



Leveraging correlated information under multivariate settings

Alencar Xavier

Breeding Analyst at Corteva
Adjunct professor at Purdue

David Habier

Sr. Research Scientist at Corteva

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

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

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

What is genomic prediction?

DATASET	GENOTYPES	PHENOTYPES
TRAINING POPULATION	YES	YES
PREDICTION TARGET	YES	NO

Purpose of GS:

- Improve selection accuracy
- Select material without phenotypes
- Selection of new parents
- Prediction of cross combinations
- Optimize resources
- Stability and genetic architecture

Rationale

- Single-trait models for genomic prediction in plant breeding are already well-established (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 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.

Practical example

Consider a dataset with multiple traits (columns) and individuals (rows)

```
> head(Y,10)
```

	TRAIT1	TRAIT2	TRAIT3	TRAIT4	TRAIT5
GENO001	NA	9.24	8.91	10.11	9.78
GENO002	13.26	11.03	13.85	NA	10.89
GENO003	12.30	10.69	12.24	NA	10.11
GENO004	10.53	10.55	9.05	7.39	NA
GENO005	10.80	11.25	NA	9.35	12.66
GENO006	10.26	10.73	NA	12.16	10.62
GENO007	9.60	8.94	8.46	6.63	10.01
GENO008	NA	9.58	10.08	9.07	12.49
GENO009	10.40	NA	NA	9.26	7.65
GENO010	12.17	10.83	9.89	11.14	12.05

