An Introduction to Scalable Multivariate & Megavariate Models

Alencar Xavier
Breeding Analyst at Corteva
Adjunct professor at Purdue
https://alenxav.github.io/

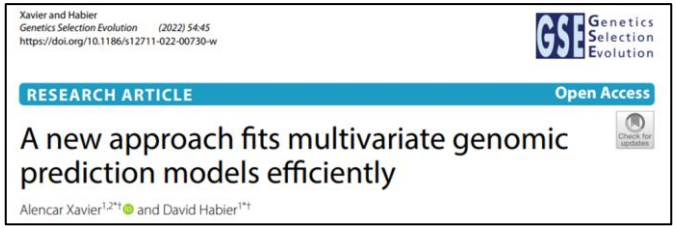
REFERENCES: https://doi.org/10.1186/s12711-022-00730-w and https://doi.org/10.1093/genetics/iyae179

Outline

1. Introduction
2. Coefficients
3. Variances
4. Simulations
5. Megavariate
6. Conclusion

1. **Introduction**
   - Rationale and statistical model
2. Coefficients
3. Variances
4. Simulations
5. Megavariate
6. Conclusion

Are these new methods any good?

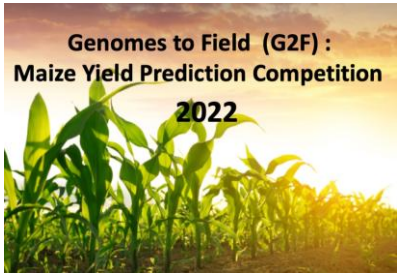

Xavier and Habier
Genetics Selection Evolution (2022) 54:45
https://doi.org/10.1186/s12711-022-00730-w

GSE Genetics Selection Evolution

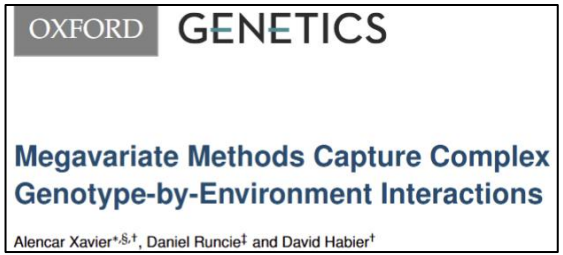
RESEARCH ARTICLE Open Access

A new approach fits multivariate genomic prediction models efficiently

Alencar Xavier[1,2*†] and David Habier[1*†]



https://doi.org/10.1093/genetics/iyae195


OXFORD GENETICS

Megavariate Methods Capture Complex Genotype-by-Environment Interactions

Alencar Xavier[*,§,†], Daniel Runcie[‡] and David Habier[†]

What is genomic prediction?

DATASET	GENOTYPES	PHENOTYPES
TRAINING POPULATION	YES	YES
PREDICTION TARGET	YES	NO

Purpose of GS:

- Improve selection **accuracy**
- Prediction of new populations
- Select non-phenotyped material
- Resource **optimization**

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

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 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

