A brief introduction to mixed models

Alencar Xavier

Quantitative Geneticist, DowDuPont

http://alenxav.wix.com/home

Outline

- **Part 1 – Concepts**
  - Mixed models in plant breeding
  - Fixed and Random terms
  - History of mixed models
  - Model notation
  - Henderson's equation
  - Variance decomposition
- **Part 2 – Applications**
  - Phenotypic selection
  - Practical example
    - Cunningham & Henderson (1986)
    - Robinson (1991)
    - Xavier et al. (2016)
  - Models with genomic data
    - G-BLUP and RR-BLUP
    - GWAS
- **References**

PART I - Concepts

Mixed models in plant breeding

- Variance components and heritability
- Genetic correlations
- Estimation of genetic values
- Estimation of breeding values
- Prediction of breeding values (unphenotyped material)
- Selection indexes
- Association analysis

Fixed and Random terms

- **Fixed effect**
  - Assumed to be invariable (you can't recollect the data)
  - Inferences are made upon the parameters
  - Results can not be extrapolated to other datasets
  - Example: Overall mean and environmental effects

- **Random effects**
  - You may not have all the levels available
  - Inference are made on variance components
  - Prior assumption – coefficients are normally distributed
  - Results can not be extrapolated to other datasets
  - Regularized (shrinkage)
  - Example: Genetic effects

