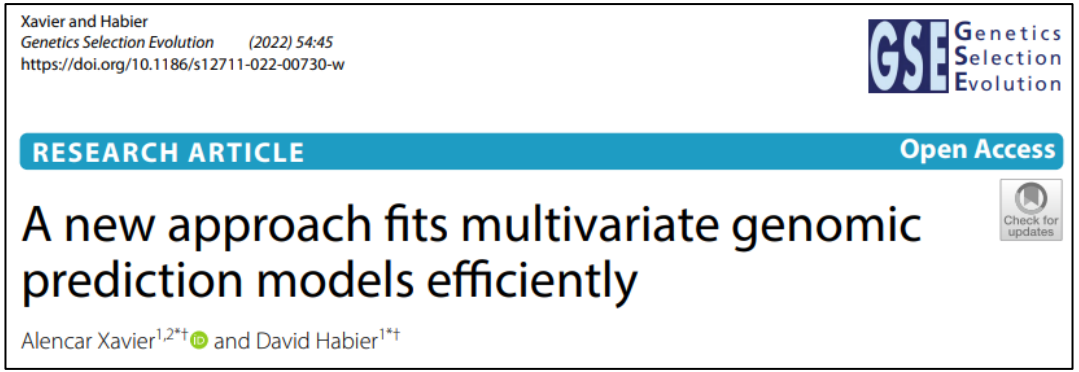
CORTEVA™
agriscience

Scalability Tricks for Genomic Prediction

Alencar Xavier
Sr. Breeding Analyst at Corteva
Adj. Assoc. Professor at Purdue
https://alenxav.github.io/

References



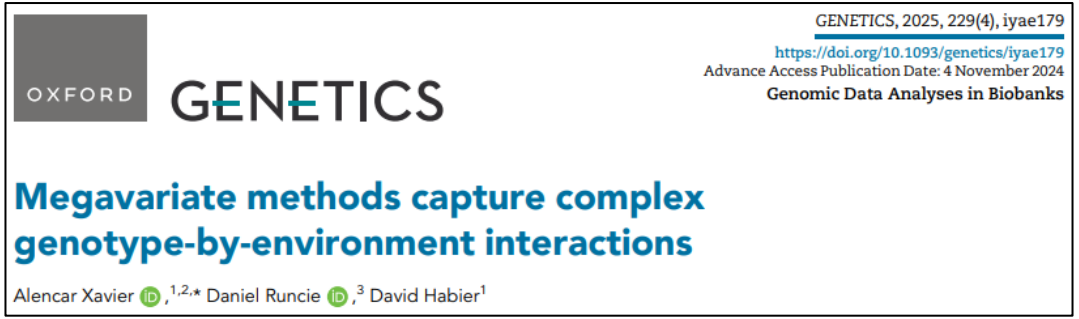
Xavier and Habier
Genetics Selection Evolution (2022) 54:45
https://doi.org/10.1186/s12711-022-00730-w

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*†]



OXFORD GENETICS

GENETICS, 2025, 229(4), iyae179
https://doi.org/10.1093/genetics/iyae179
Advance Access Publication Date: 4 November 2024
Genomic Data Analyses in Biobanks

Megavariate methods capture complex genotype-by-environment interactions

Alencar Xavier[1,2,*] Daniel Runcie[3] David Habier[1]

Rationale

Desirable features for a genomic prediction pipeline:

- Robust → Work well across crops, traits, geographies
- Interpretable → Meaningful, verifiable results
- Scalable → Fast and efficiently

Rationale

Is today's talk relevant?

2022 G2F GXE Prediction 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

Winning approach
MV-PEGS

2024 G2F GXE Prediction Competition

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

Winning approach
MegaSEM

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

How does WGP work?