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

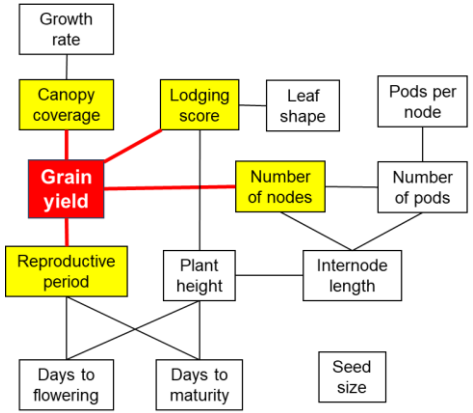
Genetic correlations

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

Multivariate models enable analysis of

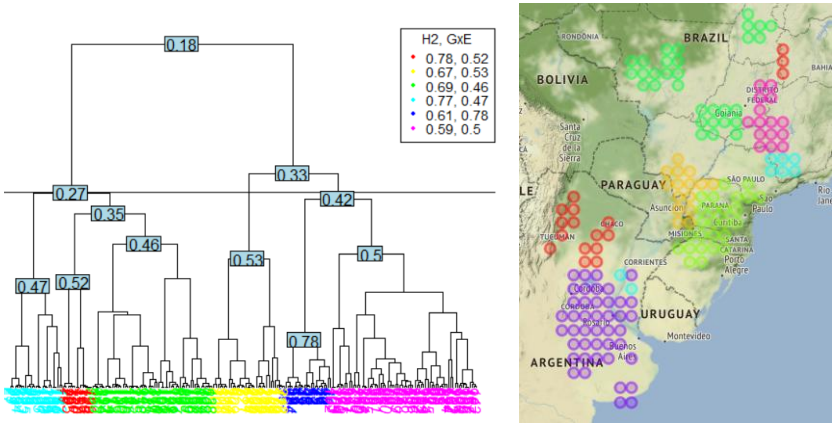
Multiple traits

Graph derived from the genetic correlation among 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

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

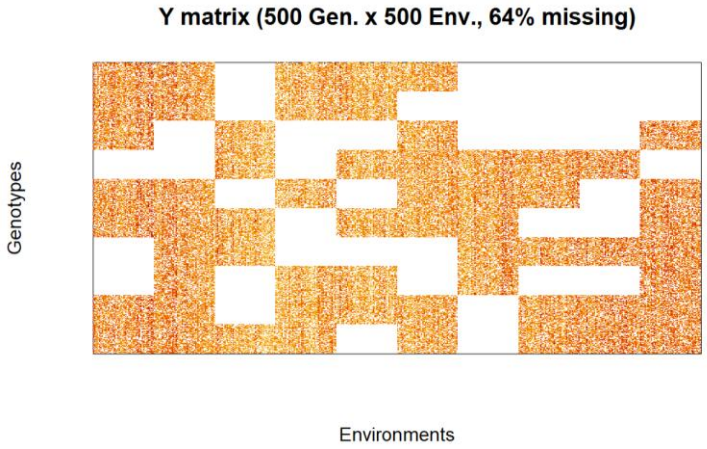
$$Var\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{bmatrix} + \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{bmatrix}$$

What is a megavariate model?

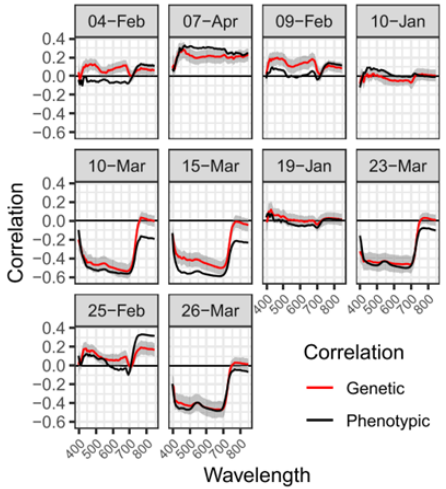
Model:

$$\mathbf{Y} = \mathbf{G} + \mathbf{E}$$
$$= \mathbf{XB} + \mathbf{E}$$
$$\mathbf{B} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{G}})$$