Fit a model using phenotypes (Y) and genotypes (X)

```
> require(bwGR)
> fit = mrr(Y,X)
> round(fit$h2,2)
[1] 0.38 0.48 0.71 0.63 0.60
> round(fit$GC,2)
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1.00	0.76	0.70	0.64	0.62
[2,]	0.76	1.00	0.56	0.65	0.39
[3,]	0.70	0.56	1.00	0.71	0.23
[4,]	0.64	0.65	0.71	1.00	0.24
[5,]	0.62	0.39	0.23	0.24	1.00

Genomic heritability

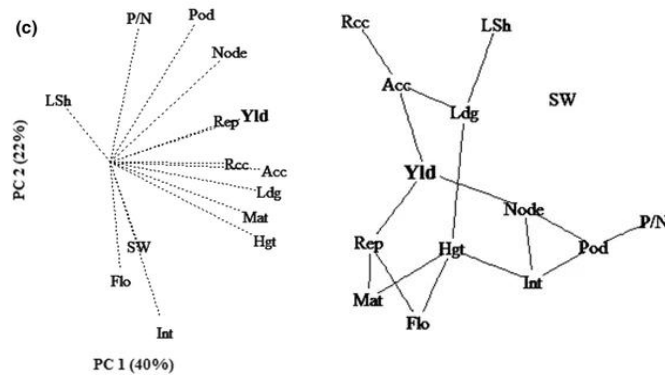
Genetic correlations

- + Genomic breeding values to make selections
- + Marker effects to predict new individuals
- + Variance components to create selection indices

Multivariate models enable analysis of

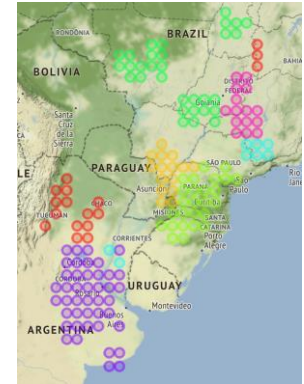
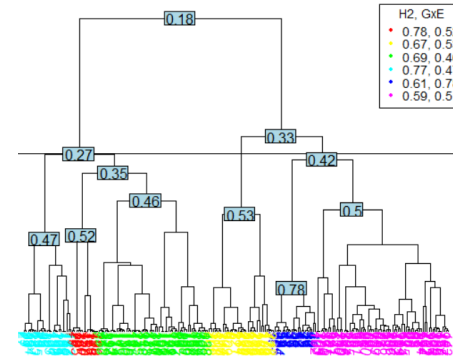
Multiple traits

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



Multiple environments

Example of environmental clustering



Why would multivariate be any better?

Simple (bivariate) model:

INFORMATION GAIN

$$y = g + e$$

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

Why would multivariate be any better?

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

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

- 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

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

- Univariate vs bivariate

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

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

INFORMATION GAIN


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

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

- Where $y = \{y_1, y_2, \dots, y_K\}$, $\mu = \{\mu_1, \mu_2, \dots, \mu_K\}$, $\beta = \{\beta_1, \beta_2, \dots, \beta_K\}$,
 $e = \{e_1, e_2, \dots, e_K\}$, $\mathbf{Z} = \text{BlockDiag}\{\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_K\}$
- Variances:

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

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:

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

