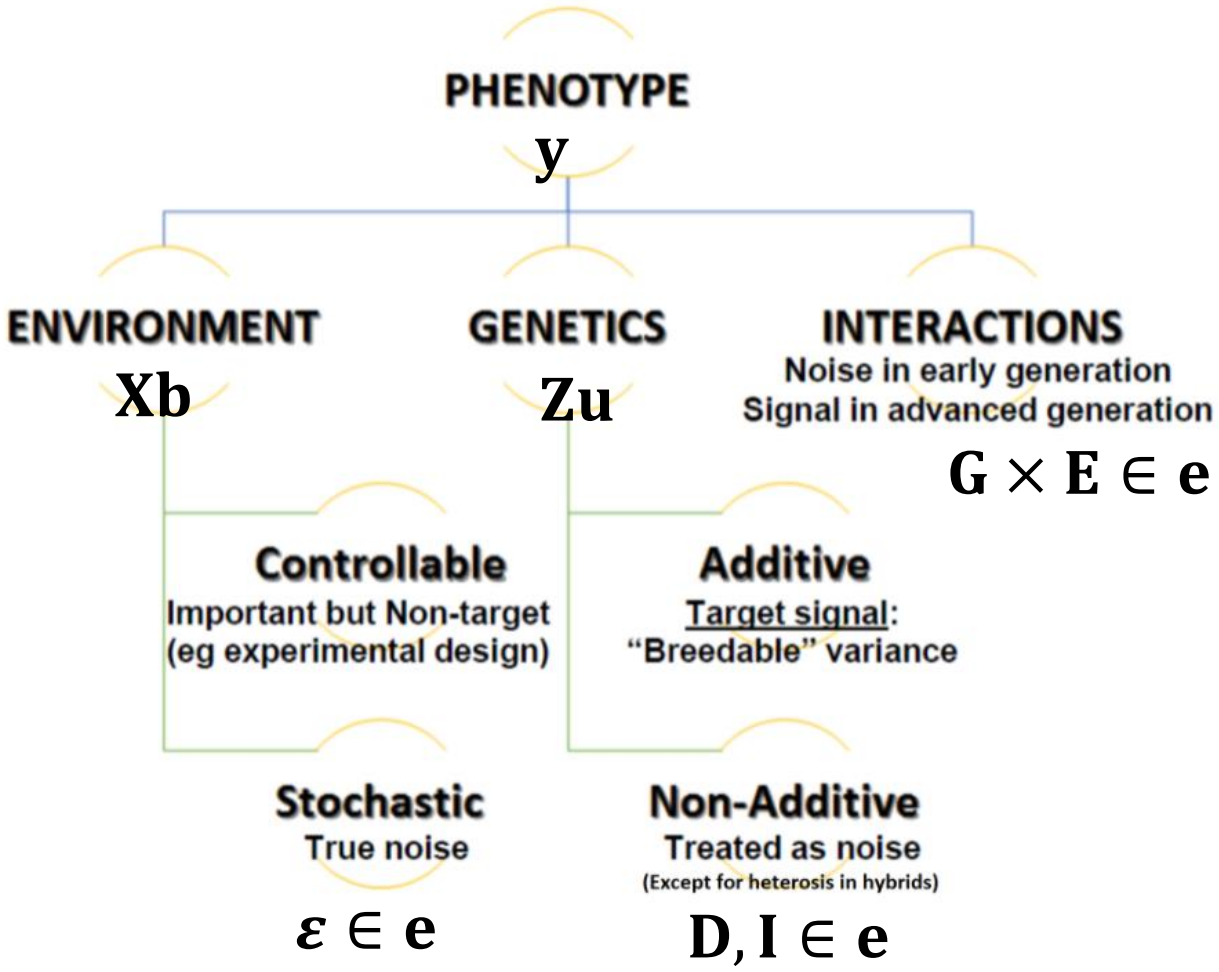
Selection & Breeding Analytics

A discussion on TPE/TPG, multiple traits, Metrics

AX10042021

A simple model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$



Model notation

$$y = Xb + Zu + e$$

$$\begin{cases} y \sim N(Xb, V) \\ y \sim N(Xb, ZAZ'\sigma_a^2 + I\sigma_e^2) \end{cases}$$

