Leveraging correlated information under multivariate settings

Alencar Xavier
Breeding Analyst at Corteva
Adjunct professor at Purdue

David Habier
Sr. Research Scientist at Corteva

Outline

1. **Introduction**
   - Rationale and statistical model
2. **Coefficients**
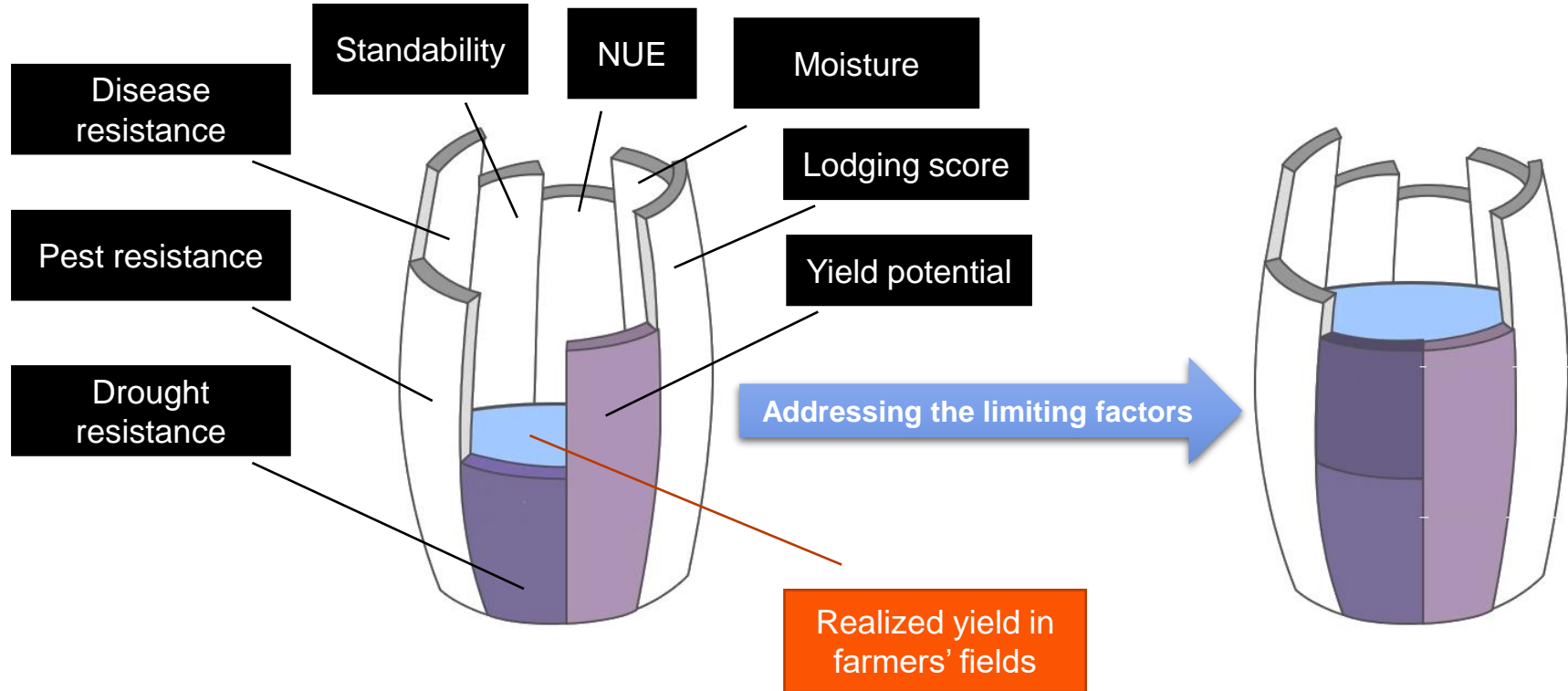   - Univariate
   - Multivariate
3. **Variances**
   - Univariate
   - Multivariate
4. **Simulations**
   - Elapsed time
   - Accuracy
   - Bias & Precision
   - Limitations and other considerations
5. **Conclusion**

Law of the minimum

Standability

NUE

Moisture

Disease resistance

Lodging score

Pest resistance

Yield potential

Drought resistance

Addressing the limiting factors

Realized yield in farmers' fields

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

CORTEVA
agriscience

Rationale

- Single-trait models for genomic prediction in plant breeding are already well-stablished (e.g. GBLUP and BayesB)

- Phenotypes come from multiple locations, years, and quantitative traits; and most traits have <mark>genetically correlated breeding values</mark>

Alencar.Xavier@Corteva.com
**Quantitative Geneticist, Breeding Analyst LAAF**

Rationale

- **Complex GxE patterns / multi-trait** = higher accuracy

- **Assess new phenomic traits** (*e.g.* canopy coverage in soy)

- **Computationally PROHIBITIVE***

* Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, *11*(4), 407-409.
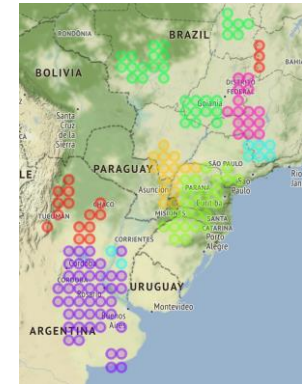
Multivariate models enable analysis of

Multiple traits

Additive genetic association among multiple soybean traits
(https://rd.springer.com/article/10.1007/s10681-017-1975-4)
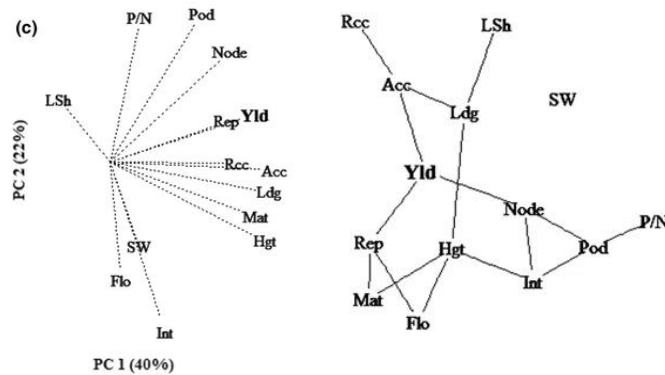
Multiple environments

Example of environmental clustering

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF
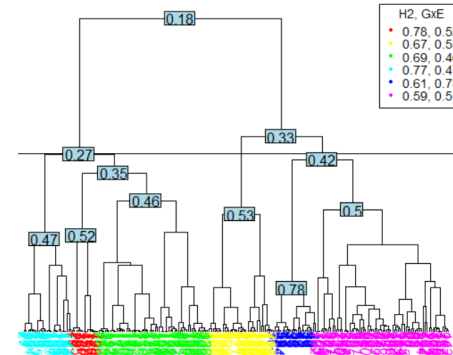
Why would multivariate be any better?

