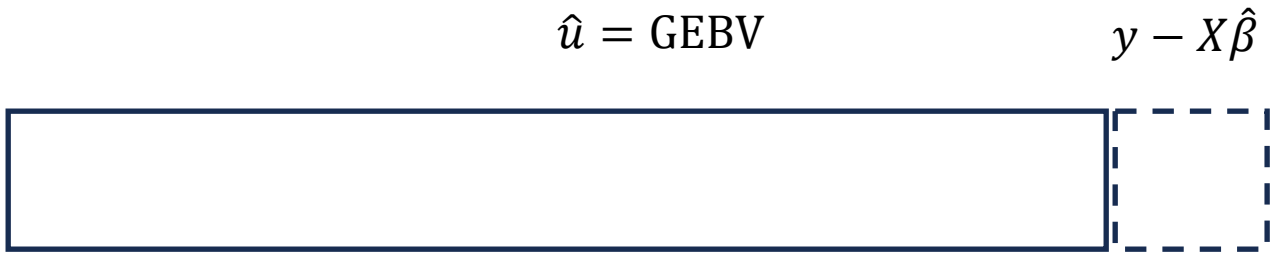
VCE for large populations I - GPP

GPP = Genetic Parameters via Predictivity

$$\widehat{h^2}: \sqrt{\frac{Nh^2}{Nh^2 + M_e}} = corr\,(y - Xb, \hat{u})/h$$

$$\widehat{h^2} = \frac{c^2 + \sqrt{c^4 + 4c^2 M_e / N_{ref}}}{2}$$

Predictivity: $c = corr\left(y - X\hat{\beta}, \hat{u}\right)$



N_{ref} – animals in reference population
Me – Independent chromosome segments, ~15k in dairy cattle

Misztal et al. (2023)
Misztal and Gowane (2025)
Gowane et al. (accepted)

VCE for large populations - GPP

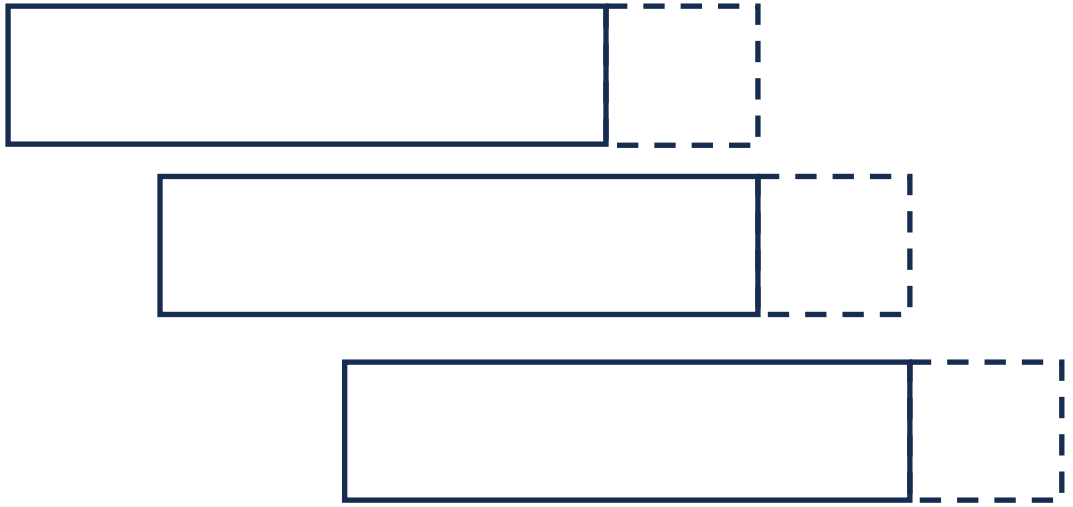
$\hat{u}_i = \text{GEBV}$ $y_i - X\hat{\beta}_i$

$\hat{u}_j = \text{GEBV}$ $y_j - X\hat{\beta}_j$

$$corr_{ij} = \frac{corr\left(y_i - X\hat{\beta}_i, \hat{u}_j\right)}{h_i \; acc_j}$$

Misztal et al. (2023)
Misztal and Gowane (2025)
Gowane et al. (accepted)

- Estimation over time



Method II: MCssGREML

RESEARCH ARTICLE Open Access

Estimation of (co)variance components for very large datasets and complex single-step genomic models

Matias Bermann[1*], Andres Legarra[1,2], Ignacio Aguilar[3], Alejandra Alvarez-Munera[1], Ignacy Misztal[1] and Daniela Lourenco[1]

- Monte-Carlo REML

Efficient Monte Carlo algorithm for restricted maximum likelihood estimation of genetic parameters