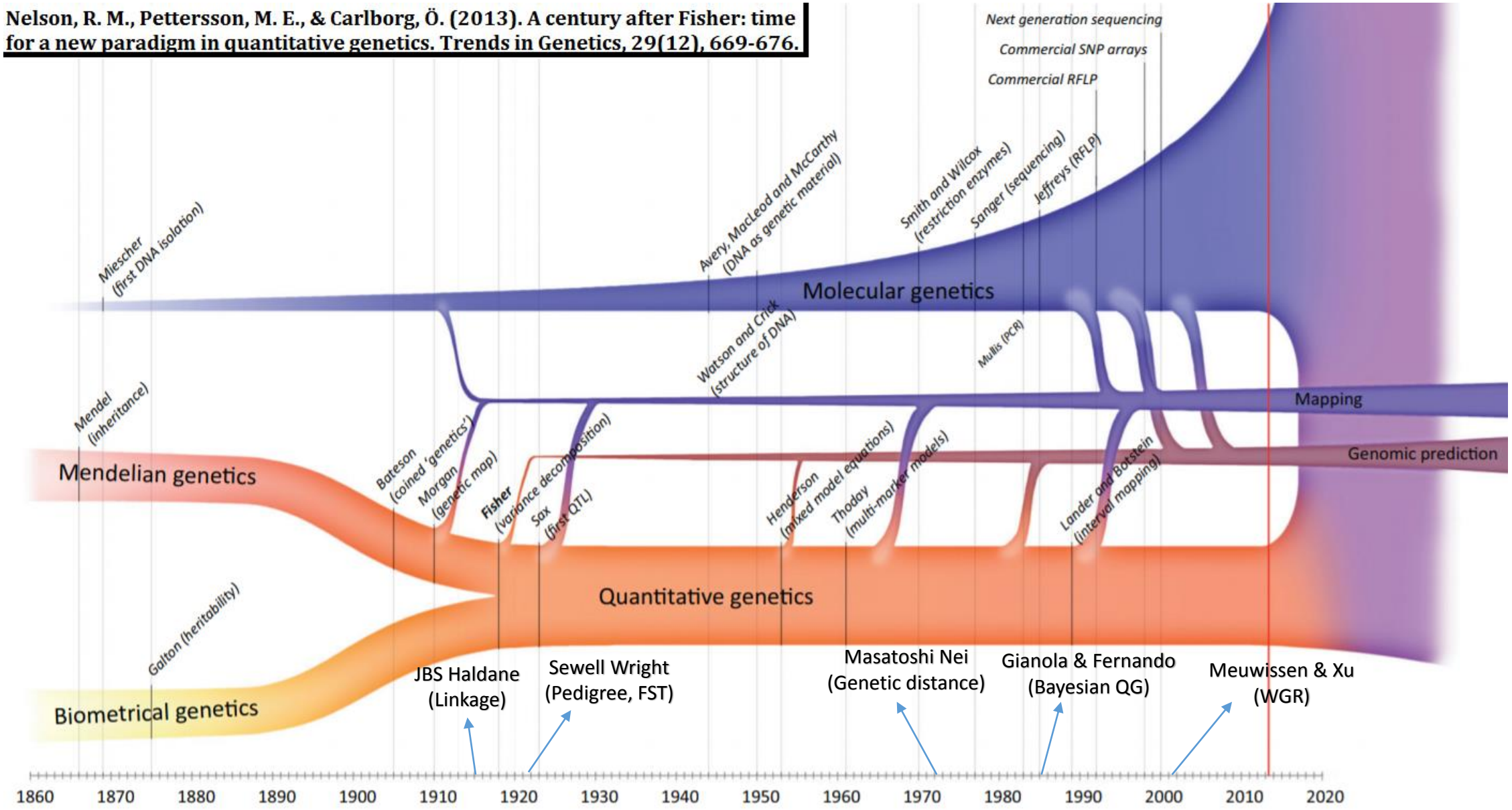
Robinson (1991) reciting Searle

> . . . when inferences are going to be confined to the effects in the model the effects are considered fixed; and when inferences will be made about a population of effects from which those in the data are considered to be a random sample then the effects are considered as random.

History of mixed models

- Francis Galton
  - **1988 –** Regression and h^2
- Ronald Fisher
  - **1918 -** Infinitesimal model (P = G + E)
- Sewall Wright
  - **1922 -** Pedigree matrix (A)
- Charles Henderson
  - **1950 -** BLUP ($u \sim A\sigma_a^2$)

Nelson, R. M., Pettersson, M. E., & Carlborg, Ö. (2013). A century after Fisher: time for a new paradigm in quantitative genetics. Trends in Genetics, 29(12), 669-676.

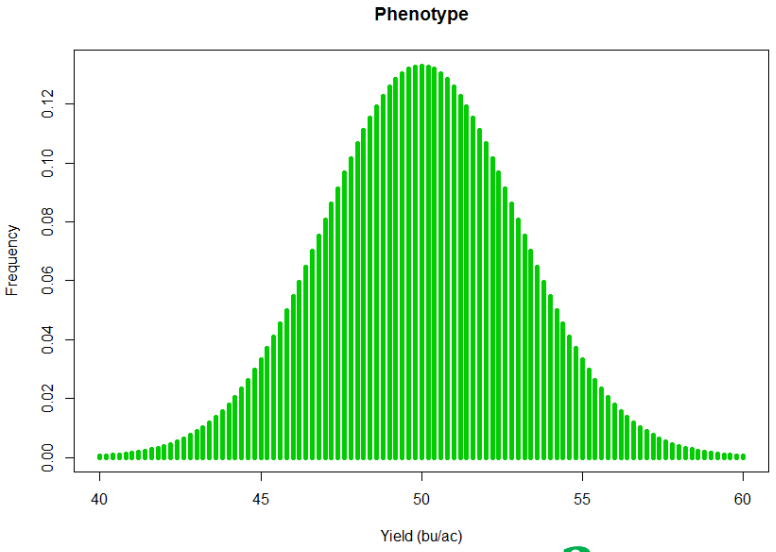


Model notation

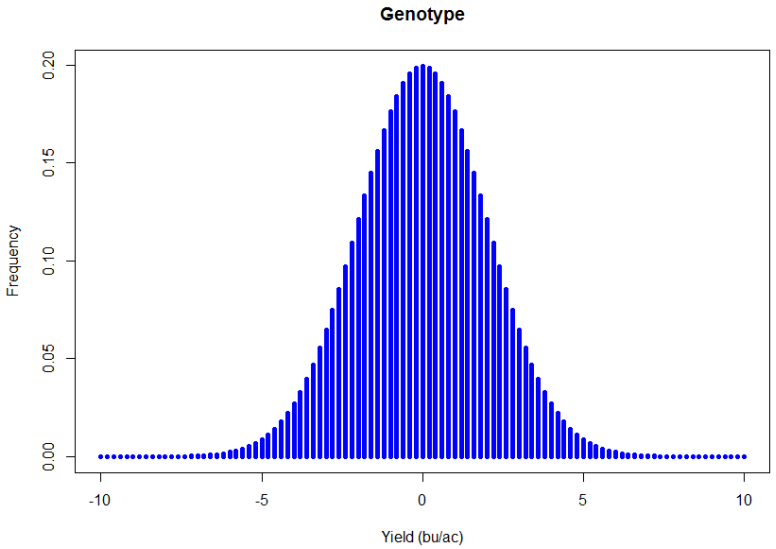
$$\mu = 50$$
$$\sigma_y^2 = 9$$
$$\sigma_g^2 = 4$$
$$\sigma_e^2 = 5$$
$$cov(g, e) = 0$$