Simple (bivariate) model:

INFORMATION GAIN

$$\mathbf{y} = \mathbf{\mu} + \mathbf{g} + \mathbf{e}$$

$$Var \begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \end{bmatrix} = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma_{a_2}^2 \end{bmatrix} + \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma_{e_2}^2 \end{bmatrix}$$

Why would multivariate be any better?

$$y = Zg + e, \qquad y \sim N(0, V)$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

- Covariance structure

$$V = G \otimes \Sigma_a + I \otimes \Sigma_e = G \otimes \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma_{a_2}^2 \end{bmatrix} + I \otimes \begin{bmatrix} \sigma_{e_1}^2 & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma_{e_2}^2 \end{bmatrix}$$

- Mixed model equation

$$\begin{bmatrix} Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11} & Z_1'\Sigma_e^{12}Z_2 + G^{-1}\Sigma_a^{12} \\ Z_2'\Sigma_e^{12}Z_1 + G^{-1}\Sigma_a^{12} & Z_2'\Sigma_e^{11}Z_2 + G^{-1}\Sigma_a^{22} \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} Z_1'(\Sigma_e^{11}y_1 + \Sigma_e^{12}y_2) \\ Z_2'(\Sigma_e^{22}y_2 + \Sigma_e^{12}y_1) \end{bmatrix}$$

- **<u>Univariate vs bivariate</u>**

INFORMATION GAIN

$$g_1 = (Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11})^{-1}(Z_1'\Sigma_e^{11}y_1)$$

$$g_1|g_2 = (Z_1'\Sigma_e^{11}Z_1 + G^{-1}\Sigma_a^{11})^{-1}(Z_1'(\Sigma_e^{11}y_1 + \Sigma_e^{12}y_2) - (Z_1'\Sigma_e^{12}Z_2 + G^{-1}\Sigma_a^{12})g_2)$$

Why marker ridge regression?

1. Regression-type models are easy to store and use for prediction

2. Compatible with the multi-stage[1,2] framework

3. Well-known properties: Gaussian, additive, and equivalent to GBLUP

4. No need to build and invert G matrix (which is not always positive definite)

5. Provides covariance components for meaningful statistics:

   • Heritability, reliability, accuracy, genetic correlations, selection indexes, correlated response

1. Smith, A., Cullis, B., and Gilmour, A. (2001). Applications: the analysis of crop variety evaluation data in Australia. Australian & New Zealand Journal of Statistics, 43(2), 129-145.
2. Mohring, J, and H-P Piepho, (2009) Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci. 49: 1977–1988.

CORTEVA agriscience

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Statistical model

$$y = \mu + \mathbf{Z}\beta + e \qquad\qquad (1)$$

- Where $y = \{y_1, y_2, \ldots, y_K\}$, $\mu = \{\mu_1, \mu_2, \ldots, \mu_K\}$, $\beta = \{\beta_1, \beta_2, \ldots, \beta_K\}$,

    $e = \{e_1, e_2, \ldots, e_K\}$, $Z = \text{BlockDiag}\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_K\}$

- Variances:

$$\Sigma_\beta = \begin{bmatrix} \sigma^2_{\beta(1)} & \cdots & \sigma_{\beta(1,K)} \\ \vdots & \ddots & \vdots \\ \sigma_{\beta(K,1)} & \cdots & \sigma^2_{\beta(K)} \end{bmatrix} \text{ and } \Sigma_e = \begin{bmatrix} \sigma^2_{e(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2_{e(K)} \end{bmatrix}$$

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Corresponding mixed model equation

Under the traditional framework, the mixed-model equations required to solve the multivariate ridge regression (eq. 1) can be written as follows:

$$\begin{bmatrix} 1_1'1_1\sigma_{e_1}^{-2} & \cdots & 0 & 1_1'Z_1\sigma_{e_1}^{-2} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1_K'1_K\sigma_{e_K}^{-2} & 0 & \cdots & 1_K'Z_K\sigma_{e_K}^{-2} \\ Z_1'1_1'\sigma_{e_1}^{-2} & \cdots & 0 & Z_1'Z_1\sigma_{e_1}^{-2}+I_m\sigma_\beta^{11} & \cdots & I_m\sigma_\beta^{1K} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \cdots \\ 0 & \cdots & Z_K'1_K'\sigma_{e_K}^{-2} & I_m\sigma_\beta^{K1} & \vdots & Z_K'Z_K\sigma_{e_K}^{-2}+I_m\sigma_\beta^{KK} \end{bmatrix} \begin{bmatrix} \hat\mu_1 \\ \vdots \\ \hat\mu_k \\ \hat\beta_1 \\ \vdots \\ \hat\beta_K \end{bmatrix} = \begin{bmatrix} \sigma_{e1}^{-2}1_1'y_1 \\ \vdots \\ \sigma_{e_K}^{-2}1_k'y_K \\ \sigma_{e1}^{-2}Z_1'y_1 \\ \vdots \\ \sigma_{e_K}^{-2}Z_k'y_K, \end{bmatrix} \quad (2)$$

where σ_β^{ij} is the element at position ij of Σ_β^{-1}. This setup involves storing K times the cross-product or marker scores ($Z_k'Z_k$), each with dimension $m \times m$.

Moreover, this **huge** matrix must be **inverted** for the estimation of covariance components: $\hat\Sigma_{\beta(i,j)} = m^{-1}[\hat{\boldsymbol{\beta}}_i'\hat{\boldsymbol{\beta}}_j + \mathrm{tr}(C^{ij})]$

Computing very large multivariate models is **impossible**

unless…
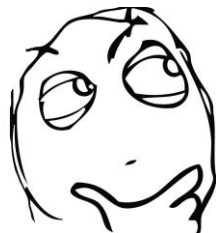
Coefficients for univariate model

1. Whole-genome regression (*e.g.* BayesA) rely on the *Gauss-Seidel* method [1]
2. GS has only two steps, whereas coordinate descent has three [2]
3. ==It avoids building the systems of equations altogether!!==
4. Estimates one marker effects, then uses residuals to update the next effect
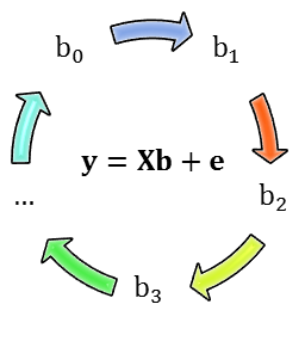
$$\hat{b}_j^{t+1} = \frac{x_j'\hat{e}^t + x_j'x_j\hat{b}_j^t}{x_j'x_j + \lambda}$$

$$\hat{e}^{t+1} = \hat{e}^t - x_j\left(\hat{b}_j^{t+1} - \hat{b}_j^t\right)$$

for j in 1:p

$y = Xb + e$

b_0 b_1 b_2 b_3 ...
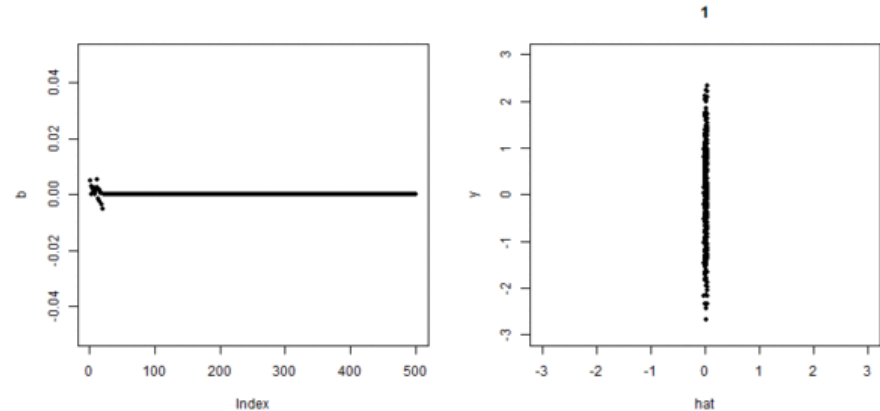
1 Legarra, A., & Misztal, I. (2008). Computing strategies in genome-wide selection. *Journal of dairy science*, *91*(1), 360-366.
2 Xavier, A. (2021). Technical nuances of machine learning. *Crop Breeding and Applied Biotechnology*, 21.

