Efficient Estimation of Polygenic Effects via Multivariate Ridge Regression

Alencar Xavier
Research Scientist at Corteva Biostatistics
Adjunct professor at Purdue University

David Habier
Sr. Research Scientist at Corteva Biostatistics

Outline

1. **Introduction**
   - Rationale and statistical model
2. **Coefficients**
   - Univariate
   - Multivariate
3. **Variances**
   - Univariate
   - Multivariate
4. **Simulations**
   - Study 1: Comparison to REML in small balanced data
   - Study 2: Performance in large unbalanced data
   - Limitations and other considerations
5. **Conclusion**

1. **Introduction**
   - Rationale and statistical model
2. **Coefficients**
   - Univariate
   - Multivariate
3. **Variances**
   - Univariate
   - Multivariate
4. **Simulations**
   - Study 1: Comparison to REML in small balanced data
   - Study 2: Performance in large unbalanced data
   - Limitations and other considerations
5. **Conclusion**

Alencar.Xavier@Corteva.com
Corteva Biostatistics, Methods group

Rationale

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

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

Rationale

- **Complex GxE / multi-trait patterns** (= 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.

Why would multivariate be any better?

Simple (bivariate) model:

INFORMATION GAIN

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

$$Var\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

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.

Statistical model

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

- Where $\mathrm{y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_K\}$, $\mu = \{\mu_1, \mu_2, \ldots, \mu_K\}$, $\beta = \{\beta_1, \beta_2, \ldots, \beta_K\}$,
  $\mathrm{e} = \{\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_K\}$, $\mathrm{Z} = \mathrm{BlockDiag}\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_K\}$
- Variances:

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

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:

$$\begin{bmatrix} \mathbf{1}_1'\mathbf{1}_1\sigma_{e_1}^{-2} & \dots & \mathbf{0} & \mathbf{1}_1'\mathbf{Z}_1\sigma_{e_1}^{-2} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{1}_K'\mathbf{1}_K\sigma_{e_K}^{-2} & \mathbf{0} & \dots & \mathbf{1}_K'\mathbf{Z}_K\sigma_{e_K}^{-2} \\ \mathbf{Z}_1'\mathbf{1}_1'\sigma_{e_1}^{-2} & \dots & \mathbf{0} & \mathbf{Z}_1'\mathbf{Z}_1\sigma_{e_1}^{-2} + \mathbf{I}_m\sigma_\beta^{11} & \dots & \mathbf{I}_m\sigma_\beta^{1K} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \dots \\ \mathbf{0} & \dots & \mathbf{Z}_K'\mathbf{1}_K'\sigma_{e_K}^{-2} & \mathbf{I}_m\sigma_\beta^{K1} & \vdots & \mathbf{Z}_K'\mathbf{Z}_K\sigma_{e_K}^{-2} + \mathbf{I}_m\sigma_\beta^{KK} \end{bmatrix} \begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_k \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_K \end{bmatrix} = \begin{bmatrix} \sigma_{e1}^{-2}\mathbf{1}_1'\mathbf{y}_1 \\ \vdots \\ \sigma_{e_K}^{-2}\mathbf{1}_k'\mathbf{y}_K \\ \sigma_{e1}^{-2}\mathbf{Z}_1'\mathbf{y}_1 \\ \vdots \\ \sigma_{e_K}^{-2}\mathbf{Z}_k'\mathbf{y}_K, \end{bmatrix} \quad (2)$$

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

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

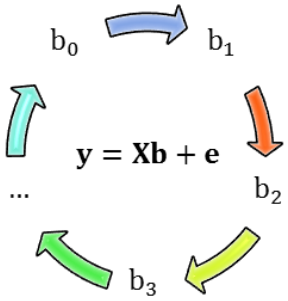
Computing very large multivariate models is **impossible**

unless…

1. Introduction
   - Rationale and statistical model
2. **Coefficients**
   - Univariate
   - Multivariate
3. Variances
   - Univariate
   - Multivariate
4. Simulations
   - Study 1: Comparison to REML in small balanced data
   - Study 2: Performance in large unbalanced data
   - Limitations and other considerations
5. Conclusion

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

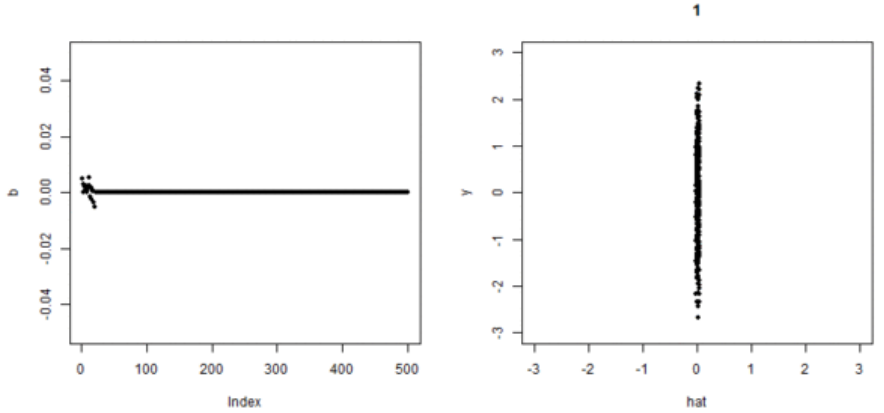


for j in 1:p {

$$\hat{b}_j^{t+1} = \frac{x_j'\hat{e}^t + x_j'x_j\hat{b}_j^t}{x_j'x_j + \lambda}$$

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

}



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, $\mathbf{Z}_j = \oplus_{k=1}^{K} \boldsymbol{z}_{jk}$ 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!!!!