Regression model: $\mathbf{y} = 1\mu + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

$\mathbf{y}$	=	$\hat{\mu}$	+	$\mathbf{X}$										$\hat{\boldsymbol{\beta}}$	+	$\boldsymbol{\epsilon}$
Pheno					SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10	SNP effect	Residuals
11				Geno1	0	0	0	-1	-1	0	0	0	0	0	0.9	-5
16				Geno2	1	-1	0	-1	-1	1	0	1	0	1	1.7	1
11				Geno3	1	1	1	0	0	1	-1	0	0	0	-3	1
14				Geno4	0	-1	1	0	1	0	0	-1	1	1	3.2	-4
15		15		Geno5	1	-1	-1	1	0	-1	1	0	0	-1	-0.4	-4
14				Geno6	1	0	1	0	-1	-1	0	-1	-1	0	-0.7	-1
5				Geno7	-1	0	0	-1	1	1	1	0	1	1	-3.8	-5
18				Geno8	-1	0	1	0	0	0	-1	1	1	1	1.2	0
12				Geno9	1	1	1	0	1	-1	0	-1	0	-1	0.5	2
12				Geno10	-1	-1	-1	-1	-1	0	-1	-1	-1	0	0.7	0
15				Geno11	1	1	0	0	-1	1	0	-1	0	0		0
8				Geno12	0	-1	-1	0	1	-1	1	0	-1	1		-3
8				Geno13	0	-1	1	0	0	-1	0	1	1	0		-4
16				Geno14	1	-1	1	0	1	0	-1	1	-1	0		3
23				Geno15	0	1	-1	0	1	-1	0	0	-1	-1		3
20				Geno16	-1	1	0	1	1	0	-1	-1	-1	0		3

-1 = AA
0 = Aa
1 = aa

How does WGP work?

Regression model: $\mathbf{y} = 1\mu + \mathbf{X\beta} + \boldsymbol{\epsilon}$

		GEBV ($\hat{\mathbf{g}}$)	
y	**1μ**	**Xb**	**e**
11	15	1	-5
16	15	0	1
11	15	-5	1
14	15	3	-4
15	15	4	-4
14	15	0	-1
5	15	-5	-5
18	15	3	0
12	15	-5	2
12	15	-3	0
15	15	0	0
8	15	-4	-3
8	15	-3	-4
16	15	-2	3
23	15	5	3
20	15	2	3

Key outputs:

1. Genomic Estimated Breeding Values (GEBV)

$$\hat{\mathbf{g}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

2. Heritability

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}$$

where $\sigma_g^2 = \sigma_\beta^2 \sum_{i=1}^{I} 2\, p_i\, (1 - p_i)$

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

How does multivariate WGP work?

Multivariate version: $\{\mathbf{y_1}, \mathbf{y_2}, \ldots\} = \boldsymbol{\mu} + \mathbf{X}\{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots\} + \{\boldsymbol{\epsilon_1}, \boldsymbol{\epsilon_2}, \ldots\}$

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10
Geno1	16	23		21			22	24		18
Geno2	20	23		24			23	25		24
Geno3	25	15		23			22	23		18
Geno4	23	20		19		18	16	23	20	21
Geno5		17		17	16	15	17	22	15	22
Geno6		16	25	25	19	16	23	21	24	25
Geno7		20	23	18	19	21		15	21	25
Geno8	21	18	20	17	21	17		21	23	
Geno9	20	18	17		15	25		15	19	
Geno10	17	17	15		16	23	23	23	19	
Geno11	21		18		20	17	20	20	21	
Geno12	24		24		17	24	18		25	20
Geno13	16		19	24	17		22		25	20
Geno14	20	22	23	17	24		15		15	19
Geno15	24	16	21	19	23		20		20	18

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Geno1	0	3	-1	2	4	2	-1	1	-3	-4
Geno2	-2	0	-4	-1	5	5	-4	-2	0	-1
Geno3	2	3	-4	-5	4	-3	0	1	-1	-1
Geno4	-3	1	0	-1	-5	-2	-3	-3	4	2
Geno5	-2	-3	-5	0	-1	5	2	0	0	-3
Geno6	2	1	-1	5	-4	-1	-3	-3	5	-3
Geno7	-2	-2	-4	2	1	2	0	4	-5	5
Geno8	-2	-3	5	5	2	2	1	5	5	2
Geno9	-1	3	5	4	4	-3	0	2	5	-4
Geno10	3	4	-1	-2	-1	0	-2	-4	1	-1
Geno11	5	-3	-1	5	0	-1	-1	-3	0	-3
Geno12	3	-1	4	-3	-4	0	-2	3	0	1
Geno13	-5	-2	5	-4	-1	-2	4	2	3	-4
Geno14	2	4	3	1	5	-4	3	2	2	-3
Geno15	-1	1	5	-2	0	4	3	5	1	2

Why would multivariate be any better?

