Analysis of Experiments using ASReml-R:

with emphasis on breeding trials ©

Salvador A. Gezan
forestats.sg@gmail.com

Patricio R. Munoz
p.munoz@ufl.edu

Melissa Pisaroglo de Carvalho
melissapisaroglo@ufl.edu

Miami, USA, April 2016

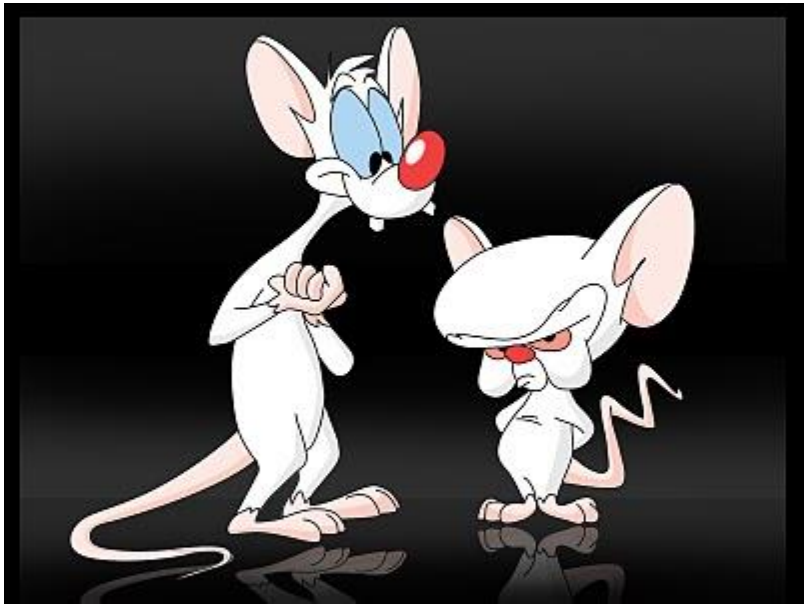
Day 1

8:30 am – 8:45 am	Introductions
8:45 am – 9:30 am	Introduction to ASReml-R
9:30 am – 10:00 am	Practical 1.1
10:00 am – 10:30 am	Introduction to Linear Mixed Models
10:30 am – 11:00 am	Coffee Break
11:00 am – 11:30 am	Job Structure in ASReml-R
11:30 am – 12:00 pm	Practical 1.2
12:00 pm – 12:30 pm	Breeding Theory
12:30 pm – 1:30 pm	Lunch Break
1:30 pm – 2:15 pm	Parental Models
2:15 pm – 2:45 pm	Practical 1.3
2:45 pm – 3:00 pm	Incorporating Pedigree
3:00 pm – 3:30 pm	Animal Models
3:30 pm – 4:00 pm	Coffee Break
4:00 pm – 4:45 pm	Practical 1.4
4:45 pm – 5:00 pm	Round Up

Day 2

8:30 am -9:00 am	Variance Structures in ASReml-R
9:00 am – 10:00 am	Multivariate Analysis / Repeated Measures
10:00 am – 10:30 am	Practical 2.1
10:30 am – 11:00 am	Coffee Break
11:00 am – 12:00 pm	Multi-environment Analysis
12:00 pm – 12:30 pm	Practical 2.2
12:30 pm – 1:30 pm	Lunch Break
1:30 pm – 2:00 pm	Spatial Analysis
2:00 pm – 2:30 pm	Practical 2.3
2:30 pm – 3:30 pm	Introduction to Genomic Selection
3:30 pm – 4:00 pm	Coffee Break
4:00 pm – 4:45 pm	Practical 2.4
4:45 pm – 5:00 pm	Round Up

Session 1



Introduction to ASReml-R

WHAT IS ASReml-R?

"ASReml-R is an statistical packages that fits linear mixed models to moderately large data sets using Residual Maximum Likelihood (REML)"

"Typical applications include the analysis of (un)balanced longitudinal data, repeated measures analysis, the analysis of (un)balanced designed experiments, the analysis of multi-environment trials, the analysis of both univariate and multivariate animal breeding, genetics data and the analysis of regular or irregular spatial data."

ASReml in R uses the *Average Information* (AI) algorithm and *sparse matrix operations* methods.

- Useful for analysis of large and complex dataset.
- Very flexible to model a wide range of variance models for random effects or error structures (however, complex to program).

HOW TO GET ASReml-R?

Distributor Page

http://www.vsni.co.uk/products/asreml (version 3)

http://www.r-project.org/ (for R)

Platforms

Windows 98/ME/2000/XP/Vista/Windows7

Linux

Apple Macintosh

Interface

ASReml-SA

- DOS (edit)
- Windows Notepad
- ASReml-W)
- Text editors (e.g. ConTEXT)

ASReml-R

- R (or S-plus)
- R-Studio

WHERE TO GET HELP?

Official Documentation

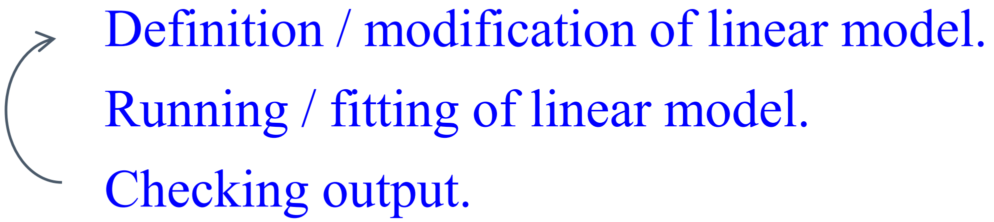
asreml-R.pdf	(use Find window for searching)
UserGuide.pdf	(for ASReml-SA)

Webpages

uncronopio.org/ASReml/HomePage (cookbook)
http://www.vsni.co.uk/software/asreml/htmlhelp/ (distributor page)
www.vsni.co.uk/forum (user forum)

STEPS FOR AN ANALYSIS

- Identify the problem and experimental design / observational study.
- Detail treatment and design structure.
- Specify hypotheses / components of interest.
- Collect and prepare data file (e.g. Excel, Access).
- Perform initial data validation and exploratory data analysis (EDA) in statistical software (e.g. R, SAS, GenStat).



Definition / modification of linear model.

Running / fitting of linear model.

Checking output.

- Extract final output.
- Report analysis.

ALFALFA EXPERIMENT

Example: `/Day1/Alfalfa/ALFALFA.txt`

An experiment was established to compare 12 alfalfa varieties (labeled A-L). These correspond to 3 different sources but the objective is to estimate heritability of varieties regardless of its source. A total of 6 plots per variety were established arranged in a RCB design. The response variable corresponds to yield (tons/acre) at harvest time.

Source	Variety	Bk1	Bk2	Bk3	Bk4	Bk5	Bk6
1	A	2.17	1.88	1.62	2.34	1.58	1.66
1	B	1.58	1.26	1.22	1.59	1.25	0.94
1	C	2.29	1.60	1.67	1.91	1.39	1.12
1	D	2.23	2.01	1.82	2.10	1.66	1.10
2	E	2.33	2.01	1.70	1.78	1.42	1.35
2	F	1.38	1.30	1.85	1.09	1.13	1.06
2	G	1.86	1.70	1.81	1.54	1.67	0.88
2	H	2.27	1.81	2.01	1.40	1.31	1.06
3	I	1.75	1.95	2.13	1.78	1.31	1.30
3	J	1.52	1.47	1.80	1.37	1.01	1.31
3	K	1.55	1.61	1.82	1.56	1.23	1.13
3	L	1.56	1.72	1.99	1.55	1.51	1.33

ALFALFA EXPERIMENT

Consider a model with block as fixed and variety as random effects.

yield = μ + block + variety + error

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effect of the i^{th} block

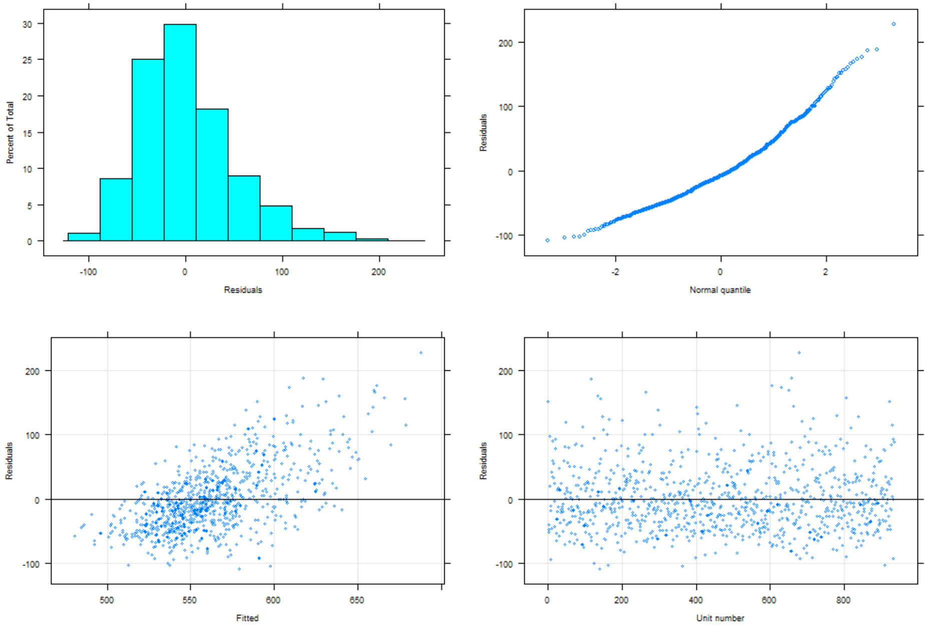
g_j random effect of the j^{th} variety, $E(g_j) = 0$, $V(g_j) = \sigma_g{}^2$

e_{ij} random error of the ij^{th} observation, $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma^2$

$i = 1, \dots, 6$ (r blocks)

$j = 1, \dots, 12$ (t treatments)

Session 2



Introduction to Linear Mixed Models

MIXED MODELS

- **Mixed models** extend the linear model by allowing a more flexible specification of the errors (and other random factors). Hence, it allows for a different type of inference and also allows to incorporate *correlation* and *heterogeneous variances* between the observations.

- **Fixed effects:** are those factors whose levels are selected by a nonrandom process or whose levels consist of the entire population of possible levels. Inferences are made *only* to those levels included in the study. Hint: all levels of interest are in your data set.

- **Random effects:** a factor where its levels consist of a random sample of levels from a population of possible levels. The inference is about the population of levels, not just the subset of levels included in the study.

- Mixed linear models contain both *random* and *fixed* effects.

MODEL FOR A RCBD

Dataset: two factors to consider: one defining the block to which each experimental unit is allocated, and the other to the treatment applied to each unit.

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

where,

y_{ij} observation belonging to the i^{th} treatment j^{th} block, $i = 1 \ldots r, j = 1 \ldots t$

μ is the population mean

α_i fixed effects of the i^{th} block

g_j random effects of the j^{th} variety, $E(g_j) = 0$, $V(g_j) = \sigma_g{}^2$

e_{ij} random error of the ij^{th} observation, $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma^2$

$$g_i \sim N[0, \sigma_g{}^2]$$
$$e_{ij} \sim N[0, \sigma^2]$$

MODEL COMPONENTS

response = systematic component + random component

response = structural component + explanatory component + random component

Structural component (or blocking structure)

- Concerned the underlying variability (heterogeneity) and structure of the experimental or measurement units.
- "Controls" different sources of natural variation amongst the units using factors (e.g. blocks) or variates (e.g. covariates).

Explanatory component (or treatment structure)

- Defines the different treatments (or treatment combinations) applied to the experimental units.
- Provides information about the differences in response caused by the different treatments and answers the questions of interest.

Multi-stratum ANOVA: makes explicit the separation between blocks (or the more general structure of units) and treatments.

MIXED MODELS

Hypothesis of interest

Fixed effects:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_t$

H_1: $\mu_i \neq \mu_j$ for some i, j in the set 1 … t

(i.e. is there a significant treatment effect)

Test statistic: F or t

Random effects:

H_0: $\sigma_g{}^2 = 0$

H_1: $\sigma_g{}^2 > 0$

(i.e. is there a significant variation due to the random effects)

Test statistic: Chi-square (likelihood ratio test)

ALFALFA EXPERIMENT

Consider a model with block as fixed and variety as random effects.

yield = μ + block + variety + error

$$y_{ij} = \mu + \alpha_i + g_j + e_{ij}$$

y_{ij} observation belonging to i^{th} treatment j^{th} block

α_i fixed effects of the i^{th} block

g_j random effects of the j^{th} variety, $\mathrm{E}(g_j) = 0$, $\mathrm{V}(g_j) = \sigma_g{}^2$

e_{ij} random error of the ij^{th} observation, $\mathrm{E}(e_{ij}) = 0$, $\mathrm{V}(e_{ij}) = \sigma^2$

$i = 1, \dots, 6$ (r blocks)

$j = 1, \dots, 12$ (t treatments)

ALFALFA EXPERIMENT

yield = μ + block + variety + error

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

$$\begin{bmatrix} y_{11} \\ . \\ . \\ . \\ y_{t1} \\ . \\ . \\ . \\ y_{1r} \\ . \\ . \\ y_{tr} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 0 \\ & & . & \\ & & . & \\ & & . & \\ 1 & 1 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ & & . & \\ & & . & \\ 1 & 0 & \ldots & 1 \\ & & . & \\ & & . & \\ 1 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ . \\ . \\ \beta_r \end{bmatrix} + \begin{bmatrix} 1 & 0 & \ldots & 0 \\ & & . & \\ & & . & \\ & & . & \\ 0 & 0 & \ldots & 1 \\ 1 & 0 & \ldots & 0 \\ & & . & \\ & & . & \\ 1 & 0 & \ldots & 0 \\ & & . & \\ & & . & \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ . \\ . \\ g_t \end{bmatrix} + \begin{bmatrix} e_{11} \\ . \\ . \\ . \\ e_{1t} \\ . \\ . \\ . \\ e_{1r} \\ . \\ . \\ e_{tr} \end{bmatrix}$$

$$\mathbf{G} = \begin{bmatrix} \sigma_g^2 & & & 0 \\ & \sigma_g^2 & & \\ & & \ldots & \\ 0 & & & \sigma_g^2 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} \sigma^2 & & & 0 \\ & \sigma^2 & & \\ & & \ldots & \\ 0 & & & \sigma^2 \end{bmatrix}$$