For(j in 1:p) {

These genetic covariances are the whole key for the MRR model

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

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

}

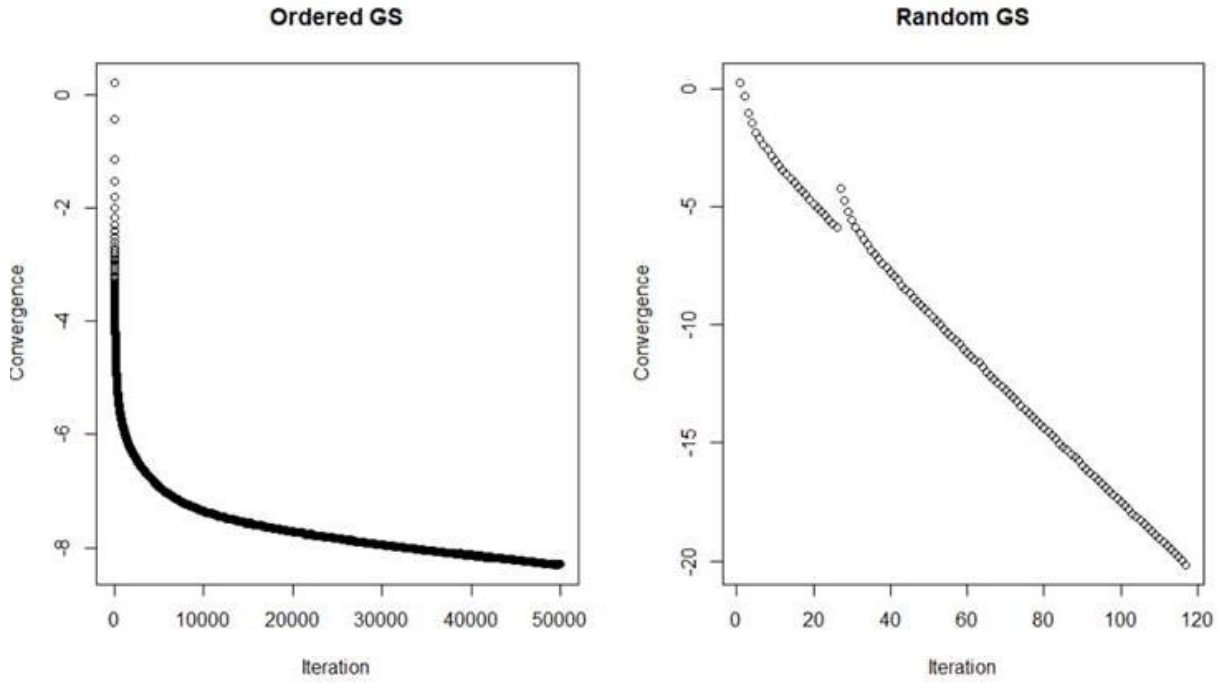
Color code

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

What is in memory?

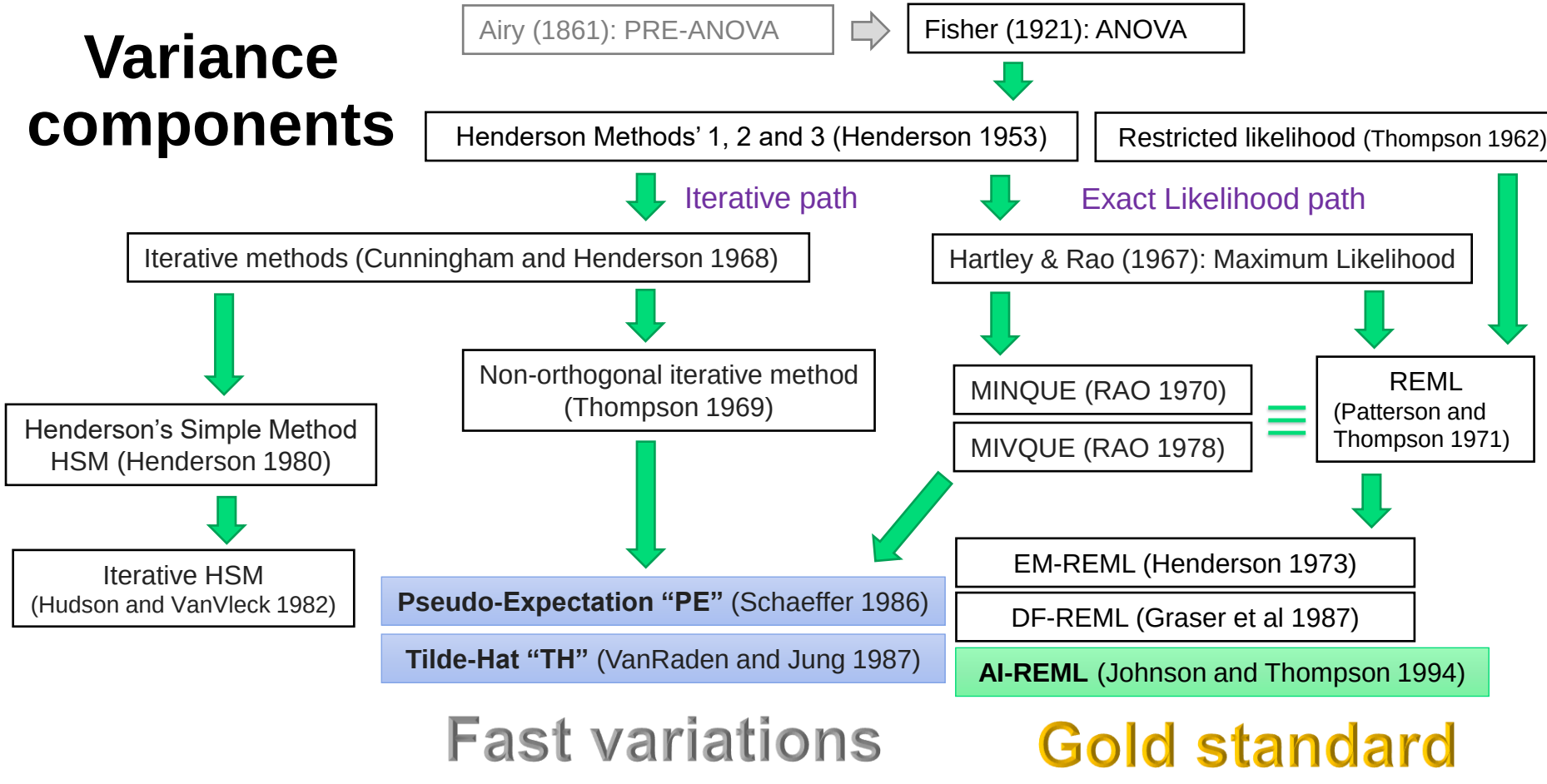
- Z (n x m)
- B (m x k)
- E (n x k)
- ZpZ (m x k)
- $\hat{\Sigma}_{\beta}^{-1}$ (k x k)
- $\hat{\Sigma}_{e}^{-1}$ (k)

