Pedigree-based genetic evaluation

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-07

Learning objectives

- Understand how to combine phenotype information from all relatives connected via pedigree

- Familiarise yourself with linear mixed models and equations

- Practice inference of breeding values with the pedigree-based model
  - simple cases using R matrix algebra
  - using other packages

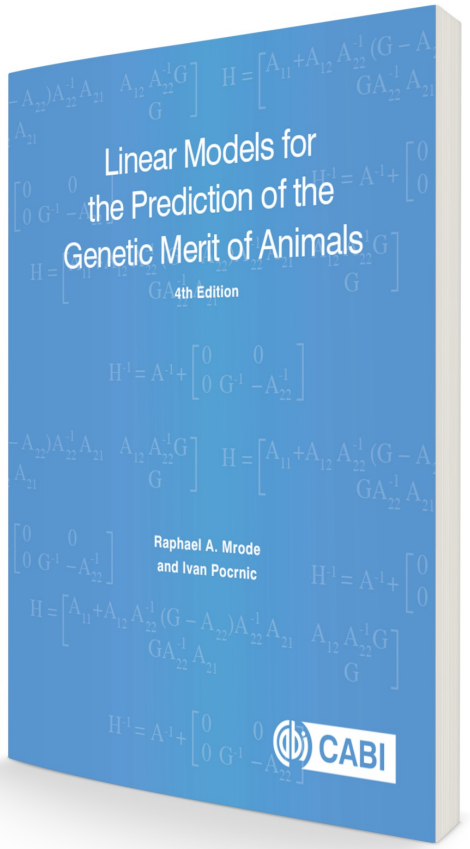
Linear Models for the Prediction of the Genetic Merit of Animals

CABI Biotechnology Series

September 2023 | 412pp

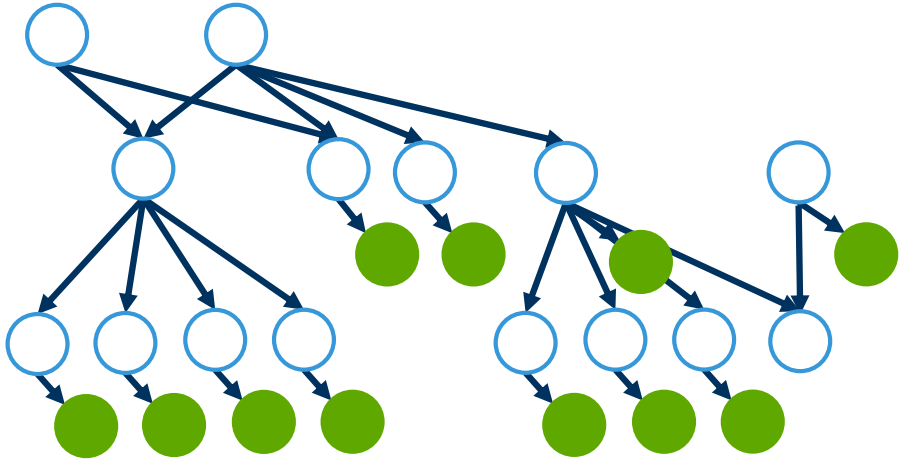
Raphael A Mrode
Ivan Pocrnic

Robin Thompson
Gregor Gorjanc



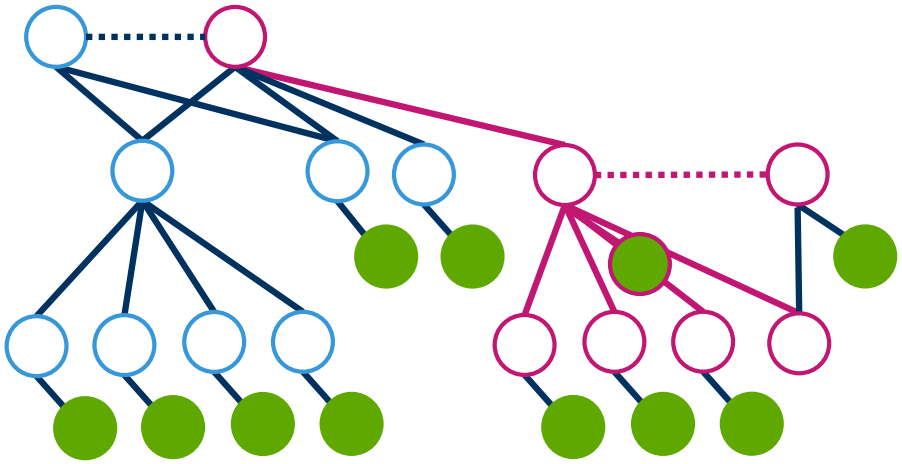
See chapters
3 & 4!