Coefficients for __multivariate__ model

For updating estimated marker effects we define, $\hat{\boldsymbol{\beta}}_j^{'(t)} = [\hat{\beta}_{j1}^{(t)} \ \hat{\beta}_{j1}^{(t)} \ \ldots \ \hat{\beta}_{jK}^{(t)}]$ to be the vector of estimated marker effects for marker j and all K environments, $\mathbf{Z}_j = \oplus_{k=1}^{K} z_{jk}$ to be a matrix containing marker scores at marker j, and $\hat{\boldsymbol{\Sigma}}_e^{(t)} = Diag\{\hat{\sigma}_{e1}^{2(t)}, \hat{\sigma}_{e2}^{2(t)}, \ \ldots \ , \hat{\sigma}_{ek}^{2(t)}\}$ to be a diagonal matrix of estimated residual variances. Effects for marker j are initialized with zero and updated as

$$\hat{\boldsymbol{\beta}}_j^{(t+1)} = (\hat{\boldsymbol{\Sigma}}_e^{-1(t)} \mathbf{Z}_j' \mathbf{Z}_j + \hat{\boldsymbol{\Sigma}}_\beta^{-1(t)})^{-1} \mathbf{Z}_j' \hat{\boldsymbol{\Sigma}}_e^{-1(t)} (\mathbf{Z}_j \hat{\boldsymbol{\beta}}_j^{(t)} + \hat{\boldsymbol{e}}^{(t)}), \tag{5}$$

and before moving to the next marker, the residual vector is updated as

$$\hat{\boldsymbol{e}}^{(t+1)} = \hat{\boldsymbol{e}}^{(t)} - \mathbf{Z}_j'(\hat{\boldsymbol{\beta}}_j^{(t+1)} - \hat{\boldsymbol{\beta}}_j^{(t)}). \tag{6}$$

Note that the computation of Kronecker products are not necessary for the multivariate Gauss-Seidel formulation (eq. 5) as long as the residual covariance $\hat{\boldsymbol{\Sigma}}_e$ is a diagonal matrix.

NO KRONECKER PRODUCTS!!!!

$$\text{For}(\,j\;\text{in}\;1\!:\!p\,)\;\{$$

1st solve for beta

$$\begin{bmatrix} \hat{\Sigma}_\beta^{11} + \mathbf{z}'_{j(1)}\mathbf{z}_{j(1)}\sigma_{e(1)}^{-2} & \hat{\Sigma}_\beta^{12} \\ \hat{\Sigma}_\beta^{21} & \hat{\Sigma}_\beta^{22} + \mathbf{z}'_{j(2)}\mathbf{z}_{j(2)}\sigma_{e(2)}^{-2} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{j(1)}^{t+1} \\ \hat{\beta}_{j(2)}^{t+1} \end{bmatrix} = \begin{bmatrix} \sigma_{e(1)}^{-2}\left(\mathbf{z}'_{j(1)}\mathbf{z}_{j(1)}\hat{\beta}_{j(1)}^{t} + \mathbf{z}'_{j(1)}\hat{e}_1^{t}\right) \\ \sigma_{e(2)}^{-2}\left(\mathbf{z}'_{j(2)}\mathbf{z}_{j(2)}\hat{\beta}_{j(2)}^{t} + \mathbf{z}'_{j(2)}\hat{e}_2^{t}\right) \end{bmatrix}$$

2nd update residuals

$$\begin{bmatrix} \hat{e}_{j(1)}^{t+1} \\ \hat{e}_{j(2)}^{t+1} \end{bmatrix} = \begin{bmatrix} \hat{e}_1^{t} + \mathbf{z}'_{j(1)}(\hat{\beta}_{j(1)}^{t+1} - \hat{\beta}_{j(1)}^{t}) \\ \hat{e}_2^{t} + \mathbf{z}'_{j(2)}(\hat{\beta}_{j(2)}^{t+1} - \hat{\beta}_{j(2)}^{t}) \end{bmatrix}$$

$$\}$$

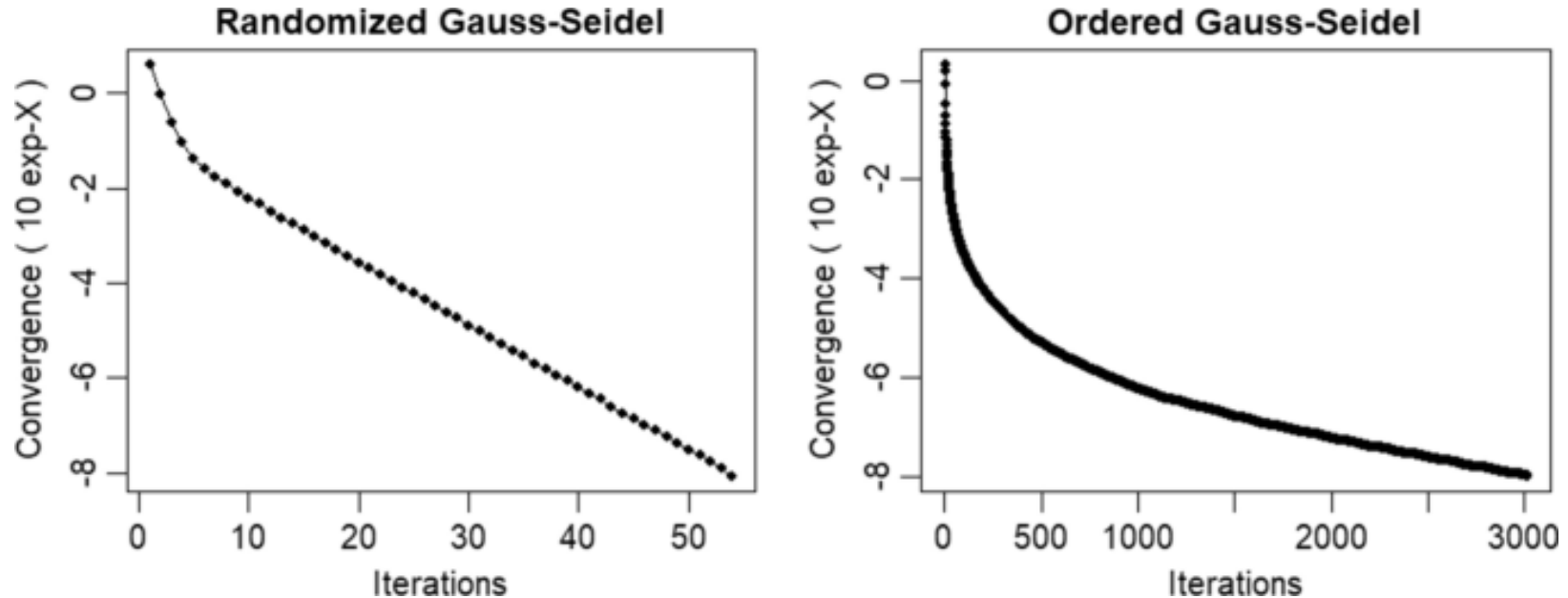**Color code**
- Computed only once, before the loop starts (ZpZ)
- Computed once every iteration
- Computed for each marker in every iteration

What is in memory?