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



1. Introduction
   - Rationale and statistical model
2. Coefficients
   - Univariate
   - Multivariate
3. **Variances**
   - Univariate
   - Multivariate
4. Simulations
   - Study 1: Comparison to REML in small balanced data
   - Study 2: Performance in large unbalanced data
   - Limitations and other considerations
5. Conclusion

Variance components



Univariate case: Variance components

- REML

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

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

All methods yield the same residual variance:

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

V is a pain to compute

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

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

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

$$\mathbf{V = ZZ'\sigma^2_{\beta} + I\sigma^2_{\beta}}$$

$$S = I - (X'X)^{-1}X'; \quad P = V^{-1}S$$

$$P = V^{-1} - V^{-1}(X'V^{-1}X)^{-1}X'V^{-1}$$

$$PX = SX = 0$$

$$Sy = \text{Centralized } y = \tilde{y}$$

$$SZ = \text{Centralized } Z = \tilde{Z}$$

- VanRaden's Tilde-Hat

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

$\tilde{\beta}$

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

Multivariate case: (co)variance components

$$\hat{\sigma}^2_{\beta(k)} = \frac{\tilde{\beta}_k \hat{\beta}_k}{\mathrm{tr}(D_k^{-1} \tilde{Z}_k{}' \tilde{Z}_k)}$$

$$\hat{\sigma}_{\beta(k,k')} = \frac{\tilde{\beta}_k \hat{\beta}_{k'} + \tilde{\beta}_{k'} \hat{\beta}_k}{\mathrm{tr}(D_k^{-1} \tilde{Z}_k{}' \tilde{Z}_k) + \mathrm{tr}(D_{k'}^{-1} \tilde{Z}_{k'}{}' \tilde{Z}_{k'})}$$

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

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?

- Z (n x m)
- Bhat (m x k)
- Btilde (m x k)
- E (n x k)
- Y (n x k)
- Ytilde (n x k)
- ZpZ (m x k)
- ZpZtilde (m x k)
- $\hat{\Sigma}_{\beta}$ (k x k)
- $\hat{\Sigma}_{e}$ (k)
- N (k)

An intuitive derivation for Schaeffer's method?

The genetic covariance is simply estimated as the cross-prediction between traits A and B normalized by mean squared genotypes (MSX)!!

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

$$\widehat{\sigma}_{\beta(A,B)} = \frac{(y_A - \mu_A)'(Z_A\beta_B) + (y_B - \mu_B)'(Z_B\beta_A)}{\mathrm{MSX_A} + \mathrm{MSX_B}}$$

$$*\mathrm{MSX} = \mathrm{Tr}\left(\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}\right) = \mathrm{n}\sum_{\mathrm{j}=1}^{\mathrm{P}}\widehat{\sigma}^2_{\mathrm{Z_j}}$$

The key parameters from multivariate models

- Genetic variance

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

- Heritability

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

- Genetic correlations

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

1. Introduction
   - Rationale and statistical model
2. Coefficients
   - Univariate
   - Multivariate
3. Variances
   - Univariate
   - Multivariate
4. **Simulations**
   - Study 1: Comparison to REML in small balanced data
   - Study 2: Performance in large unbalanced data
   - Limitations and other considerations
5. Conclusion

Metrics

1. Breeding values:

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

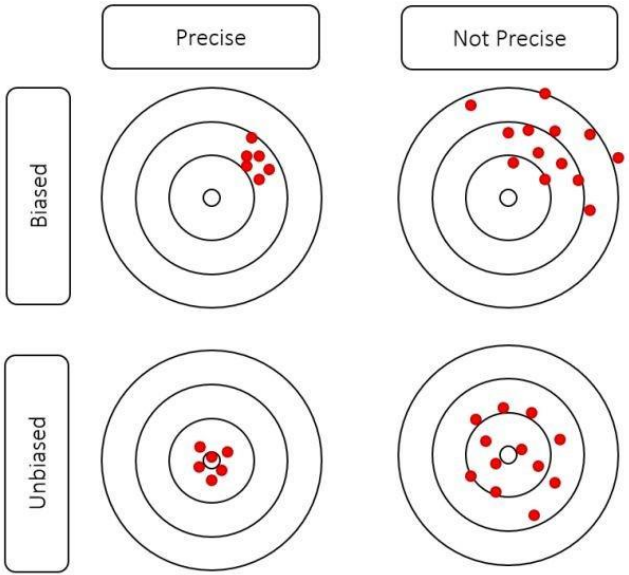
2. Heritability (h^2) and genetic correlations (ρ):