- Large number of environments
- High-throughput phenotyping



Simulated example of sparse multi-environmental trials



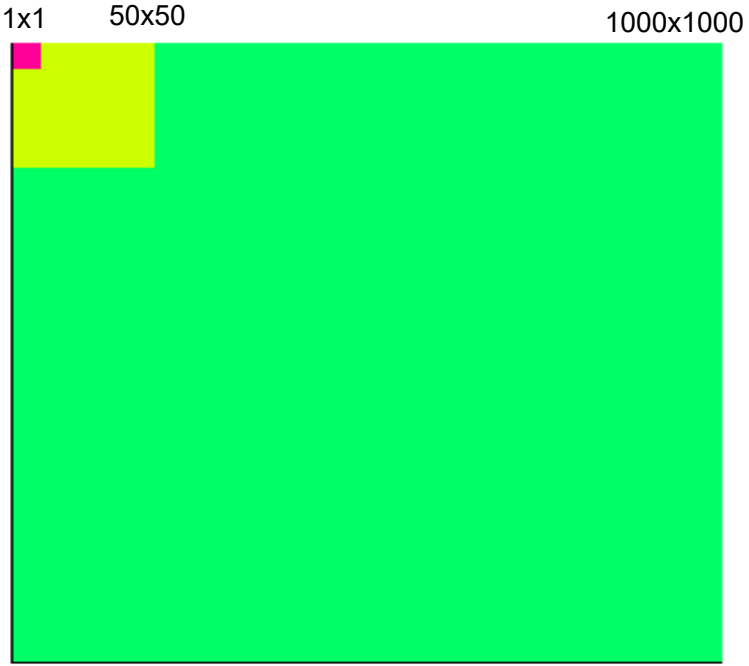
Runcie et al. (2021) Mega-scale linear mixed models for genomic predictions with thousands of traits. https://doi.org/10.1186/s13059-021-02416-w

Scale of Σ_β

Megavariate
50+ traits

Multivariate
2-50 traits

Univariate



Multivariate computational complexity is exponential k^7 (Zhou and Stephens 2014)

Multivariate Statistical Model

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

- 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:

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

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: $\hat{\Sigma}_{\beta(i,j)} = m^{-1}[\hat{\boldsymbol{\beta}}_i'\hat{\boldsymbol{\beta}}_j + tr(\mathbf{C}^{ij})]$

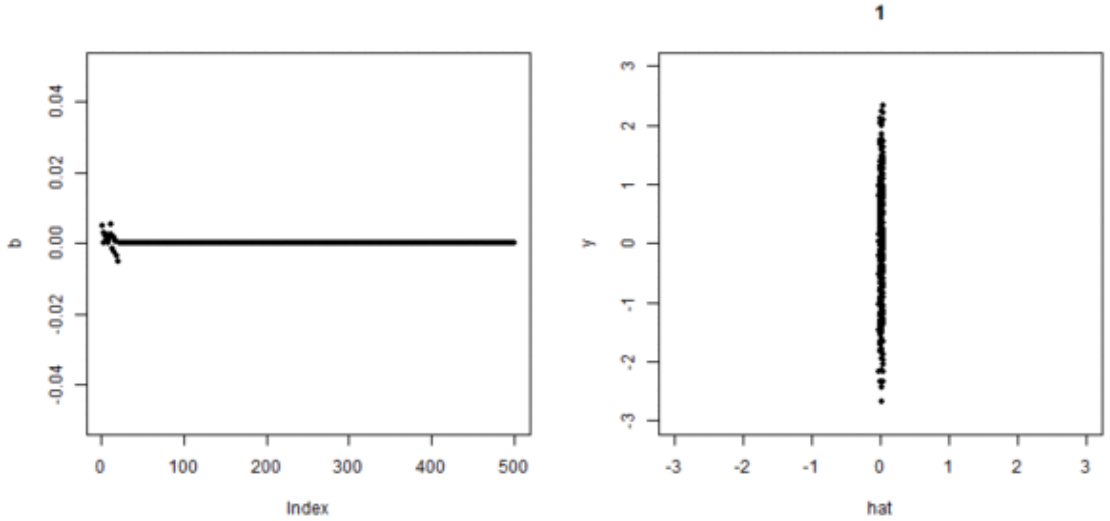
Computing very large multivariate models is **impossible**

unless…

1. Introduction
2. **Coefficients**
3. Variances
4. Simulations
5. Megavariate
6. Conclusion

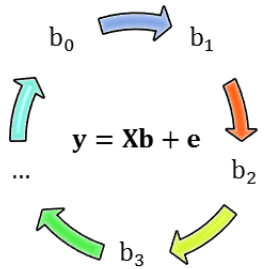
Efficient coefficient estimation

The *Gauss-Seidel* method avoids building the systems of equations by one marker at the time