$$\boldsymbol{u \sim N(0, A\sigma_a^2)}$$

$$\boldsymbol{e \sim N(0, I\sigma_e^2)}$$

$$\boldsymbol{cov(Zu, e) = 0}$$

n = number of observations
p = number of parameters
q = number of individuals

y = vector of observations (n)
X = design matrix of fixed effects (n × p)
b = vector of fixed effect coefficients (p)
Z = incidence matrix of random effects (n × q)
u = vec. of random effects − genetics values (q)
e = vector of residuals (n)
σ_a^2 = random effect variance (1)
σ_e^2 = residual variance (1)
A = random effect correlation matrix (q × q)
R = residual correlation matrix (n × n)
$\lambda = \sigma_e^2 : \sigma_a^2$ = regularization parameter (1)

Simple breeding model using relationship

- Simplest model - Additive
  - $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$
    - $\mathbf{u} \sim \mathrm{N}(\mathrm{O}, \mathbf{G}),\ \ \mathbf{G} = \mathbf{A}\sigma_a^2$
    - $\mathbf{e} \sim \mathrm{N}(\mathrm{O}, \mathbf{R}),\ \ \mathbf{R} = \mathbf{I}\sigma_e^2$ (homoscedastic)
  - $\mathbf{y}$ = Pheno; $\mathbf{Xb}$ = Fixed Env. effect; $\mathbf{Zu}$= Genetics; $\mathbf{e}$=residuals

Equivalently, GCA model for hybrid crops

$$y = Xb + Z_1u_1 + Z_2u_2 + e$$
$$u_1 \sim N(O, A_1\sigma_{a_2}^2)$$
$$u_2 \sim N(O, A_2\sigma_{a_2}^2)$$
$$e \sim N(O, I\sigma_e^2)$$

- Phenotypic variance: $\mathbf{Var(y)} = \mathbf{ZGZ'} + \mathbf{R}$
- Genetic variance: $\mathbf{V_G} = \mathbf{ZGZ'}$

Example from **Cunningham & Henderson 1968**

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Xa} + \mathbf{Zb} + \mathbf{e}.$$

DATA AND INCIDENCE MATRICES

y	μ	a_1	a_2	b_1	b_2	b_3
3	1	1	0	1	0	0
2	1	1	0	0	1	0
3	1	1	0	0	0	1
2	1	1	0	1	0	0
3	1	1	0	0	1	0
5	1	1	0	0	1	0
6	1	1	0	0	1	0
7	1	1	0	0	1	0
2	1	0	1	1	0	0
8	1	0	1	0	1	0
4	1	0	1	0	0	1
3	1	0	1	1	0	0
8	1	0	1	0	1	0
4	1	0	1	0	0	1
9	1	0	1	0	1	0
3	1	0	1	0	0	1
2	1	0	1	0	0	1
5	1	0	1	0	0	1

$$\mathbf{y} = \mathbf{Xa} + \mathbf{Zb} + \mathbf{e}$$

The least squares equations (ignoring $\boldsymbol{\mu}$) are

$$\left[\begin{array}{cc|ccc} 8 & 0 & 2 & 5 & 1 \\ 0 & 10 & 2 & 3 & 5 \\ \hline 2 & 2 & 4 & 0 & 0 \\ 5 & 3 & 0 & 8 & 0 \\ 1 & 5 & 0 & 0 & 6 \end{array}\right] \left[\begin{array}{c} a_1 \\ a_2 \\ \hline b_1 \\ b_2 \\ b_3 \end{array}\right] = \left[\begin{array}{c} 31 \\ 48 \\ \hline 10 \\ 48 \\ 21 \end{array}\right]$$