General setting



- Evaluation/Estimation
  - ancestors phenotypes
  - own phenotypes
  - sib… phenotypes
  - descendants phenotypes
- Prediction
  - progeny
- Population/group means

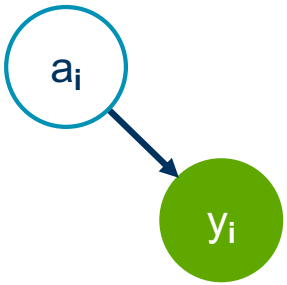
Information for an individual



Pedigree-based model

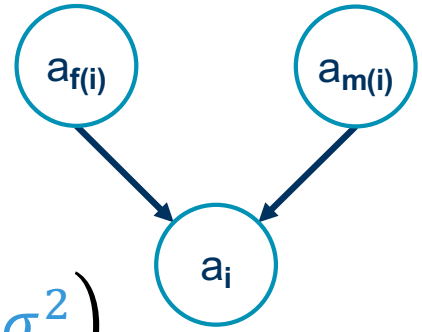
Phenotype model

$$y_i = \mu + a_i + e_i$$
$$e_i \sim N(0, \sigma_e^2)$$



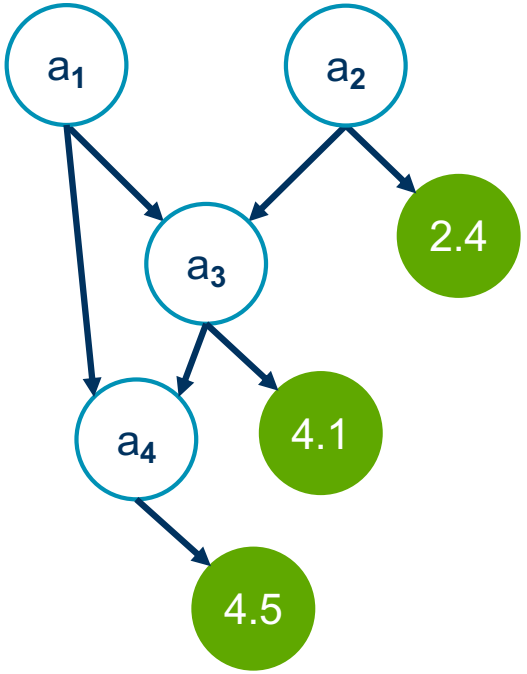
Pedigree model

$$a_i \sim N(0, \sigma_a^2)$$
$$a_i = \frac{1}{2}a_{f(i)} + \frac{1}{2}a_{m(i)} + r_i$$
$$a_i | a_{f(i)}, a_{m(i)} \sim N\left(\frac{1}{2}a_{f(i)} + \frac{1}{2}a_{m(i)}, k\sigma_a^2\right)$$
$$r_i \sim N(0, k\sigma_a^2)$$
$$k = \frac{1}{2} - \frac{1}{4}\left(F_{f(i)} + F_{m(i)}\right)$$



Galton (1886)
Wright (1920+)

Pedigree-based model - example



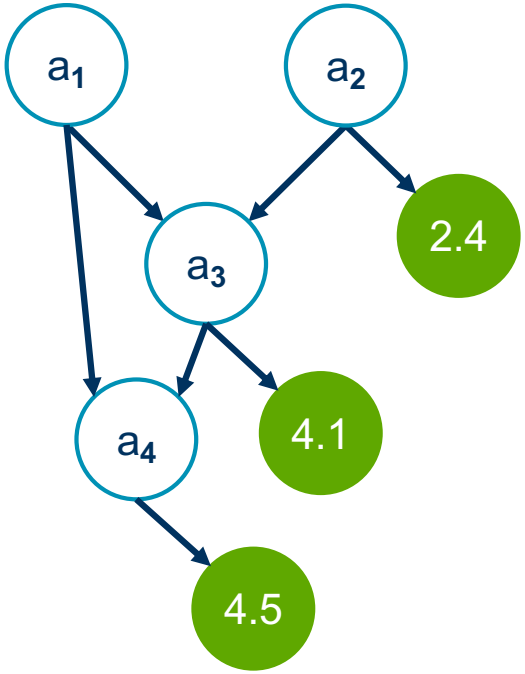
Phenotype model

$$y_2 = 2.4 = \mu + a_2 + e_2$$
$$y_3 = 4.1 = \mu + a_3 + e_3 \qquad e_i \sim N(0, \sigma_e^2)$$
$$y_4 = 4.5 = \mu + a_4 + e_4$$