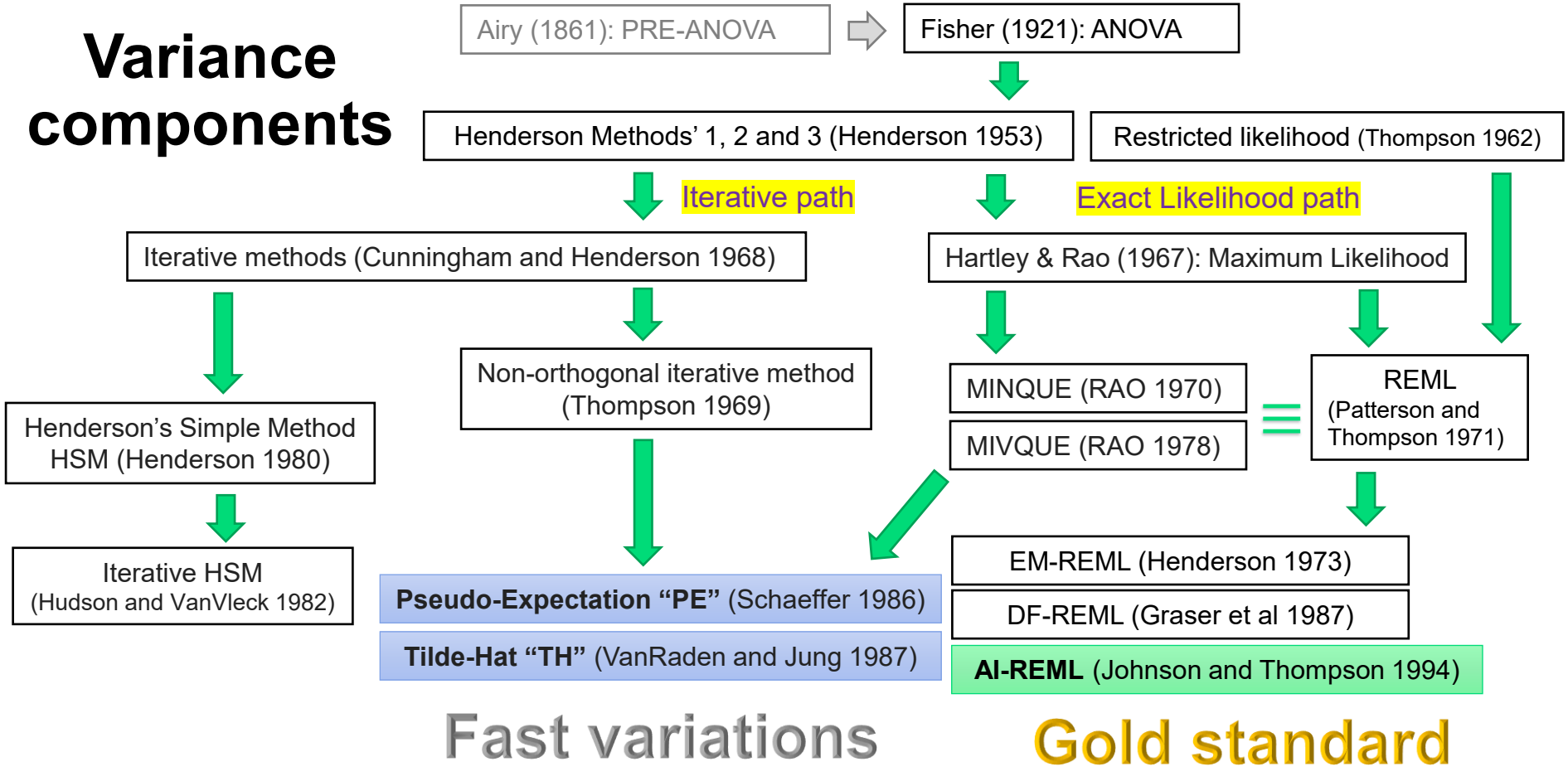
$$\hat{\beta}_j^{(t+1)} = (\hat{\Sigma}_e^{-1(t)} \mathbf{Z}_j' \mathbf{Z}_j + \hat{\Sigma}_\beta^{-1(t)})^{-1} \mathbf{Z}_j' \hat{\Sigma}_e^{-1(t)} (\mathbf{Z}_j \hat{\beta}_j^{(t)} + \hat{e}^{(t)}),$$