where σ_{β}^{ij} is the element at position ij of Σ_{β}^{-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: $\hat{\Sigma}_{\beta(i,j)} = m^{-1}[\hat{\beta}'_i \hat{\beta}_j + \text{tr}(\mathbf{C}^{ij})]$

Computing very large multivariate models is **impossible**

unless...



1. Introduction

- Rationale and statistical model

2. Coefficients

- Univariate
- Multivariate

3. Variances

- Univariate
- Multivariate

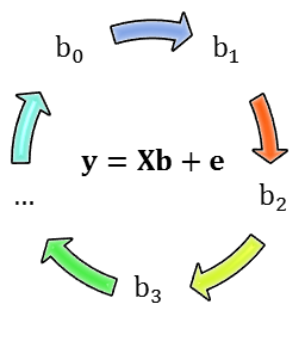
4. Simulations

- Elapsed time
- Accuracy
- Bias & Precision
- 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

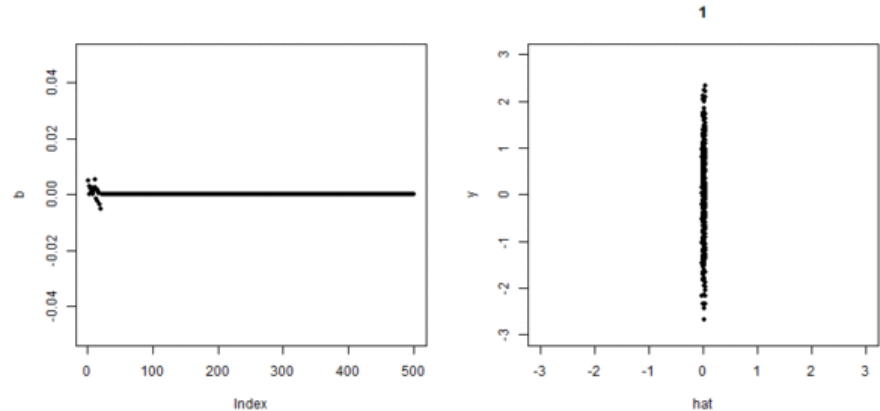


for j in $1:p$ {

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

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

}

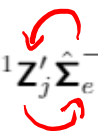


¹ Legarra, A., & Misztal, I. (2008). Computing strategies in genome-wide selection. *Journal of dairy science*, 91(1), 360-366.

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

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


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

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

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

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

2nd update residuals

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

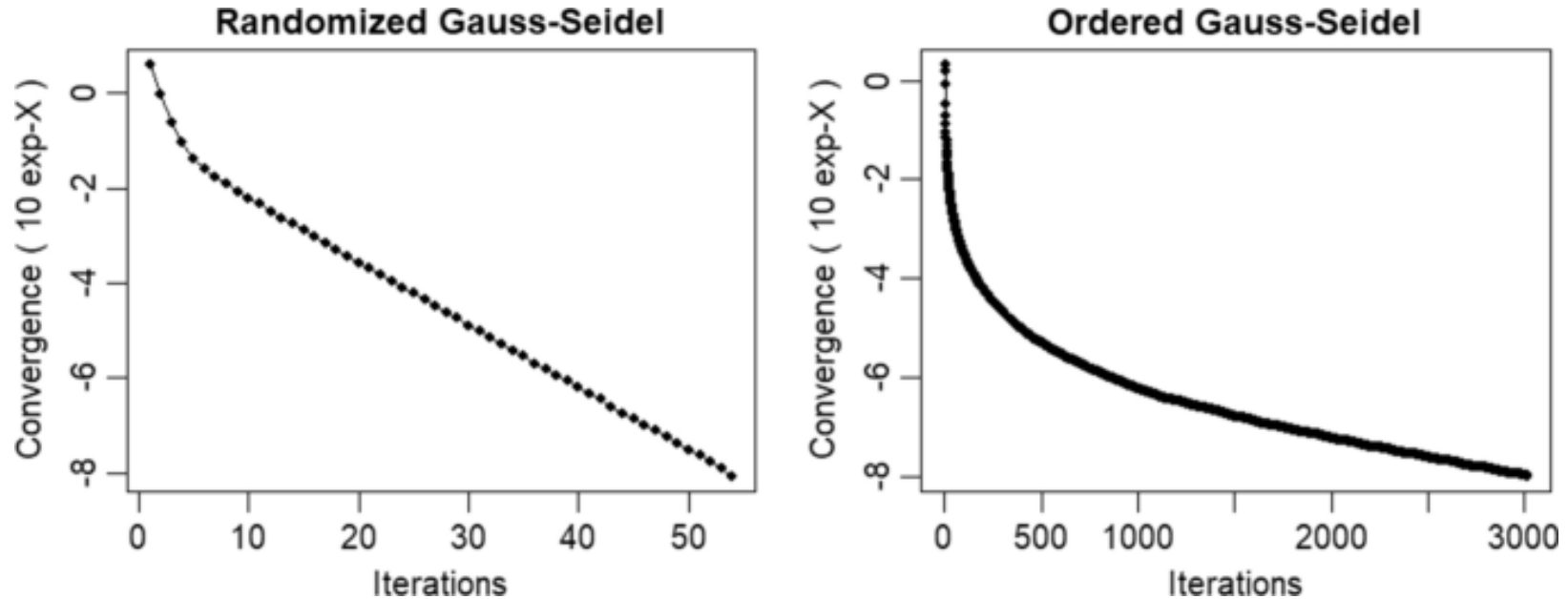
Color code

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

1. Introduction

- Rationale and statistical model

2. Coefficients

- Univariate
- Multivariate

3. Variances

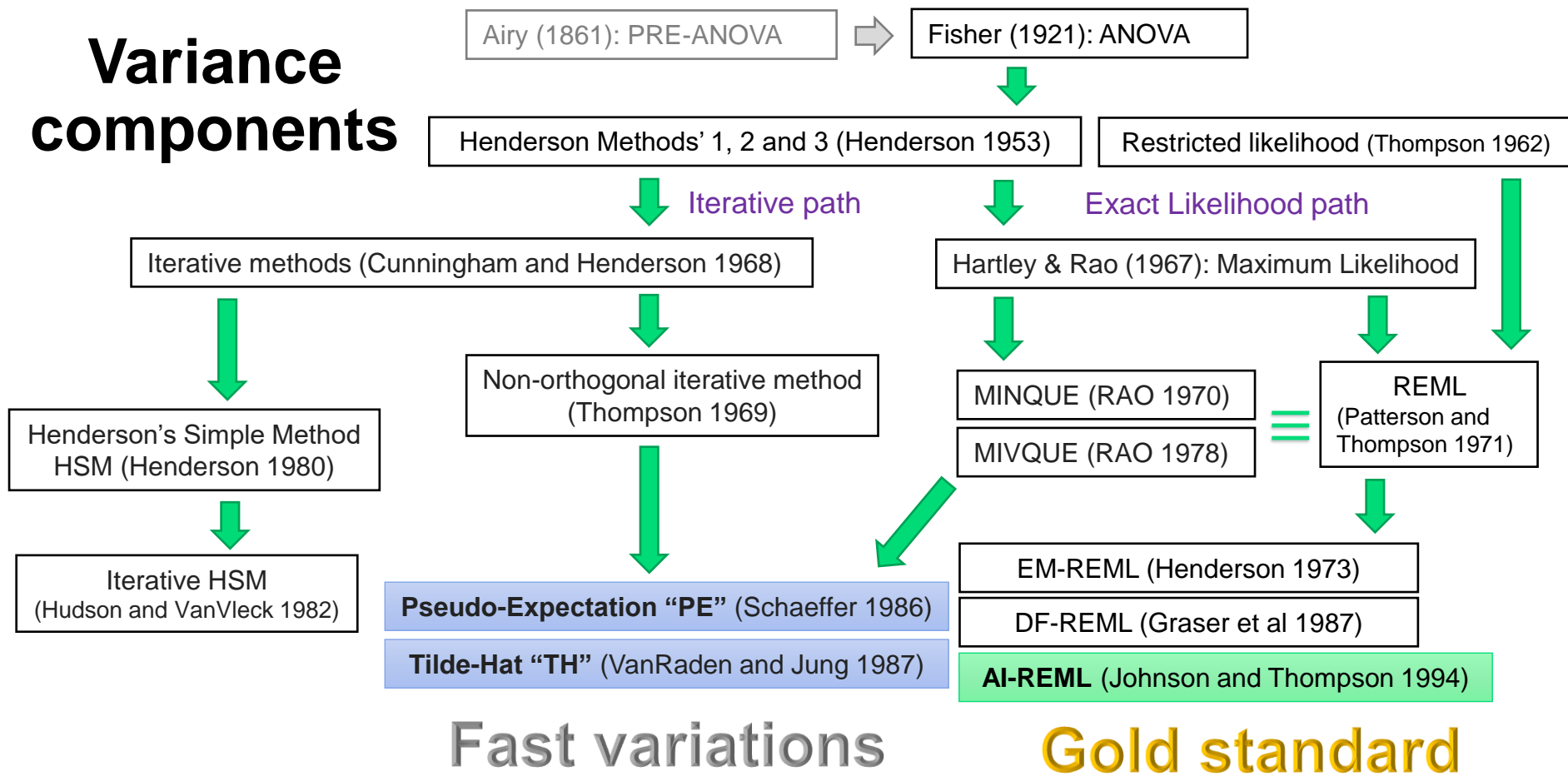
- Univariate
- Multivariate

4. Simulations

- Elapsed time
- Accuracy
- Bias & Precision
- Limitations and other considerations

5. Conclusion

Variance components



Univariate case: Variance components

- REML

$$\frac{\partial LL}{\partial \hat{\sigma}_\beta^2} = 0 \rightarrow \hat{\sigma}_\beta^2 = \frac{y'S'V^{-1}ZZ'V^{-1}Sy}{\text{tr}(V^{-1}ZZ')} = \frac{\hat{\beta}'\hat{\beta}}{\text{tr}(V^{-1}\tilde{Z}'\tilde{Z})}$$