In algebraic terms, these equations are

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

$$\lambda = 0.5721$$

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{I}k \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

$$\left[\begin{array}{cc|ccc} 8 & 0 & 2 & 5 & 1 \\ 0 & 10 & 2 & 3 & 5 \\ \hline 2 & 2 & 4.5721 & 0 & 0 \\ 5 & 3 & 0 & 8.5721 & 0 \\ 1 & 5 & 0 & 0 & 6.5721 \end{array}\right] \left[\begin{array}{c} a_1 \\ a_2 \\ \hline b_1 \\ b_2 \\ b_3 \end{array}\right] = \left[\begin{array}{c} 31 \\ 48 \\ \hline 10 \\ 48 \\ 21 \end{array}\right]$$

```
> solve(C, g)
 [1] 2.9371 4.8684 -1.2272 2.1826 -0.9554
```

a1 a2 b1 b2 b3

Example from **Robinson 1991**

Data

Herd	Sire	Yield
1	A	110
1	D	100
2	B	110
2	D	100
2	D	100
3	C	110
3	C	110
3	D	100
3	D	100

Design matrices

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \quad Z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Solving

$$\left(\begin{array}{ccc|cccc} 2 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 4 & 0 & 0 & 2 & 2 \\ \hline 1 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 12 & 0 \\ 1 & 2 & 2 & 0 & 0 & 0 & 15 \end{array}\right) \begin{pmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \hat{h}_3 \\ \hat{s}_A \\ \hat{s}_B \\ \hat{s}_C \\ \hat{s}_D \end{pmatrix} = \begin{pmatrix} 210 \\ 310 \\ 420 \\ 110 \\ 110 \\ 220 \\ 500 \end{pmatrix}$$

which has solution

$$\hat{\beta} = (105.64, 104.28, 105.46)^T, \qquad (1.4)$$
$$\hat{u} = (0.40, 0.52, 0.76, -1.67)^T.$$

MME

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{Z'X} \\ \mathbf{X'Z} & \mathbf{Z'Z} + \lambda \mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

$$(\lambda = \sigma_e^2 / \sigma_a^2)$$

- Linear model: $y = Xb + Zu + e$
- Genetic variance: $V(u) = G = A\sigma_a^2$
- Residual variance: $V(e) = R = I\sigma_e^2$
- Henderson's equation ($Cg = r$)

$$\begin{bmatrix} X'R^{-1}X & Z'R^{-1}X \\ X'R^{-1}Z & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

- We know (data): $x = \{y, X, Z, A\}$
- We want (parameters): $\theta = \{b, u, \sigma_a^2, \sigma_e^2\}$
- Parameter estimation based on Gaussian likelihood: $L(x|\theta)$

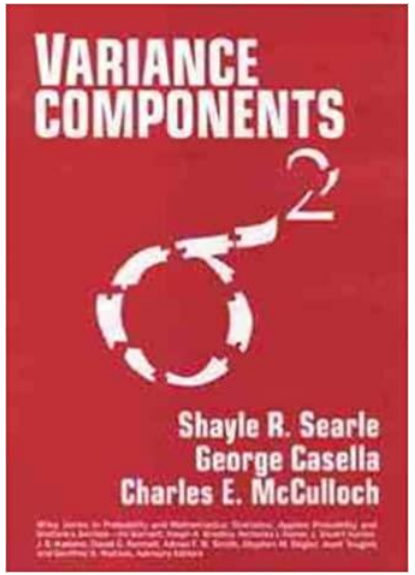
Note

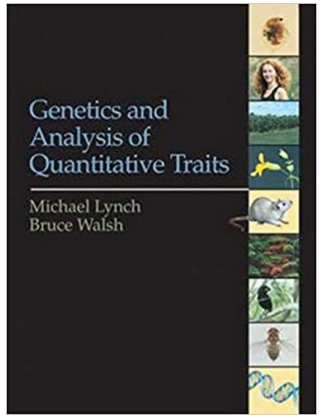
$$\mathrm{E}(\mathbf{y}) = \mathbf{Xb}$$
$$\mathrm{E}(\mathbf{y}|\mathbf{u}) = \mathbf{Xb} + \mathbf{Zu}$$
$$\mathrm{V}(\hat{\mathbf{b}}) = \mathbf{C}^{11}$$
$$\mathrm{E}(\mathbf{g}) = 0$$
$$\mathrm{V}(\mathbf{g}) = \mathrm{G}$$
$$\mathrm{E}(\hat{\mathbf{g}}) = \mathbf{g}$$
$$\mathrm{V}(\hat{\mathbf{g}}) = \mathbf{G} - \mathbf{C}^{22}$$