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



1. Introduction
2. Coefficients
3. **Variances**
4. Simulations
5. Megavariate
6. Conclusion

Variance components

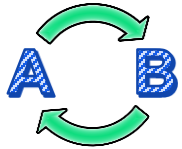


An intuitive derivation for Schaeffer's PE method

The genetic covariance is simply estimated as the cross-prediction between traits A and B

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

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



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

and residual variance

$$\hat{\sigma}_{e(A,B)} = \frac{\mathbf{y}_A'\mathbf{e}_A}{n_A - 1}$$

1. Introduction
2. Coefficients
3. Variances
4. **Simulations**
   - Elapsed times
   - Benchmarks
5. Megavariate
6. Conclusion

Proposed Efficient Solutions

PEGS = **P**seudo **E**xpectation (variances) + **G**auss **S**eidel (coefficients)

Metrics

1. Computation efficiency:

   Elapsed time to fit the model

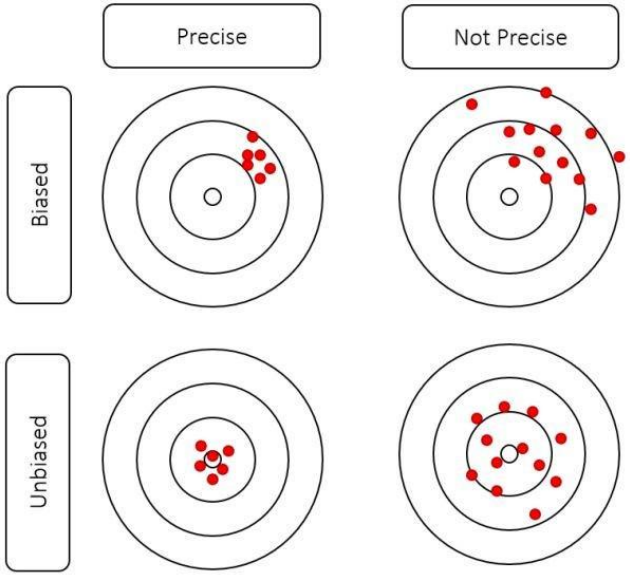
2. Breeding values:

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

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

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

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



Picture source

Datasets

	Small Balanced	Large 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 capable of estimating covariance components** without overlapping individuals