- Z (n x m)
- B (m x k)
- E (n x k)

- ZpZ (m x k)
- $\hat{\Sigma}_\beta^{-1}$ (k x k)
- $\hat{\Sigma}_e^{-1}$ (k)

Side note: Updating markers in random order can speed up convergence

Convergence of the Gauss–Seidel solver with (left) and without (right) randomizing the order in which marker effects were updated for one replicate of the simulation of scenario 2

Alencar.Xavier @Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Variance components

Airy (1861): PRE-ANOVA → Fisher (1921): ANOVA

Henderson Methods' 1, 2 and 3 (Henderson 1953) | Restricted likelihood (Thompson 1962)

Iterative path | **Exact Likelihood path**

Iterative methods (Cunningham and Henderson 1968) | Hartley & Rao (1967): Maximum Likelihood

Henderson's Simple Method HSM (Henderson 1980)

Non-orthogonal iterative method (Thompson 1969)

MINQUE (RAO 1970)

MIVQUE (RAO 1978)

REML (Patterson and Thompson 1971)

Iterative HSM (Hudson and VanVleck 1982)

Pseudo-Expectation "PE" (Schaeffer 1986)

Tilde-Hat "TH" (VanRaden and Jung 1987)

EM-REML (Henderson 1973)

DF-REML (Graser et al 1987)

AI-REML (Johnson and Thompson 1994)

Fast variations   Gold standard

CORTEVA agriscience
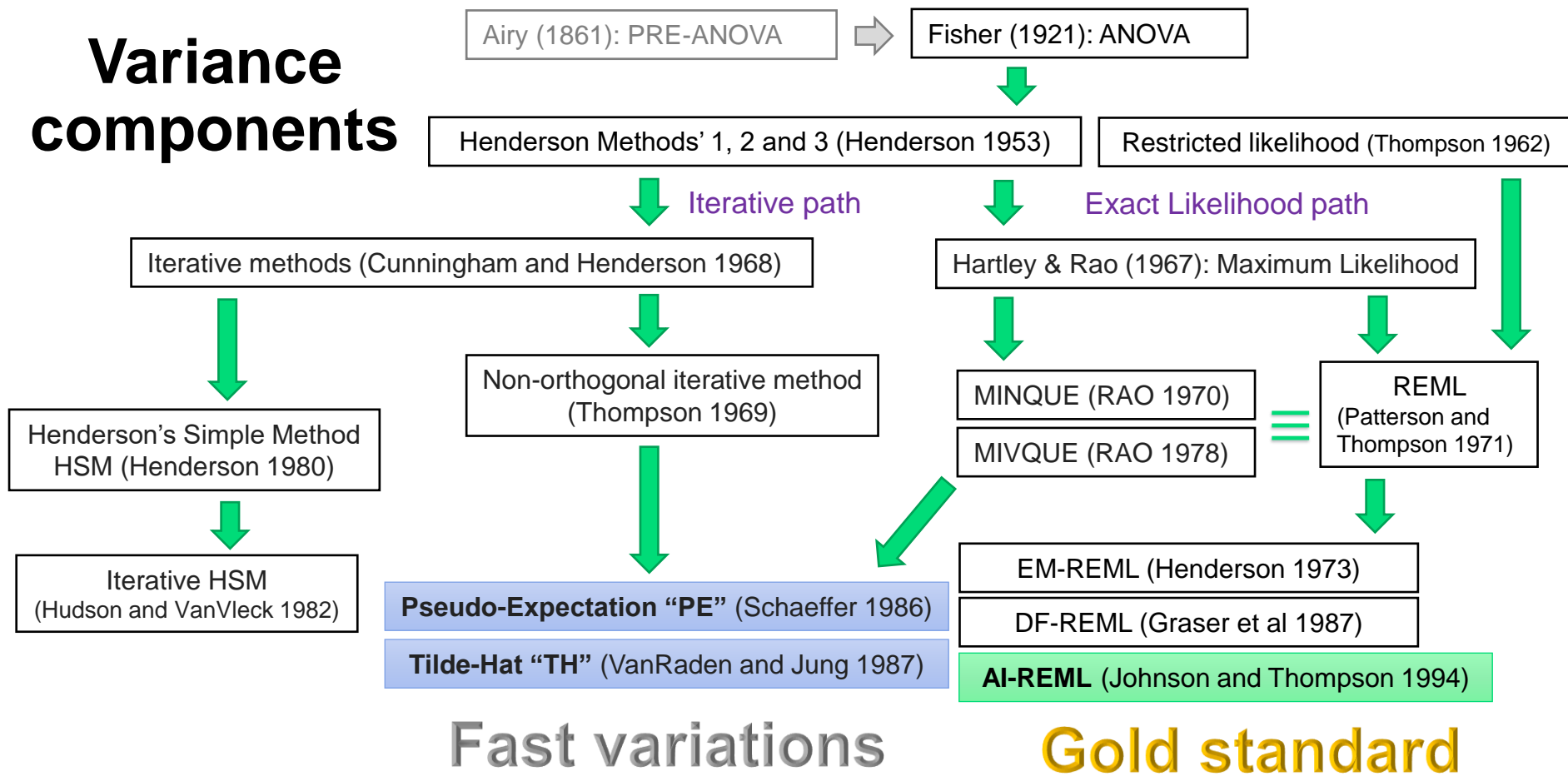
Univariate case: Variance components

- REML

$$\hat{\sigma}_\beta^2 = \frac{y'P'V_iPy}{\text{tr}(PV_i)} = \frac{y'S'V^{-1}ZZ'V^{-1}Sy}{\text{tr}(V^{-1}SZZ')} = \frac{\hat{\beta}\hat{\beta}}{\text{tr}(V^{-1}\tilde{Z}'\tilde{Z})}$$

"Let's get rid of this V^{-1}!"

- Schaffer's (Thompson's) Pseudo-Expectation

$$\hat{\sigma}_\beta^2 = \frac{y'S'\cancel{V^{-1}}ZZ'V^{-1}Sy}{\text{tr}(\cancel{V^{-1}}SZZ')} = \frac{\tilde{y}'Z\hat{\beta}}{\text{tr}(\tilde{Z}'\tilde{Z})}$$

"Let's replace this V^{-1} by something similar, but easier to compute!"

- VanRaden's Tilde-Hat

$$\hat{\sigma}_\beta^2 = \frac{y'S'D^{-1}ZZ'V^{-1}Sy}{\text{tr}(D^{-1}SZZ')} = \frac{\overbrace{\tilde{y}D^{-1}Z\hat{\beta}}^{\tilde{\beta}}}{\text{tr}(D^{-1}\tilde{Z}'\tilde{Z})} = \frac{\tilde{\beta}\hat{\beta}}{\text{tr}(D^{-1}\tilde{Z}'\tilde{Z})}$$

$$D = \text{Diag}(Z'Z\hat{\sigma}_e^{-2} + I\hat{\sigma}_\beta^{-2})$$

All methods yield the same residual variance:

$$\hat{\sigma}_e^2 = \frac{y'e}{n-1}$$

DA Harville 1977

V is a pain to compute

$$V = ZZ'\sigma_\beta^2 + I\sigma_\beta^2$$
$$S = I - (X'X)^{-1}X'; \quad P = V^{-1}S$$
$$P = V^{-1} - V^{-1}(X'V^{-1}X)^{-1}X'V^{-1}$$
$$PX = SX = 0$$
$$Sy = \text{Centralized } y = \tilde{y}$$
$$SZ = \text{Centralized } Z = \tilde{Z}$$

Multivariate case: (co)variance components

$$\widehat{\sigma}^2_{\beta(k)} = \frac{\tilde{\boldsymbol{\beta}}_k \widehat{\boldsymbol{\beta}}_k}{\text{tr}\left(\mathbf{D}_k^{-1} \tilde{\mathbf{Z}}_k{}' \tilde{\mathbf{Z}}_k\right)}$$

$$\widehat{\sigma}_{\beta(k,k')} = \frac{\tilde{\boldsymbol{\beta}}_k \widehat{\boldsymbol{\beta}}_{k'} + \tilde{\boldsymbol{\beta}}_{k'} \widehat{\boldsymbol{\beta}}_k}{\text{tr}\left(\mathbf{D}_k^{-1} \tilde{\mathbf{Z}}_k{}' \tilde{\mathbf{Z}}_k\right) + \text{tr}\left(\mathbf{D}_{k'}^{-1} \tilde{\mathbf{Z}}_{k'}{}' \tilde{\mathbf{Z}}_{k'}\right)}$$

$$\widehat{\sigma}^2_{e(k)} = \frac{y_k' \widehat{e}_k}{n_k - 1}$$

Note: Schaffer's is obtained by assuming $\mathbf{D} = \mathbf{I}$

**No V, No C, No LHS,
No determinants,
No dense inversions**

Color code
➢ **Computed only once, before the loop starts (ZpZ)**
➢ **Computed once every iteration**
➢ **Computed once for PE, and every iteration for TH**

What is in memory?	-	Y (n x k)	-	$\widehat{\Sigma}_{\boldsymbol{\beta}}$ (k x k)
-	Z (n x m)	-	Ytilde (n x k)	-	$\widehat{\Sigma}_e$ (k)
-	Bhat (m x k)	-	ZpZ (m x k)	-	N (k)
-	Btilde (m x k)	-	ZpZtilde (m x k)	
-	E (n x k)			

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

An intuitive derivation for Schaeffer's method?