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

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

3. Computation efficiency:

Elapsed time to fit the model



Picture source

Study 1

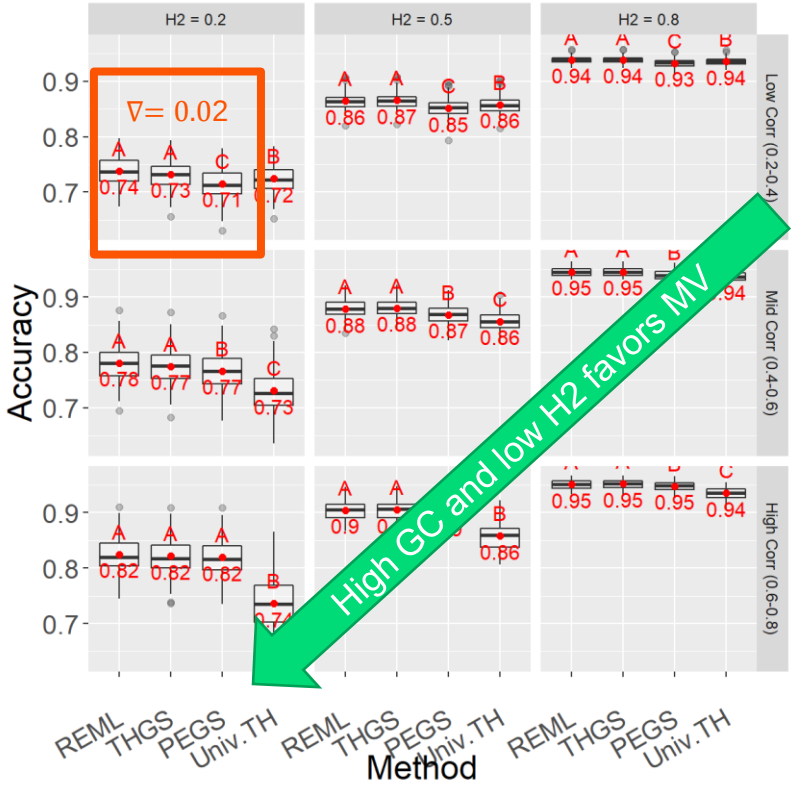
- Wheat data (CYMMIT)
- 599 Individuals
- 1299 Markers
- Scenario: 10 environments, all individuals observed in all locations
- Methods: REML, PEGS, THGS, Univariate

Elapsed time

Method	Time in minutes (SD)	
REML	**256.9 (60.57)**	= 4 hours and 17 minutes
PEGS	0.27 (0.02)	= 16 seconds
THGS	0.27 (0.02)	
Univariate	0.23 (0.03)	= 13 seconds

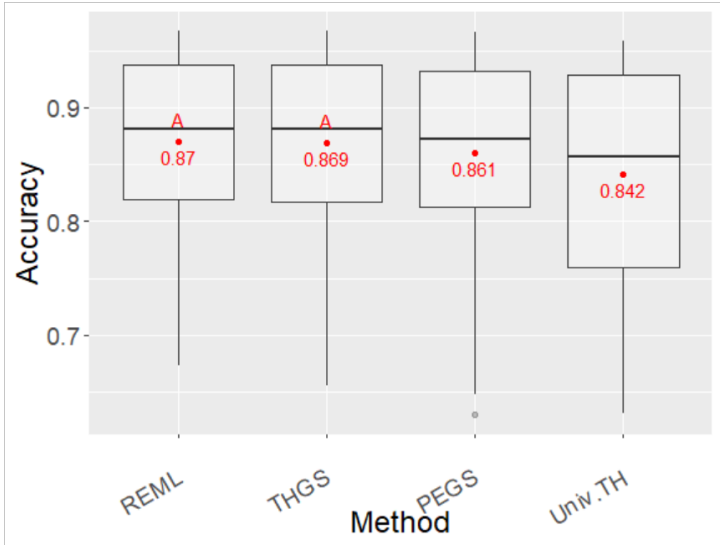
Wheat dataset: 10 traits, 599 individuals, 1299 markers
(available in the BGLR package)

Accuracy of breeding values

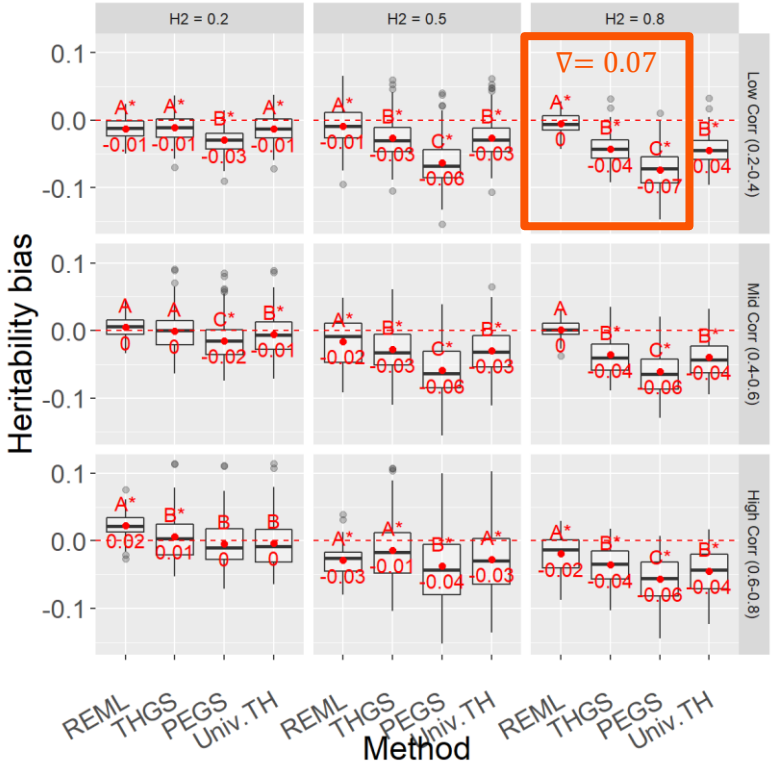


$$Acc = cor(GEBV, TBV)$$

(Higher is better)