Thus, REML was **not** used in the unbalanced scenario 😓

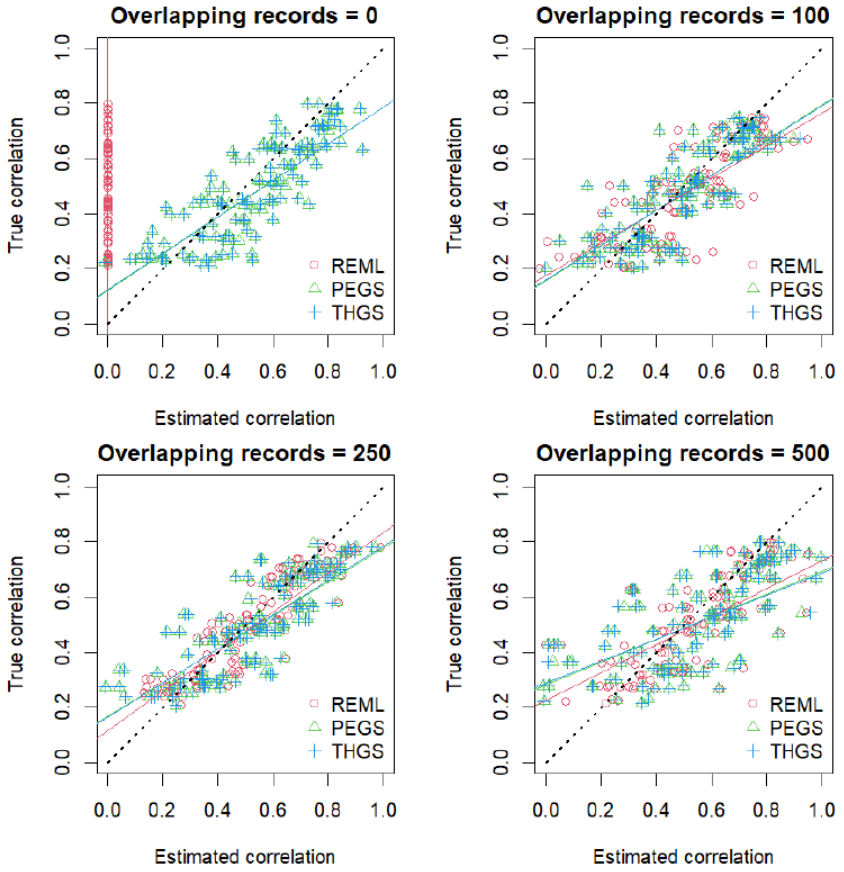


Figure 1: Scatter plot between true and estimated genetic correlations using the soybean dataset with varying number of overlapping individuals across environments.

Elapsed time in small balanced dataset

Method	Time (seconds)
ASREML 4.2	272.6
AIREMLF90	109.8
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

# Traits	Time (minutes)
10	0.2
50	3.5
200	80.5

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

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
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

1. Introduction
2. Coefficients
3. Variances
4. Simulations
5. **Megavariate**
   - Framework
   - Benchmarks
6. Conclusion

Megavariate solvers

Key elements

- Scalable for number of response variables
- Covariance components not estimated explicitly

Statistical framework

- Latent spaces for managing dimensionality
- Tricks: Structural Equations (SEM), Factor analytics (XFA)

Models: MegaLMM (2021), MegaSEM (2024), Canonical Transformation (1980's)

MegaSEM

- Fit each trait as a function of SNPs (genomic info)
- Compute principal components*
- Fit each trait as a function of PCs (genomic+GxE info)

*SVD a complete GEBV matrix with genotypes as rows and locations as columns

MegaSEM

Step 1: univariate by trait

$$y_k = \mu_k + \mathbf{Z}_k\boldsymbol{\beta}_k^{UV} + e_k$$

Step 2: fit and decompose GEBV matrix (single-value decomposition)

$$\hat{\mathbf{G}}^{UV} = \mathbf{Z}\hat{\mathbf{B}}^{UV}$$
$$\hat{\mathbf{G}}^{UV} = \underbrace{\mathbf{UD}}\mathbf{V}'$$
$$= \mathbf{F}\ \mathbf{V}'$$

Step 3: refit each using principal components

$$y_k = \mu_k + \mathbf{F}_k\,\hat{\boldsymbol{\lambda}}_k + e_k$$

Bonus step: recover coefficients

$$\hat{\mathbf{B}}^{SEM} = \hat{\mathbf{B}}^{UV}\mathbf{V}\hat{\boldsymbol{\Lambda}}$$

Runtime benchmark

Table 1 Average runtime in minutes (s.e.) for the balanced experimental design based on 10 simulated replicates. Six scenarios vary in terms of the number of environments and individuals (No. environments / No. individuals). Models are ordered based on computational performance. Standard error shown in parenthesis.