Kaarina Matilainen | Esa A. Mäntysaari | Ismo Strandén

- Simulation for single-step GBLUP

How pedigree errors affect genetic evaluations and validation statistics

E. C. G. Pimentel,* C. Edel, R. Emmerling, and K.-U. Götz
Institute of Animal Breeding, Bavarian State Research Center for Agriculture, Grub, 85586 Germany

- Approximation of log-determinants for the likelihood

RESEARCH ARTICLE Open Access

Monte Carlo approximation of the logarithm of the determinant of large matrices with applications for linear mixed models in quantitative genetics

Matias Bermann[1*], Alejandra Alvarez-Munera[1], Andres Legarra[2], Ignacio Aguilar[3], Ignacy Misztal[1] and Daniela Lourenco[1]

Algorithms + computing efficiency

- Update algorithms to avoid bottlenecks
  - APY
  - Iteration on memory
  - Parallel computing
  - Memory mapping

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

- 53M animals in pedigree
- 50M records – milk, fat, and protein
- 1.9M genotyped animals
- APY ssGBLUP (41k core)
- > 350M equations

Program	Start Time	Finish Time	Duration	Iterations	Max Memory Used
renumf90	02-03-2026 09:46:25	02-03-2026 11:43:32	1h 57m 7s	n.a.	91.27G
preGSf90	02-03-2026 12:01:40	02-03-2026 15:44:59	3h 43m 19s	n.a.	679.68G
blup90iod3	02-03-2026 15:45:13	02-04-2026 01:58:30	10h 13m 17s	511	93.42G
accf90GS3	02-04-2026 02:00:08	02-04-2026 02:30:36	0h 30m 28s	n.a.	255.55G
postGSf90	02-04-2026 02:30:38	02-04-2026 02:55:11	0h 24m 33s	n.a.	368.29G

Cesarani et al. (2022): 72 hours for blup90iod2

16h 48m

679.68 GB

UNIVERSITY OF GEORGIA

Take-home messages

- Adapted to virtually any model and data (including very large populations)
  - Reasonable computing cost

- Multi-trait threshold (threshold-linear) models' issue is solved
  - Feasible for routine evaluations
  - GEBV and reliability

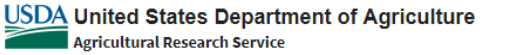
- Variance components for large genotyped populations
  - MCssGREML and GPP for linear models

- More developments are coming soon…

It takes a team of dedicated people

UNIVERSITY OF GEORGIA

and collaborators



Solving a 30-year-old issue

- Solution: Maximum a posteriori (MAP):
  - Bayes' Theorem
  - Find the most likely value for an unknown parameter given the data we just observed

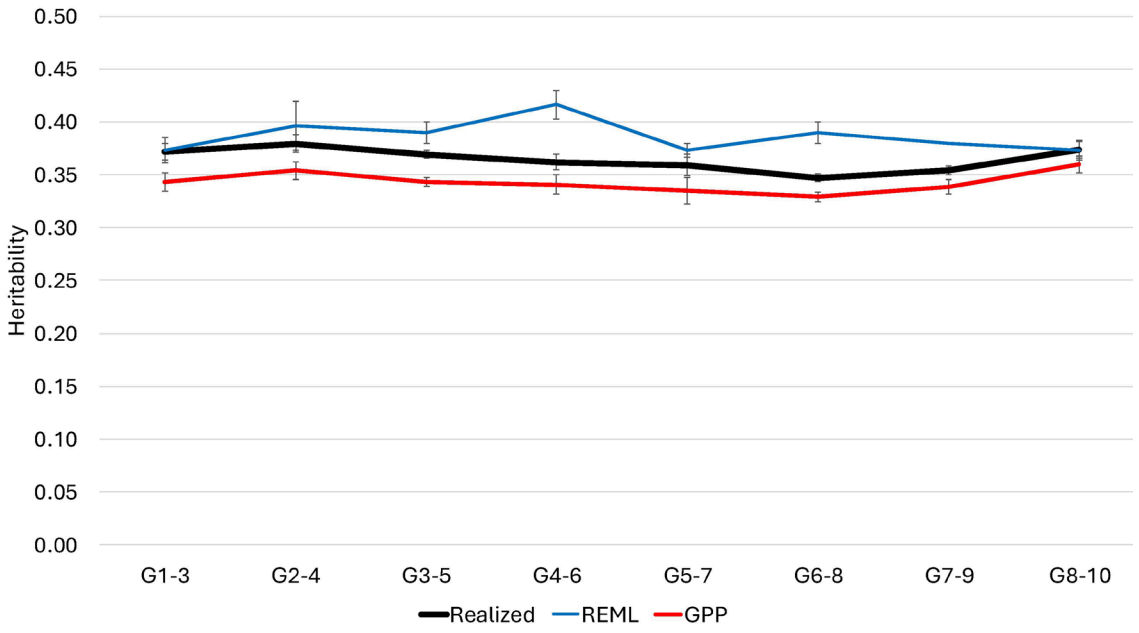
Feature	Maximum A Posteriori (MAP)	Gibbs Sampling
The Goal	**Optimization:** Find the single point value $\hat{\theta}$ that maximizes the posterior probability.	**Estimation/Simulation:** Generate a representative collection of data points from the posterior distribution.
The Output	A **single value** (a vector of parameters).	A **large collection of samples** (thousands of parameter sets).
Analogy	Finding the highest peak on a mountain range.	Dropping a drone that wanders around the mountains, spending more time in the high valleys but mapping the whole range.