LINEAR MIXED MODEL

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \qquad E\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

X (*n* x *r*) design matrix for fixed effects

β (*r* x 1) vector of fixed effects

Z (*n* x *t*) design matrix for random effects

g (*t* x 1) vector of random effects

e (*n* x 1) vector of random errors

G (*t* x *t*) matrix of variance-covariance of random effects

R (*n* x *n*) matrix of variance-covariance of random errors

LINEAR MIXED MODEL

$$\mathbf{G} = \begin{array}{c} \\ g_1 \\ g_2 \\ \cdots \\ g_t \end{array} \begin{array}{c} \begin{array}{cccc} g_1 & g_2 & \cdots & g_t \end{array} \\ \begin{bmatrix} \sigma_g^2 & & & 0 \\ & \sigma_g^2 & & \\ & & \cdots & \\ 0 & & & \sigma_g^2 \end{bmatrix} \end{array} = \sigma_g^2 \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & \cdots & \\ 0 & & & 1 \end{bmatrix} = \sigma_g^2 \mathbf{I}_t$$

$$\mathbf{R} = \begin{array}{c} \\ e_{11} \\ e_{12} \\ \\ e_{tr} \end{array} \begin{array}{c} \begin{array}{cccc} e_{12} & e_{12} & \cdots & e_{tr} \end{array} \\ \begin{bmatrix} \sigma^2 & & & 0 \\ & \sigma^2 & & \\ & & \cdots & \\ 0 & & & \sigma^2 \end{bmatrix} \end{array} = \sigma^2 \mathbf{I}_{tr}$$

LINEAR MIXED MODEL

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \qquad E\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

Assumptions

- Random effects: $E(\mathbf{g}) = \mathbf{0}$, $\mathbf{V}(\mathbf{g}) = \mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$
- Deviations: $E(\mathbf{e}) = \mathbf{0}$, $\mathbf{V}(\mathbf{e}) = \mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$
- **g** and **e** independent.

hence, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R}$$

Note: normality assumptions can be made about **g** and **e**.

$$\mathbf{g} \sim \text{MVN}(0, \mathbf{G}) \quad \text{and} \quad \mathbf{e} \sim \text{MVN}(0, \mathbf{R})$$

VARIANCE COMPONENTS

- Henderson (1950) derived the Mixed Model Equations (MME) to obtain the solutions of all effects:

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

hence,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'\hat{V}^{-1}X})^{-1}\mathbf{X'\hat{V}^{-1}y}$$ BLUE → EBLUE

$$\hat{\mathbf{g}} = \mathbf{\hat{G}Z'\hat{V}^{-1}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$ BLUP → EBLUP

with

$$\mathbf{\hat{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{Z\hat{G}Z'+\hat{R}}$$

VARIANCE COMPONENTS

- Variance components need to be estimated before obtaining estimates of fixed/random effects and performing any type of inference.

$$\hat{\mathbf{G}} = \mathbf{G}(\hat{\theta}) \qquad \Rightarrow \hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}}) = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}' + \hat{\mathbf{R}}$$

$$\overset{\frown}{\mathbf{R}} = \mathbf{R}(\hat{\theta})$$

- Restricted/residual maximum likelihood (REML) is a likelihood-based method used to estimate these variance components and is based assuming that both **g** and **e** follow a multivariate normal distribution.

- The REML variance component estimates are later used to estimate the solutions of fixed and random effects.

- Approximated t-tests and F-tests are based on these variance components.

VARIANCE STRUCTURES

ASReml®

id: identity

$$\sigma^2\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}=\begin{bmatrix}\sigma^2 & 0 & 0 & 0\\ 0 & \sigma^2 & 0 & 0\\ 0 & 0 & \sigma^2 & 0\\ 0 & 0 & 0 & \sigma^2\end{bmatrix}$$

ar1v: autocorrelation 1st order

$$\sigma^2\begin{bmatrix}1 & \rho^1 & \rho^2 & \rho^3\\ \rho^1 & 1 & \rho^1 & \rho^2\\ \rho^2 & \rho^1 & 1 & \rho^1\\ \rho^3 & \rho^2 & \rho^1 & 1\end{bmatrix}$$

diag: diagonal

$$\begin{bmatrix}\sigma_1^2 & 0 & 0 & 0\\ 0 & \sigma_2^2 & 0 & 0\\ 0 & 0 & \sigma_3^2 & 0\\ 0 & 0 & 0 & \sigma_4^2\end{bmatrix}$$

corh: uniform heterogeneous

$$\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4\\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4\\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4\\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2\end{bmatrix}$$

corv: uniform correlation

$$\sigma^2\begin{bmatrix}1 & \rho & \rho & \rho\\ \rho & 1 & \rho & \rho\\ \rho & \rho & 1 & \rho\\ \rho & \rho & \rho & 1\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \sigma_2^2 & \sigma_2^2 & \sigma_2^2\\ \sigma_2^2 & \sigma_1^2 & \sigma_2^2 & \sigma_2^2\\ \sigma_2^2 & \sigma_2^2 & \sigma_1^2 & \sigma_2^2\\ \sigma_2^2 & \sigma_2^2 & \sigma_2^2 & \sigma_1^2\end{bmatrix}$$

us: unstructured

$$\begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2\\ \sigma_{12}^2 & \sigma_{22}^2 & \sigma_{23}^2 & \sigma_{24}^2\\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_{33}^2 & \sigma_{34}^2\\ \sigma_{14}^2 & \sigma_{24}^2 & \sigma_{34}^2 & \sigma_{44}^2\end{bmatrix}$$

CORRELATION STRUCTURES

cor: uniform correlation

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

corb: banded correlation

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

ar1: autocorrelation 1st order

$$\begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$

corg: general correlation

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

PROPERTIES OF EBLUE (optional)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

- $\mathrm{V}(\boldsymbol{\beta}) = (\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X})^{-1}$
- $\mathrm{V}(\mathbf{L}\boldsymbol{\beta}) = \mathbf{L}(\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{L}'$
- $\mathbf{L}\boldsymbol{\beta}$ is the best linear unbiased estimate of $\mathbf{L}\boldsymbol{\beta}$

- Test of H_0: $\mathbf{L}\boldsymbol{\beta} = \mathbf{0}$

$\boldsymbol{\beta}'\,\mathbf{L}'\,(\mathbf{L}\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X}\mathbf{L}'\,)^{-1}\mathbf{L}\boldsymbol{\beta} \sim$ F (approx) with $df_1 = r(\mathbf{L})$ and df_2 (Satterthwaite or Kenward-Roger)

- $100(1-\alpha)\%$ confidence interval for $l'\,\boldsymbol{\beta}$

$$l'\,\boldsymbol{\beta} \pm z_{\alpha/2}\, l'\,(\mathbf{X}'\,\mathbf{V}^{-1}\mathbf{X})^{-1} l$$

PROPERTIES OF EBLUP (optional)

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Z} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{X} & \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z}+\hat{\mathbf{G}}^{-1} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{Y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix}$$

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^{\mathbf{xx}} & \mathbf{C}^{\mathbf{xz}} \\ \mathbf{C}^{\mathbf{zx}} & \mathbf{C}^{\mathbf{zz}} \end{bmatrix}^{-} \begin{bmatrix} \mathbf{X}'\hat{\mathbf{R}}^{-1}\mathbf{y} \\ \mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{y} \end{bmatrix}$$

$$Var(\hat{\boldsymbol{\beta}}) = \mathbf{C}^{\mathbf{xx}}$$

$$Var(\hat{\mathbf{g}}) = \hat{\mathbf{G}} - \mathbf{C}^{\mathbf{zz}}$$

$$Var(\mathbf{g} - \hat{\mathbf{g}}) = \mathbf{C}^{\mathbf{zz}}$$

Predictions

- Linear Combination of a function of fixed and random effects:

$$\hat{\mathbf{P}} = \mathbf{L}'\hat{\boldsymbol{\beta}} + \mathbf{M}'\hat{\mathbf{g}}$$

$$\mathrm{Var}(\hat{\mathbf{P}}) = \mathbf{L}'\mathbf{C}^{\mathbf{xx}}\mathbf{L} + \mathbf{M}'(\hat{\mathbf{G}} - \mathbf{C}^{\mathbf{zz}})\mathbf{M}$$

PROPERTIES OF EBLUP (optional)

- **SE(BLUP)**: standard error of a random effect

$$\mathrm{SD}(\hat{\mathrm{g}}_i) = \sqrt{c^{ii}} = \sqrt{\mathrm{PE}V(\hat{\mathrm{g}}_i)}$$

- **PEV**: predictor error variance

$$\mathrm{PEV}(\hat{\mathrm{g}}_i) = c^{ii} = [\mathrm{SD}(\hat{\mathrm{g}}_i)]^2$$

- **r²**: reliability (correlation between true and predicted genetic values)

$$r^2(\hat{\mathrm{g}}_i) = 1 - \frac{\mathrm{PEV}(\hat{\mathrm{g}}_i)}{\sigma_g^2}$$

- **r**: accuracy

$$r(\hat{\mathrm{g}}_i) = \sqrt{r^2(\hat{\mathrm{g}}_i)} = \sqrt{1 - \frac{\mathrm{PEV}(\hat{\mathrm{g}}_i)}{\sigma_g^2}}$$

HOW GOOD IS H^2 / h^2 ESTIMATION

- *Inferences* with respect to h^2 are done in in terms of:
  - Confidence intervals
  - Hypothesis testing.

Heritability confidence interval

- Approximate 95% CI is: $\hat{h}^2 \pm 2se(\hat{h}^2)$
- The estimate of h^2 is a random variable resulting from a ratio of two random variables which are correlated.
- These two variables are approximately chi-square.
- Two (approximation) methods are in general use to estimate the standard error:
  - Dickerson's Method.
  - Delta Method.

DELTA METHOD

- Asymptotic Covariance of Variance Component Estimates and Taylor Series Approximation of the Variance of a Ratio – REML Estimation

- Let **V** = the covariance matrix for the variance components (nxn) where n equals the number of variance components.
- Let **l** be the matrix containing the weights for the numerator and denominator of h^2 (2xn).
- Then the variance of the numerator (1,1) and denominator (2,2) and their covariance (1,2 or 2,1) is contained in **l`Vl** (2x2).
- The approximation use is:

$$Var(\hat{h}^2) = (\frac{1}{D})^2 Var(N) - 2(\frac{N}{D^3})Cov(N,D) + (\frac{N^2}{D^4})Var(D)$$

where N is the numerator and D the denominator.

TESTING VAR. COMPONENTS

LRT: likelihood ratio test

- ○ Based on asymptotic derivations.
- ○ Used to compare nested models and is valid if the fixed effects are the same (under REML).
- ○ Examples: $H_0\colon \rho = 0$ against $H_0\colon \rho \neq 0$
  $H_0\colon \sigma^2_g = 0$ against $H_0\colon \sigma^2_g > 0$
- ○ Test Statistic: $d = 2\,[\,\log L_2 - \log L_1] \sim \chi^2_{r2\text{-}r1}$

Hypothesis	**P-value**
Two-sided	$\mathrm{Prob}(\chi^2_{r2\text{-}r1} > d)$
One-sided	$0.5(1 - \mathrm{Prob}(\chi^2_1 \leq d))$

[Self and Liang (1987, JASS 82:605–610)]

TESTING VAR. COMPONENTS

Critical values

r_2 - r_1	$\alpha = 0.05$		$\alpha = 0.01$	
Δdf	Two-sided	One-sided	Two-sided	One-sided
1	3.84	2.71	6.63	5.41
2	5.99	4.61	9.21	7.82
3	7.81	6.25	11.34	9.84
4	9.49	7.78	13.28	11.67
5	11.07	9.24	15.09	13.39

Goodness-of-fit statistics

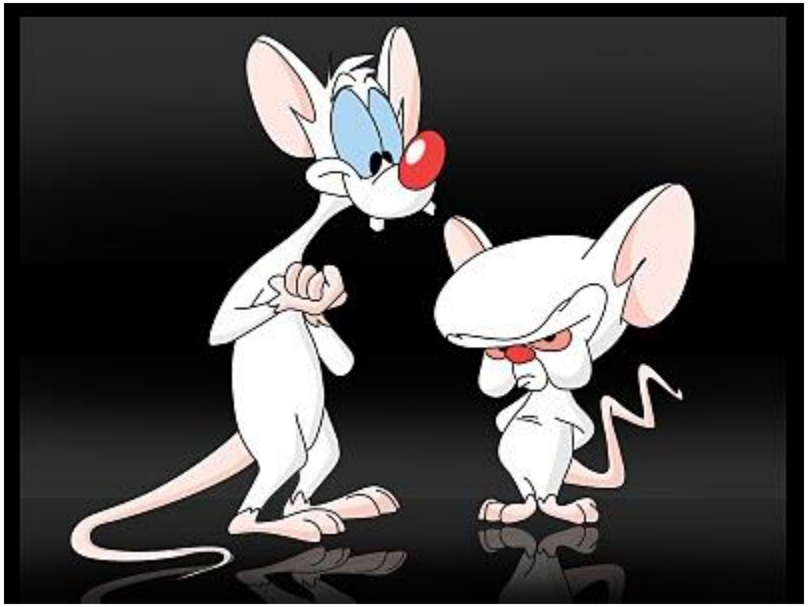
○ AIC and BIC can be used to select/rank non-nested models

$$\mathrm{AIC} = -2 \times \mathrm{logL} + 2 \times t$$
$$\mathrm{BIC} = -2 \times \mathrm{logL} + 2 \times t \times \log(v)$$

t number of variance parameters in the model

v residual degrees of freedom, $v = n - p$

Session 3



Job Structure in ASReml-R

JOB FILE (.R)

PART A: Data definition and reading of data set.