$$y = 1\mu + g + e$$

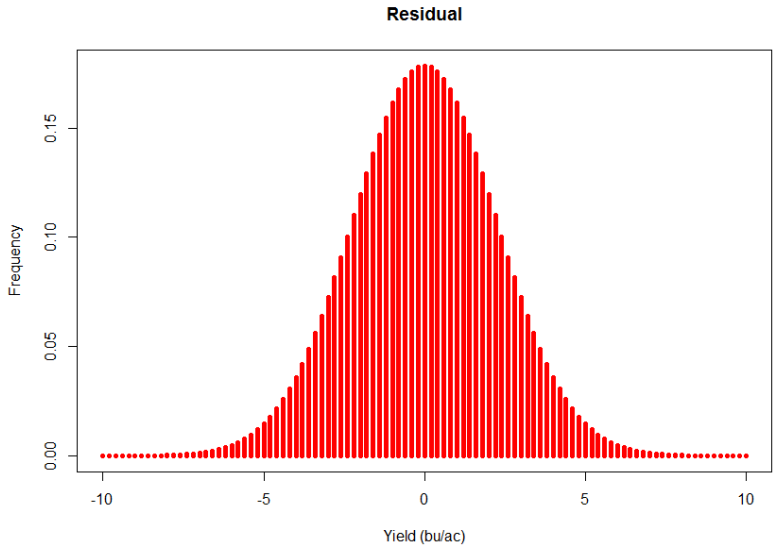
$$\sigma_y^2 = \sigma_g^2 + \sigma_e^2$$