Pedigree model

$$a_1 = r_1 \qquad r_i \sim N(0, k\sigma_a^2)$$
$$a_2 = r_2$$
$$a_3 = \frac{1}{2}a_1 + \frac{1}{2}a_2 + r_3 = \frac{1}{2}r_1 + \frac{1}{2}r_2 + r_3$$
$$a_4 = \frac{1}{2}a_1 + \frac{1}{2}a_3 + r_4 = \frac{1}{2}r_1 + \frac{1}{2}\left(\frac{1}{2}r_1 + \frac{1}{2}r_2 + r_3\right) + r_4$$
$$= \frac{3}{4}r_1 + \frac{1}{4}r_2 + \frac{1}{2}r_3 + r_4$$

Pedigree-based model - example



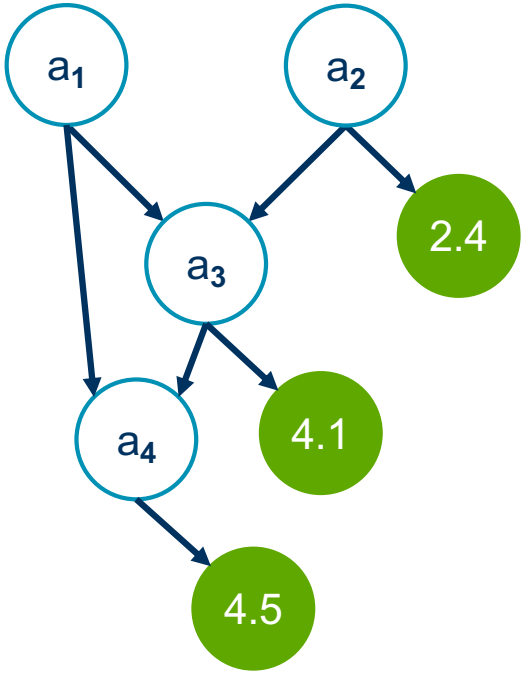
Phenotype model

$$\begin{pmatrix} y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 4.1 \\ 4.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} + \begin{pmatrix} e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$

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

Pedigree-based model - example



Phenotype model

$$\begin{pmatrix} y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 4.1 \\ 4.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} + \begin{pmatrix} e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}$$

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

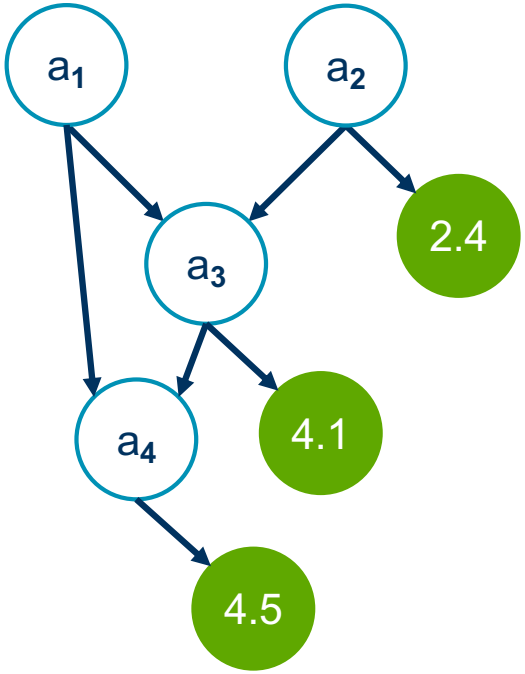
Pedigree model

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{3}{4} & \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix}$$

$$\boldsymbol{a} = \boldsymbol{T}\boldsymbol{r}$$

$$\boldsymbol{r} \sim N(\boldsymbol{0}, \boldsymbol{R}\sigma_a^2)$$

Pedigree-based model – example & INTERPRETATION



Phenotype model

$$\begin{pmatrix} y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 4.1 \\ 4.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} + \begin{pmatrix} e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$

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

μ

PHENO. POPULATION

(often not clearly definable)