PART B: Definition of analysis (options, linear model).

```
asreml(fixed=~1,random,sparse,
       rcov=~units,G.param,R.param,
       predict=predict.asreml(),
       constraints=asrem.constraints(),
       data=sys.parent(),
       subset,family=asreml.gaussian(),
       weights=NULL,offset=NULL,
       na.method.Y=“include”,na.method.X=“fail”,
       keep.order=F,fixgammas=F,
       asmultivariate=NULL,
       model.frame=F,start.values=F,
       dump.model=F,model=F,
       control=asreml.control(…),…)
```

PART C: Extraction of output (options, linear model, output).

```
model<-asreml(fixed=yield~treatment+sex,
              random=~Variety+Dose+mother,
              data=fish)
```

JOB FILE (.R)

Reading Data

- ASCII file (delimited by: tab, comma or space) (R formatting).
- "NA" identify missing values, `na.method.Y=c(‘omit’,’include’)`
- Factors need to be defined, `na.method.X=c(‘omit’,’include’)`
- Labels are stored in the order on which they are read.

General Relevant File Syntax

~ separates *response* from the list of fixed and random terms.

comment following (skips rest of line).

, model specification continues on next line.

$ specifies an user-input option from commands.

Basic Model Syntax Operators

: interaction or nested effects (e.g. `A:B`).

+ sum of two factors in the model

JOB FILE

Relevant Options `asreml.control()`

`workspace` size of workspace for the REML routines in double precision words (Groups of 8 bytes). Default `workspace=8e6` (64,000,000 bytes).

`pworkspace` size of workspace for forming predictions of linear functions of variables in the model, measured in double precision words (Group of 8 bytes)

`maxiter` indicates a maximum number iterations (default 10)

`Csparse` non-zero elements of the inverse of the C matrix (of coefficient) are stored in this data frame (row, column, value).

`Cfixed` part of the C-inverse matrix is returned in component `Cfixed` of the ASReml object.

JOB FILE

Specification of Linear Models

Univariate case

```
model<-asreml(fixed=y~<fixed effects>,
              random=~<random effects>,
              rcov=~<error structure>,
              data=<dataset>)
```

Examples

```
asreml(fixed=yield~Variety,random=~Block,data=potato)
asreml(fixed=volume~Site+Site:Block,
       random=~Mother+Mother:Site,data=MET)
```

JOB FILE

Specification of Linear Models

- ASReml-R uses the Wilkinson and Rogers (1973) notation.

`A:B` indicates crossed factors

Interaction	`A*B = A + B + A:B`	SAS: `A + B + A*B`
Nested	`A/B = A + A:B`	SAS: `A + B(A)`

- Hence, the model term `A:B` denotes interaction or nested effects depending on which other terms are previously included in the model.

Examples

```
asreml(fixed=volume~Site,
       random=~Genotype+Site:Genotype,data=MET)
asreml(fixed=volume~Site,
       random=~Site:Genotype,data=MET)
asreml(fixed=yield~A:B,random=~Block,data=potato)
```

JOB FILE

Model Functions

`and()`	overlays a design matrix over the previous one
`at()`	creates a binary variable for the condition specified in a factor
`factor()`	forms a factor with the values of a continuous variable
`lin()`	treats a factor as variates. The `lin()` does not center or scale the variables
`units`	creates a factor with level of each experimental unit; allows a second error term to be explicity fitted
`id()`	fits an additional factor without its genetic relationship matrix
`inv(v)`	calculates inverse of variable `v`
`log(v)`	calculates the natural logarithm of `v`
`pow(y,p)`	calculates the variable `y` to power `v`
`sqrt(v)`	calculates the square root of `v`
`spl(v,n)`	fits a spline for variable `v` with `n` knots
`pol(y,n)`	forms a set of orthogonal polynomials of order `n`

JOB FILE

Model Functions

`random` specifying the random effects part of the model with the terms

`sparse` specifying the fixed effects to be absorbed with the terms. This argument has the same general characteristics as fixed but there will be no left side to the expression

`rcov` specifying the error structure of the model

`G.param` representing variance structures of random terms of the model to hold initial parameter estimates and constraints.

`R.param` representing the error structure of the model to hold initial parameter estimates and constraints

`predict` named by classifying terms where each element is in term list with components `pvals`, `sed`, `cov` and `avsed`.

`constraits` a matrix specifying constraints among the variance components with the same row and columns as there are variance parameters.

`family` this option is under development and currently only gaussian with an identity link function is supported via the `asreml()`

JOB FILE

Model functions

`weigths` character or name identifying the column of data to use as weights in the fit

`offset` character or name identifying the column of data to include as an offset in the model

`na.method.Y` character to control filtering of missing values data in the response. Possibles values are `include`, `omit` and `fail`.

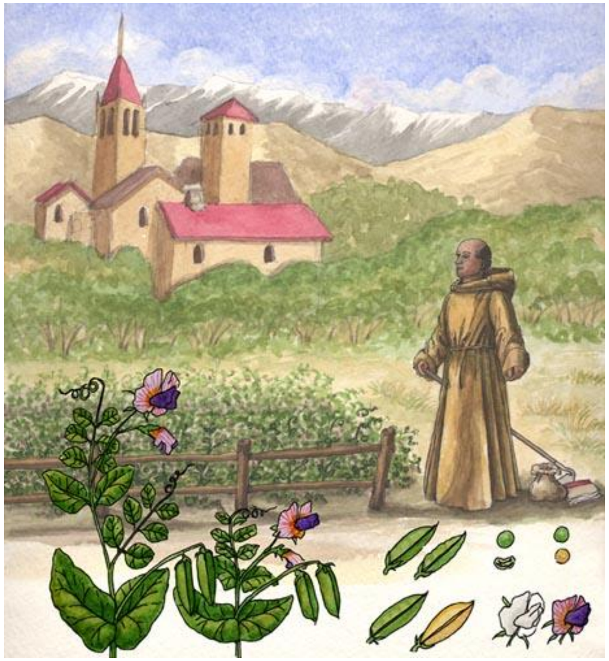
`na.method.X` character to control filtering of missing values data in the explanatory variates. Possibles values are `include`, `omit` and `fail`.

`keep.order` terms in the fixed formula will be keep in the order they are specified.

`model.frame` if `TRUE`, the model frame used in the fit is returned in the asreml object

`start.values` if `TRUE`, `asreml()` exits prior to the fitting process.

Session 4



Breeding Theory

PHENOTYPIC VALUE (Optional)

$$\mathbf{p = \mu + g + e}$$

- Phenotypic value (**p**) deviates from the mean (**μ**) because the genotypic component (**g**) and the environmental deviation (**e**).
- To isolate **g** we need to test the progeny!!!

$$\mathbf{g = a + d + i}$$

$$\mathbf{p = \mu + a + d + i + e}$$

a is the additive component, i.e. cumulative effect of the genes or breeding value (also known as GCA).

d is the dominance deviation, i.e. interaction between alleles or within-locus interaction (also known as SCA).

i is the epistatic deviation, i.e. between-loci interaction and higher order interactions.

e is the random deviation o residual.

VARIANCE COMPONENTS

- Partition of the variance is central to quantitative genetics and breeding, because is the way we *quantify* the relative importance of genetic and environmental influences (e.g. heritability).
- Partition is possible with data where the *resemblance* among relatives can be used to estimate genetic variance components.

$$\mathbf{V_p = V_g + V_e}$$
$$\mathbf{V_p = V_a + V_{na} + V_e}$$

where, $\mathbf{V_{na} = V_d + V_i}$ is the non-additive variance.

- In the statistical analysis (MM) the genetic variance estimates (e.g. $\mathbf{V_a}$) are obtained by relating them to the *causal component* (e.g. σ_a^2)

HERITABILITY

Broad sense heritability or degree of genetic determination

$\mathbf{H^2 = V_g / V_p}$ How much of the total variation is due to genetic causes (g). Important when working with clonally replicated individuals.

Narrow sense heritability

$\mathbf{h^2 = V_a / V_p}$ Extent to which phenotypes are determined by the genes transmitted from parents. Determines the degree of resemblance among relatives. The most important measure for breeding programs.

Heritabilities vary from 0 to 1 (e.g. 0.5 could be considered high).

Other definitions: family, plot-mean heritabilities and clonal repeatability

NON-ADDITIVE RATIOS

Dominance ratio

$d^2 = V_d / V_p$ How much of the total variation is due to dominance effects (d). Relevant when crosses are going to be deployed.

Epistatic ratio

$i^2 = V_i / V_p$ How much of the total variation is due to epistatic effects (i). Corresponds to the other portion of the non-additive genetic variance that is important when deploying clones or RILs.

`

BREEDING VALUE (BLUP)

Definition

- The **average effect** of the parental *alleles* passed to the offspring determine the mean genotypic value of its offspring, or
- The **genetic value** of an individual (or cross) judged by mean value of its progeny.

  - Sum of average effects across loci (theoretical, now molecular).
  - Mean value of offspring (practical).

- Not equivalent concepts if interaction between loci is present or if mating is not at random.

Estimation

- By **BLUP** (Best Linear Unbiased Predictor), i.e. the *prediction* of the random effects from linear mixed models.

BLUP (or EBLUP)

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$\hat{\mathbf{g}}$ vector of random effect predictions.

$\hat{\mathbf{G}}\mathbf{Z}' = \mathbf{C}'$ covariance matrix between observations and random (genetic) effects to be predicted.

$\hat{\mathbf{V}}$ variance-covariance matrix for the observations.

$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ individual observations ‘corrected’ by fixed effects.

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

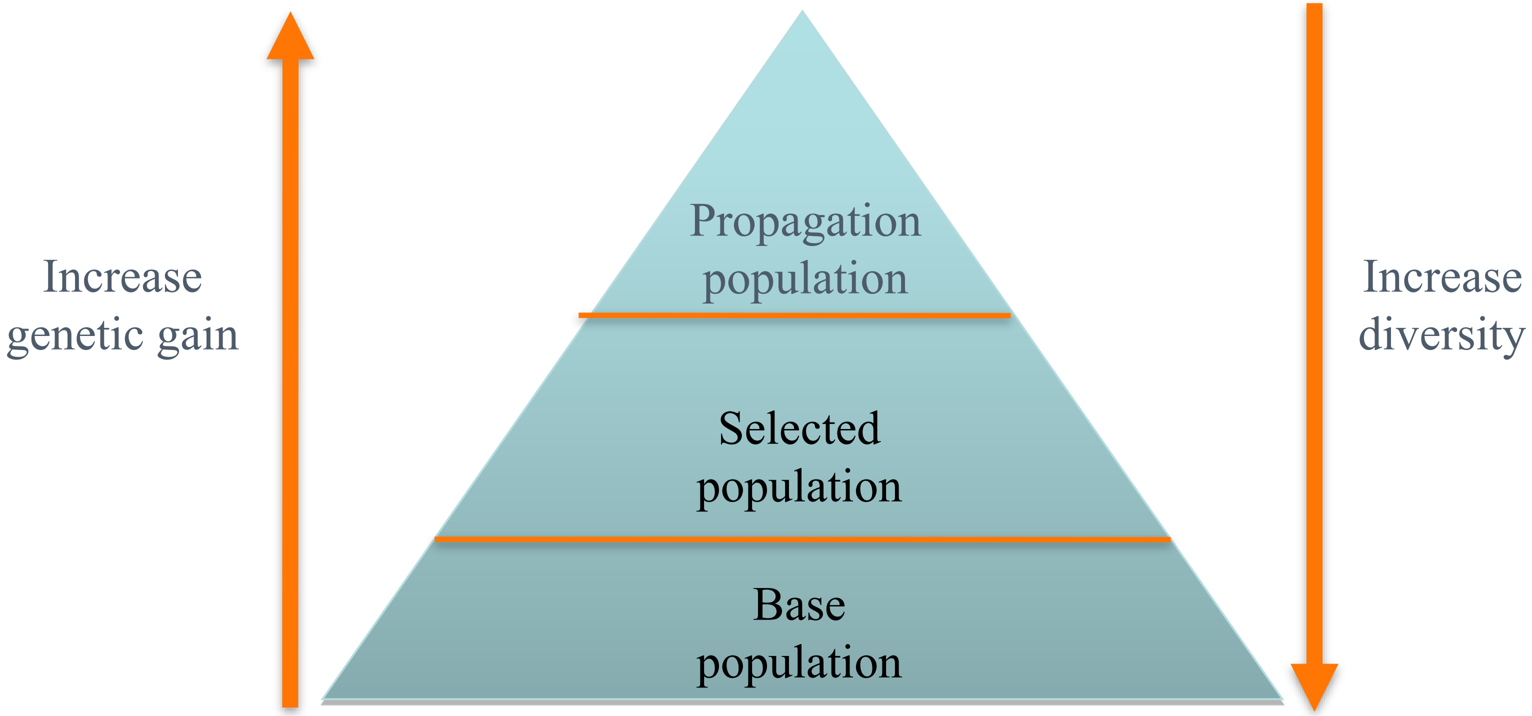
$$\hat{\mathrm{g}}_i = [\sigma_a^2 / \sigma_p^2] \times (y_i - \bar{y})$$

$$\hat{\mathrm{g}}_i = h^2 \times (y_i - \bar{y}) \qquad \rightarrow \ \Delta\text{Gain}$$

Note: the expression changes depending of what trait is being evaluated (**y**).

SELECTION

- All kind of selection have by aim to increase frequency of favourable alleles at loci influencing the selected trait(s).
- Types: mass, parental, family, combined, indirect, forward, backward.