The genetic covariance is simply estimated as the <u>cross-prediction between traits A and B</u> normalized by the scale of Zs

Centered phenotype of **A** — A predicted from **B** — Centered phenotype of **B** — **B** predicted from **A**

$$\widehat{\sigma}_{\beta(A,B)} = \frac{(\mathbf{y}_A - \mu_A)'(\mathbf{Z}_A\boldsymbol{\beta}_B) + (\mathbf{y}_B - \mu_B)'(\mathbf{Z}_B\boldsymbol{\beta}_A)}{\mathrm{Tr}(\tilde{\mathbf{Z}}_A{}'\tilde{\mathbf{Z}}_A) + \mathrm{Tr}(\tilde{\mathbf{Z}}_B{}'\tilde{\mathbf{Z}}_B)}$$

The key parameters from multivariate models

- Genetic variance

$$\hat{\sigma}^2_{a(k)} = \hat{\sigma}^2_{\beta(k)} \text{tr}(\mathbf{D}_k^{-1} \tilde{\mathbf{Z}}_k' \tilde{\mathbf{Z}}_k)$$

- Heritability

$$\hat{h}^2_{(k)} = \frac{\hat{\sigma}^2_{a(k)}}{\hat{\sigma}^2_{a(k)} + \hat{\sigma}^2_{e(k)}}$$

- Genetic correlations

$$\hat{\rho}_{(k,k')} = \frac{\hat{\sigma}_{\beta(k,k')}}{\sqrt{\hat{\sigma}^2_{\beta(k)} \hat{\sigma}^2_{\beta(k')}}}$$

CORTEVA
agriscience

Metrics

1. <u>Computation efficiency</u>:

   Elapsed time to fit the model

2. <u>Breeding values</u>:

   Accuracy $= \mathrm{cor}(\mathrm{GEBV}, \mathrm{TBV})$

3. <u>Heritability (h^2) and genetic correlations</u> (ρ):

   Bias $= \mathrm{E}(\hat{\theta} - \theta)$

   Precision $= \mathrm{SD}(\hat{\theta} - \theta)$

Datasets

	Balanced	Unbalanced
	Scenario 1	**Scenario 2**
Number of environments (traits)	10	10
Number of environments per line	10	1
Number of lines per environment	599	514
% of lines per environment	100%	10%
Number of phenotypic records	5990	51,420
Number of markers	1279	4311
Species	Wheat	Soy

Unbalancedness

- REML implementations (ASREML, REMLF90, AIREMLF90) were not suitable to estimate covariance components without overlapping individuals

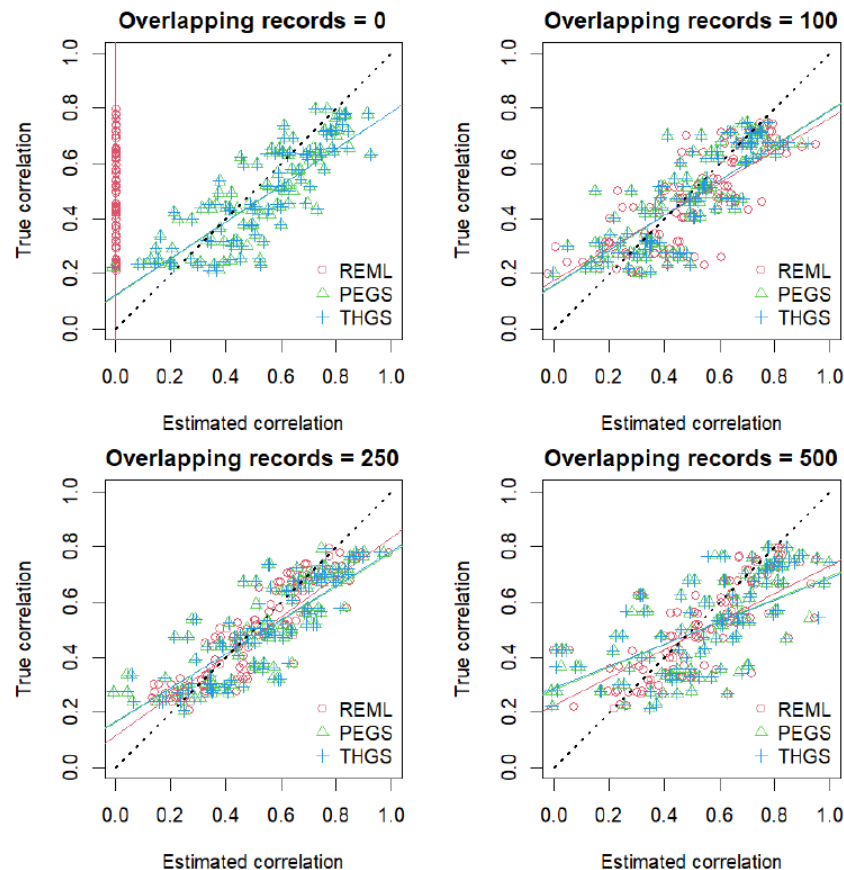- Thus, REML was not used in the unbalanced scenario

Figure 1: Scatter plot between true and estimated genetic correlations using the soybean dataset with varying number of overlapping individuals across environments.

Elapsed time in small balanced dataset

Method	Time (seconds)
ASREML 4.2	272.6
AIREMLF90	109.8
GIBB3F90	559.8
PEGS, THGS	0.27
Univariate THGS	0.23

Wheat dataset: 10 traits, 599 individuals, 1299 markers
(data available in the BGLR package)

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Elapsed time in large unbalanced dataset

Elapsed time to fit multivariate PEGS, THGS

# Markers	# Traits	Time (minutes)
4,311	10	0.2
4,311	50	3.5
4,311	200	80.5
42,034	10	0.8
42,034	50	9.9
42,034	200	123

Soybean dataset: 4628 Individuals
(data available in the SoyNAM package)