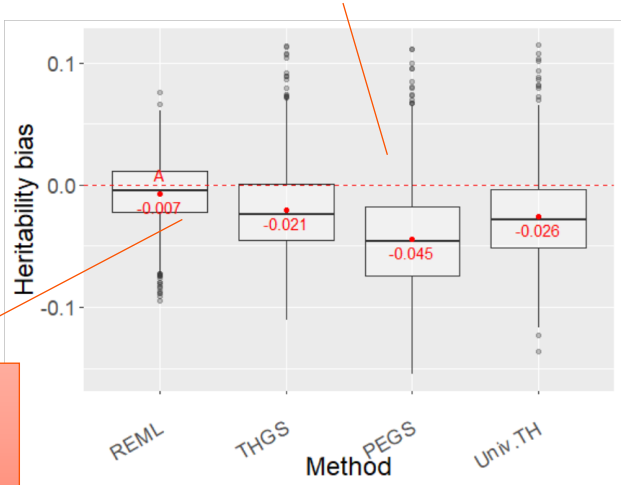
Bias of heritability estimates



$$\text{Bias } h^2 = E(\hat{h}^2 - h^2)$$

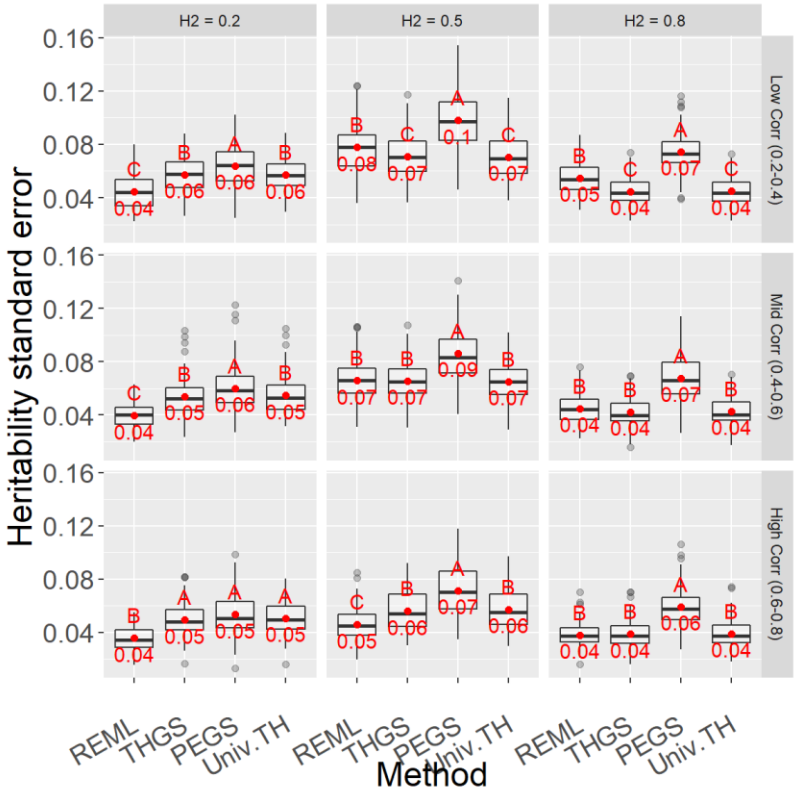
(Closer to zero is better)

PEGS underestimated h2 when true h2 was mid-high



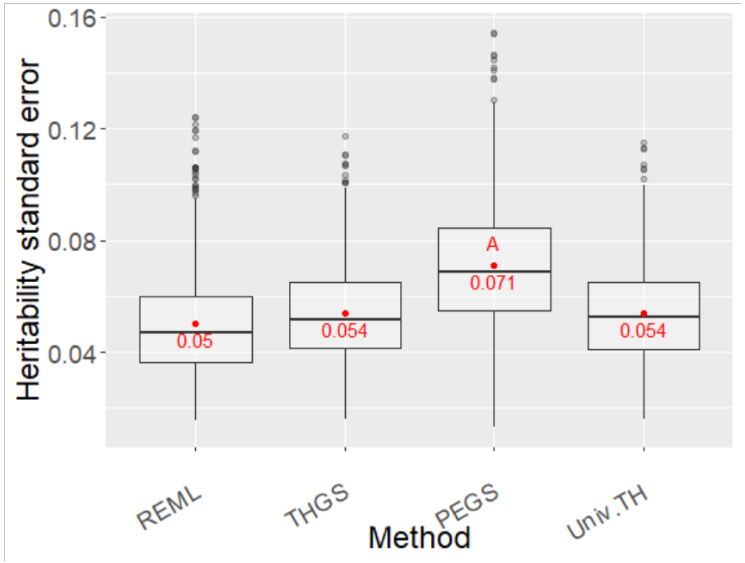
REML also shows large variation…

Precision of heritability estimates



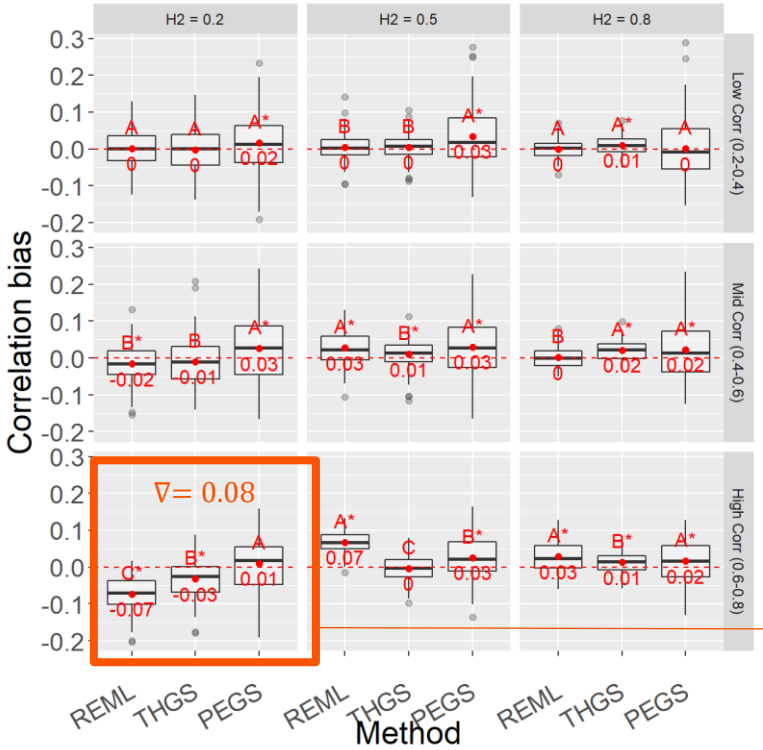
$$\text{Prec } h^2 = SE(\hat{h}^2 - h^2)$$