GENETIC GAIN (ΔG_A)

ASReml®

- In mass selection, genetic gain can be quantified as the difference between the average breeding (e.g. additive) values from the selected and original population, i.e.

$$\Delta G_a = \bar{a}_S - \bar{a}_P = h^2 S$$

But $i = S / \sigma_p$ then

$$\Delta G_a = h^2 S = i\, h^2\, \sigma_p$$

- Genetic gain depends of the selection intensity (i), heritability (h^2) and the phenotypic standard deviation.
- Here i corresponded to the selection differential

($S = \mu_{\text{selected}} - \mu_{\text{population}}$) expressed in terms of phenotypic standard deviations.

TYPE-A CORRELATIONS

Definition: **Correlation between traits (pleitrophy)**

- Property of genes of influencing more than one phenotypic trait.
- It could be negative or positive (-1 to 1).
- Informs about the biological relationships among traits.
- Assists in the selection of 'good' individuals by looking into two traits simultaneously.

$$rg_{A(p)} = \frac{Cov(p_1, p_2)}{\sqrt{Var(p_1) \times Var(p_2)}} \qquad rg_{A(g)} = \frac{Cov(g_1, g_2)}{\sqrt{Var(g_1) \times Var(g_2)}}$$

Indirect Selection

$$\Delta G_{a1} = i_2 \times h_1 \times h_2 \times rg_{A(a)} \times \sigma_{p1}$$

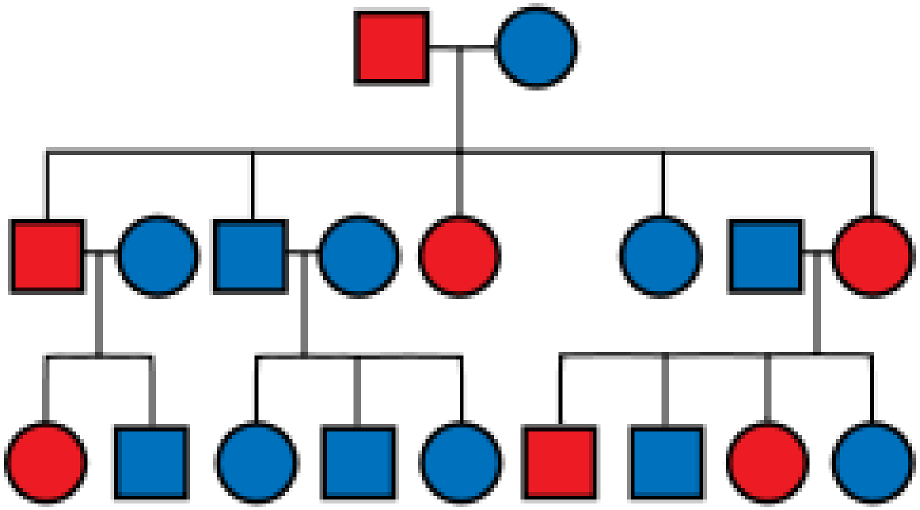
TYPE-B CORRELATIONS

Definition: **Correlation between sites**

- Is a relative expression of ***genotype-by-environment*** interaction.
- It could be zero or positive (0 to 1).
- A value close to 0 indicates that the rank in one environment is very different than the rank in another environment (i.e. low stability)
- A value close to 1 indicates that a single ranking can be used across all environments without loss of information (i.e. high stability).
- $\mathbf{V_{axs}}$ is the variance estimation of the site by genotype interaction.
- The following expressions represent the average correlation between sites (if more than 2 sites are analyzed).

$$rg^2_{B(a)} = \frac{\mathbf{V_a}}{\mathbf{V_a} + \mathbf{V_{axs}}} \qquad rg^2_{B(g)} = \frac{\mathbf{V_g}}{\mathbf{V_g} + \mathbf{V_{gxs}}}$$

Session 5



Parental Models

GENETIC MODELS

Parental Models

- **Half-sib crosses / sire model.**
  - One parent known. Parent selection.
- **Full-sib crosses model.**
  - Both parents known. Parent/cross selection. Add and Dom effects estimable.
- **Family model.**
  - Both parents known. Cross selection. Add and Dom effects confounded.
- **Clonal model.**
  - Clonally replicated individuals. Parent/cross/individual selection.

Individual Models

- **Animal model.**
  - One or two parents known. Individual/parent selection.
- **Reduced animal model.**
  - One or two parents known. Individual/parent selection (only individuals with records).

HALF-SIB / SIRE MODEL

General aspects

- One parent is known (mother, sire, variety).
- The other parent is assumed to be unknown and to mate at random.
- Only additive component (**Va**) can be estimated.
- Useful for selection of parents (backward selection).
- Parental pedigree can (and should) be incorporated.
- Runs faster than other models (e.g. animal model).

Difficulties

- Concern about situations under non-random mating.
- Selection does not capture non-additive genetic variability.

HALF-SIB / SIRE MODEL

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{s} + \mathbf{e}$$

y vector of observations

β vector of fixed effects

b vector of random design effects (e.g. block or plot effect), ~ $N(0, \mathbf{I}\sigma^2_b)$

s vector of random sire effects (i.e. ½ breeding value), ~ $N(0, \mathbf{A}\sigma^2_s)$

e vector of random residual effects, ~ $N(0, \mathbf{I}\sigma^2)$

X, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are incidence matrices

A is the numerator relationship matrix for sires. Replace by **I** if no pedigree.

I is an identity matrix

$$\mathbf{V_a} = 4\,\sigma^2_s \qquad \mathbf{V_p} = \sigma^2_b + \sigma^2_s + \sigma^2$$

$$h^2 = \mathbf{V_a} / \mathbf{V_p} = 4\,\sigma^2_s / [\sigma^2_b + \sigma^2_s + \sigma^2]$$

OPEN POLLINATION

Example: `/Day1/OpenPol/OPENPOL.txt`

A tree genetic study consisting on seeds from a total of 28 female parents were collected from mass selection and tested in a RCBD together with 3 control female parents. The experiment consisted in 10 replicates with 34 plots each of size 2 x 3. The response variables of interest are total height (HT, cm) and diameter at breast height (DBH, cm). For now we will concentrate in the response HT. The objective is to rank the female parents for future selections and seed production. *In this analysis parental pedigree will be ignored*. Note that a model can be fitted with and without the controls included as parents.

```
ID      REP     PLOT    FEMALE  TYPE    DBH     HT
1       1       1       FEM1    Test    23.8    12.4
2       1       1       FEM1    Test    24.4    12.1
3       1       1       FEM1    Test    25.4    10.9
4       1       1       FEM1    Test    28.0    12.7
5       1       1       FEM1    Test    20.9    11.9
6       1       1       FEM1    Test    22.6    11.2
7       1       2       FEM15   Test    22.4    10.7
8       1       2       FEM15   Test    21.9    11.6
9       1       2       FEM15   Test    20.8    11.3
...
```

FULL-SIB MODELS

General Aspects

- Both parents are known (mother, father, family or cross).
- Mating is often planned (e.g. diallels).
- Additive and dominance component (**V_a** and **V_d**) can be estimated.
- Some studies allow to obtain common environment, reciprocals, etc.
- Useful for selection of parents (backward selection) or specific crosses.
- Increased gain as dominance effects can be 'captured'.
- Parental pedigree can be incorporated.

Difficulties

- Dominance effects usually estimated with low precision, or confounded with other effects.
- Better results obtained with a proper planning of crosses (e.g. connected diallels).
- Need to check connectivity and number of crosses per parent (male and female) otherwise this model cannot be fitted.

FULL-SIB: CLASSIC APPROACH

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 b} + \mathbf{Z_2 m} + \mathbf{Z_3 f} + \mathbf{Z_4 mf} + \mathbf{e}$$

β vector of fixed effects (e.g. μ, replicate)

b vector of random design effects (e.g. block or plot effect), ~ N(0, $\mathbf{I}\sigma^2_b$)

m vector of random male effects (i.e. ½ BV), ~ N(0, $\mathbf{A}\sigma^2_m$)

f vector of random female effects (i.e. ½ BV), ~ N(0, $\mathbf{A}\sigma^2_f$)

mf vector of random interaction male by female effects, ~ N(0, $\mathbf{I}\sigma^2_{mf}$)

e vector of random residual effects, ~ N(0, $\mathbf{I}\sigma^2$)

$$\mathbf{V_a} = 2\,(\sigma^2_m + \sigma^2_f) \qquad \text{or} \qquad \mathbf{V_a} = 4\,\sigma^2_m \quad (\text{when } \sigma^2_m = \sigma^2_f)$$

$$\mathbf{V_d} = 4\,\sigma^2_{mf}$$

$$\mathbf{V_p} = \sigma^2_b + \sigma^2_m + \sigma^2_f + \sigma^2_{mf} + \sigma^2$$

$$h^2 = \mathbf{V_a} / \mathbf{V_p} = [2\,(\sigma^2_m + \sigma^2_f)] / [\sigma^2_b + \sigma^2_m + \sigma^2_f + \sigma^2_{mf} + \sigma^2]$$

$$d^2 = \mathbf{V_d} / \mathbf{V_p} = 4\,\sigma^2_{mf} / [\sigma^2_b + \sigma^2_m + \sigma^2_f + \sigma^2_{mf} + \sigma^2]$$

FULL-SIB: CLASSIC

Example: `/Day1/ContPol/CONTPOL.txt`

A total of 177 families and 8 checklots were planted in a test using a RCBD with 25 blocks. For all families planted both parents are known. *In this analysis a dummy parental pedigree will be considered*. The objective is to estimate the different variance components, and calculate heritabilities for the response variable *YIELD*.

```
REP     FAMILY    FEMALE    MALE      YIELD     CHECKLOT
1       FAM007    PAR0001   PAR0024   128.68    0
1       FAM163    PAR0059   PAR0041   119.462   0
1       C10       C10       C10       NA        1
1       FAM040    PAR0020   PAR0053   103.641   0
1       FAM114    PAR0051   PAR0001   NA        0
1       FAM053    PAR0032   PAR0032   NA        0
1       FAM048    PAR0031   PAR0018   NA        0
1       FAM057    PAR0033   PAR0035   155.226   0
1       FAM120    PAR0051   PAR0051   NA        0
1       FAM165    PAR0059   PAR0059   193.982   0
1       FAM133    PAR0053   PAR0009   184.308   0
1       FAM057    PAR0035   PAR0033   NA        0
1       C30       C30       C30       141.912   1
1       FAM082    PAR0044   PAR0006   288.692   0
1       FAM060    PAR0034   PAR0037   NA        0
1       FAM169    PAR0015   PAR0024   245.664   0
1       FAM047    PAR0031   PAR0016   NA        0
...
```

FAMILY MODEL (Optional)

General Aspects

- More common in animal breeding
- Occurs when parents are only present in a single cross.
- Parents might, or might not, be known.
- Additive and dominance component (V_a and V_d) can not be separated, unless there is a well connected parental pedigree.
- Useful for family selection or forward selection.
- Of practical use when dominance variance is known to be negligible.

Difficulties

- Dominance effects are confounded with additive effects.
- Potentially it could over-estimate future genetic gain.

FAMILY MODEL (Optional)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2 F + \mathbf{e}$$

β vector of fixed effects (e.g. μ, replication)

b vector of random design effects (e.g. block or plot effect), ~ N(0, $\mathbf{I}\sigma^2_{\mathrm{b}}$)

F vector of random family effects, ~ N(0, $\mathbf{A}\sigma^2_F$) or N(0, $\mathbf{I}\sigma^2_F$)

e vector of random residual effects, ~ N(0, $\mathbf{I}\sigma^2$)

$$\sigma^2_F = \mathbf{V_a}/2 + \mathbf{V_d}/4$$

$$\mathbf{V_p} = \sigma^2_{\mathrm{b}} + \sigma^2_F + \sigma^2$$

$$h^2_{cross} = \mathbf{V_{family}} / \mathbf{V_p} = \sigma^2_F \;/\; [\sigma^2_{\mathrm{b}} + \sigma^2_F + \sigma^2]$$

$\mathbf{V_a}$ and $\mathbf{V_d}$ can not be separated unless we assumed that $\mathbf{V_d} = 0$

If $\mathbf{V_d} = 0$ then $\mathbf{V_a} = 2\ \sigma^2_F$

$$h^2 = \mathbf{V_a} / \mathbf{V_p} = 2\ \sigma^2_F \;/\; [\sigma^2_{\mathrm{b}} + \sigma^2_F + \sigma^2]$$

FAMILY MODEL (Optional)

Example: `/Day1/FamilyM/FISHF.txt`

A total of 459 fish were derived from single parental crosses composed of 32 sires and 32 females to generate 32 families. Number of individuals per family varied form 2 to 40. The idea is to rank the families and progeny for selection by using the variable `Weight`.

```
ID        SireID    DamID     Family    Weight
1001      120       125       22        88.3
1002      120       125       22        84.9
1003      120       125       22        76.8
1004      121       114       23        95.4
1005      121       114       23        85.4
1006      121       114       23        74.8
1007      121       114       23        103.4
1008      121       114       23        78.7
1009      121       114       23        109.5
1010      121       114       23        113.1
1011      121       114       23        95.4
1012      121       114       23        91.1
1013      121       114       23        85.4
1014      121       114       23        85.4
1015      121       114       23        86.0
...
```

CLONAL MODEL (Optional)

General aspects

- It can estimated total genetic variability ($\mathbf{V_g}$).
- If both parents are known (mother, father, family or cross) then the additive, dominance and epistasis components ($\mathbf{V_a}$, $\mathbf{V_d}$ and $\mathbf{V_i}$) can be reasonably estimated.
- Useful for selection of parents (backward selection), crosses or specific genotypes.
- Allows to capture, in new generations, additive, dominance and epistasis effects.