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Accuracy

$\text{Acc} = \text{cor}(\text{GEBV}, \text{TBV})$
(Higher is better)
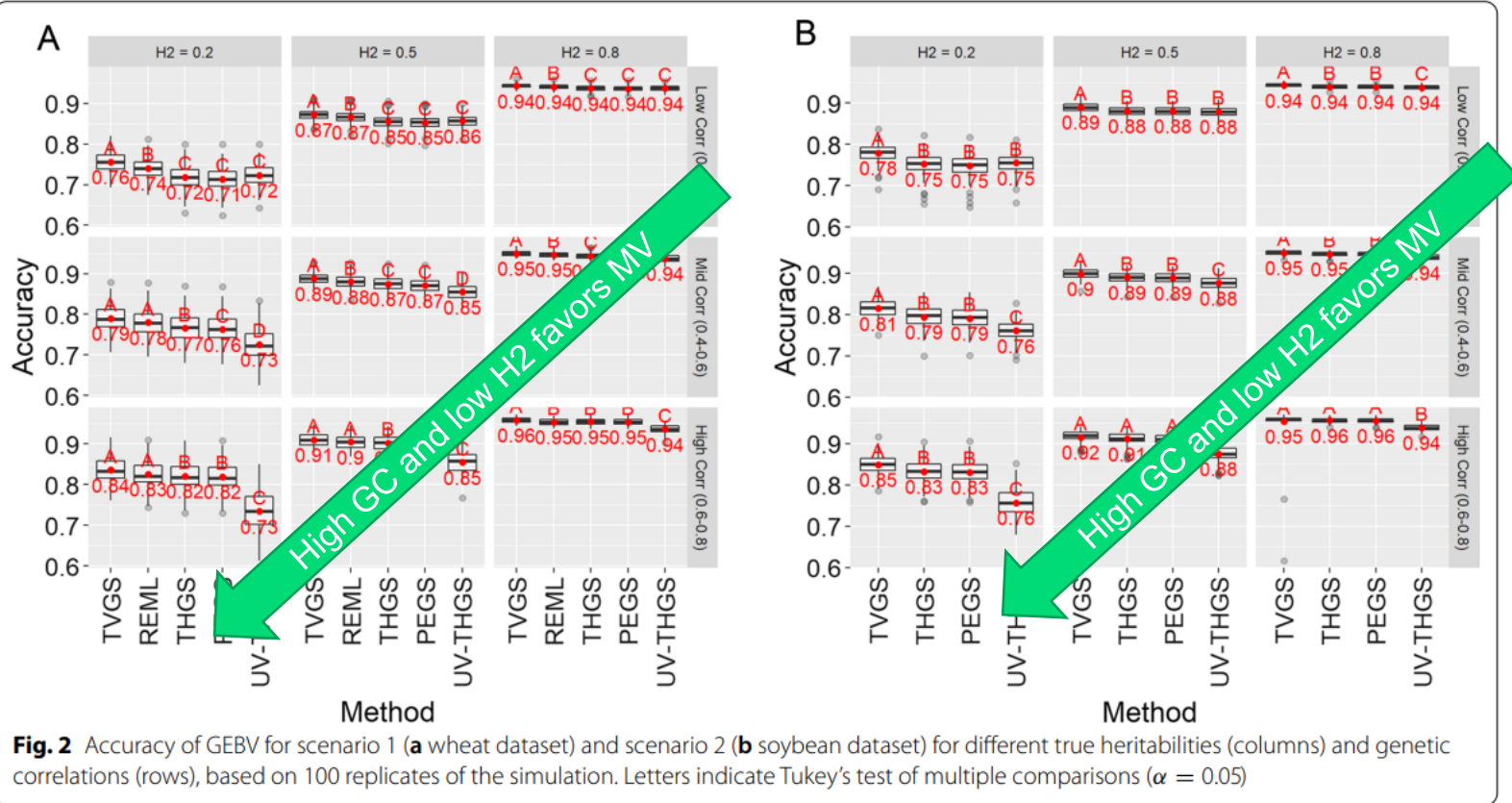
Fig. 2 Accuracy of GEBV for scenario 1 (**a** wheat dataset) and scenario 2 (**b** soybean dataset) for different true heritabilities (columns) and genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$)

Alencar.Xavier @Corteva.com
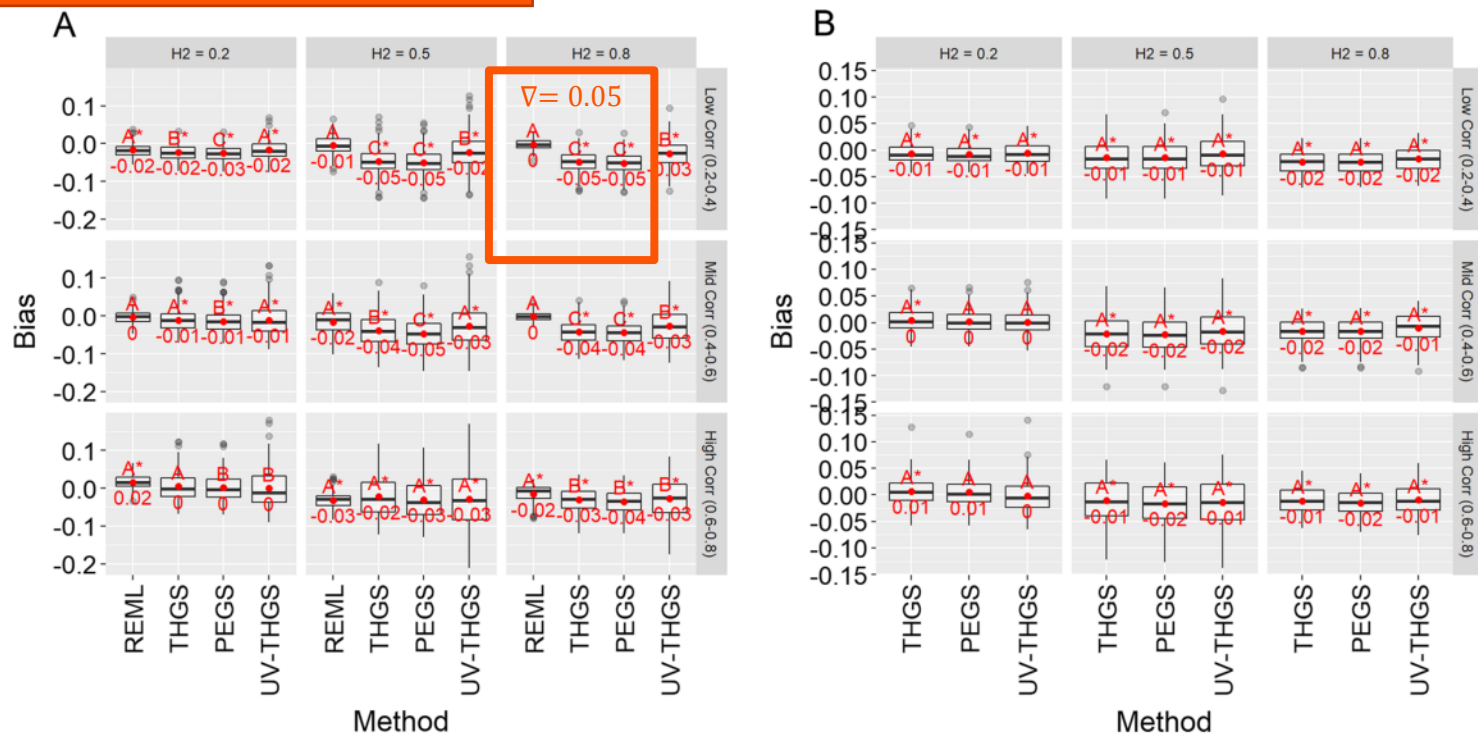Quantitative Geneticist, Breeding Analyst LAAF

Fig. 4 Bias of estimates of heritability for scenario 1 (**a** wheat dataset) and scenario 2 (**b** soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$)

Alencar.Xavier @Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Bias GC

$$\text{Bias } \rho = E(\hat{\rho} - \rho)$$

(Closer to zero is better)