$$y \sim N(\mu, I\sigma_y^2)$$



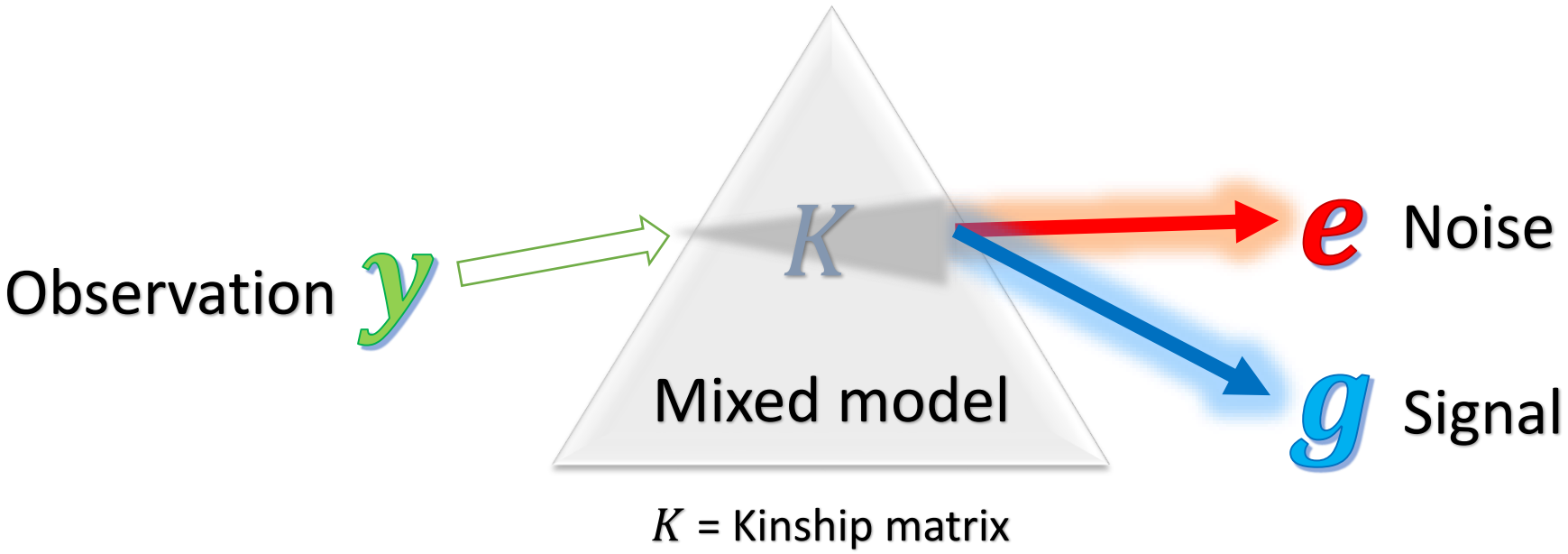
$$u \sim N(0, I\sigma_g^2)$$



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

Model notation

$$y = \mu + g + e$$



Model notation

$$y = Xb + Zu + e$$

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

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

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

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

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)

Searle 1997

Henderson's equation

Model statement

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

$$y \sim N(Xb, V)$$

$$\mathbf{V} = \mathbf{ZDZ}' + \mathbf{R}$$

$$D = A\sigma_a^2$$
$$R = I\sigma_e^2$$

Generalized Mixed Linear Model Equation

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{D}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

BLUE & BLUP solutions

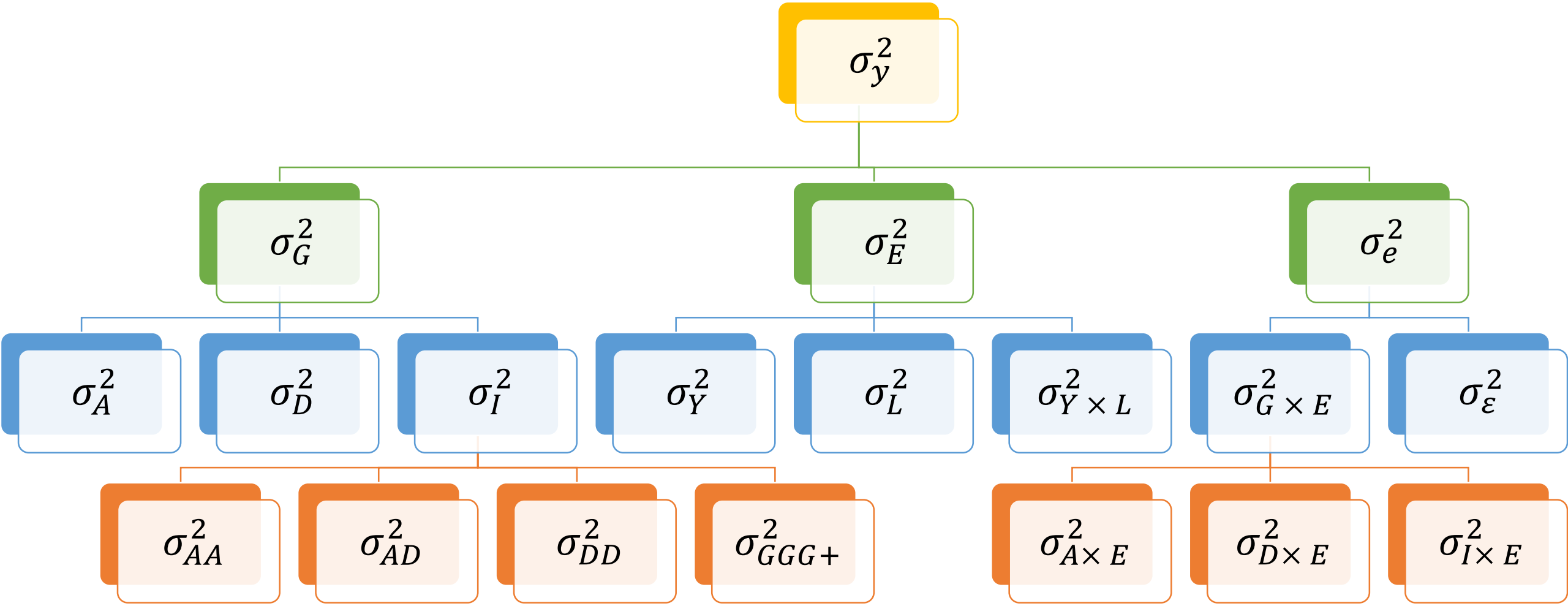
$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \quad \text{with} \quad \text{BLUE}(\mathbf{X\beta}) = \mathbf{X}\tilde{\boldsymbol{\beta}}$$

$$\tilde{\mathbf{u}} = \mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \text{BLUP}(\mathbf{u})$$