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

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

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

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

- VanRaden's Tilde-Hat

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

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

$$V = ZZ'\sigma_\beta^2 + I\sigma_e^2$$

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

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

$$PX = SX = 0$$

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

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

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

Multivariate case: (co)variance components

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

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

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

**No \mathbf{V} , No \mathbf{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? - \mathbf{Y} (n x k) - $\hat{\Sigma}_{\beta}$ (k x k)

- \mathbf{Z} (n x m)	- $\mathbf{Y}_{\text{tilde}}$ (n x k)	- $\hat{\Sigma}_e$ (k)
- \mathbf{B}_{hat} (m x k)	- \mathbf{ZpZ} (m x k)	- \mathbf{N} (k)
- $\mathbf{B}_{\text{tilde}}$ (m x k)	- $\mathbf{ZpZ}_{\text{tilde}}$ (m x k)	
- \mathbf{E} (n x 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 the scale of Zs

$$\hat{\sigma}_{\beta(A,B)} = \frac{\overset{\text{Centered phenotype of A}}{(\mathbf{y}_A - \mu_A)'} \overset{\text{A predicted from B}}{(\mathbf{Z}_A \boldsymbol{\beta}_B)} + \overset{\text{Centered phenotype of B}}{(\mathbf{y}_B - \mu_B)'} \overset{\text{B predicted from A}}{(\mathbf{Z}_B \boldsymbol{\beta}_A)}}{\text{Tr}(\tilde{\mathbf{Z}}_A' \tilde{\mathbf{Z}}_A) + \text{Tr}(\tilde{\mathbf{Z}}_B' \tilde{\mathbf{Z}}_B)}$$



The key parameters from multivariate models

- Genetic variance

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

- Heritability

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