VCE for large populations - GPP

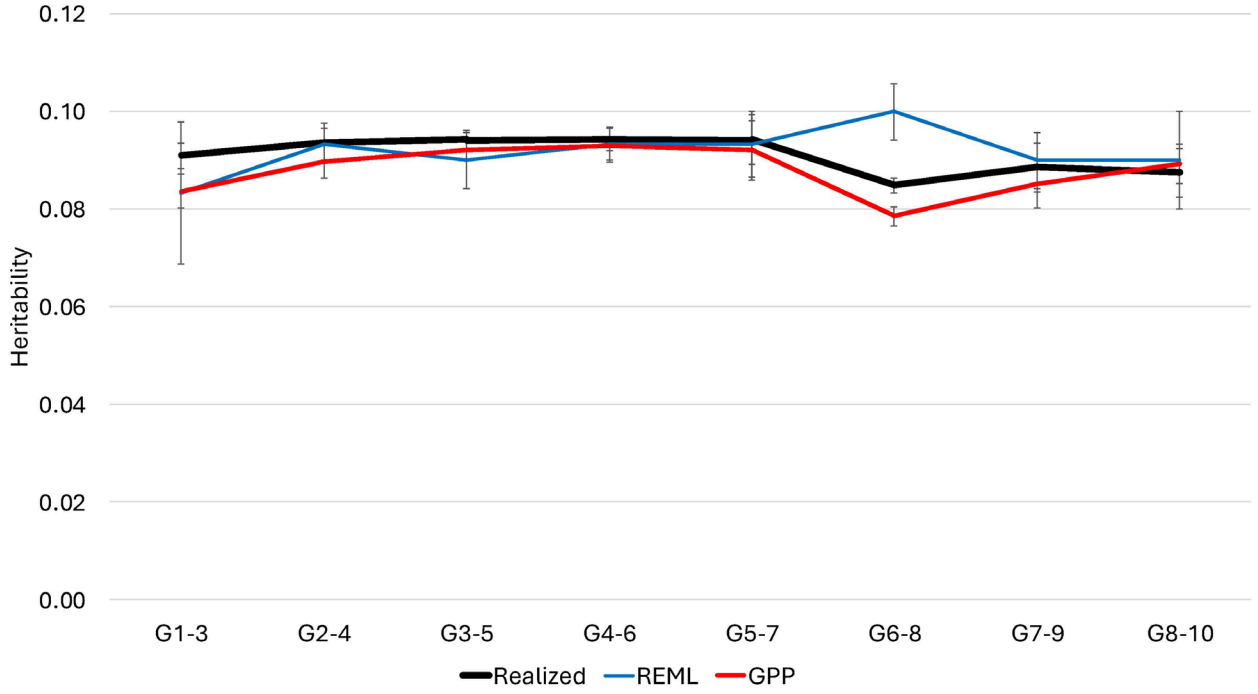
- GPP: Genetic parameters via predictivity

Genetic parameters via predictivity in large populations under strong genomic selection

Gopal Gowane[1,2*], Jorge Hidalgo[1], Mary Kate Hollifield[1], Ignacy Misztal[1], Daniela Lourenco[1]