Pedigree model

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{3}{4} & \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix}$$

$$\boldsymbol{a} = \boldsymbol{Tr}$$

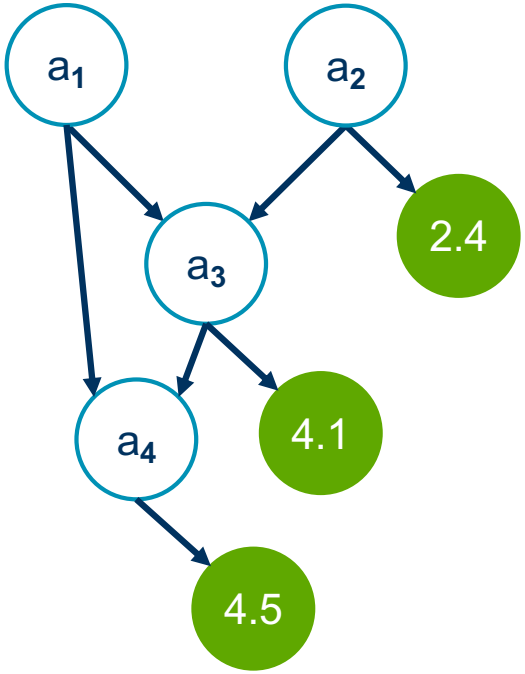
$$\boldsymbol{r} \sim N(\boldsymbol{0}, \boldsymbol{R}\sigma_a^2)$$

σ_a^2

BASE POPULATION!!!!

(often not clearly definable)

Pedigree-based model - example



Phenotype model

$$\begin{pmatrix} y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 4.1 \\ 4.5 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} (\mu) + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} + \begin{pmatrix} e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$

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

Pedigree model

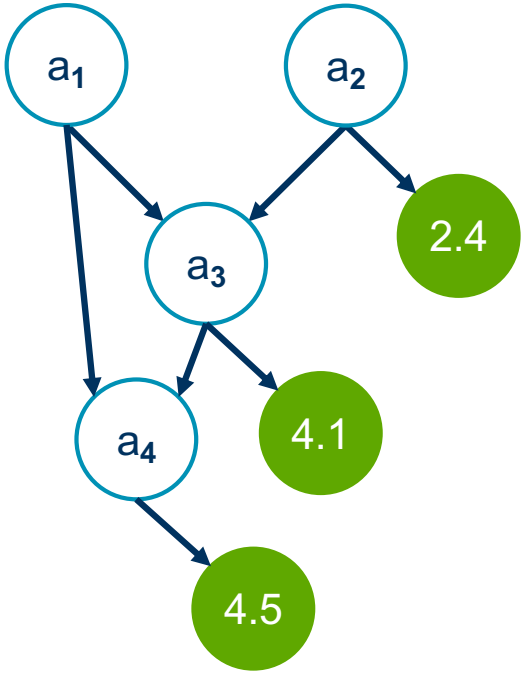
$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \\ \frac{3}{4} & \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix}$$

$$\boldsymbol{a} = \boldsymbol{Tr}$$

$$\boldsymbol{r} \sim N(\mathbf{0}, \boldsymbol{R}\sigma_a^2)$$

$$\begin{aligned} Var(\boldsymbol{a}) &= Var(\boldsymbol{Tr}) \\ &= \boldsymbol{T}Var(\boldsymbol{r})\boldsymbol{T}^T \\ &= \boldsymbol{TRT}^T\sigma_a^2 \\ &= \boldsymbol{A}\sigma_a^2 \end{aligned}$$

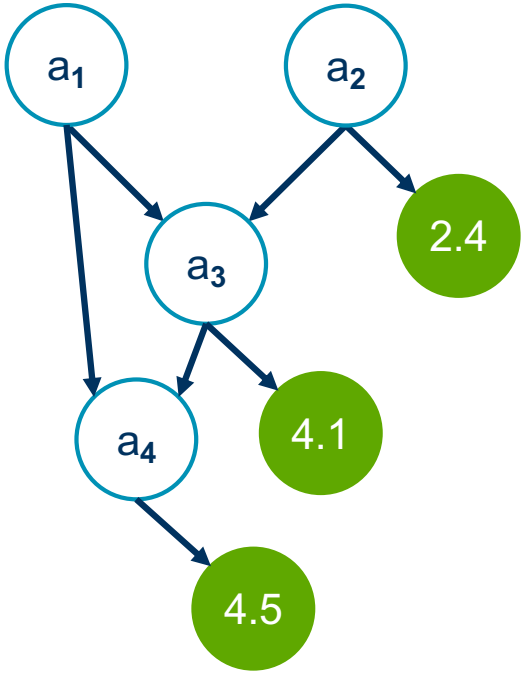
Pedigree-based relationship matrix



$$Var(\boldsymbol{a}) = \boldsymbol{A}\sigma_a^2$$

- Elements of **A** are covariance coefficients!
- Wright's relationships were correlations
  - $R_{i,j}=\mathbf{A}_{1,2}/\text{sqrt}(\mathbf{A}_{i,i}\,\mathbf{A}_{j,j})$
  - **A** → Numerator Relationship Matrix (NRM)
- $\mathbf{A}_{i,i}=1+F_i$, F_i – inbreeding coefficient (are alleles of the individual i IBD)
- $F_i=K_{f(i),m(i)}=1/2\mathbf{A}_{f(i),m(i)}$
- $\mathbf{A}_{i,j}=2K_{i,j}$, $K_{i,j}$ – kinship coefficient (are alleles of the individuals i and j IBD)
- $\mathbf{A}_{i,j}=1/2\mathbf{A}_{i,f(j)}+1/2\mathbf{A}_{i,m(j)}$
- $\mathbf{A}=2\mathbf{K}$