- Genetic correlations

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

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

Metrics

1. Computation efficiency:

Elapsed time to fit the model

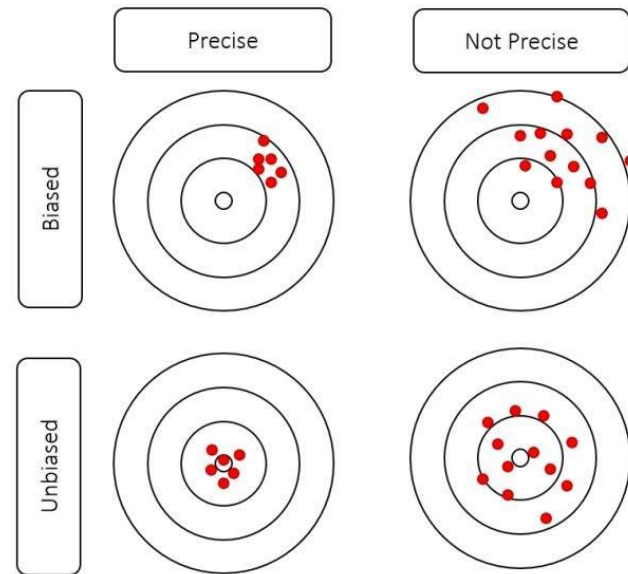
2. Breeding values:

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

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

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

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



[Picture source](#)

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, AIREMLF90) were not suitable to estimate covariance components without overlapping individuals
- Herein REML is just used in scenario A (i.e., genotypes with overlapping records)

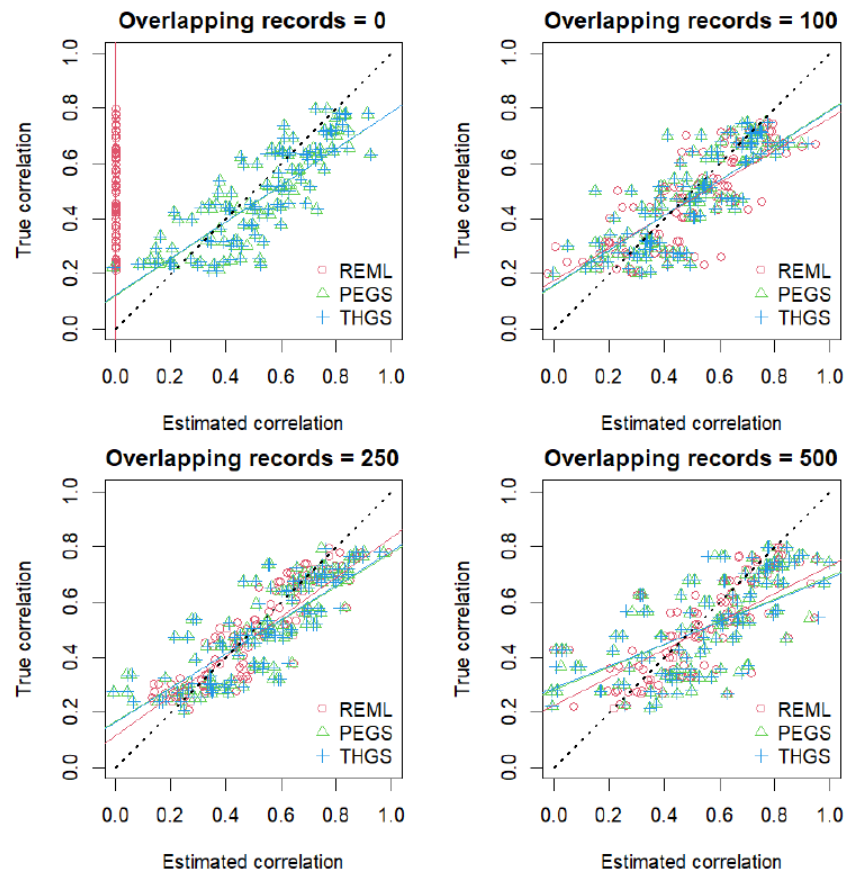


Figure 1: Scatter plot between true and estimated genetic correlations using the soybean dataset with varying number of overlapping individuals across environments.

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)

Elapsed time in large unbalanced dataset

Elapsed time to fit multivariate PEGS, THGS

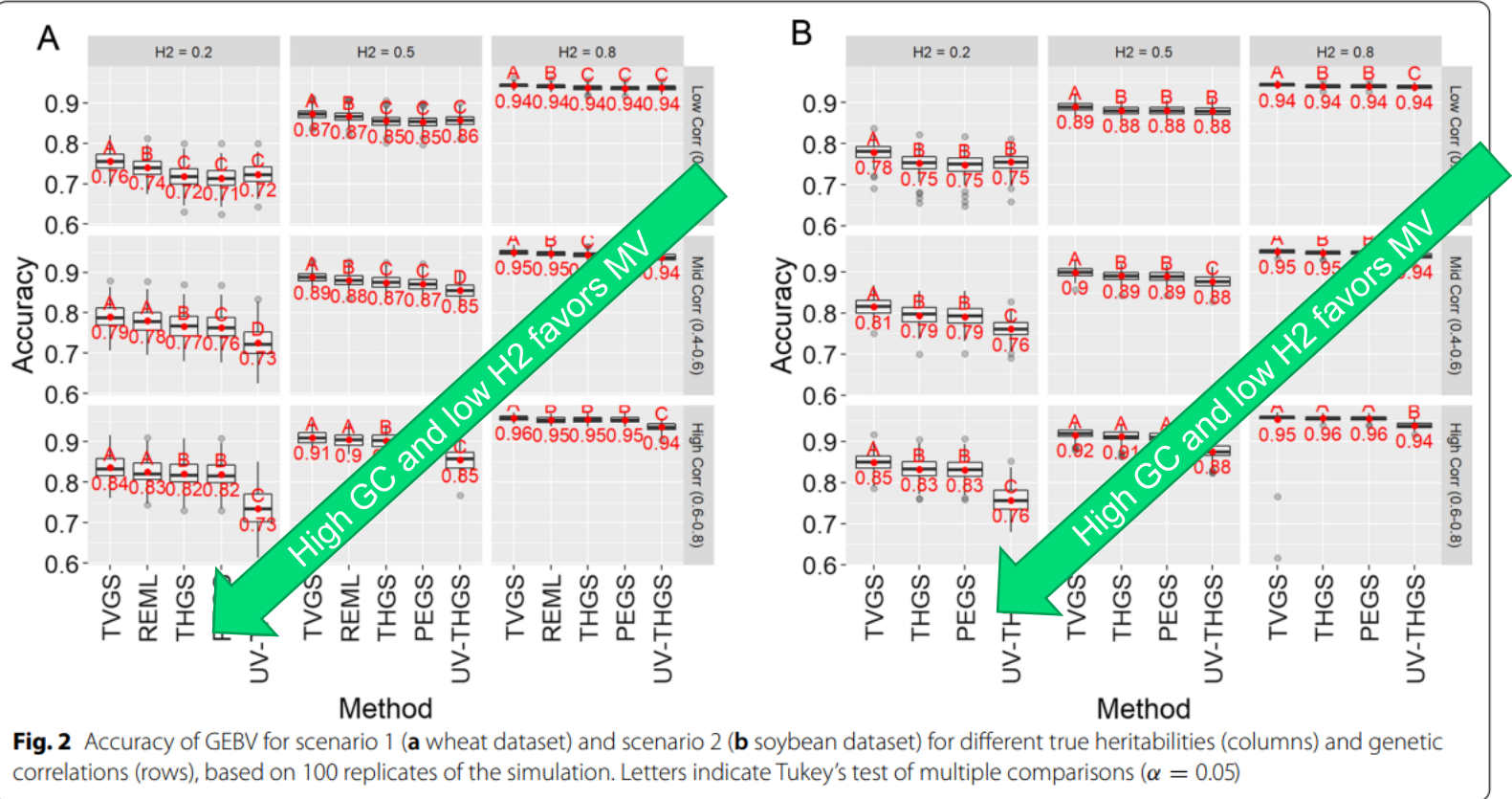
# Markers	# Traits	Time (minutes)
4,311	10	0.2
4,311	50	3.5
4,311	200	80.5
42,034	10	0.8
42,034	50	9.9
42,034	200	123

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

Accuracy

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

(Higher is better)



PEGS/THGS underestimated H2 for **highly-heritable traits with low correlations**

Bias H2

Bias $h^2 = E(\hat{h}^2 - h^2)$
(Closer to zero is better)

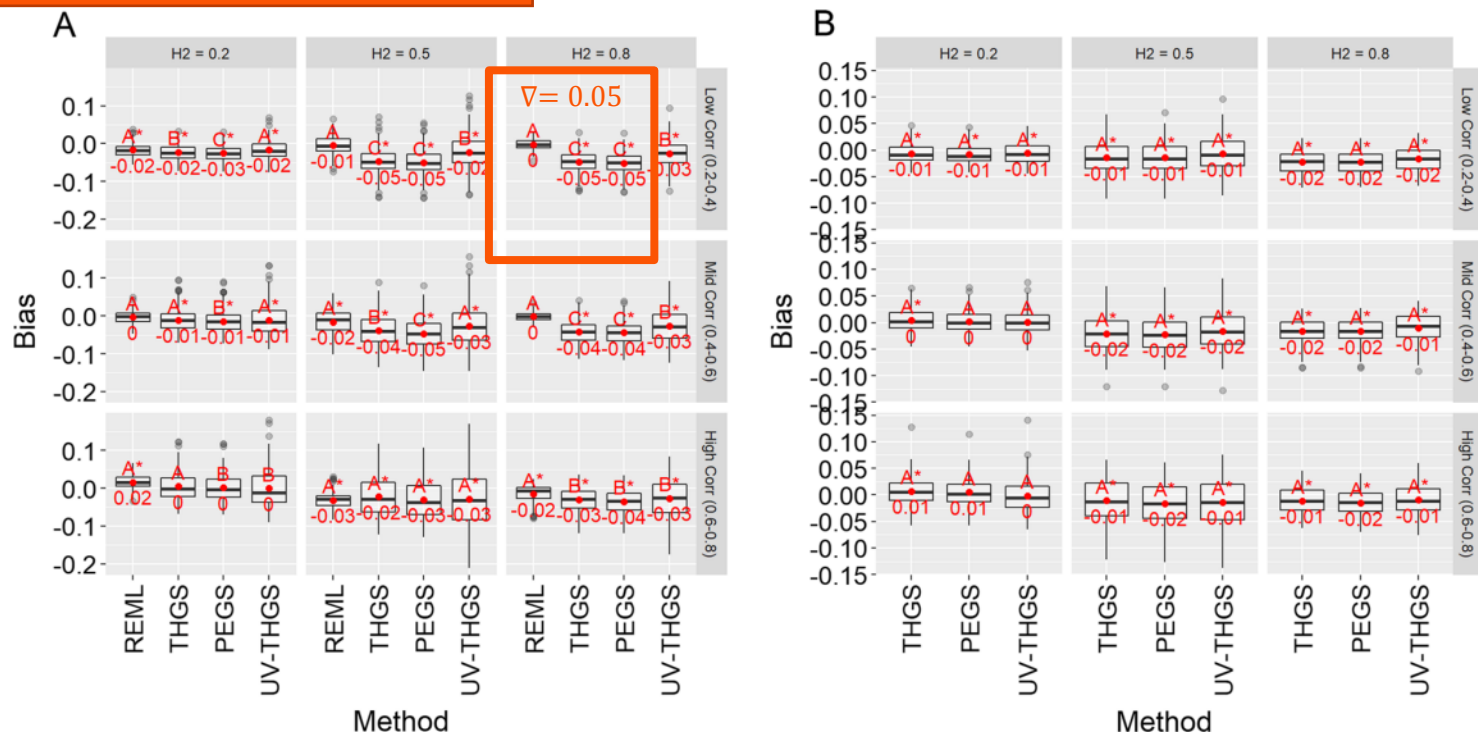


Fig. 4 Bias of estimates of heritability for scenario 1 (a wheat dataset) and scenario 2 (b soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$).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

Differences are large because REML is doing a poor job

Bias GC

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

(Closer to zero is better)

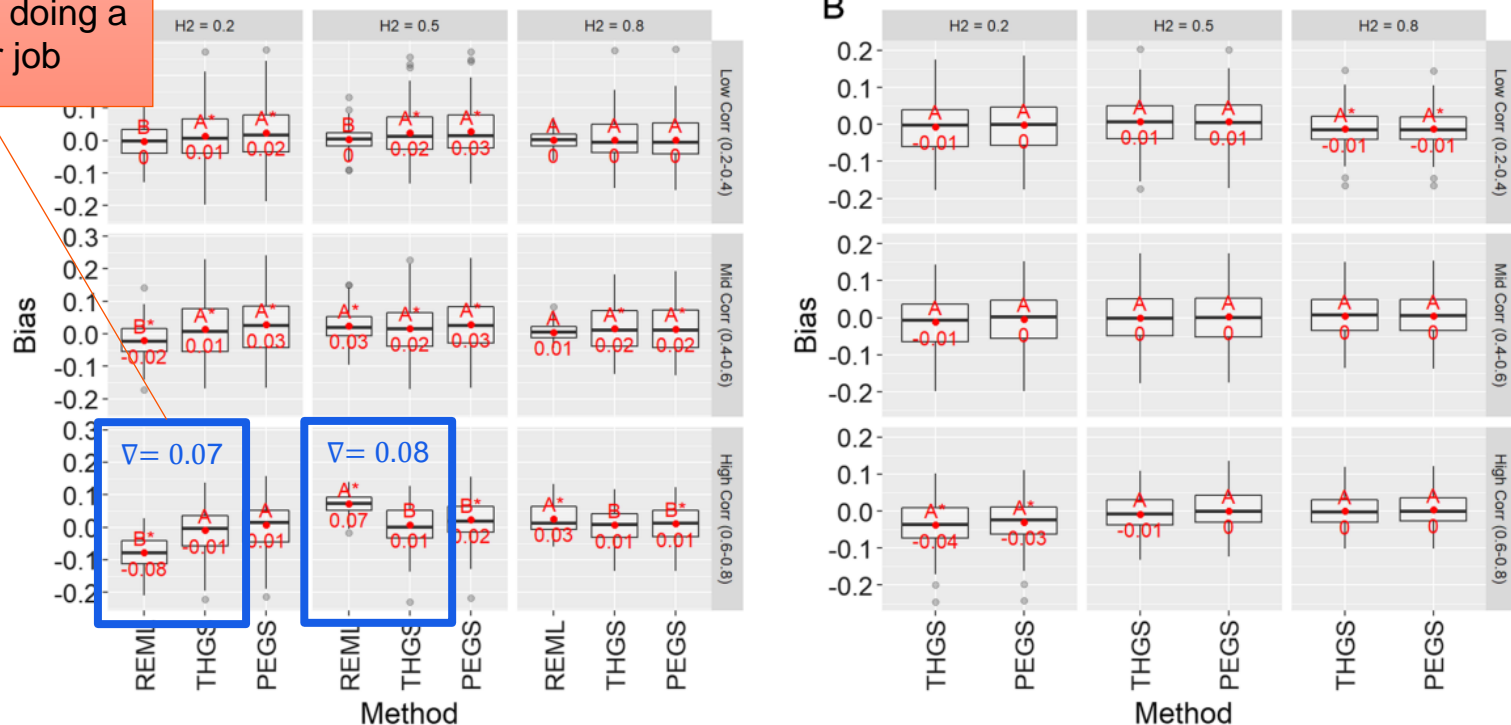