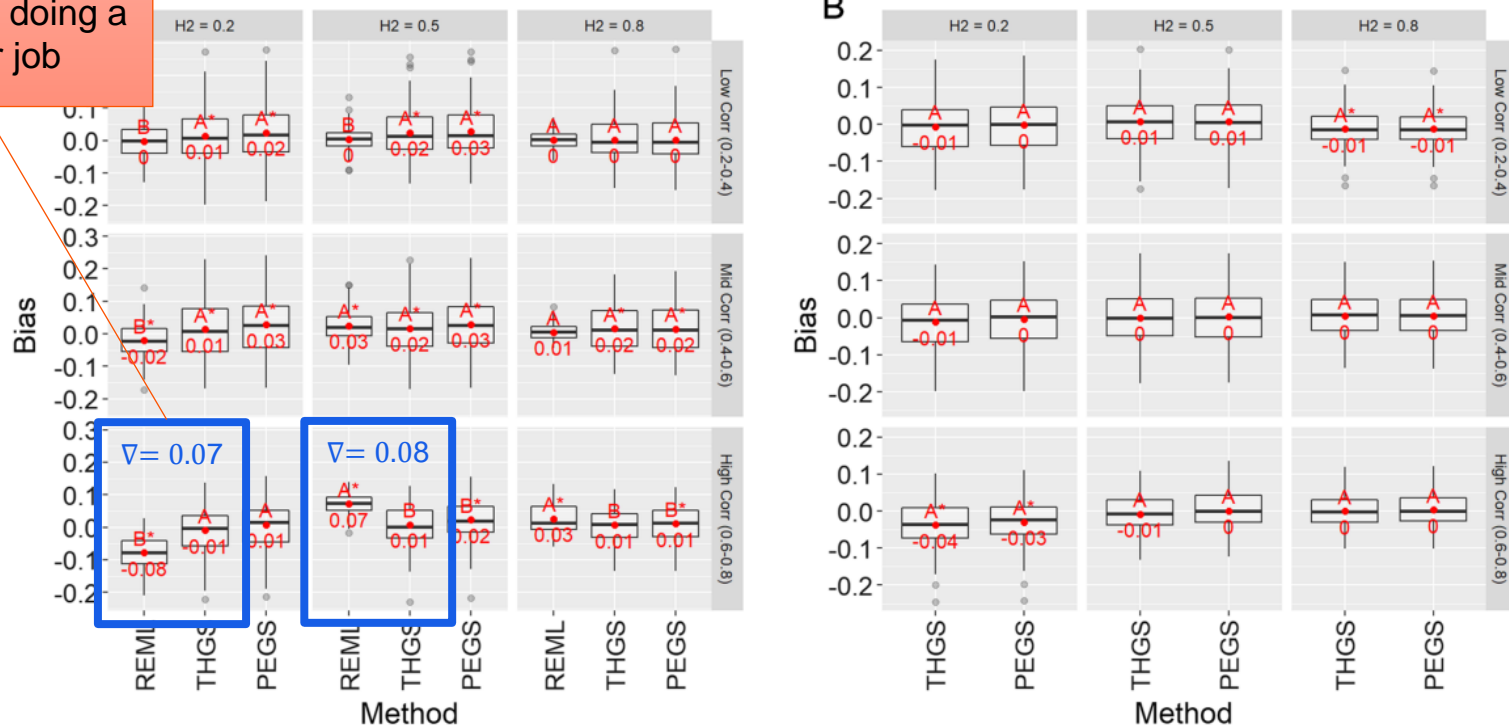Differences are large because REML is doing a poor job

Fig. 6 Bias of estimates of genetic correlation for scenario 1 (**a** wheat dataset) and scenario 2 (**b** soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), and based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparison ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$)

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Precision H2

$$Acc = cor(GEBV, TBV)$$

(Higher is better)

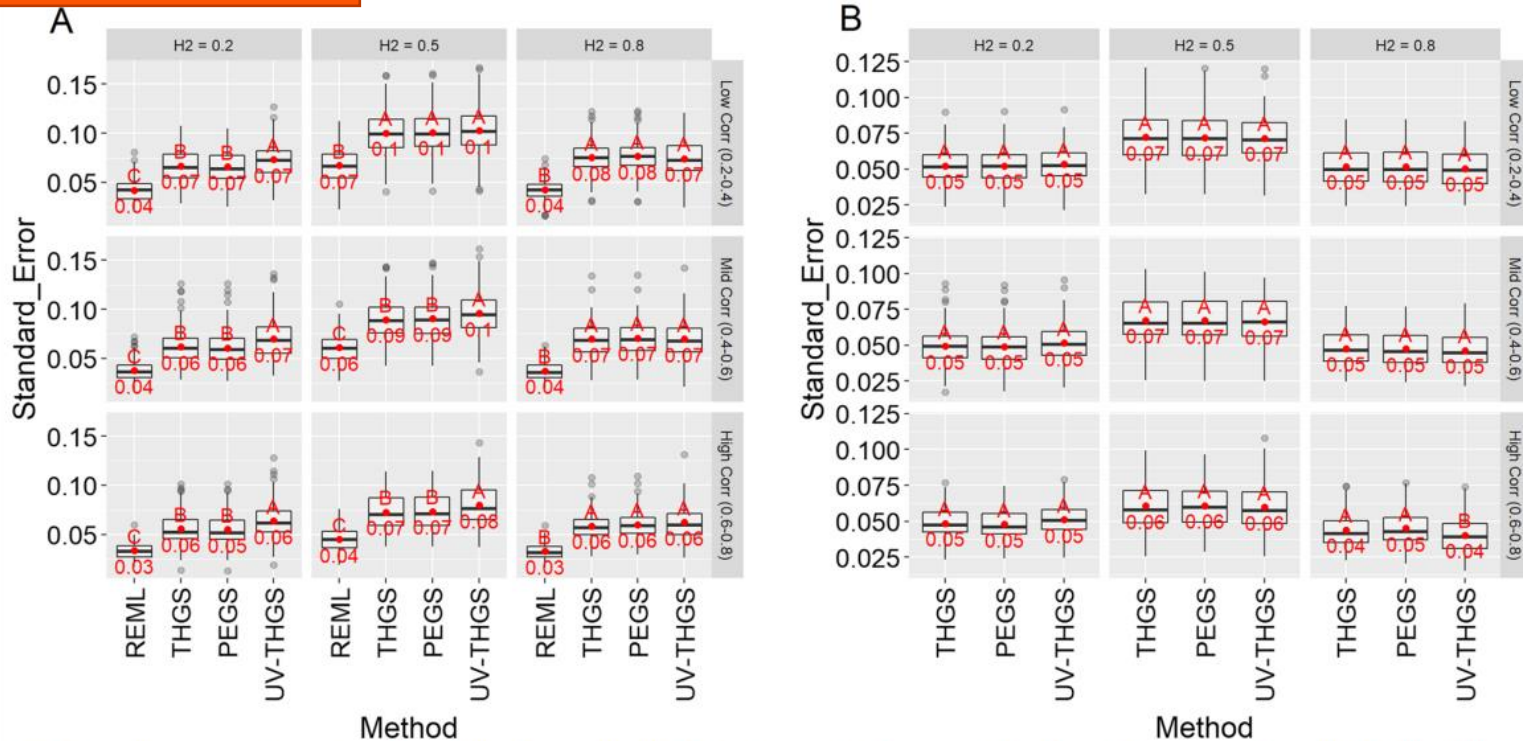PEGS/THGS provided S.E. 0.03 larger than REML in small dataset

Fig. 5 Standard error of estimates of heritability for scenario 1 (**a** wheat dataset) and scenario 2 (**b** soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$)

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

Precision GC

$$Acc = cor(GEBV, TBV)$$

(Higher is better)