Simple (bivariate) model:

COVARIANCES:
INFORMATION GAIN

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

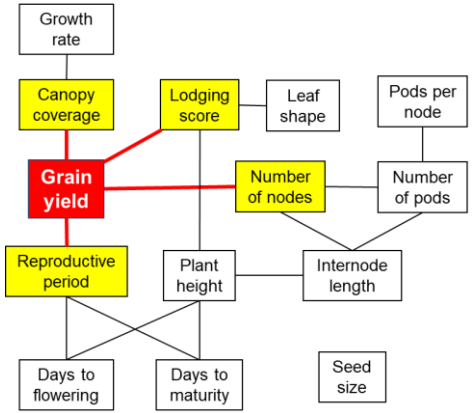
$$Var\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\sigma^2_{g_1} & \sigma_{g_{12}}\\ \sigma_{g_{12}} & \sigma^2_{g_2}\end{bmatrix} + \begin{bmatrix}\sigma^2_{e_1} & \sigma_{e_{12}}\\ \sigma_{e_{12}} & \sigma^2_{e_2}\end{bmatrix}$$

Why would multivariate be any better?

Multivariate problem: What can I do with a genetic correlations?

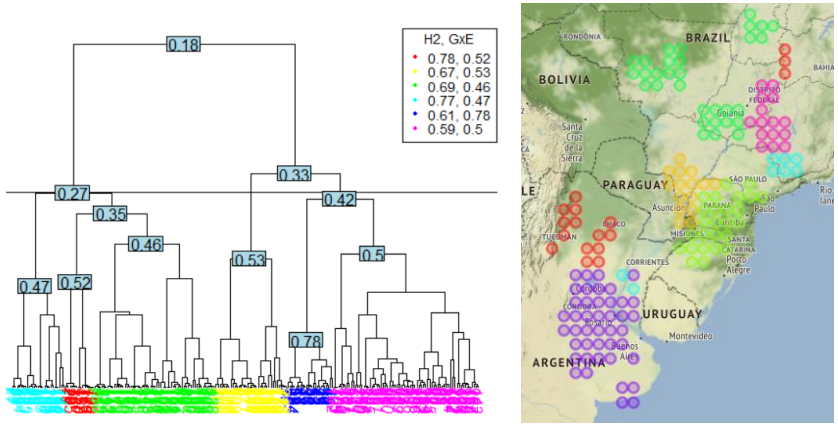
Multiple traits

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



Multivariate system of equations

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

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

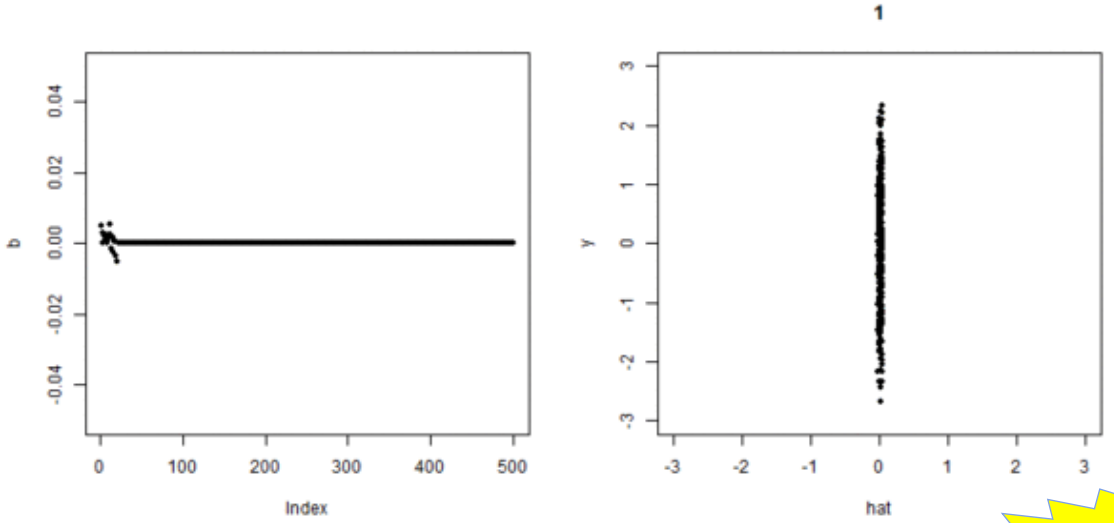
where σ_β^{ij} is the element at position ij of $\mathbf{\Sigma}_\beta^{-1}$. This setup involves storing K times the cross-product or marker scores ($\mathbf{Z}_k'\mathbf{Z}_k$), each with dimension $m \times m$.