Fig. 6 Bias of estimates of genetic correlation for scenario 1 (a) wheat dataset and scenario 2 (b) soybean dataset for different true heritabilities (columns) and true genetic correlations (rows), and based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparison ($\alpha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

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

Precision H2

$Acc = cor(GEBV, TBV)$
(Higher is better)

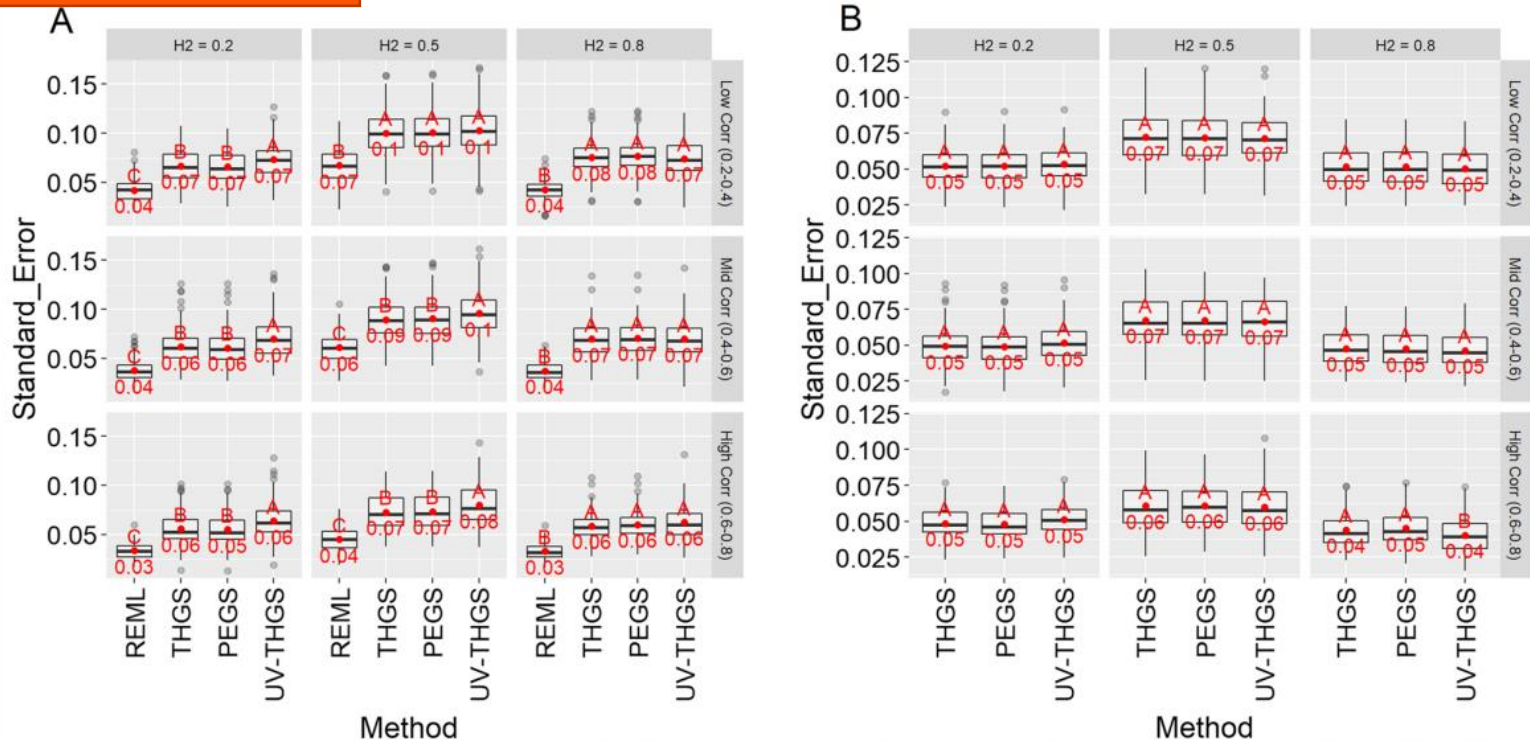


Fig. 5 Standard error of estimates of heritability for scenario 1 (**a** wheat dataset) and scenario 2 (**b** soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons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

PEGS/THGS provided S.E. 0.07
larger than REML in small dataset
under low-heritability settings

Precision GC

$Acc = cor(GEBV, TBV)$
(Higher is better)

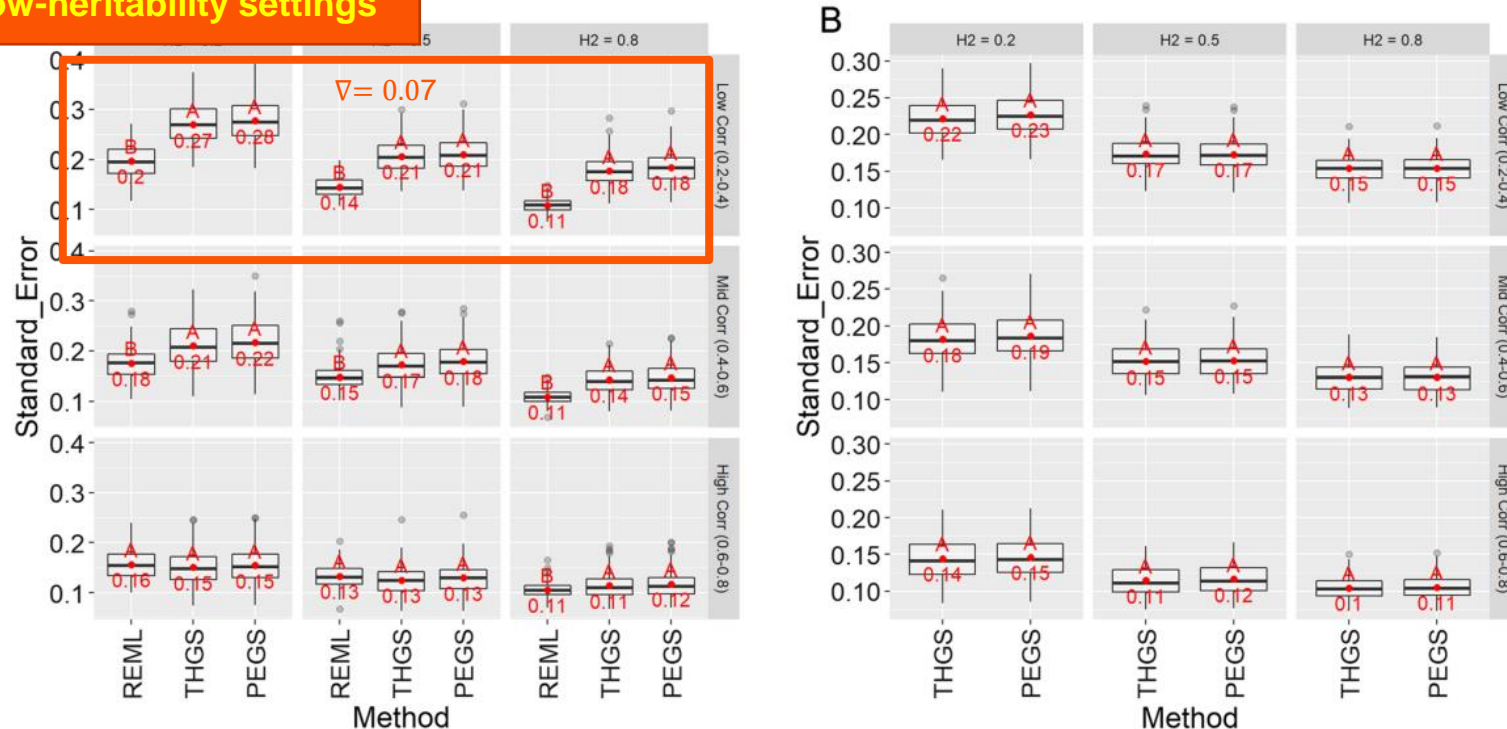


Fig. 7 Standard error of estimates of genetic correlations for scenario 1 (a wheat dataset) and scenario 2 (b soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$)

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 can slightly increase computational cost. One start the analysis with phenotypic BLUEs, BLUPs or deregressed BLUPs¹.
- **Correlated residuals**: Explicit modeling of residual covariances may offset savings in computation time. Instead, an additional random effect with identity design can be used to get residual correlations.
- **Kernels & SVD**: When $P \gg N$, Gauss-Seidel may be costly. When feasible, a solution comes from regress Eigenvectors² 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.

General recommendation

- REML for balanced sets, small datasets with few traits, or pairwise covariance estimations
- Bayesian 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 1-1000 traits and/or any size datasets**
 - <https://gsejournal.biomedcentral.com/articles/10.1186/s12711-022-00730-w>
- XFA-SEM (e.g., MegaLMM-like) for 1000+ traits