PEGS/THGS provided S.E. 0.07 larger than REML in small dataset **under low-heritability settings**
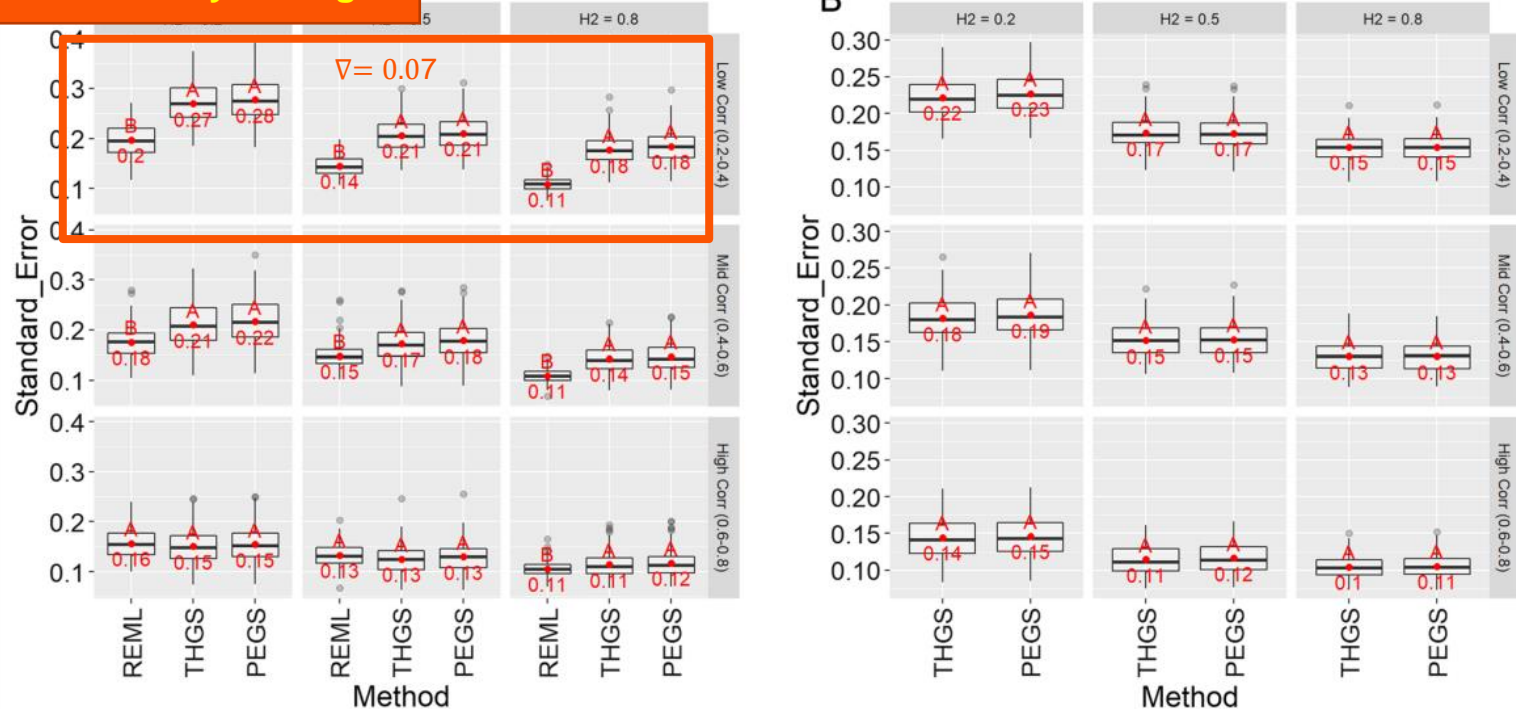
Fig. 7 Standard error of estimates of genetic correlations for scenario 1 (**a** wheat dataset) and scenario 2 (**b** soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$)

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

More data = less bias = more precision

Table 5 Accuracy of GEBV, regression of TBV on GEBV (Slope), and bias and standard error (SE) of estimates of heritabilities ($\hat{h}^2$) and genetic correlations (GC) with increasing numbers of observations per environment (Obs/Env) in scenario 3, based on 100 replicates of the simulation

Method	Obs/Env	Accuracy	Slope	Bias of $\hat{h}^2$	SE of $\hat{h}^2$	Bias of GC	SE of GC
PEGS	250	0.82 (0.03)	0.98 (0.03)	− 0.01 (0.03)	0.07 (0.01)	− 0.01 (0.06)	0.17 (0.02)
PEGS	3000	0.96 (0.03)	1.00 (0.03)	− 0.01 (0.03)	0.04 (0.01)	0.00 (0.06)	0.13 (0.02)
THGS	250	0.82 (0.03)	0.98 (0.04)	0.00 (0.03)	0.07 (0.01)	− 0.02 (0.06)	0.17 (0.02)
THGS	3000	0.96 (0.03)	1.00 (0.03)	− 0.01 (0.03)	0.04 (0.01)	0.00 (0.06)	0.13 (0.02)
UV-THGS	250	0.79 (0.03)	1.04 (0.03)	− 0.01 (0.03)	0.07 (0.01)	–	–
UV-THGS	3000	0.95 (0.03)	1.00 (0.04)	− 0.01 (0.03)	0.04 (0.01)	–	–

Standard errors of statistics are in parenthesis

PEGS pseudo expectation Gauss–Seidel, *THGS* tilde-hat Gauss–Seidel, *UV-THGS* univariate-tilde-hat Gauss–Seidel

Limitations and other considerations

- **More fixed effects?** The absorption of fixed effects beyond the intersect can create a large computational burden. But it is OK to work with pre-adjusted phenotypes like BLUEs, BLUPs and deregressed BLUPs[1].

- **Correlated residuals**: Modeling residual covariances may offset most saving in computation time because of the need for $n \times n$ Kronecker products.

- **Kernels & SVD**: When P>>N, Gauss-Seidel may be costly. When feasible, a solution comes from regress Eigenvectors[2] instead (Z=UDV, solve the MRR using Z*=UD, back solve coefficients $\beta = \beta*$V).

- **Bending[3]**: The covariance $\hat{\Sigma}_\beta$ may not be inversible with too many correlated traits. One may need to shrink the covariance until $\hat{\Sigma}_\beta$ can be inverted. Alternatively, use of simpler covariances: CS and XFA.

- **Balanced data**: REML can be efficiently computed when all phenotypes are collected in all individuals using *canonical transformation[4]* or *kernel diagonalization via eigendecomposition[5]*

1 Garrick et al (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution, 41(1), 1-8.
2 Ødegård et al (2018). Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genetics Selection Evolution, 50(1), 1-12.
3 Jorjani et al (2003). A simple method for weighted bending of genetic (co) variance matrices. Journal of dairy science, 86(2), 677-679.
4 Meyer, K. (1985). Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. Biometrics, 153-165.
5 Lee and Van der Werf (2016). MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics, 32(9), 1420-1422.

Alencar.Xavier@Corteva.com
Quantitative Geneticist, Breeding Analyst LAAF

General recommendation

- REML for balanced sets, small datasets with few traits, or pairwise covariance estimations

- Gibbs sampling for 5-20 traits, small to moderate size datasets

  - https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1439-0388.2008.00774.x

- PEGS / THGS for 5-1000 traits and or moderate to large datasets

  - https://gsejournal.biomedcentral.com/articles/10.1186/s12711-022-00730-w

- MegaLMM for 1000+ traits for moderate size datasets

  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02416-w

Thank you for your attention!

Final remarks:

1) Multivariate models are valuable, but these have been computationally unfeasible

2) Efficient estimation of coefficients (RGS) and variances (PE/TH) enable large MRR

3) THGS/PEGS are suitable replacements to REML, especially for larger dimensionalities

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com