Production Trait
$h^2 = 0.4$
Large data



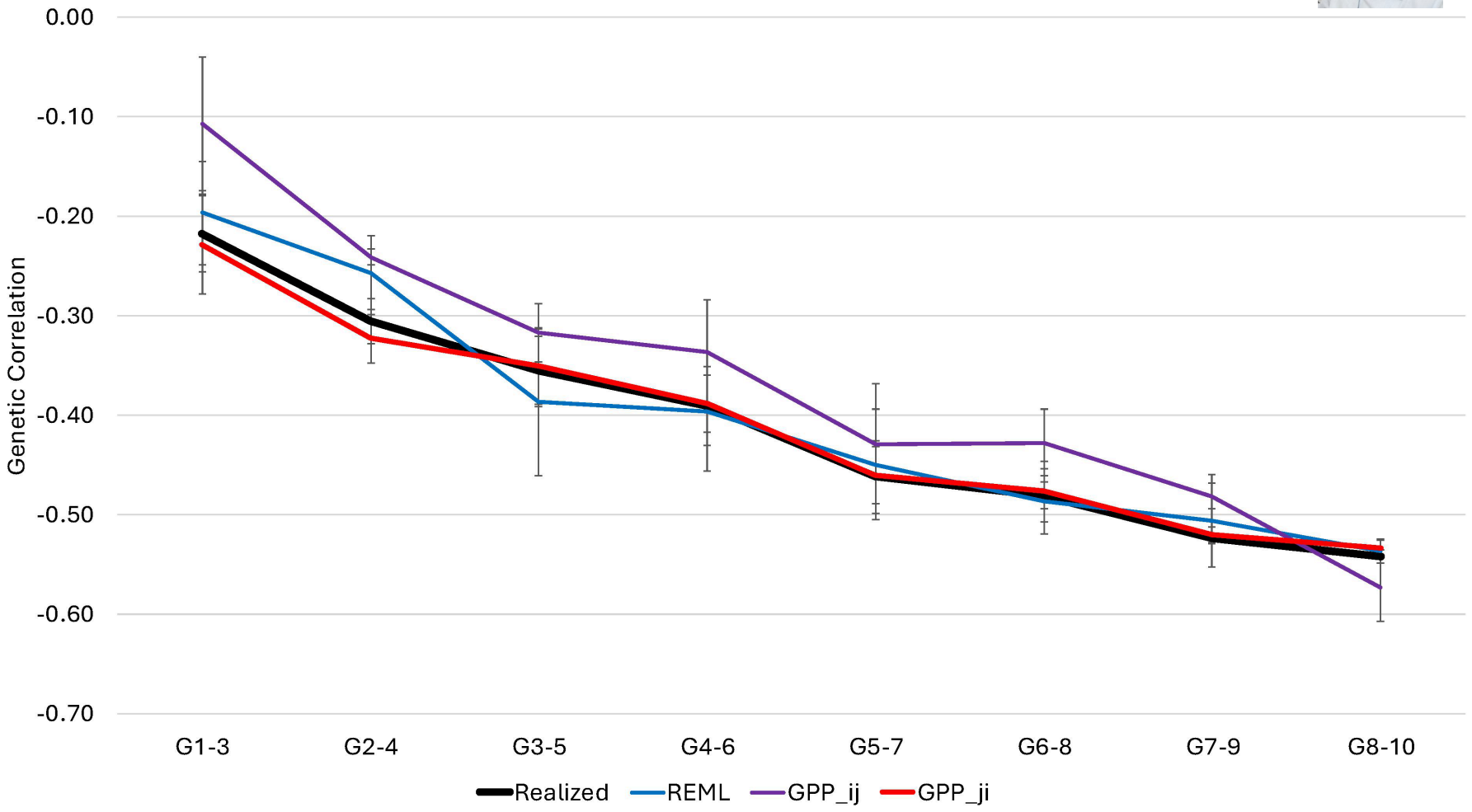
Fitness Trait
$h^2 = 0.1$
Large data

VCE for large populations - GPP

- GPP: Genetic parameters via predictivity

Genetic parameters via predictivity in large populations under strong genomic selection

Gopal Gowane[1,2*], Jorge Hidalgo[1], Mary Kate Hollifield[1], Ignacy Misztal[1], Daniela Lourenco[1]



Correlation (prod, fit) Large data

GPP seems to work well with large data!

MC-ss-GREML Algorithm

1. Solve the ssGBLUP MME using iteration on data (Schaeffer and Kennedy, 1986; Misztal and Gianola, 1986)
2. Calculate quadratic forms
3. For s Monte Carlo samples, simulate random effects, residuals, and phenotypes, and solve the MME using such data
4. Approximate traces
5. For AI-MC-REML, calculate the Average Information matrix
6. Obtain new estimates for variance components
7. Approximate the log-likelihood and calculate $\mathrm{CV}(\mathrm{logL})$
8. Finish the iteration if $\Delta\boldsymbol{\theta} < t_1$ or $\mathrm{CV}(\mathrm{logL}) < t_2$

MC-ss-GREML tests

RESEARCH ARTICLE Open Access

Estimation of (co)variance components for very large datasets and complex single-step genomic models

Matias Bermann[1*], Andres Legarra[1,2], Ignacio Aguilar[3], Alejandra Alvarez-Munera[1], Ignacy Misztal[1] and Daniela Lourenco[1]

- 100k animals in the pedigree
- 10k genotyped
- Direct and maternal effects
- 33k phenotypes
  - BW, WW, PWG
- 14 parameters

	ssGREML	MC-ssGREML
Rounds	21	19
Memory (GB)	105.5	1.1
Running time	6.5 days	22 hours

- 7M animals in the pedigree
- 331k genotyped
- Direct and maternal effects
- 5.8M phenotypes
  - BW
- 4 parameters

	ssGREML	MC-ssGREML
Rounds	NA	11
Memory (GB)	NA	53.5
Running time	NA	5 days