(Lower is better)



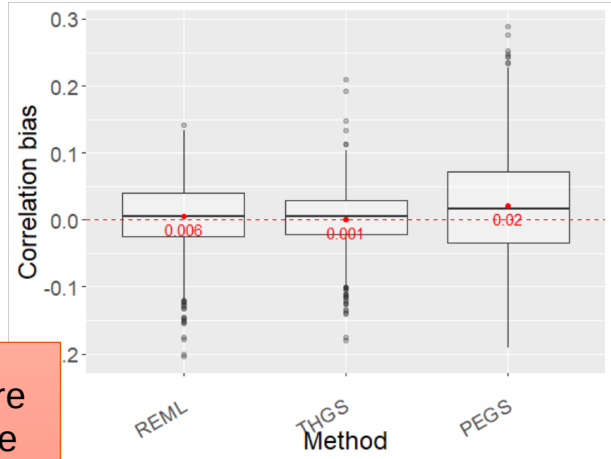
REML / THGS / Univ > PEGS

Bias of genetic correlation estimates



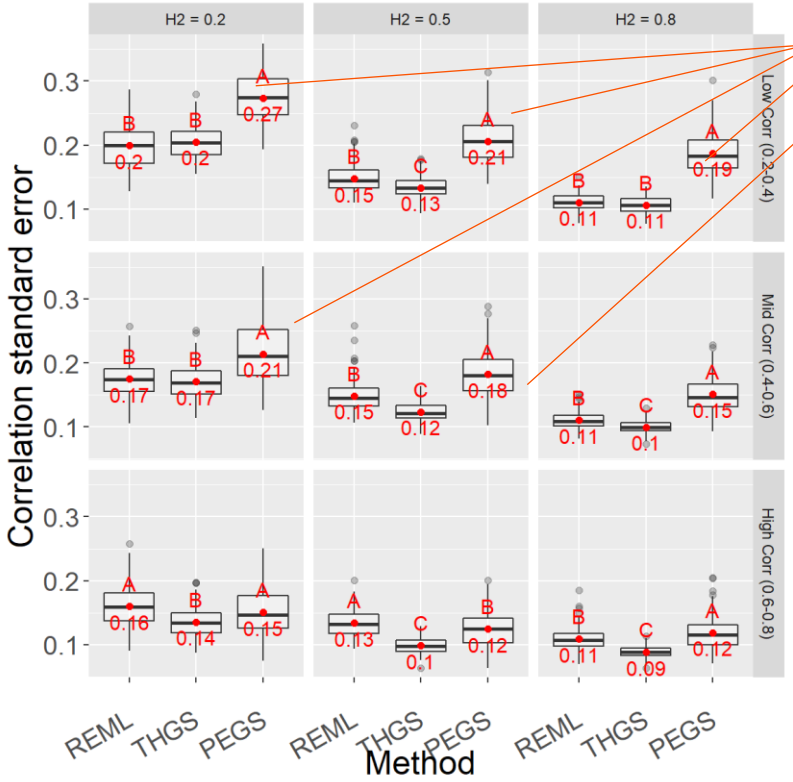
$$\text{Bias } \rho = \mathrm{E}(\hat{\rho} - \rho)$$

(Closer to zero is better)



Differences are large because REML is doing a poor job

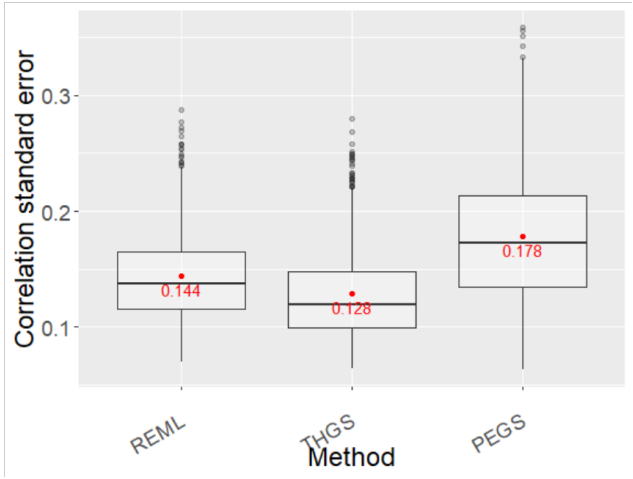
Precision of genetic correlation estimates



PEGS has a hard time to estimate **correlations** when heritability is low, possibly because it **underestimates Genetic Variances**

$$\text{Precision } \rho = \text{SE}(\hat{\rho} - \rho)$$

(Lower is better)



THGS > REML > PEGS

Summary of the smaller & balanced (wheat) dataset

Method	Accuracy	Bias H2	Precision H2	Bias GC	Precision GC
REML	**0.88 (0.01)**	-0.00 (0.03)	0.04 (0.02)	0.01 (0.05)	0.15 (0.03)
PEGS	0.87 (0.02)	-0.03 (0.02)	0.04 (0.01)	0.01 (0.08)	**0.18*** (0.04)
THGS	**0.88 (0.01)**	-0.01 (0.01)	**0.03 (0.01)**	**-0.01 (0.04)**	**0.13 (0.02)**
Univariate	0.85 (0.03)	-0.01 (0.01)	**0.03 (0.01)**	-	-

* PEGS correlations were less precise than THGS, but not statistically different than REML in small balanced datasets

Study 2

- Soybean data (SoyNAM)
- 5000 Individuals
- 4300 Markers
- Scenario: 10 environments, no overlapping individuals
  - **Each individual is observed in a single environment!**
- Methods: PEGS, THGS, Univariate