Difficulties

- Presents same difficulties as full-sib models.
- Some confounding of the epistasis component occurs (higher order terms).
- Occasionally produces negative causal variance components.

CLONAL MODEL (Optional)

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{f} + \mathbf{Z}_4\mathbf{mf} + \mathbf{Z}_5\mathbf{mf.c} + \mathbf{e}$$

β and **b** as defined before

m vector of random male effects, ~ N(0, $\mathbf{A}\sigma^2_m$)

f vector of random female effects, ~ N(0, $\mathbf{A}\sigma^2_f$)

mf vector of random interaction male by female effects, ~ N(0, $\mathbf{I}\sigma^2_{mf}$)

mf.c vector of random clonal within family effects, ~ N(0, $\mathbf{I}\sigma^2_c$)

e vector of random residual effects, ~ N(0, $\mathbf{I}\sigma^2$)

$V_a = 2\ (\sigma^2_m + \sigma^2_f)$ or $V_a = 4\ \sigma^2_m$ (when $\sigma^2_m = \sigma^2_f$)

$V_d = 4\ \sigma^2_{mf}$ $V_i = \sigma^2_c - (\sigma^2_m + \sigma^2_f) - 3\ \sigma^2_{mf}$ (approx.)

$$V_g = V_a + V_d + V_i$$

$$V_p = \sigma^2_b + \sigma^2_m + \sigma^2_f + \sigma^2_{mf} + \sigma^2_c + \sigma^2$$

$H^2 = V_g / V_p$ $h^2 = V_a / V_p$ $d^2 = V_d / V_p$ $i^2 = V_i / V_p$

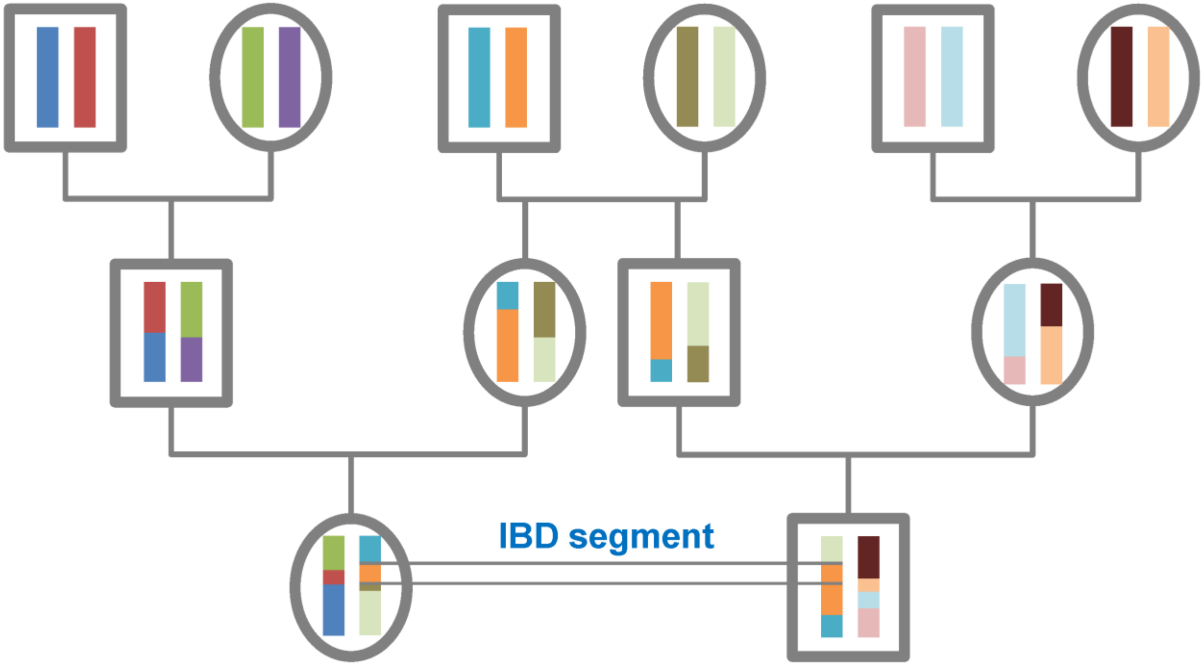
CLONAL MODEL (Optional)

Example: `/Day1/Clonal/CLONES.txt`

A clonal test derived from a total of 61 families crossed in a circular mating design were established in a field trial with 3 repetitions and incomplete blocks. Each family has several clones. The objective of this study is to estimate all variance components (additive, dominance and epistasis).

```
IDSORT    FamilyID  Female    Male      cloneid   Rep       IncBlock  Tree      VOL
1         46        Par927    Par931    677       1         1         1         537.7436
2         33        Par908    Par914    476       1         1         2         492.1155
3         53        Par924    Par907    775       1         1         3         704.826
4         41        Par913    Par917    608       1         1         4         494.6012
6         27        Par923    Par905    391       1         2         1         622.0541
7         14        Par925    Par908    192       1         2         2         425.1107
8         22        Par913    Par923    304       1         2         3         298.8255
9         11        Par929    Par920    144       1         2         4         513.8072
11        23        Par901    Par924    320       1         3         1         457.7191
12        60        Par929    Par904    838       1         3         2         709.3598
15        12        Par917    Par921    162       1         3         5         NA
16        53        Par924    Par907    763       1         4         1         392.4941
17        13        Par901    Par916    179       1         4         2         463.7218
19        24        Par915    Par904    340       1         4         4         445.3584
20        40        Par922    Par917    592       1         4         5         623.984
21        30        Par904    Par903    424       1         5         1         439.2273
...
```

Session 6



Incorporating Pedigree

GENETIC MODELS

Parental Models

- **Half-sib crosses / sire model.**
  - One parent known. Parent selection.
- **Full-sib crosses model.**
  - Both parents known. Parent/cross selection. Add and Dom effects estimable.
- **Family model.**
  - Both parents known. Cross selection. Add and Dom effects confounded.
- **Clonal model.**
  - Clonally replicated individuals. Parent/cross/individual selection.

Individual Models

- **Animal model.**
  - One or two parents known. Individual/parent selection.
- **Reduced animal model.**
  - One or two parents known. Individual/parent selection (only individuals with records).

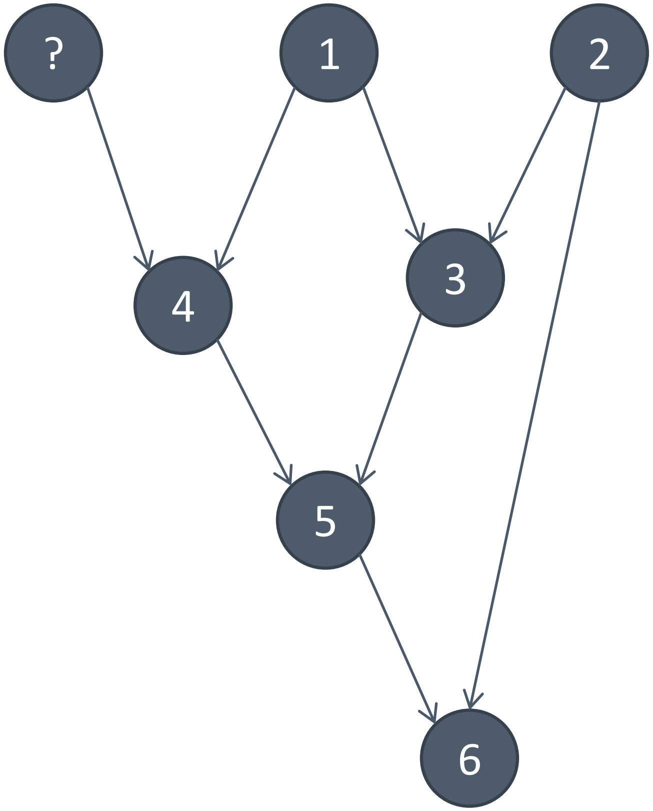
INCORPORATING PEDIGREE

- Why worry about the pedigree in genetic analyses?
  - Statistically, random genetic effects (i.e. BLUPs) are not independent and their matrix of correlations or co-variances (**G** or **A**) needs to be specified.
  - Genetically, it is important to consider information about relatives as they will share some alleles, and therefore their response is correlated.
- How to incorporate this information?
  - *Genetic relationships* can be calculated using genetic theory (expected values) or molecular information (e.g. SNPs), and included into the linear mixed model by specifying a pedigree file,
- Are there other benefits?
  - Many. It is a more efficient use of the information about individuals, but also genetic values of individual not tested, but with relatives tested, can be *predicted* and selected.

PEDIGREE

Example

Pedigree of a group of individuals:



Individual	Male	Female
3	1	2
4	1	Unknown
5	4	3
6	5	2

PEDIGREE

Numerator relationship matrix (A)

$$\mathbf{A} = \begin{array}{cc} & \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \begin{bmatrix} 1.00 & 0.00 & 0.50 & 0.50 & 0.50 & 0.25 \\ & 1.00 & 0.50 & 0.00 & 0.25 & 0.625 \\ & & 1.00 & 0.25 & 0.625 & 0.563 \\ & & & 1.00 & 0.625 & 0.313 \\ & & & & 1.125 & 0.688 \\ & & & & & 1.125 \end{bmatrix} \end{array}$$

- Linked to the concept of **identity by descent**.
- **Diagonal** $a_{ii} = 1 + F_i$ (inbreeding coefficient on individual *i*)
- Twice the probability that two gametes taken at random from animal *i* will carry identical alleles by descent.
- **Off-diagonal** a_{ij} numerator of the coefficient of relationship between animal *i* and *j*.
- Several algorithms are available to obtain this matrix.

PEDIGREE

Obtaining the A matrix

- Let $\mathbf{A} = \{a_{ij}\}$ be the relationship matrix.
- Let $a_{i,-j}$ the the i-th row of **A** except for the j-th element.
- Assume the relationship matrix for the base animals is known (e.g. unrelated, non inbred). This will for a base matrix (e.g. identity)
- The row of the relationship matrix for the progeny of two parents is generates as the average of the relationship matrix rows for the parents:

$$a_{i,-j} = (a_{s,-i} + a_{d,-i})/2$$

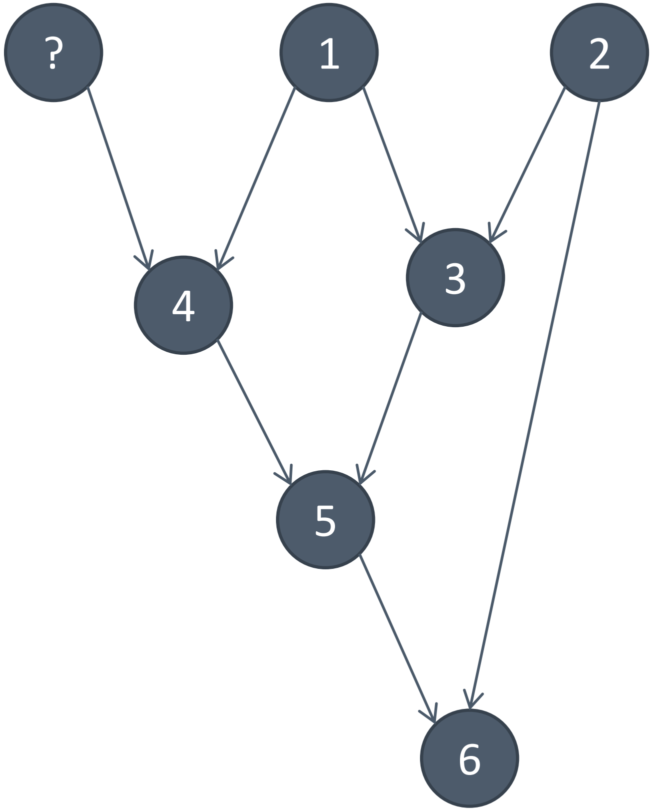
- The diagonal element, $a_{i,i}$ of this new individual is:

$$a_{i,i} = 1 + a_{s,d}/2 = 1 + F_i$$

where F_i is the inbreeding coefficient.

PEDIGREE FILE

Graphically



In ASReml

Indiv	Male	Female
1	0	0
2	0	0
3	1	2
4	1	0
5	4	3
6	5	2

```
pedind<-read.table("PEDIND.txt",h=T)
ainv<-asreml.Ainverse(pedind)$ginv

asreml(...,ginverse=list(Indiv=ainv=ainv),...)
```

PEDIGREE FILE

In ASReml-R

- Pedigree file can be part of the data file

  (first 3 columns: *individual*, *parent1* and *parent2*).
- Method used to construct the **A** inverse s based on the algorithm of Meuwissen and Luo (1992).
- Genetic groups can be defined here and there are many other options.

Some Useful Options

`ginv`	data frame with 3 columns holding the lower triangle of the inverse of relationship matrix in sparse form.
`inbreeding`	the inbreeding coefficient for each individual. .
`ainv`	the diagonal elements of the inverse relationship matrix
`det`	the determinant.
`selfing`	allows for partial selfing according to variable when the third field of pedigree is unknown.
`groups`	includes genetic groups in the pedigree according to variable `g`.
`msg`	if TRUE, the third identity in the pedigree file is the male parent of the female parent rather than female parent.

PEDIGREE FILE

Construction / Check

- Pedigree information is associated with proper management and validation/check of data.
- Individuals need to be ordered by generation (e.g. parents need to be defined before progeny).
- All parents need to be defined in pedigree file (the inclusion of founder parents is optional).
- All individuals present in dataset (i.e. levels associated with pedigree file) need to be defined in pedigree file.
- Individuals can be defined as male or female parents (but this should be checked if is not biologically possible).

Session 7

Animal Models

ANIMAL / INDIVIDUAL MODEL

General aspects

- Requires defining individual and parental pedigree.
- A breeding value (or GCA) is obtained for each individual in the dataset, and for all individuals (e.g. parents) in pedigree file.
- Typically used to estimates additive component (V_a) only, but it can be extended to non-additive and maternal effects.
- Useful for selection of individuals based on additive values (forward selection) but can be also used to select parents.
- GCA values (or EBV) of parents will be proportional to a parental model.