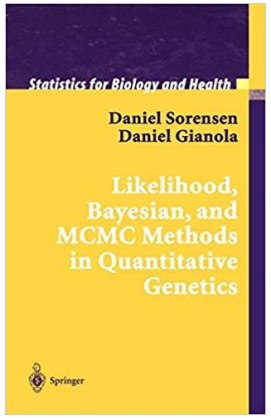
Solutions

- **ANOVA**? Only for perfectly balanced experiments without relationship
- Common solution (EM-REML)
  - $\hat{\sigma}_u^2 = \frac{\hat{u}'A^{-1}\hat{u}+C^{22}}{q}$
  - $\hat{\sigma}_e^2 = \frac{y'e}{n-r}$
    - q number of levels of random effect = No. of individuals
    - C^{22} comes from inverting LHS (aka. C), get block of random term
    - r number of levels of fixed effect(s) / rank of X
    - n number of observations
- More generally, the first derivative solution comes from
  - $\hat{\sigma}_i^2 = \frac{\mathbf{yPV_iPy}}{\mathbf{tr(PV_i)}} = \frac{(\mathbf{y-X\hat{b}})'\mathbf{V^{-1}V_iV^{-1}(y-X\hat{b})}}{\mathbf{tr(PV_i)}}$

$$\boxed{V_i = \frac{\partial V}{\partial \sigma_i^2} = Z_i G_i Z_i'}$$







Multiple traits/environments

$$y = \{y_1, y_2, \ldots, y_k\}$$

With multiple traits, the relation among traits is modeled

$$V(u) = A \otimes \Sigma_a = \begin{bmatrix} A\sigma^2_{a_1} & A\sigma_{a_{12}} \\ A\sigma_{a_{21}} & A\sigma^2_{a_2} \end{bmatrix}$$

$$V(e) = I \otimes \Sigma_e = \begin{bmatrix} I\sigma^2_{e_1} & I\sigma_{e_1 e_2} \\ I\sigma_{e_2 e_1} & I\sigma^2_{e_2} \end{bmatrix}$$

Why does it matter? Covariances ($\sigma_{a_{12}}$, $\sigma_{e_{12}}$) are extra information!!

Selection index (Smith-Hazel)

- Historically
  - H^2 ?
  - $index = \boldsymbol{V_g V_y^{-1}} \alpha = \Sigma_a (\Sigma_a + \Sigma_e)^{-1} \alpha$
  - $\alpha = economic\ value$
  - Estimated from multi-variate models:
    - $\Sigma_a = genetic\ covariance\ among\ traits\ (k \times k)$
    - $\Sigma_e = residual\ covariance\ among\ traits\ (k \times k)$

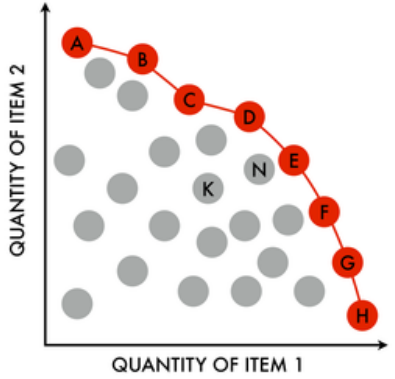
- It also is possible to use the matrix version of V_g and V_y for a more information-rich index