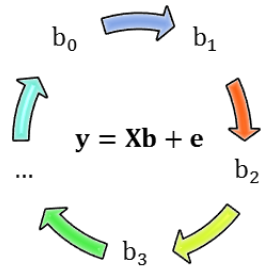
Moreover, this **huge** matrix must be **inverted** for the estimation of covariance components: $\hat{\Sigma}_{\beta(i,j)} = m^{-1}[\hat{\boldsymbol{\beta}}_i'\hat{\boldsymbol{\beta}}_j + \mathrm{tr}(\mathbf{C}^{ij})]$

impossible to compute!!

Regression coefficients

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





$\forall j \in \{1, \ldots, P\}$:

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

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

No equations!!
No Kronecker!!

Variance components

An intuitive idea for the Pseudo-Expectation (PE) method

The genetic (co)variance 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

$$\hat{\sigma}^2{}_{e(A)} = \frac{\mathbf{y}'_A\mathbf{e}_A}{n_A - 1}$$

“Pedigree” of PEGS solver

PE variance components

Cunningham and Henderson 1968
https://doi.org/10.2307/2528457

Thompson 1969
https://doi.org/10.2307/2528574

Schaeffer 1986
https://doi.org/10.3168/jds.S0022-0302(86)80743-3

VanRaden 1988
https://doi.org/10.3168/jds.S0022-0302(88)79541-7

Efficiency, single iteration

Unbiasedness, invariance

Generalized to multiple effects

Equivalence to REML

GS coefficients

Algorithm fitting one feature at a time

Tibshirani 1996
https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Residual update (for SNP-BLUP)

Legarra and Misztal 2008
https://doi.org/10.3168/jds.2007-0403

Randomization for fast convergence

Leventhal and Lewis 2010
https://doi.org/10.1287/moor.1100.0456

Xavier and Habier 2022
https://doi.org/10.1186/s12711-022-00730-w

Combines PE+GS, multivariate generalization

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

Contrasting solvers' properties

	PEGS	REML	GIBBS
Solution (σ_i^2)	$\frac{\tilde{\mathbf{u}}_i'\hat{\mathbf{u}}_i}{\mathrm{Tr}(\mathbf{Z}_i'\mathbf{S}\mathbf{Z}_i)}$	$\frac{\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i + \mathbf{C}^{ii}}{q_i}$	$\frac{(\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i + S_0\nu_0)}{\chi^2(q_i + \nu_0)}$
Unbiased, Invariant	✔	✔	✘
Provide S.E.	✘	✔	✔
Numerical stability	✔	✘	✔

Contrasting methods' runtime

Runtime benchmark with small, real dataset

Implementation	Time (in seconds)
GIBBS3F90	559.8
ASREML 4.2	272.6
AIREMLF90	109.8
PEGS	**0.27**

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

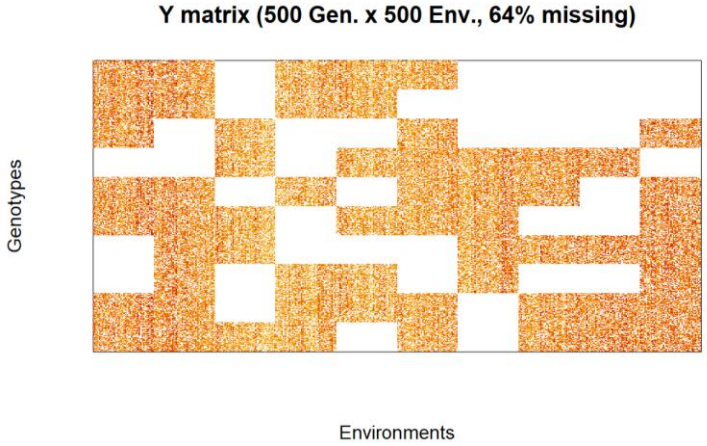
Megavariate models

What is a megavariate model?