Reduce Model ($R = I$)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \{_d\, \lambda_i \mathbf{I}_{q_i}\}_{i=1}^{r} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix}$$

Variance decomposition: Multiple random effects



PART II - Applications

Phenotypic selection

1. Genetic values (image on right)
   - From BLUP or BLUE
   - Non-additive
   - Require replicated trials

2. Breeding values
   - Use pedigree information
   - Additive genetics
   - Not necessarily replicated

3. Genomic Breeding values
   - Genotypes replace pedigree
   - Not necessarily additive

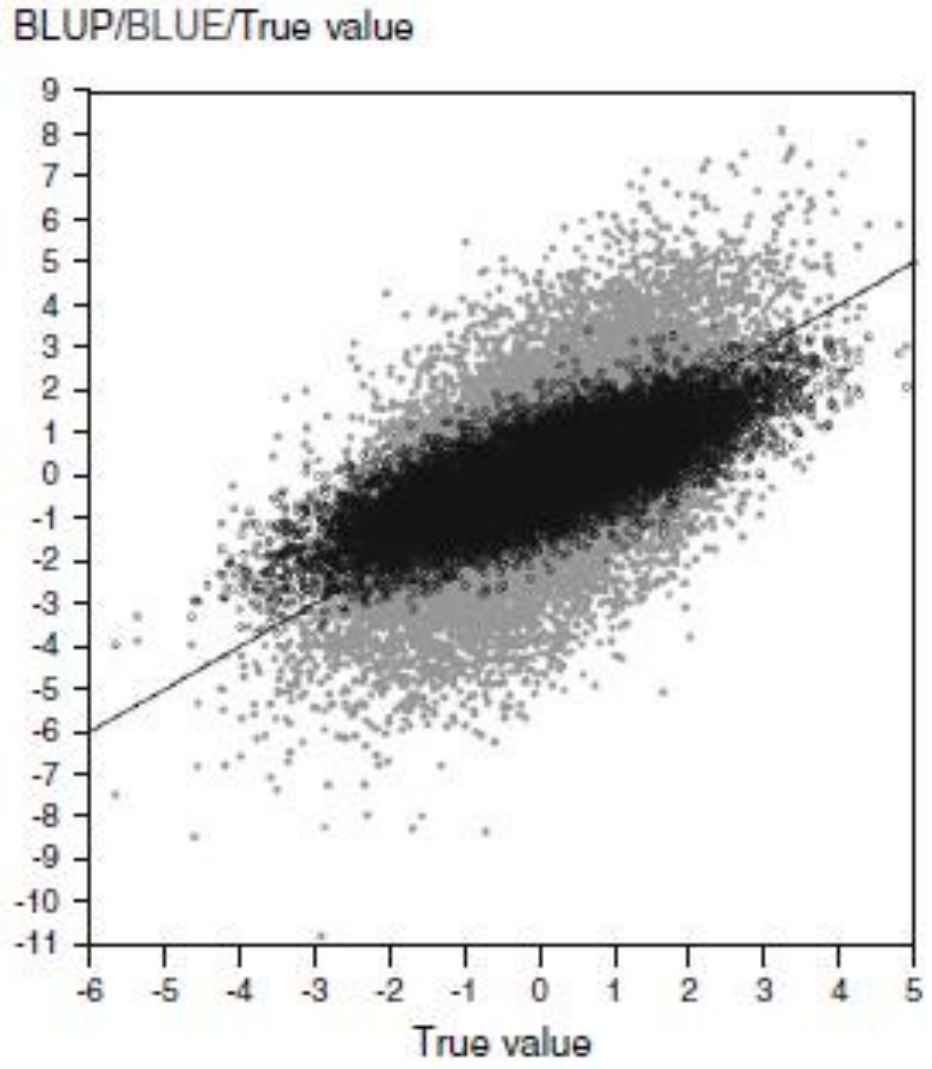


Fig. 4 Plot of simulated BLUP (black circles), BLUE (grey dots), and true genetic values (solid line) versus true values