Difficulties

- For large datasets it can be computationally costly.
- Pedigree file could be difficult to construct/maintain and it needs to be checked carefully.

ANIMAL / INDIVIDUAL MODEL

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{a} + \mathbf{e}$$

β vector of fixed effects

b vector of random design effects (e.g. block effect), ~ N(0, $\mathbf{I}\sigma^2_b$)

a vector of random additive effects (i.e. BV), ~ N(0, $\mathbf{A}\sigma^2_a$)

e vector of random residual effects, ~ N(0, $\mathbf{I}\sigma^2$)

$$\mathbf{V_a} = \sigma^2_a$$

$$\mathbf{V_p} = \sigma^2_b + \sigma^2_a + \sigma^2$$

$$h^2 = \mathbf{V_a} / \mathbf{V_p} = \sigma^2_a / [\sigma^2_b + \sigma^2_a + \sigma^2]$$

Note: any individual that are included in the pedigree file will have a prediction of its breeding values (those that without phenotypes).

ANIMAL / INDIVIDUAL MODEL

Example: `/Day1/Fish/FISH.txt`

The dataset for a fish breeding program contains a total of 933 records of fish. The objective is to fit an animal model that considers the complete pedigree. The parental pedigree is found in the file `PEDPAR.txt`, *but an individual pedigree needs to be constructed*. For fitting the model consider the factor `SEX` as a covariate. The response of interest is days to market size (`DAYSM`).

```
INDIV    Sire     Dam      FAM      DaysM    Sex      Market
1001     564      727      564-727  741.46   1        1
1002     564      727      564-727  500.09   2        1
1003     564      727      564-727  495.07   1        1
1004     564      727      564-727  506.25   2        0
1005     564      727      564-727  593.21   2        1
1006     564      727      564-727  671.1    1        1
1007     564      727      564-727  523.48   1        1
1008     564      727      564-727  531.33   1        1
1009     564      727      564-727  446.02   2        0
1010     564      727      564-727  599.2    1        1
1011     564      727      564-727  509.38   2        1
1012     564      727      564-727  643.45   2        1
1013     607      707      607-707  711.68   1        1
...
```

ANIMAL / INDIVIDUAL MODEL

Additional Aspects

- When pedigree is available from several generations, usually more than 3 generations does not produce a significant improvement on precision of estimates.

- Incorporation of genetic groups is critical in order to consider previous achieved genetic gains, and to describe the proper structure of the data.

- Reduced animal model (RAM), it is an alternative that runs faster as only animals with records are considered.

- Other variants exist of the animal model exist that consider:
  - **Environmental effects.**
  - **Maternal effects**
  - **Genetic maternal effects**
  - **Model with non-additive genetic effects (mainly dominance)**
  - **Common environment (CE) effects**

CE EFFECTS (Optional)

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1 b} + \mathbf{Z_2 a} + \mathbf{Z_3 ce} + \mathbf{e}$$

β vector of fixed effects

b vector of random design effects (e.g. block effect), $\sim N(0, \mathbf{I}\sigma^2_b)$

a vector of random additive effects (i.e. BV), $\sim N(0, \mathbf{A}\sigma^2_a)$

ce vector of random common environmental effects, $\sim N(0, \mathbf{I}\sigma^2_{ce})$

e vector of random residual effects, $\sim N(0, \mathbf{I}\sigma^2)$

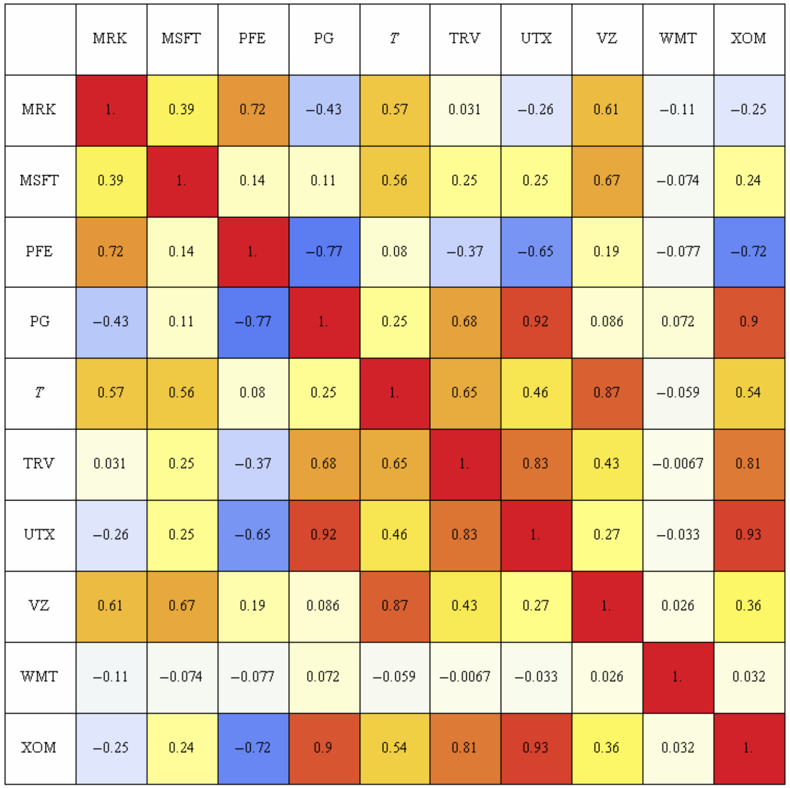
$$\mathbf{V_a} = \sigma^2_a$$

$$\mathbf{V_p} = \sigma^2_b + \sigma^2_a + \sigma^2_{ce} + \sigma^2$$

$$h^2 = \mathbf{V_a} / \mathbf{V_p} = \sigma^2_a / [\sigma^2_b + \sigma^2_a + \sigma^2_{ce} + \sigma^2]$$

Note: common environment effects are non-genetic effects that causes resemble between members of the same family.

Session 8



	MRK	MSFT	PFE	PG	*T*	TRV	UTX	VZ	WMT	XOM
MRK	1.	0.39	0.72	−0.43	0.57	0.031	−0.26	0.61	−0.11	−0.25
MSFT	0.39	1.	0.14	0.11	0.56	0.25	0.25	0.67	−0.074	0.24
PFE	0.72	0.14	1.	−0.77	0.08	−0.37	−0.65	0.19	−0.077	−0.72
PG	−0.43	0.11	−0.77	1.	0.25	0.68	0.92	0.086	0.072	0.9
T	0.57	0.56	0.08	0.25	1.	0.65	0.46	0.87	−0.059	0.54
TRV	0.031	0.25	−0.37	0.68	0.65	1.	0.83	0.43	−0.0067	0.81
UTX	−0.26	0.25	−0.65	0.92	0.46	0.83	1.	0.27	−0.033	0.93
VZ	0.61	0.67	0.19	0.086	0.87	0.43	0.27	1.	0.026	0.36
WMT	−0.11	−0.074	−0.077	0.072	−0.059	−0.0067	−0.033	0.026	1.	0.032
XOM	−0.25	0.24	−0.72	0.9	0.54	0.81	0.93	0.36	0.032	1.

Variance Structures in ASReml-R

VARIANCE STRUCTURES

Direct Sum

- The desired matrix is specified by several square matrices in a block diagonal matrix.

Example

$$\mathbf{R} = \oplus_{\mathrm{j}=1}^{3} \mathbf{R}_j = diag(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3) = \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_3 \end{bmatrix}$$

VARIANCE STRUCTURES

Direct Product

- Variance structures are specified by using direct products or two or more matrices (⊗, or Kronecker product).

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix}$$

Example

$$\mathbf{A} = \begin{array}{c} g_1 \\ g_2 \\ g_3 \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{B} = \begin{array}{c} t_1 \\ t_2 \end{array} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

$$\mathbf{A} \otimes \mathbf{B} = \begin{array}{c} g_1 t_1 \\ g_1 t_2 \\ g_2 t_1 \\ g_2 t_2 \\ g_3 t_1 \\ g_3 t_2 \end{array} \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_{12} & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

DISEASE RESISTANCE

Example: `/Day2/VarStruct/LEAFAREA.TXT`

This trial investigates the resistance of 12 varieties of a plant to a soil borne disease. The trial was done in a glasshouse based on a randomized complete block design (RCBD) with 10 blocks. Each of these blocks consisted in 24 pots, with a single plant per pot, which had randomly assigned one of the 12 varieties and one of the two types of soil: healthy (`H`) or infected (`I`). Therefore, we have a 12 x 2 factorial experiment with the two treatment factors: `variety` and `disease`. In this study, the response variable corresponds to total leaf area (in cm) of each plant, and `variety` will be considered random.

```
id block pot variety disease trt leafarea
   1    1    1 P         H       H_P    147.7
   2    2    1 P         H       H_P    110.6
   3    3    1 P         H       H_P     93.9
   4    4    1 P         H       H_P     89.6
   5    5    1 P         H       H_P     98.5
   6    6    1 P         H       H_P     88.9
   7    7    1 P         H       H_P    107.4
...
```

VARIANCE STRUCTURES

id: identity

$$\sigma^2\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix}=\begin{bmatrix}\sigma^2&0&0&0\\0&\sigma^2&0&0\\0&0&\sigma^2&0\\0&0&0&\sigma^2\end{bmatrix}$$

ar1v: autocorrelation 1st order

$$\sigma^2\begin{bmatrix}1&\rho^1&\rho^2&\rho^3\\\rho^1&1&\rho^1&\rho^2\\\rho^2&\rho^1&1&\rho^1\\\rho^3&\rho^2&\rho^1&1\end{bmatrix}$$

diag: diagonal

$$\begin{bmatrix}\sigma_1^2&0&0&0\\0&\sigma_2^2&0&0\\0&0&\sigma_3^2&0\\0&0&0&\sigma_4^2\end{bmatrix}$$

corh: uniform heterogeneous

$$\begin{bmatrix}\sigma_1^2&\rho\sigma_1\sigma_2&\rho\sigma_1\sigma_3&\rho\sigma_1\sigma_4\\\rho\sigma_1\sigma_2&\sigma_2^2&\rho\sigma_2\sigma_3&\rho\sigma_2\sigma_4\\\rho\sigma_1\sigma_3&\rho\sigma_2\sigma_3&\sigma_3^2&\rho\sigma_3\sigma_4\\\rho\sigma_1\sigma_4&\rho\sigma_2\sigma_4&\rho\sigma_3\sigma_4&\sigma_4^2\end{bmatrix}$$

corv: uniform correlation

$$\sigma^2\begin{bmatrix}1&\rho&\rho&\rho\\\rho&1&\rho&\rho\\\rho&\rho&1&\rho\\\rho&\rho&\rho&1\end{bmatrix}=\begin{bmatrix}\sigma_1^2&\sigma_2^2&\sigma_2^2&\sigma_2^2\\\sigma_2^2&\sigma_1^2&\sigma_2^2&\sigma_2^2\\\sigma_2^2&\sigma_2^2&\sigma_1^2&\sigma_2^2\\\sigma_2^2&\sigma_2^2&\sigma_2^2&\sigma_1^2\end{bmatrix}$$

us: unstructured

$$\begin{bmatrix}\sigma_{11}^2&\sigma_{12}^2&\sigma_{13}^2&\sigma_{14}^2\\\sigma_{12}^2&\sigma_{22}^2&\sigma_{23}^2&\sigma_{24}^2\\\sigma_{13}^2&\sigma_{23}^2&\sigma_{33}^2&\sigma_{34}^2\\\sigma_{14}^2&\sigma_{24}^2&\sigma_{34}^2&\sigma_{44}^2\end{bmatrix}$$

CORRELATION STRUCTURES

cor: uniform correlation

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

corb: banded correlation

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

ar1: autocorrelation 1st order

$$\begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}$$

corg: general correlation

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

VARIANCE STRUCTURES

Variance models (VCODE)

Common structures

`id`	Identity	**1**
`diag`	Diagonal	w
`us`	Unstructured	$w(w + 1)/2$
`ainv`	Numerator relationship matrix (**A**)	**0** or **1**
`cor`	Uniform correlation	**1**

Correlation/Spatial structures

`corb`	Banded correlation	w-1
`ar1`	First order autoregressive	**1**
`ar2`	Second order autoregressive	**2**
`arma`	Autoregressive and moving average	**2**
`corg`	General correlation (homogeneous)	$w(w - 1)/2$
`ante1`	Antedependence of order 1	$w(w - 1)/2$
`lvr`	Linear variance	**1**

VARIANCE STRUCTURES

Correlation-variance structures (homogeneous)

`arv1`	First order autoregressive (homog.)	**2**
`corv`	Uniform correlation (homogenoeus)	**2**
`corbv`	Banded correlation (homogeneos)	w
`corgv`	general correlation (homogeneous)	$w(w - 1)/2 + 1$

Heterogeneous structures

`idh = diag`	Identity (heterogenoeus)	w
`ar1h`	First order autoregressive (heterog.)	$1 + w$
`corh`	Uniform correlation (heterogeneous)	$1 + w$
`corbh`	Banded correlation (heterogeneos)	$2w - 1$
`corgh = us`	general correlation (heterogeneous)	$w(w - 1)/2 + w$

Special structures

`iexp`	Isotropic Exponential	**1**
`aexp`	Anisotropic Exponential	**2**
`giv`	User supplied General (Inverse) matrix	**0 or 1**

VARIANCE STRUCTURES

Direct Product

```
random=~ped(Genotype):us(Site)
rcov=~units:us(trait)
```

- Specifies a different variance structure for each factor term.
- Default is identity (`id`).
- Units is used as a counting factor.