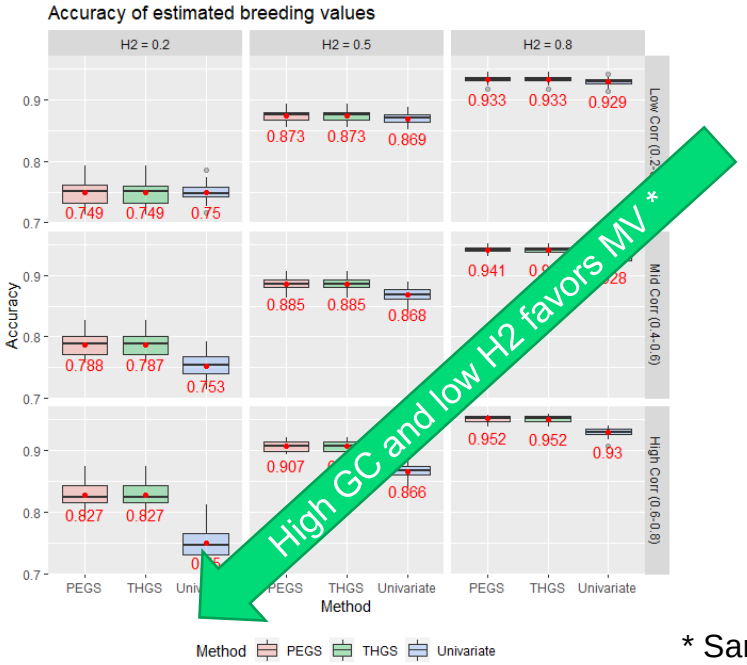
Elapsed time

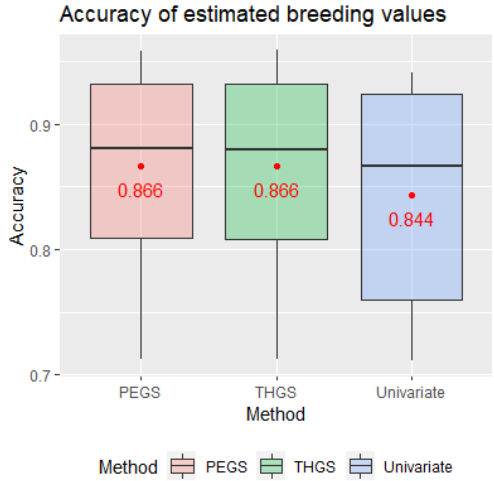
No. of environments		PEGS	THGS	Univariate-THGS
10		0.7 (0.2)	0.7 (0.2)	0.2 (0.0)
50		12 (3)	12 (3)	1.0 (0.1)
100		43 (13)	44 (14)	2.0 (0.3)
200	~3h	168 (48)	165 (44)	4.0 (0.4)
400	~10h	568 (47)	560 (53)	8.0 (1.9)
500	~14h	807 (39)	832 (49)	10.0 (0.6)

(Time in minutes)