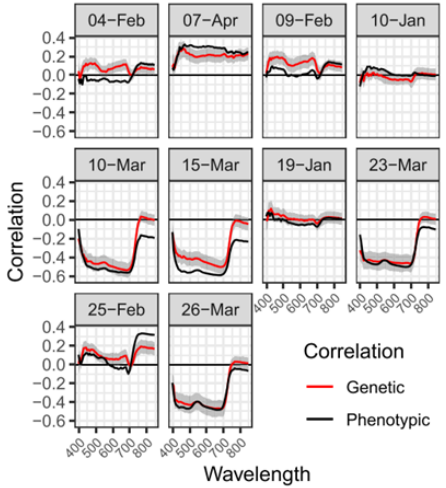
Model:

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

- Large number of environments
- High-throughput phenotyping

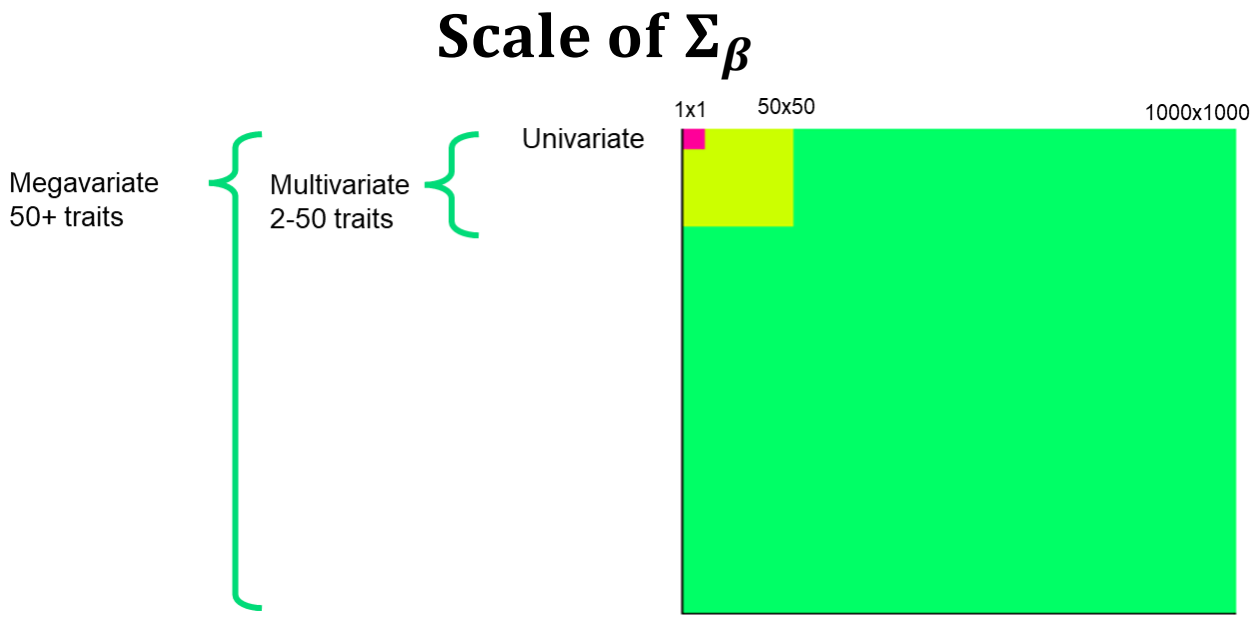


Simulated example of sparse multi-environmental trials



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

Megavariate models



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

Megavariate solvers

Key elements

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

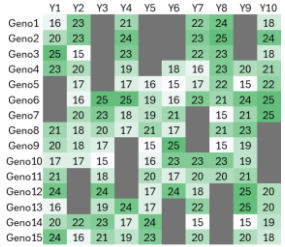
Statistical framework

- Latent spaces for managing dimensionality
- Tricks: Structural equations (SEM), latent spaces (LS/PC), and factor analytics (XFA)

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

How does MegaSEM work?