 - <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02416-w>
 - https://github.com/alencxav/miscellaneous/blob/master/Toy_MegaLmm_XFA.R

1. Introduction

- Rationale and statistical model

2. Coefficients

- Univariate
- Multivariate

3. Variances

- Univariate
- Multivariate

4. Simulations

- Elapsed time
- 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!

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

Supplementary material

Deriving PE

$$\begin{aligned}y &= Xb + Wu + e \\ Sy &= SXb + SWu + e \\ \tilde{y} &= Zu + e\end{aligned}$$

← Initial model
← Absorption, $SX=0$
← Final model

- $LL = \ln|P| + y'Py$
- $= \ln|V| + \tilde{y}'V^{-1}\tilde{y}$ (fixed effects were absorbed)
- $\frac{\partial LL}{\partial \sigma_u^2} = \text{tr}(V^{-1}ZZ_i') - \tilde{y}'V^{-1}Z\tilde{y}$
- $= \text{tr}(V^{-1}ZZ_i') - \tilde{y}'V^{-1}Z\hat{u}\sigma_u^{-2}$
- $\text{tr}(V^{-1}ZZ_i')\sigma_i^2 = \tilde{y}'V^{-1}Z\hat{u}$

$$\sigma_u^2 = \frac{\tilde{y}'V^{-1}Z\hat{u}}{\text{tr}(V^{-1}ZZ')}$$

Schaeffer's trick

$$\sigma_u^2 = \frac{\tilde{y}'V^{-1}Z\hat{u}}{\text{tr}(V^{-1}ZZ')} = \frac{\tilde{y}'Z\hat{u}}{\text{tr}(ZZ')} = \frac{\tilde{u}\hat{u}}{\text{tr}(Z'Z)}$$

$$\tilde{u} = ZSy$$

$$\begin{aligned}\text{tr}(Z'Z) &= \text{tr}(W'SW) \\ &= \sum_{j=1}^J w_j' S w_j\end{aligned}$$

Column-wise computation

$$\begin{aligned}S w_j &= w_j - X(X'X)^{-1}X'w_j \\ &= w_j - Xb_w\end{aligned}$$

SVD for RRBLUP

$$y = Xb + Zu + e$$

SVD

$$\begin{aligned} Z &= UDV' \\ &= QV' \end{aligned}$$

Model is reparametrized as

$$y = Xb + Qa + e$$

and

$$u = V'a$$

EVD for GBLUP

$$y = Xb + u + e$$

$$V[u] = G\sigma_u^2$$

EVD

$$\begin{aligned} G &= UDU' \\ &= UD^{0.5}D^{0.5}U' \\ &= QQ' \end{aligned}$$

Model is reparametrized as

$$y = Xb + Qa + e$$

and

$$u = Qa$$

Founder rotation

EVD-GBLUP model

$$y = \mu + u + e$$

$$V[u] = G\sigma_a^2$$

$$G = UDU'$$

$$Q = UD^{0.5}$$

$$u = Qa$$

$$y = \mu + Qa + e$$

Qa is GSRU friendly!

Rotation matrix

Use

$$R = UD^{-0.5}$$