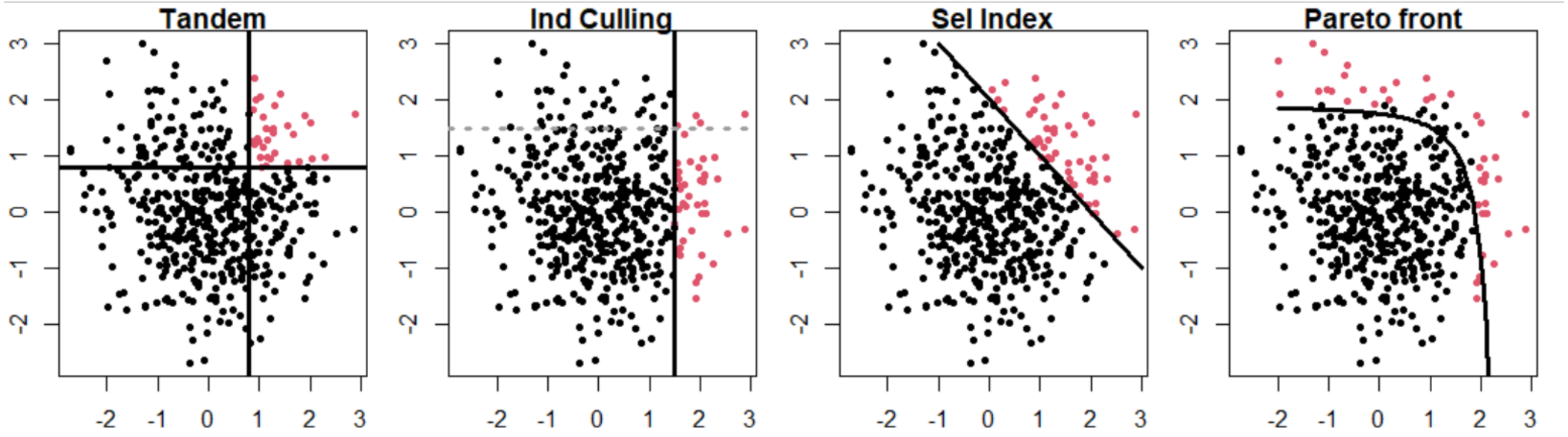
Are selection index and BLUPs the same??

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{u} + \mathbf{e}$$
$$\mathbf{u} = \mathbf{V_g V_y^{-1}} (\mathbf{y} - \boldsymbol{\mu})$$
$$\mathbf{u} = \mathbf{H^2} \times \mathbf{BLUE}$$

Alternative to selection index

- Tandem selection
- Independent culling
- **<u>Multi-objective selection (MOOB)</u>**
  - See https://www.nature.com/articles/s41437-018-0147-1





Covariance Σ among traits/environments

- Unstructured $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$ $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \mathrm{G} \times \mathrm{E}$ correlation between I and J

- Compound symmetry $\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_2 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_2 & \sigma_3^2 \end{bmatrix}$

  $\rho =$ same $\mathrm{G} \times \mathrm{E}$ correlation for all pairs of environments

- XFA - $\Sigma_{un} = \mathrm{UDU}' \rightarrow \Sigma_{XFA} = \mathrm{U}_* \mathrm{D}_* \mathrm{U}_*'$, where *= less PCs

- Diagonal (ignoring associations) $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$

  no $\mathrm{G} \times \mathrm{E}$ correlation

Key metrics

- **Heritability** (plot level): $\mathbf{H_p = V_y^{-1} V_G}$
- **Heritability** (entry level): $\mathbf{H_e = V_u V_{\hat{u}}^{-1} = G\left(G - C^{22}\right)^{-1} = \left(I - C^{22} G^{-1}\right)^{-1}}$
- **Accuracy**: $\mathbf{a} = \mathrm{cor}(u, \hat{u}) = \dfrac{\mathbf{cov}(\hat{u}, u)}{\sqrt{\mathbf{var}(\hat{u})\mathbf{var}(u)}} = \sqrt{\dfrac{\mathbf{GZ'V^{-1}ZG}}{\mathbf{G}}}$
- **Reliability**: $\mathbf{r} = \sqrt{\mathrm{diag}(\mathbf{H_e})}$

Heritability on balanced populations without relationship:

$$\frac{\sigma_a^2}{\sigma_a^2 + n^{-1}\sigma_e^2}$$

Accuracy on (observed) balanced population without relationship:

$$\sqrt{\frac{\sigma_a^2}{\sigma_a^2 + n^{-1}\sigma_e^2}}$$

Reliability of observed individuals from population without relationship:

$$\sqrt{\frac{\sigma_a^2}{\sigma_a^2 + n_i^{-1}\sigma_e^2}}$$

Key metrics

- Heritability
  - Used for
    - Direct measure of genetic control
    - Assess statistical models, experimental designs
- Accuracy
  - Used for
    - Check how well we can predict something
    - Optimize TPE/TPG, experimental designs, training sets
    - Response to selection ($R \propto i \times r_{g,\hat{g}} \times \sigma_a$)
- Reliability
  - Used for
    - Direct measure of confidence
    - Deregression = Unshkring BLUPs for GWAS and multistage analysis
      - Mitigate Bulmer effect (changes in relative ranking)

TPE/TPG

- Target population of environments (TPE)
  - Influences accuracies via GxE correlation
  - Which environments should I be able to predict?