Accuracy of breeding values

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

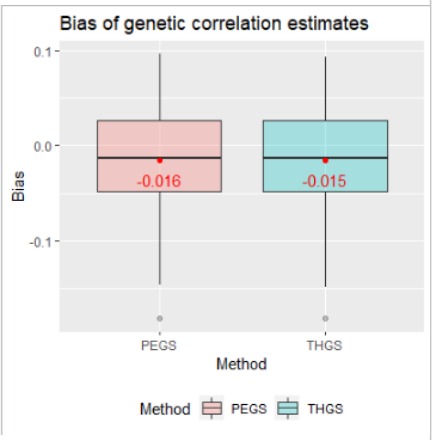


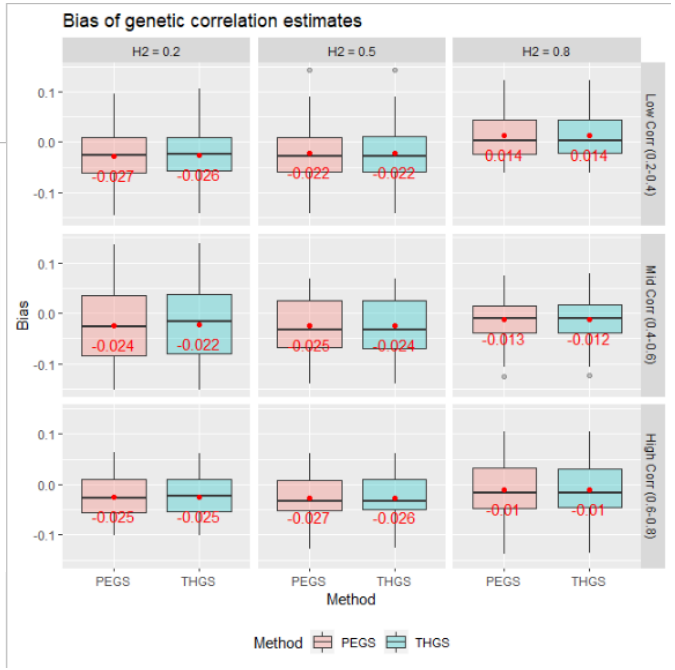


* Same trend observed in wheat

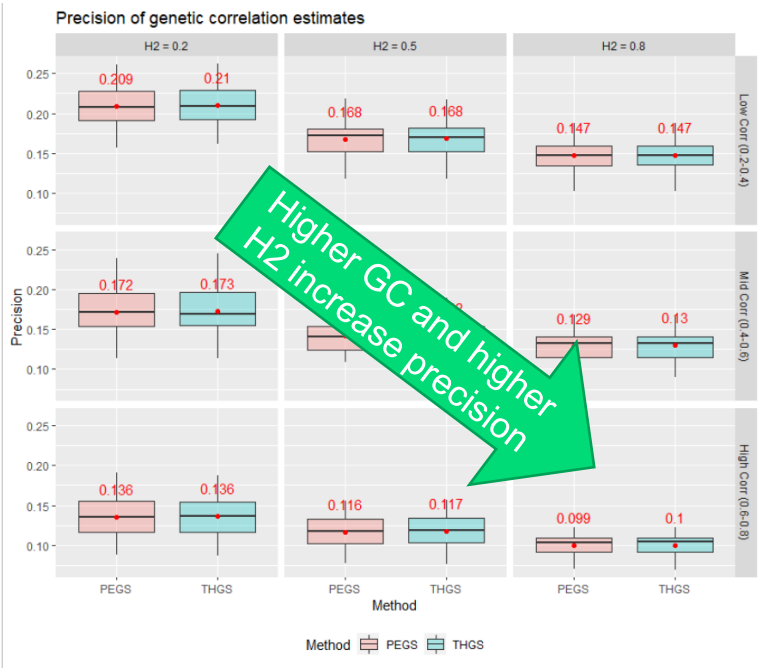
Bias of genetic correlation estimates

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



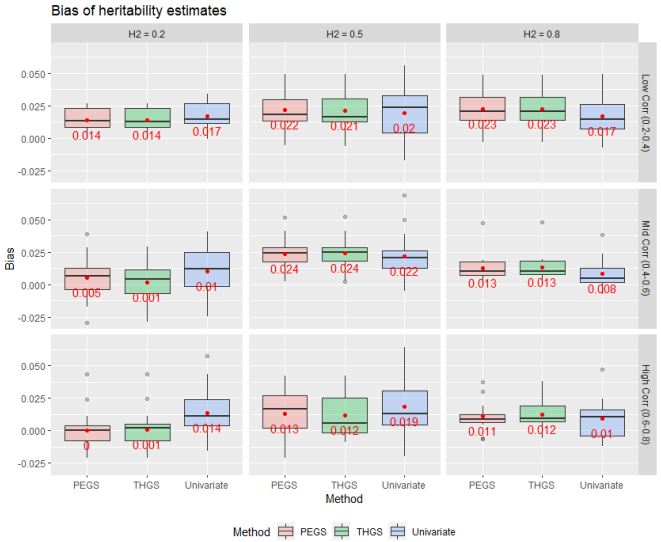


$$\text{Precision } \rho = \mathrm{SE}(\hat{\rho} - \rho)$$

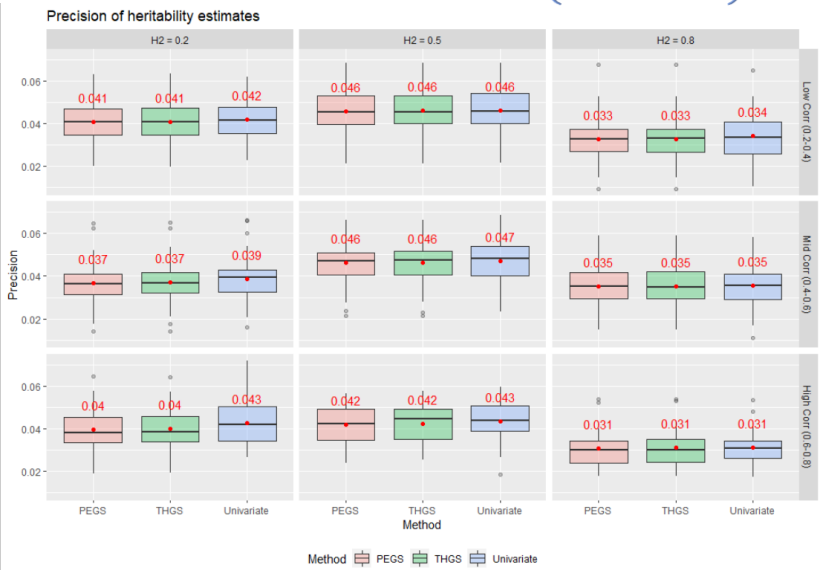


Bias of heritability estimates

$$\text{Bias } h^2 = \mathrm{E}(\widehat{h}^2 - h^2)$$



$$\text{Precision } h^2 = \mathrm{SE}(\widehat{h}^2 - h^2)$$



All roughly the same ~ bias 0.01, S.E. 0.04

Summary in smaller balanced dataset (wheat)

Method	Time (in min.)	Accuracy	Bias H2	Precision H2	Bias GC	Precision GC
REML	256.90 (60.57)	0.88 (0.01)	-0.00 (0.03)	0.04 (0.02)	0.01 (0.05)	0.15 (0.03)
PEGS	0.27 (0.02)	0.87 (0.02)	-0.03 (0.02)	0.04 (0.01)	0.01 (0.08)	0.18 (0.04)
THGS	0.27 (0.02)	0.88 (0.01)	-0.01 (0.01)	0.03 (0.01)	-0.01 (0.04)	0.13 (0.02)
Univariate	0.23 (0.03)	0.85 (0.03)	-0.01 (0.01)	0.03 (0.01)	-	-

THGS ≥ REML ≥ PEGS > Univ

Summary in larger unbalanced dataset (soy)

Method	Accuracy	Bias H2	Prec. H2	Bias GC	Prec. GC
PEGS	0.87 (0.01)	-0.01 (0.01)	0.04 (0.01)	-0.02 (0.06)	0.14 (0.02)
THGS	0.87 (0.01)	-0.01 (0.01)	0.04 (0.01)	-0.02 (0.06)	0.14 (0.02)
Univariate	0.85 (0.02)	-0.02 (0.02)	0.04 (0.01)	-	-

PEGS≅THGS > Univ

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 β= β*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.

1. **Introduction**
   - Rationale and statistical model
2. **Coefficients**
   - Univariate
   - Multivariate
3. **Variances**
   - Univariate
   - Multivariate
4. **Simulations**
   - Study 1: Comparison to REML in small balanced data
   - Study 2: Performance in large unbalanced data
   - Limitations and other considerations
5. **Conclusion**

Thank you for your attention!

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 have some limitations but are suitable replacements to REML

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com