Pedigree-based relationship matrix



$$Var(\boldsymbol{a}) = \boldsymbol{A}\sigma_a^2$$

- Some colleagues are extremely good in understanding and tweaking NRMs directly (long tradition in genetics since Wright)
- I find it more “illuminating” to focus on the statistical model and its assumptions and then NRM are “just” a by-product!

Pedigree based model

- The usual notation for pedigree-based linear mixed model

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$
$$\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$$

- Estimator/Predictor (summarising the conditional distribution):

$$E(\boldsymbol{a}|\boldsymbol{y}) = \hat{\boldsymbol{a}} = Cov(\boldsymbol{a}, \boldsymbol{y})Var(\boldsymbol{y})^{-1}\big(\boldsymbol{y} - E(\boldsymbol{y})\big)$$

$$Var(\boldsymbol{a}|\boldsymbol{y}) = Var(\boldsymbol{a}) - Cov(\boldsymbol{a}, \boldsymbol{y})Var(\boldsymbol{y})^{-1}Cov(\boldsymbol{y}, \boldsymbol{a})$$

$$Cor(\boldsymbol{a}, \hat{\boldsymbol{a}})^2 = \boldsymbol{1} - Var(\boldsymbol{a}|\boldsymbol{y})/Var(\boldsymbol{a})$$

Henderson (1950+)

Pedigree based model

- The usual notation for pedigree-based linear mixed model

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Za} + \boldsymbol{e}$$
$$\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{E}\sigma_e^2)$$
$$\boldsymbol{a} \sim N(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$$

- Estimator/Predictor (summarising the conditional distribution):

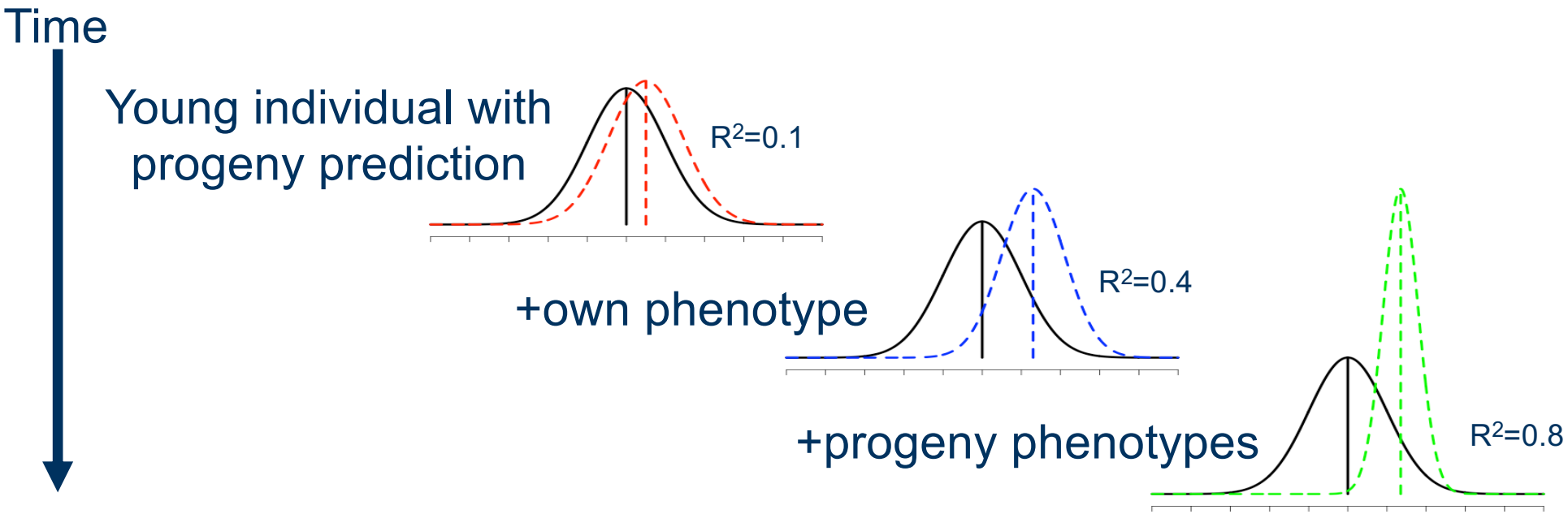
$$\begin{pmatrix} \boldsymbol{X^T E^{-1} X} & \boldsymbol{X^T E^{-1} Z} \\ \boldsymbol{Z^T E^{-1} X} & \boldsymbol{Z^T E^{-1} Z} + \boldsymbol{A^{-1}} \frac{\sigma_e^2}{\sigma_a^2} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X^T E^{-1} y} \\ \boldsymbol{Z^T E^{-1} y} \end{pmatrix}$$