Model	Solver	10 / 500	10 / 2,000	50 / 2,000	200 / 2,000	2,000 / 2,000	200 / 20,000
GREML	REML	46.75 (0.37)	172.61 (17.93)	-	-	-	-
MegaSEM	PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.14 (<0.01)	2.92 (0.02)	5.26 (0.07)
MV (PEGS)	PEGS	<0.01 (<0.01)	<0.1 (<0.01)	0.02 (<0.01)	9.12 (1.62)	97.14 (1.29)	82.22 (5.71)
UV	PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.14 (<0.01)	1.44 (0.01)	5.20 (0.06)

Accuracy benchmark

Table 2 Within environment accuracy for the balanced experimental design based on 10 simulated replicates. Six scenarios vary in terms of the number of environments and individuals (No. environments / No. individuals). Models are ordered based on computational performance. Standard error shown in parenthesis.

Model	Solver	10 / 500	10 / 2,000	50 / 2,000	200 / 2,000	2,000 / 2,000	200 / 20,000
GREML	REML	0.81 (0.03)	0.89 (<0.01)	-	-	-	-
MegaSEM	PEGS	0.79 (0.04)	0.88 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	0.96 (<0.01)
MV	PEGS	0.81 (0.03)	0.89 (<0.01)	0.89 (<0.01)	0.90 (<0.01)	0.88 (<0.01)	0.96 (<0.01)
UV	PEGS	0.78 (0.04)	0.87 (<0.01)	0.87 (<0.01)	0.87 (<0.01)	0.87 (<0.01)	0.95 (<0.01)

2022 Genomes-to-Fields dataset

Table 3 Predictive ability from the 2022 G2F GxE prediction competition. Corn grain yield observed in 4,836 hybrids across 217 locations (2014-2021) predicting 548 hybrids observed across 21 environments (2022). Models are ordered based on the pairwise metric. Standard error shown in parenthesis.

Model	Pairwise	Region	Overall
UVW	0.08 (0.03)	0.22 (0.14)	0.27 (0.11)
MV	0.12 (0.05)	0.27 (0.12)	0.30 (0.11)
MegaSEM	0.13 (0.05)	0.25 (0.15)	0.27 (0.11)
MegaLMM	0.18 (0.06)	0.24 (0.19)	0.27 (0.10)

Model that won the 2022 competition

Rank	Participant team	Mean_RMSE (↓)
1	CLAC	2.328863
2	igorkf	2.345147
3	phenomaize	2.374471
4	UCD_MegaLMM	2.387404
5	CGM	2.390754
6	breedingteam	2.39849
7	Purdue	2.4018
8	SmAL	2.424722
9	ML_APT	2.471617
10	MPB_Group	2.543666

2024 Genomes-to-Fields dataset

Rank	Participant team	Mean_r
Baseline	Model 2022 (not competing)	0.437658
1	PARaBra	0.437097
2	transform(Base)	0.426054
3	The Cornquerors	0.425742
4	GxE4GoodY	0.414302
5	LisbonBio	0.414063
6	fortunehy	0.412884
7	G2Amours(G_et_E)	0.405235
8	UFEED	0.403289
9	Demeter	0.393626
10	ihaveadream(not competing)	0.385322

MegaSEM

1. Introduction
2. Coefficients
3. Variances
4. Simulations
5. **Conclusion**

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

Thank you for your attention!

Final remarks:

1) Multivariate models are more accurate, but traditional methods are not feasible
2) Efficient estimation of coefficients and variances enable large multivariate models
3) Specialized parametrizations enable even larger dimensionalities (***mega***variate)

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com

General recommendation

- REML for balanced sets, small datasets with few traits, or pairwise covariance estimations
- Bayesian Gibbs Sampling when REML is not stable enough (5-20 traits)
- Efficient multivariate solver (PEGS, THGS) for up to 100 traits, mid-to-large datasets
- Megavariate methods (MegaLMM, MegaSEM) for 100+ traits