Direct Sum

```
random=~at(Site):incblock
rcov=~at(Site):units
```

- Defines a block diagonal variance structure.
- For residual terms, it requires that the data is sorted by the factor of interest.
- Default is identity (`id`).

VARIANCE STRUCTURES

- Order of starting values for variance and correlation matrices is important

Variance Matrices

$$\begin{bmatrix} 1 & 2 & 4 & 7 \\ - & 3 & 5 & 8 \\ - & - & 6 & 9 \\ - & - & - & 10 \end{bmatrix}$$

Correlation Matrices

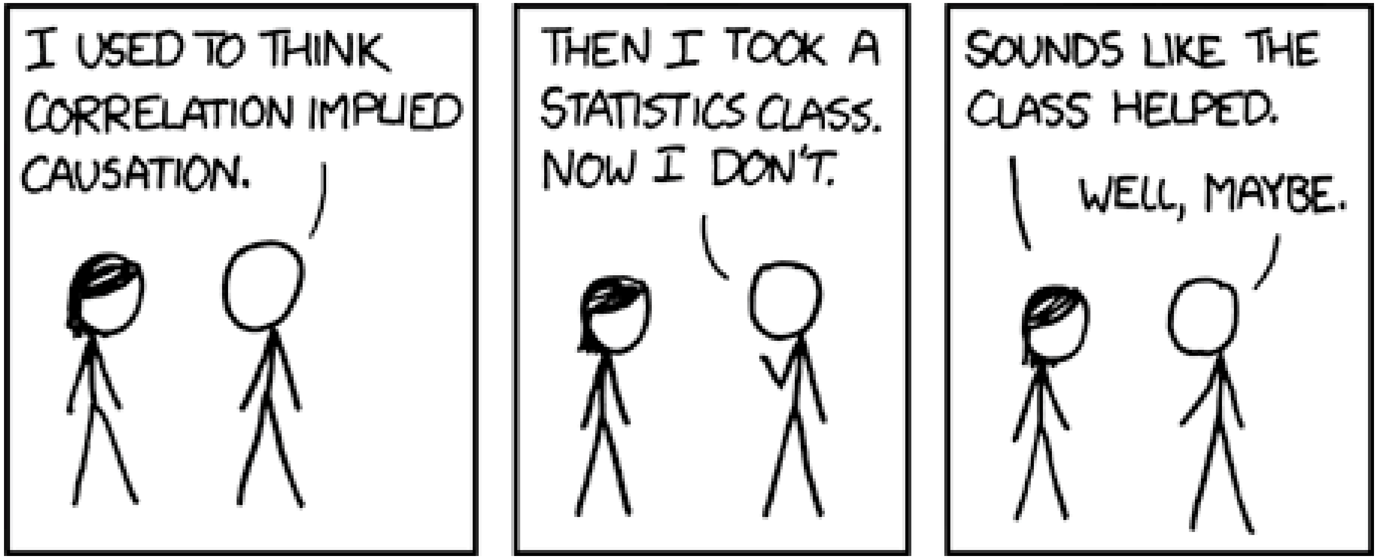
$$\begin{bmatrix} 7 & 1 & 2 & 4 \\ - & 8 & 3 & 5 \\ - & - & 9 & 6 \\ - & - & - & 10 \end{bmatrix}$$

Note: for most complex variance structures it is critical to specify starting values.

Examples

```
random=~ped(Genotype,init=5)
random=~id(Block,init=1)
rcov=~units:us(trait,init=c(12,3.5,7))
```

Session 9


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Multivariate Analysis / Repeated Measures

MULTIVARIATE ANALYSIS

General Uses

- More *efficient* analysis that combines information on two or more response variables.
- Produces an improvement on the precision of the breeding values (BLUPs).
- Allows to estimate *correlations* among traits (e.g. phenotypic and genetic correlations).
- Assists in *predicting* individual breeding values for traits that were not measured (but they need to be correlated).
- Relevant to assess importance of *indirect selection*.
- Can be used to combine different sources of, complete or incomplete, sources of data.
- Generates the required matrices to construct a *selection index*.
- Recommended analysis for cases where a prior selection was done based in a trait.

BIVARIATE ANALYSIS

- Considers a 2 x 2 matrix for each effect, e.g. $\mathbf{V}(\mathbf{g}_i) = \begin{array}{c} \\ g_1 \\ g_2 \end{array} \begin{array}{cc} g_1 & g_2 \\ \left[\begin{matrix} \sigma^2_{t1} & \sigma_{t1t2} \\ \sigma_{t1t2} & \sigma^2_{t2} \end{matrix}\right] \end{array}$

In ASReml-R

- Uses individual stacked responses: $y_i = [y_{i(1)}\ y_{i(2)}]'$ (for all i).
- The word `Trait` is used to defined the stacked response vector.
- Typically genetic and error effects are defined with a `un` variance structure.
- Other effects can be defined as `us` or `diag` structures.
- It is also recommended to use some of the correlation to maintain parameter space.

$$\mathbf{y} = \begin{bmatrix} y_{1t1} \\ y_{2t2} \\ y_{2t1} \\ y_{2t2} \\ . \\ . \\ . \\ y_{nt1} \\ y_{nt2} \end{bmatrix}$$

BIVARIATE ANALYSIS

Strategy for fitting models in ASReml-R

- Sensible to initial starting values (for any multivariate analysis).
- Strategy: start with univariate analysis and add one variable at the time.
- Get rough estimates: estimate phenotypic or genetic correlations / covariances using univariate solutions, or prior knowledge.
- Favour simple correlation structures if you have problems, e.g. coru, diag.

OPEN POLLINATION

Example: `/Day2/Bivar/OPENPOL.txt`

A tree genetic study consisting on seeds from a total of 28 female parents were collected from mass selection and tested in a RCBD together with 3 control female parents. The experiment consisted in 10 replicates with 34 plots each of size 2 x 3. The response variables of interest are total height (`HT`, cm) and diameter at breast height (`DBH`, cm). For now we will concentrate in the response `HT`. The objective is to rank the female parents for future selections and seed production. Note that a model can be fitted with and without the controls included as parents.

```
ID      REP     PLOT    FEMALE  TYPE    DBH     HT
1       1       1       FEM1    Test    23.8    12.4
2       1       1       FEM1    Test    24.4    12.1
3       1       1       FEM1    Test    25.4    10.9
4       1       1       FEM1    Test    28.0    12.7
5       1       1       FEM1    Test    20.9    11.9
6       1       1       FEM1    Test    22.6    11.2
7       1       2       FEM15   Test    22.4    10.7
8       1       2       FEM15   Test    21.9    11.6
9       1       2       FEM15   Test    20.8    11.3
10      1       2       FEM15   Test    21.6    13.3
...
```

MULTIVARIATE ANALYSIS

Strategy for fitting models in ASReml

- For fitting model use same strategies as for bivariate analysis.
- Standardized responses, particularly when variables have different scales.
- Implement simple structures first (e.g. `id`, `diag`, `corv`, `corgv`).
- Correlation variance structures (`corh`, `corbh`, `corgh`) tend to give better results.
- Be aware that it might not fit at all!

Extensions

- Consider different sites (or years) as different traits (e.g. helps to classify sites).
- Variance-covariance matrices can be used to 'study' genetic structure (e.g. evaluating / separating genetic groups).

REPEATED MEASURES

- Very similar to multivariate analysis but every measurement point (time) is considered as a different trait.
- Requires modelling of the mean effects (patterns) and variance structures.
- Additional modelling of fixed effects of time points is possible (e.g. polynomials or splines).
- Convergence conflicts are still present, but to a lesser extent.
- Two modelling approaches:
  - **Multiple vectors**: parallel vectors with, typically, `us` error structure.
  - **Single vector**: stacked responses with, typically, `ar1v` correlations.

Relevant functions in ASReml-R

`pol(y,n)`	forms a set of orthogonal polynomials of order `n`
`lin(f)`	transform the factor `f` into a covariate
`spl(v,k,points)`	defines a spline model term for the variable `v` with `k` knots

REPEATED MEASURES: AS MV

Example: `/Day2/RepMeas/MVCOLS.txt`

A total of 824 individuals were measured at 4 equally spaced time points. These correspond to offspring of 26 parents that were planted as a RCBD with 4 blocks at 2, 4, 6 and 8 years after establishment.

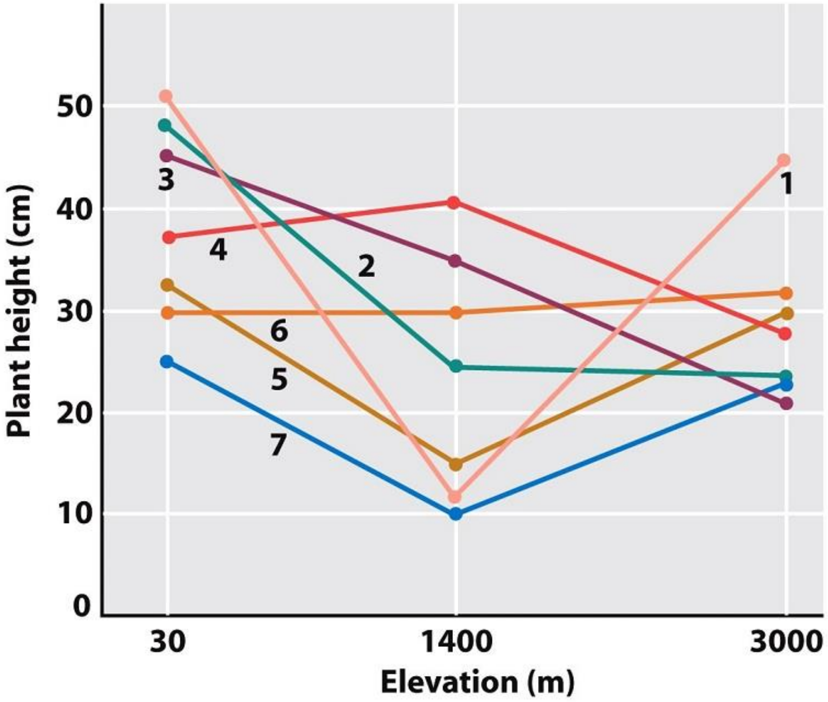
```
IDD      Indiv    Female   Rep      HT1      HT2      HT3      HT4
1        1        F09      1        62.0     108.0    240.0    411.5
2        2        F02      1        66.0     154.0    275.0    442.0
3        3        F21      1        65.0     116.0    245.0    323.1
4        4        F25      1        68.0     102.0    225.0    350.5
5        5        F13      1        58.0     170.0    325.0    457.2
6        6        F14      1        117.0    265.0    445.0    588.3
7        7        F14      1        NA       NA       NA       NA
8        8        F15      1        75.0     162.0    315.0    484.6
9        9        F18      1        74.0     182.0    340.0    493.8
10       10       F03      1        100.0    230.0    350.0    518.2
11       11       F07      1        72.0     148.0    310.0    313.9
12       12       F14      1        69.0     164.0    310.0    469.4
13       13       F11      1        87.0     208.0    340.0    493.8
14       14       F24      1        50.0     148.0    290.0    454.2
15       15       F02      1        66.0     173.0    350.0    521.2
16       16       F21      1        75.0     164.0    305.0    469.4
17       17       F15      1        78.0     166.0    315.0    493.8
...
```

REPEATED MEASURES: AS UNIV

Example: /Day2/RepMeas/REPCOLS.txt

```
IDD      Indiv    Female   Rep      Time     HT
1        1        F09      1        1        62
2        1        F09      1        2        108
3        1        F09      1        3        240
4        1        F09      1        4        411.5
5        2        F02      1        1        66
6        2        F02      1        2        154
7        2        F02      1        3        275
8        2        F02      1        4        442
9        3        F21      1        1        65
10       3        F21      1        2        116
11       3        F21      1        3        245
12       3        F21      1        4        323.1
13       4        F25      1        1        68
14       4        F25      1        2        102
15       4        F25      1        3        225
16       4        F25      1        4        350.5
17       5        F13      1        1        58
18       5        F13      1        2        170
19       5        F13      1        3        325
20       5        F13      1        4        457.2
...
```

Session 10



Multi-Environment Analysis

MET ANALYSIS

General Uses

- Incorporates information from several experiments (over different sites or years) to obtain overall BVs.
- Allows to estimate *Genotype-by-Environment* (or *Genotype-by-Year*) effects, and their variance structure. Hence, it separates genetic effects into their pure component and their interaction with site (or year).
- Provides with unbiased estimates of heritability and Type-B correlations.
- Critical to understand the genotypes structure of the population and to define breeding strategies.

Difficulties

- Every site (or year) has its own 'personality' (i.e. error structure, design effects, etc.) that needs to be combined into a single analysis.
- Amount of data can large with difficulties in fitting and convergence.
- Requires additional prior checks (e.g. EDA, coding, etc.).

SIRE / HALF-SIB MODEL

Single Site

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{f} + \mathbf{e}$$

Multiple Sites

$$\mathbf{y} = \mathbf{X}_1\mathbf{s} + \mathbf{X}_2\boldsymbol{\beta}_\mathbf{s} + \mathbf{Z}_1\mathbf{b}_\mathbf{s} + \mathbf{Z}_2\mathbf{f} + \mathbf{Z}_3\mathbf{fs} + \mathbf{e}$$

s vector of fixed environment effects (e.g. sites)

β vector of fixed design effects (e.g. replicates)

$\boldsymbol{\beta}_\mathbf{s}$ vector of fixed design effects within site

b vector of random design effects (e.g. blocks, plots), ~ N(0, $\mathbf{I}\sigma^2{}_b$)

$\mathbf{b}_\mathbf{s}$ vector of random design effects within site (e.g. blocks, plots), ~ N(0, **D**)

f vector of random sire or female effects (i.e. ½BV), ~ N(0, $\mathbf{A}\sigma^2{}_f$)

fs vector of random interaction effects (i.e. BV), ~ N(0, $\mathbf{I}\sigma^2{}_{fs}$)

e vector of random residual effects, ~ N(0, $\mathbf{I}\sigma^2$) or N(0, $\oplus\mathbf{R}$)

VAR-COV G MATRIX

- The challenge is to model a **G** matrix that has the genetic (additive, dominant, etc.) correlations between all pairs of sites.
- Uniform correlation (`cor` or `cs`) is the traditional and simplest approach, but non-optimal under most situations.
- Ideally an unstructure (or general heterogeneous correlation) is the best alternative.
- However, with large number of sites (s > 5) convergence is difficult and other models should be used (e.g. factor analytic or `fa`).