$$\boldsymbol{C} \begin{pmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{a}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X^T E^{-1} y} \\ \boldsymbol{Z^T E^{-1} y} \end{pmatrix}$$

Henderson (1950+)

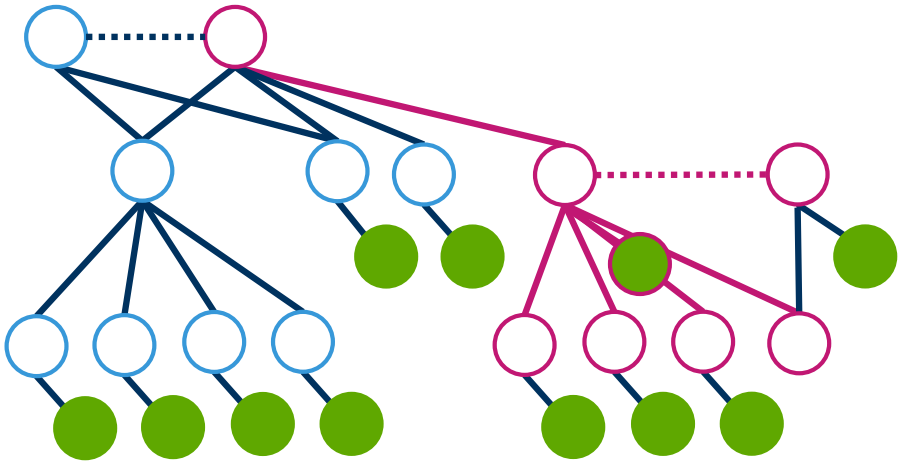
$$Var(\boldsymbol{a}|\boldsymbol{y}) = diag(\boldsymbol{C^{-1}})_{\boldsymbol{a}}\, \sigma_e^2$$

Accumulation of information over time



- Response = f(accuracy, intensity, diversity, time)
- Accuracy vs. Time conflict

Information for an individual



- We take all information into account, though recursively
- Markov blanket for an individual breeding value
  - graphical “parents”, “progeny”, and “mates” (=non-zero elements in $\boldsymbol{C}$)

Pedigree-based linear-mixed model

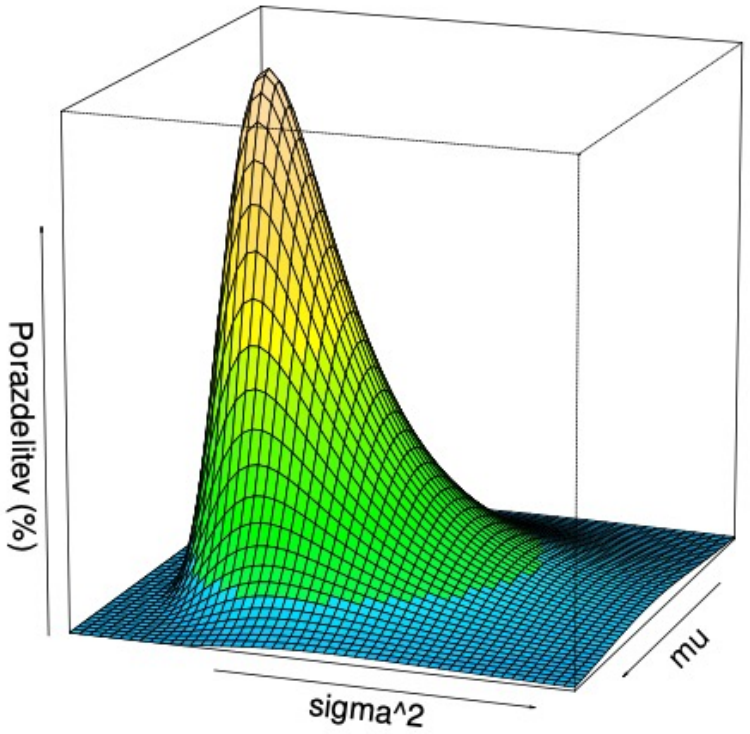
- Phenotype extensions
  - multiple traits
  - different distributions
    (binary, threshold, ordinal, survival, …)
- Genetic extensions
  - sire/father, sire/father-dam/mother, animal/individual model
  - genetic groups, uncertain parentage
  - gametic & imprinting effects
  - sex & cytoplasmic inheritance
  - “social” genetic effects
  - individual QTL & genes