to get

$$Q = GR$$

Need SNP effects?

If $G = ZZ'$

Then $\beta = Z'Ra$

Conditional rotation

EVD of founders (F)

$$G_F = U_F D_F U_F'$$

$$R_F = U_F D_F^{-0.5}$$

$$Q_F = G_F R_F$$

Rotate sample pop (S)

$$Q_{S|F} = G_{S,F} R_F$$

Single-Step model?

$$Q_H = A_{SF} R_F$$

fit

$$y = \mu + Q_H a + e$$

where

$$A_{SF} = \begin{bmatrix} A_F \\ A_{S,F} \end{bmatrix}$$

$$A = \begin{bmatrix} A_F & A_{F,S} \\ A_{S,F} & A_S \end{bmatrix}$$

Simplifying covariance structures

- Heteroskedastic compound symmetry (HCS)

$$\sigma_{ij} = \rho \sigma_i \sigma_j$$

$$E[\rho] = \bar{\rho}$$

- Extended factor analytics 2 (XFA2)

$$\Sigma_b^t = U D U'$$

$$\Sigma_b^{t+1} = U^* D^* U^{*'} \quad (* = \text{Only first two eigenpairs})$$

$$\text{Diag}(\Sigma_b^{t+1}) = \Sigma_b^t$$

Non-linear factor

For the model

$$y = \mu + Z\beta + e$$

$$V(\beta_k) = \hat{\sigma}_{\beta(k)}^2 \Lambda_k$$

$$V(\beta_{k,k'}) = \hat{\sigma}_{\beta(k,k')}^2 I_m$$

$$V(e) = \bigoplus_{k=1}^K \sigma_e^2 W_k^{-1}$$

We got to find SNP weights (Λ_k)!

Let ξ be a non-linear factor

- $\Omega_k = \xi \frac{|\beta_k| - \min |\beta_k|}{\max |\beta_k| - \min |\beta_k|} + (1 - \xi)$
- $\Lambda_k = \Omega + (1 - \bar{\Omega})$

For(j in 1:p) {

1st solve for beta

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

2nd update residuals

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

}

NOTE: SNP weights
always go hand-to-hand
with the residuals

SEM

Univariate model

$$y_k = X_k b_k + Z_k \beta_k + e_k$$

SEM

$$y_k = X_k b_k + Z_k \beta_k + e_k$$

$$y_k = X_k b_k + [G_{-k} b_{-k} + Z_k \beta_k^*] + e_k$$

$$G = Z [\beta_1 \quad \dots \quad \beta_K]$$

Final estimator

$$\beta_k = b_{-k} \beta_{-k} + \beta_k^*$$

SEM-XFA

SEM

$$y_k = X_k b_k + Z_k \beta_k + e_k$$

$$y_k = X_k b_k + [G^* \alpha + Z_k \beta_k^{**}] + e_k$$

where

$$G = Z [\beta_1 \quad \dots \quad \beta_K]$$

$$G^* = U^* D^*$$

(* = Few PCs from SVD of G)

Final estimator

$$\beta_k = \alpha V^* + \beta_k^{**}$$