Piepho et al. 2008

Practical examples

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

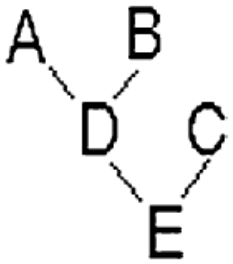
$$\begin{bmatrix} \mathbf{X'X} & \mathbf{Z'X} \\ \mathbf{X'Z} & \mathbf{Z'Z} + \lambda \mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix}$$

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

Example from **Xavier et al. 2016**

Field map (line and its yield)		
A = 27	Missing	E = 21
E = 27	C = 20	B = 27
B = 21	A = 25	Missing

Pedigree



INPUT

$$\mathbf{y} = \begin{array}{c} \text{Yield} \\ \begin{bmatrix} 25 \\ 27 \\ 27 \\ 21 \\ 20 \\ 21 \\ 27 \end{bmatrix} \end{array} \quad \mathbf{X} = \begin{array}{c} \text{Intercept} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array} \quad \mathbf{Z} = \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \quad \mathbf{K} = \begin{array}{c|ccccc} & A & B & C & D & E \\ \hline A & 1 & 0 & 0 & 0.5 & 0.25 \\ B & 0 & 1 & 0 & 0.5 & 0.25 \\ C & 0 & 0 & 1 & 0 & 0.5 \\ D & 0.5 & 0.5 & 0 & 1 & 0.5 \\ E & 0.25 & 0.25 & 0.5 & 0.5 & 1 \end{array}$$

OUTPUT

$$\mathbf{b} = [23.812] \quad \mathbf{u} = \begin{bmatrix} 1.191 \\ 0.172 \\ -1.291 \\ 0.799 \\ -0.060 \end{bmatrix} \quad \sigma_a^2 = 4.004 \quad \sigma_e^2 = 6.987$$

Models with genomic data

G-BLUP and RR-BLUP

- **Purpose**: Estimate or predict breeding values using genomic data
- **Method 1 (G-BLUP)**: MME using genomic relationship matrix (**K=G**)
- **Method 2 (RR-BLUP)**: MME using centralized markers as the random design matrix (**Z=M**)

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

- Environment
- Block effect
- Covariates

- Breeding values (1): $\mathbf{u} \sim N(0, \mathbf{G}\sigma_a^2)$
- Breeding values (2): $\mathbf{u} = \mathbf{M}\boldsymbol{\alpha}, \boldsymbol{\alpha} \sim N(0, \mathbf{I}\sigma_u^2)$

$$G = \frac{MM'}{2\sum p_j(1-p_j)}$$

$$\sigma_a^2 = \sigma_u^2\, 2\sum_{j=1}^{P} p_j(1-p_j)$$

Genome-Wide Association Studies (GWAS)

- **Purpose**: Identify markers associated to the trait of interest
- **Method**: Likelihood Ration Test (LRT) between of model with and without markers

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

- Intercept
- Sub-population (**Q**)
- Covariates
- Marker (**with** vs. **without**)

- Breeding values: $\mathbf{u} \sim \mathrm{N}(0, \mathbf{K}\sigma_a^2)$

REFERENCES

1. Cunningham, E. P., & Henderson, C. R. (1968). An iterative procedure for estimating fixed effects and variance components in mixed model situations. *Biometrics*, 13-25.

2. Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical science*, 15-32.

3. Searle, S. R. (1997). The matrix handling of BLUE and BLUP in the mixed linear model. *Linear algebra and its applications*, *264*, 291-311.

4. Piepho, H. P., Möhring, J., Melchinger, A. E., & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, *161*(1-2), 209-228.

5. Xavier, A., Muir, W. M., Craig, B., & Rainey, K. M. (2016). Walking through the statistical black boxes of plant breeding. *Theoretical and Applied Genetics*, *129*(10), 1933-1949.