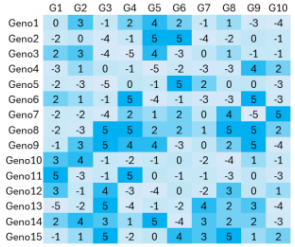
Step 1



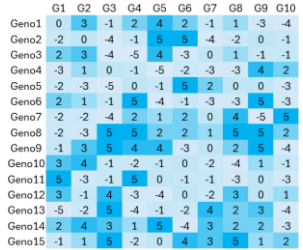
For all traits, fit one at a time

$$\mathbf{y} = 1\mu + \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$$

$$\mathbf{G}_0 = \mathbf{X}\boldsymbol{\beta}_0$$

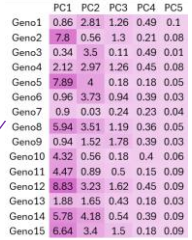


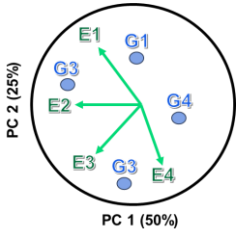
Step 2



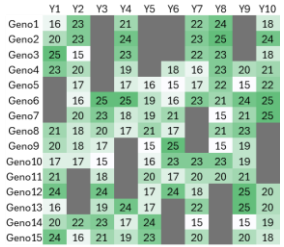
Get PCs ($\mathbf{Q}$) of GEBV matrix

$$\mathbf{G}_0 = \mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{F}\mathbf{V}'$$



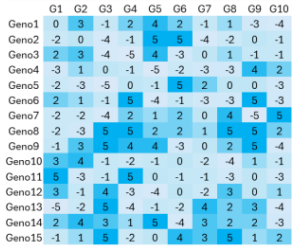


Step 3



For all traits, fit one at a time

$$\mathbf{y} = 1\mu + \mathbf{F}\boldsymbol{\lambda} + \boldsymbol{\epsilon}$$



Final estimators:

$$\boldsymbol{\beta}_{\mathbf{SEM}} = \boldsymbol{\beta}_0\mathbf{V}_0'\boldsymbol{\Lambda}$$

GenCor = $\mathrm{cor}(\boldsymbol{\beta}_{\mathbf{SEM}})$

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

Contrasting methods' runtime

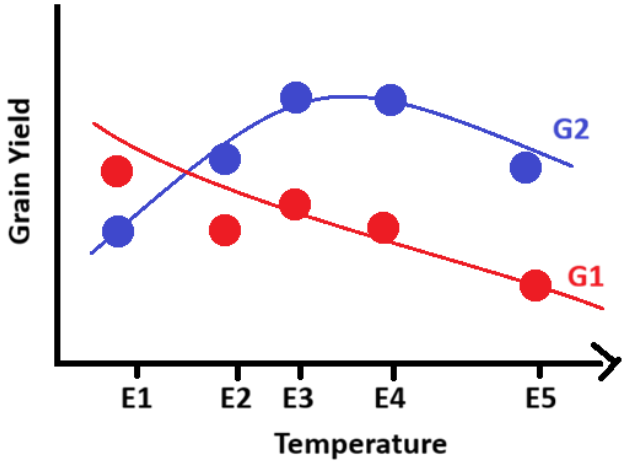
Runtime benchmark with simulated dataset

Scenario	# Envir.	# Geno.	MV REML	CT-MV REML	MV PEGS	MegaSEM PEGS	UV PEGS
1	10	500	46.75	0.06	**<0.01**	**<0.01**	<0.01
2	10	2 000	172.61	0.19	**<0.01**	**0.01**	0.01
3	50	2 000	-	4.38	**0.02**	**0.04**	0.04
4	200	2 000	-	-	9.12	**0.14**	0.14
5	2 000	2 000	-	-	97.14	**2.92**	1.44
6	200	20 000	-	-	82.22	**5.26**	5.20

Minutes to fit the model. Average across 10 runs.

What about reaction norm (GxE) modeling?

- One can parameterize GxE with environmental covariate (EC)
- How do reaction norms look like?
  1. Models with interaction terms
  2. Analyze ***post hoc***