Computational aspects

- Standard programs (ASReml, blupf90, Mix99, WOMBAT, …)
- Write your own! :) :(
- Sparse vs. dense matrices ($\boldsymbol{C}$ and $\boldsymbol{A}^{-1}$ are sparse!)
- Tasks
  - Estimate “fixed” and “random” effects (=location parameters)
    → $p(\boldsymbol{b}, \boldsymbol{a} | \boldsymbol{y}, \sigma_a^2, \sigma_e^2)$
  - Estimate variance components (=dispersion/hyper-parameters)
    → $p(\sigma_a^2, \sigma_e^2 | \boldsymbol{y})$ MAP/REML or full distribution
    (”Hill-climbing” (EM, NR, AI, …) or “Hill-exploring” (MCMC, MC-EM) algorithms)

Simple mean & variance problem

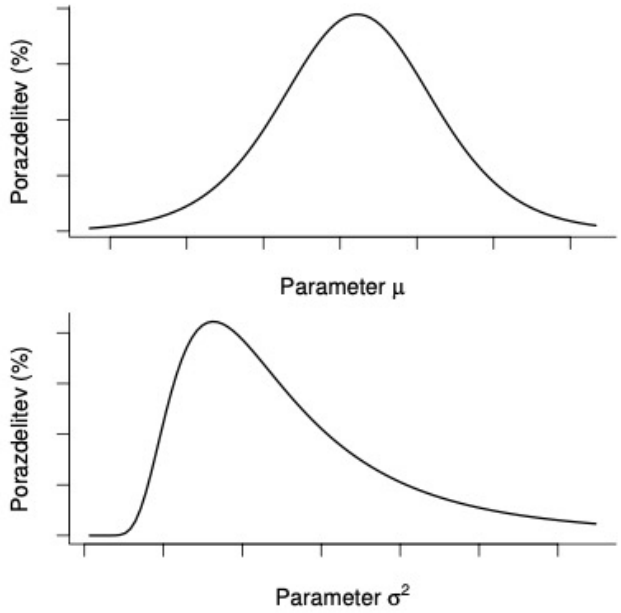
$$
\begin{aligned}
p\left(y|\mu,\sigma^2\right) &= N\left(\mu,\sigma^2\right) \\
p\left(\mu\right) &= N\left(\ldots\right) \\
p\left(\sigma^2\right) &= IG\left(\ldots\right) \\
p\left(\mu,\sigma^2|y\right) &\propto p\left(y|\mu,\sigma^2\right) p\left(\mu\right) p\left(\sigma^2\right)
\end{aligned}
$$