- Target population of genotypes (TPG)
  - Influences accuracies via genetic relationship
    - Connected to the size and quality of estimation set
  - Which genetics should I be able to predict?

From QTLs to Adaptation Landscapes: Using Genotype-To-Phenotype Models to Characterize G×E Over Time

Daniela Bustos-Korts[1], Marcos Malosetti[1], Karine Chenu[2], Scott Chapman[3,4], Martin P. Boer[1], Bangyou Zheng[3] and Fred A. van Eeuwijk[1]**

What Should Students in Plant Breeding Know About the Statistical Aspects of Genotype × Environment Interactions?

Fred A. van Eeuwijk,* Daniela V. Bustos-Korts, and Marcos Malosetti

An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments FREE

Yvonne C J Wientjes ✉, Piter Bijma, Roel F Veerkamp, Mario P L Calus

Genetics, Volume 202, Issue 2, 1 February 2016, Pages 799–823,
https://doi.org/10.1534/genetics.115.183269

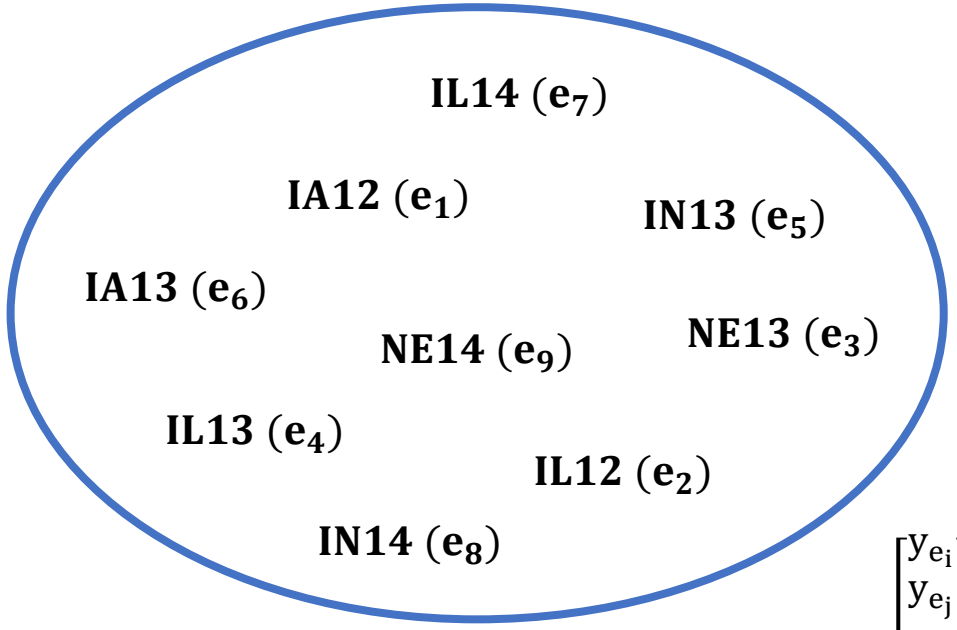
Multiple environments

- Any given breeding trial happens in each environment that is sample of many environments:

$$e_i \in E$$

That is:

$$\mathbf{TPE}\ (\mathbf{E}) =$$



$$\begin{bmatrix} y_{e_i} \\ y_{e_j} \\ g_E \end{bmatrix} = \begin{bmatrix} \sigma^2_{g(e_i)} + \sigma^2_{\epsilon(e_i)} & \sigma_{g(e_i,e_j)} & \sigma_{g(e_i,E)} \\ \sigma_{g(e_j,e_i)} & \sigma^2_{g(e_j)} + \sigma^2_{\epsilon(e_j)} & \sigma_{g(e_j,E)} \\ \sigma_{g(E,e_i)} & \sigma_{g(E,e_j)} & \sigma^2_{g(E)} \end{bmatrix}$$

TPE can be data-driven