MIXED MODEL INTERACTION TERMS

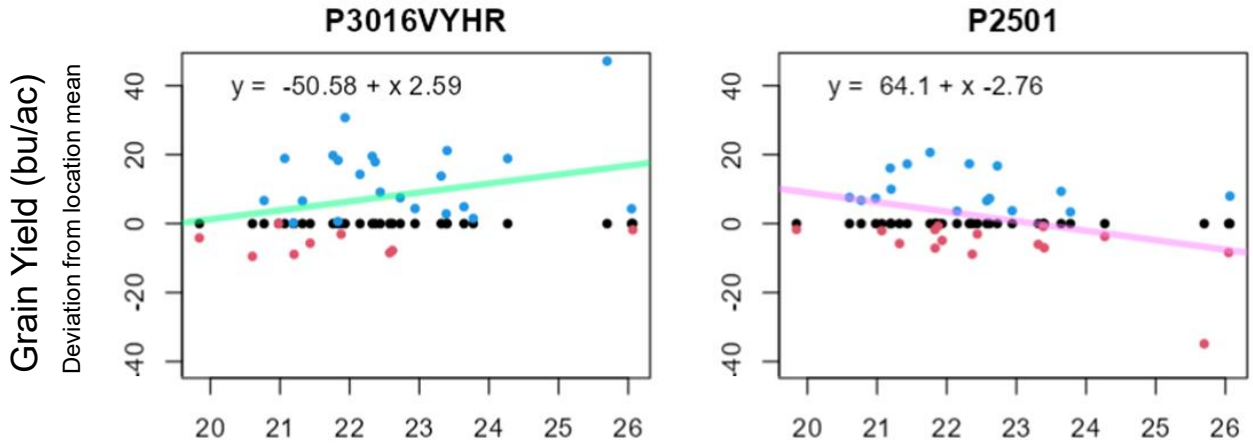
Approach	Parametrization	Dimension
Kernelized interactions	$\mathbf{g}_{\mathrm{G\times E}} \sim N(0, [\mathbf{XX}' \odot \boldsymbol{WW}']\sigma^2_{G\times E})$	$n \times n$
Explicit interaction	$\mathbf{g}_{\mathrm{G\times E}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \mathrm{x_i w_j \beta_{ij}}$	
Random regression	$\mathbf{g}_{\mathrm{G\times E}} = \sum_{j=1}^{J} \mathbf{Z_j g_j}$, $\mathbf{g} \sim \mathrm{N}(0, \boldsymbol{\Sigma}_{\mathrm{G\times E}} \otimes \mathbf{G})$	

Crazy expensive !!!

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

Reaction norms
POST HOC ANALYSIS

- Summary of location-level GEBVs from multivariate analysis
- At it simplest, **Finlay-Wilkinson**:



Reaction norms

POST HOC ANALYSIS

Multiple features? Conditional expectation

	Observed data	Prediction target
Population of Genetics	**A**	**B**
Population of Environments	**X**	**Z**

Model is trained on AX → Target is to predict BZ

$$\widehat{\mathbf{G}}_{\mathrm{BZ|AX}} = \boldsymbol{\Sigma}_{\mathrm{ZX}} \widehat{\boldsymbol{\Sigma}}_{\mathrm{AX}}^{-1} (\mathbf{X}_{\mathrm{B}} \widehat{\mathbf{B}}_{\mathrm{AX}})$$

We use the estimates from the multivariate model that fits the AX dataset ($\widehat{\mathbf{B}}_{\mathrm{AX}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{AX}}$)

Reaction norms

POST HOC ANALYSIS

$$\widehat{\mathbf{G}}_{\mathrm{BZ|AX}} = \boldsymbol{\Sigma}_{\mathrm{ZX}}\widehat{\boldsymbol{\Sigma}}_{\mathrm{AX}}^{-1}(\mathbf{X}_{\mathrm{B}}\widehat{\mathbf{B}}_{\mathrm{AX}})$$

where do we get the covariance between observed ($\mathbf{X}$) and unobserved ($\mathbf{Z}$) environments ($\boldsymbol{\Sigma}_{\mathrm{ZX}}$)??

Step 1. Get PCs of $\boldsymbol{\Sigma}_{\mathrm{X}}$ → $\boldsymbol{\Sigma}_{\mathrm{X}} = \mathbf{Q}_{\mathrm{X}}\mathbf{Q}'_{\mathrm{X}}$

Step 2. Model PCs of X → $\mathbf{Q}_{\mathrm{X}} = \mathbf{W}_{\mathrm{X}}\boldsymbol{\Lambda} +$

Step 3. Predict PCs of Z → $\widehat{\mathbf{Q}}_{\mathrm{Z}} = \mathbf{W}_{\mathrm{Z}}\widehat{\boldsymbol{\Lambda}}$

Step 4. Build covariance → $\widehat{\boldsymbol{\Sigma}}_{\mathrm{ZX}} = \widehat{\mathbf{Q}}_{\mathrm{Z}}\mathbf{Q}'_{\mathrm{X}}$

Thank you for your attention!

Final remarks:

1) Efficient computation depends on the model, parametrization, and solver
2) We presented a new solver and a megavariate model
3) We introduced a post-hoc (inexpensive) way to model GxE

Questions??

Alencar Xavier

Alencar.Xavier@Corteva.com