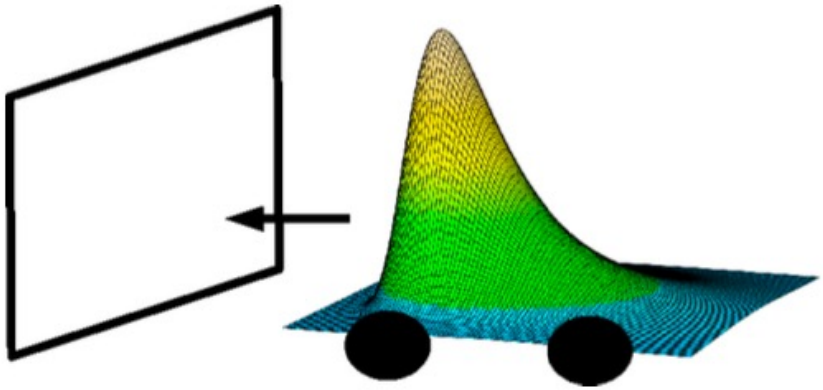


Simple mean & variance problem

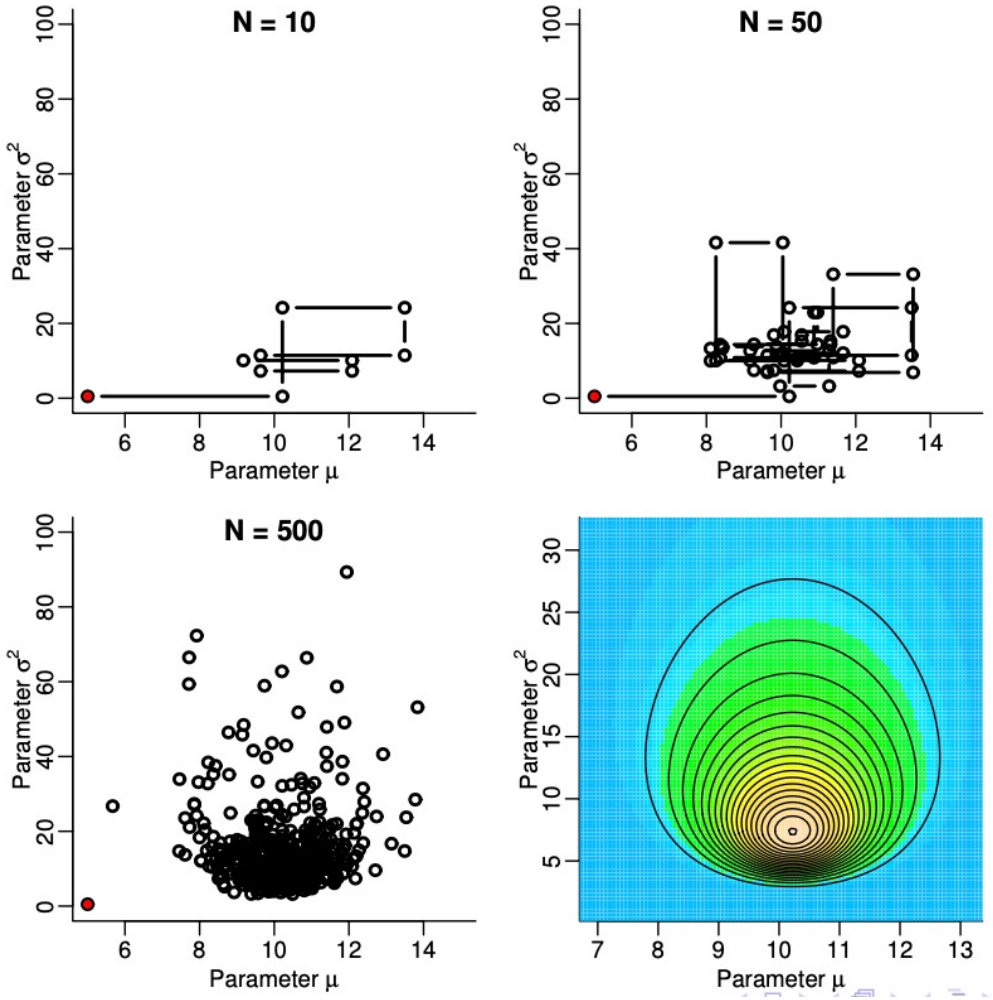
$$p(\mu|y) = \int_0^{\infty} p(\mu,\sigma^2|y)\, d\sigma^2 \qquad p(\sigma^2|y) = \int_{-\infty}^{+\infty} p(\mu,\sigma^2|y)\, d\mu$$

$$\hat{\mu} = \sum_{i=1}^{n} y_i / n \qquad \hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{\mu})^2 / (n-1)$$



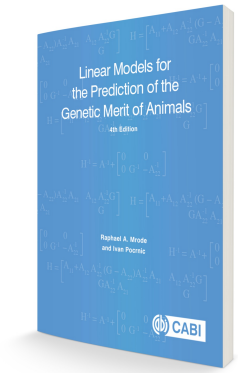


Simple mean & variance problem



Computational aspects

- Sparse vs. dense matrices ($\boldsymbol{C}$ and $\boldsymbol{A}^{-1}$ are sparse!)
- Tasks
  - Estimate “fixed” and “random” effects (=location parameters)
    → $p(\boldsymbol{b}, \boldsymbol{a} | \boldsymbol{y}, \sigma_a^2, \sigma_e^2)$
  - Estimate variance components (=dispersion/hyper-parameters)
    → $p(\sigma_a^2, \sigma_e^2 | \boldsymbol{y})$ MAP/REML or full distribution
    (”Hill-climbing” (EM, NR, AI, …) or “Hill-exploring” (MCMC, MC-EM) algorithms)



See chapters
3, 17-18, &
Appendix A!

Learning objectives

- Understand how to combine phenotype information from all relatives connected via pedigree

- Familiarise yourself with linear mixed models and equations

- Practice inference of breeding values with the pedigree-based model
  - simple cases using R matrix algebra
  - using other packages

Questions?!

Pedigree-based genetic evaluation

Gregor Gorjanc, Chris Gaynor, Jon Bancic, Daniel Tolhurst

UNE, Armidale
2024-02-07