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_1\sigma_4 & \rho\sigma_2\sigma_4 & \rho\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}$$

$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \sigma_{14}^2 \\ \sigma_{12}^2 & \sigma_{22}^2 & \sigma_{23}^2 & \sigma_{24}^2 \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_{33}^2 & \sigma_{34}^2 \\ \sigma_{14}^2 & \sigma_{24}^2 & \sigma_{34}^2 & \sigma_{44}^2 \end{bmatrix}$$

$$\begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \rho_{14}\sigma_1\sigma_4 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \rho_{24}\sigma_2\sigma_4 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ \rho_{14}\sigma_1\sigma_4 & \rho_{24}\sigma_2\sigma_4 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}$$

MET ANALYSIS

Strategy for fitting MET models in ASReml

- Careful cleaning process (same factors, values, etc.).
- Start analyzing every site *individually* determining all necessary (and significant) design effects and error structure.
- Evaluate which sites to consider for full analysis (sites with low heritability contribute little to ranking).
- Consider implementing a data standardization.
- Incorporate and evaluate which variables or factors will act as '*covariates*' through all trials.
- Combine all trials into a simple single analysis (e.g. heterogeneous error variances but with common additive variance).
- Progress *slowly* to more complex variance structure for different model terms (e.g. `diag` for additive).
- Considering favouring the simplest model that suits your requirements (practical, operational).

MET ANALYSIS

MET in ASReml-R

- Flexible and fast enough to incorporate many datasets.
- Each site will have its own model specification (fixed effects, random components and error structure).

Complex Variance Structures

- Ideal objective: to fit a `us` structure to the GxE matrix to understand the genetic structure and evaluate stability of genotypes and breeding zones.
- A `us` structure is difficult to fit, but other simpler (approximate) structures are available.
- ASReml-R allows other structures based in multivariate techniques (e.g. factor analytic covariance).

TYPE-B CORRELATIONS

Definition: Correlation between sites

- Type B Genetic Correlation (Yamada) treats the same trait measured in two environments as different traits
- It is a relative expression of ***genotype-by-environment*** interaction.
- It could be zero or positive (0 to 1).
- A value close to 0 indicates that the rank in one environment is very different than the rank in another environment (i.e. low stability)
- A value close to 1 indicates that a single ranking can be used across all environments without loss of information (i.e. high stability).
- $\mathbf{V}_{\mathbf{axs}}$ is the variance estimation of the site by genotype interaction.
- The following expressions represent the average correlation between sites (if more than 2 sites are analyzed).

$$rg^2_{B(a)} = \frac{\mathbf{V_a}}{\mathbf{V_a} + \mathbf{V_{axs}}} \qquad rg^2_{B(g)} = \frac{\mathbf{V_g}}{\mathbf{V_g} + \mathbf{V_{gxs}}}$$

MET ANALYSIS

Option 1: *Simple GxE structure*

- Aims at modelling a common GxE correlation.
- Common structures are: `diag`, `corh`.
- Correlation corresponds to an average value across all sites.
- It is simpler to fit, easy to converge.
- It does not allow for a better understanding of the GxE.

Option 2: *Complex GxE structure*

- Aims at modelling the ‘full’ GxE correlation structure.
- Common structures are: `corgh`, `us`, `fa`.
- Provides with a different GxE correlation for each pair of sites.
- It is difficult to fit, particularly for several sites.
- Simplifications are usually required, e.g. standardization.

MET ANALYSIS

Variant 1: *Explicit GxE*

```
asreml(fixed=yield~Site,
       random=~Genotype+Site.Genotype,
       rcov=~at(Site):units,data=trials)
```

- Provides with average genetic values across all sites, together with *GxE deviations* for each site.
- Useful for generating ranking across all sites.
- Allows for simplification of GxE term.

Variant 2: *Implicit GxE*

```
asreml(fixed=yield~Site,random=Site.Genotype,data=trials)
```

- Provides with a different genetic value for each site.
- Useful for generating rankings for each site.
- It could make use of the full correlation structure of the GxE.
- Typically used to understand the dynamics of GxE.

MET HALF-SIB / SIRE MODEL

ASReml®

Explicit GxE

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_s + \mathbf{X}_2\mathbf{s} + \mathbf{Z}_1\mathbf{b}_s + \mathbf{Z}_2\mathbf{f} + \mathbf{Z}_3\mathbf{fs} + \mathbf{e}$$

y vector of observations

$\boldsymbol{\beta}_{\mathbf{s}}$ vector of fixed design (within site) or covariate effects

s vector of fixed location (sites or years) effects

$\mathbf{b}_{\mathbf{s}}$ vector of random design effects within site (e.g. block effect), ~ $N(0, \mathbf{D}_s)$

f vector of random sire effects (i.e. ½ breeding value), ~ $N(0, \mathbf{A}\sigma^2_f)$

fs vector of random sire-by-location interactions, ~ $N(0, \mathbf{I}\sigma^2_{fs})$

e vector of random residual effects, ~ $N(0, \mathbf{D}_{\mathbf{e}})$ or $N(0, \bigoplus_{i=1}^{s} R_i)$

$$\mathbf{V}_{\mathbf{a}} = 4\ \sigma^2_f \qquad \mathbf{V}_{\mathbf{axs}} = 4\ \sigma^2_{fs}$$

$$\mathbf{V}_{\mathbf{p}} = \overline{\sigma}^2_{bs} + \sigma^2_f + \sigma^2_{fs} + \overline{\sigma}^2$$

$$h^2 = \mathbf{V}_{\mathbf{a}} / \mathbf{V}_{\mathbf{p}} = 4\ \sigma^2_f / [\overline{\sigma}^2_{bs} + \sigma^2_f + \sigma^2_{fs} + \overline{\sigma}^2]$$

$$rg_{B(a)} = \mathbf{V}_{\mathbf{a}} / [\mathbf{V}_{\mathbf{a}} + \mathbf{V}_{\mathbf{axs}}] = \rho_s$$

MET HALF-SIB / SIRE MODEL

Example: /Day2/MultiEnv/TRIALS4.txt

A set of 4 trials were established as part of a breeding program. A total of 61 unrelated parents were considered (i.e. half-sib model). All trials corresponded to IBD with 4 full replicates. The response variable of interest is HT. We are interested in obtaining an analysis using all four sites simultaneously.

```
IDD      Test     Genotype  Surv     DBH      HT
10001    1        G41       1        736.6    557.8
10002    1        G33       1        685.8    588.3
10003    1        G22       1        838.2    551.7
10004    1        G31       1        660.4    539.5
10005    1        G18       1        406.4    411.5
10006    1        G01       1        508.0    417.6
10007    1        G05       1        711.2    518.2
10008    1        G54       1        609.6    463.3
10009    1        G30       1        482.6    466.3
10010    1        G17       1        736.6    527.3
10011    1        G58       1        584.2    472.4
10012    1        G37       1        431.8    442.0
10013    1        G07       1        736.6    600.5
10014    1        G42       1        711.2    566.9
10015    1        G38       1        711.2    518.2
10016    1        G33       1        736.6    606.6
10017    1        G50       1        736.6    576.1
10018    1        G20       1        660.4    539.5
...
```

MET HALF-SIB / SIRE MODEL

Implicit GxE

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_s + \mathbf{X}_2\mathbf{s} + \mathbf{Z}_1\mathbf{b}_s + \mathbf{Z}_3\mathbf{fs} + \mathbf{e}$$

y vector of observations

$\boldsymbol{\beta}_s$ vector of fixed design or covariate effects

s vector of fixed location (sites or years) effects

$\mathbf{b}_s$ vector of random design effects within site (e.g. block effect), ~ N(0, $\mathbf{D}_s$)

fs vector of random sire effect within location, ~ N(0, $\mathbf{A}\otimes\mathbf{G}$)

e vector of random residual effects, ~ N(0, $\mathbf{D}_e$)

G matrix of variance-covariances of genotype-by-location

A numerator relationship matrix

D diagonal matrix of dimension s

MET ANALYSIS

Factor Analytic models

- Useful approximations for modelling an **U** matrix on GxE or multivariate analyses.
- Flexible models that require fewer variance-components than `us`, and tend to converge better and quicker.
- Allow for additional interpretation of underlie environmental factors associated with the matrix of correlations.
- Finding solutions for FA models can be difficult requiring proper specification of initial values.
- Several alternative models are available within ASReml-R: `fa(,k)`.
- Based on the parameterization:

$$\mathbf{G} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Psi}$$

$\boldsymbol{\Gamma}$ is a matrix of loadings on the covariance scale

$\boldsymbol{\Psi}$ is a diagonal matrix.

PLANNING FOR MET

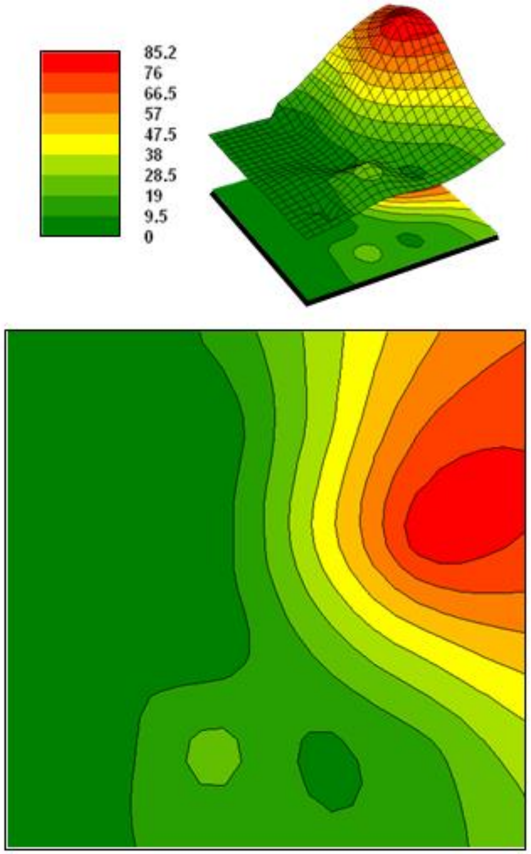
A priori

- Proper definition of experimental unit and measurement unit.
- Use of the basic elements of design:
  - Randomization: eliminate potential sources of bias
  - Replication: determine proper sample size for each genetic level!
  - Control: use more sophisticated designs for control of spatial variability (e.g. IB, Row-Col, Latinized).
- Connectivity among genotypes and sites.
- Determine blocking for each stage of the experiment (confounding?).

A posteriori

- Add covariates as required (uncorrelated with response of interest).
- Specify correct blocking structure.
- Implement post-hoc blocking if needed.
- Combine repeated measures into analysis (e.g. two years of data).
- Incorporate spatial analysis of field trials.

Session 11



Spatial Analysis

SPATIAL ANALYSIS

General Uses

- It corresponds to an extension to the single vector repeated measures analysis.
- Incorporates information from physical positions (x and y coordinates).
- Effect: improves estimates (BLUPs) and allows for a better control of errors. Hence, it will increase heritability and genetic gains.
- More efficient analysis (under presence of correlation) as it 'borrows' information from neighbours.
- ASReml can handle regular or irregular grids.
- Can be used for unreplicated trials!

Difficulties

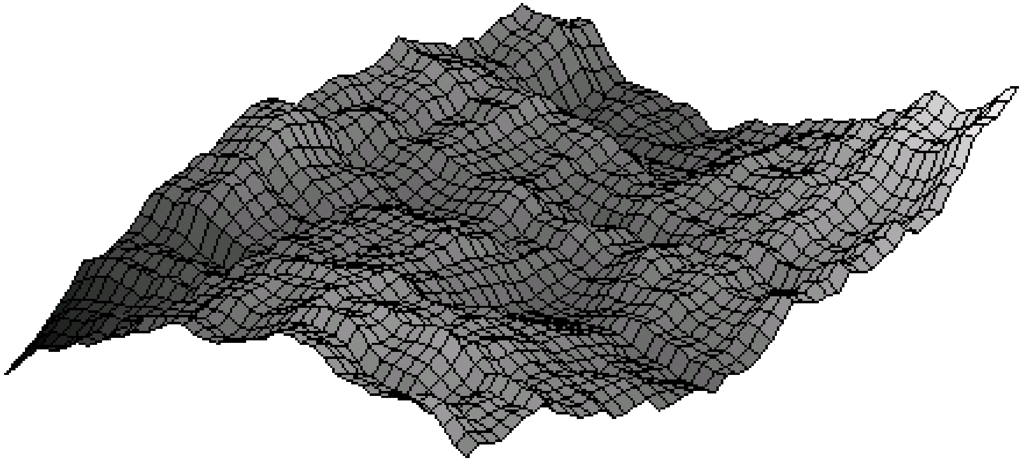
- At the present is more like an 'art' that requires to evaluate several options.
- Requires the knowledge of the position of each individual experimental unit (e.g. plant or plot).
- Additional variance components need to be estimated (i.e. convergence problems).

SPATIAL ANALYSIS

- **Gradients or Trends**
  - Linear trends
  - Polynomial functions, e.g. $f(x_c, y_c) = \alpha + \beta_1 x_c + \beta_2 y_c + \beta_3 x_c^2 y_c + \beta_4 x_c y_c^2$
  - Row or Column effects (random).
- **Patches**
  - Incomplete Blocks
  - Spatial Error Structures, e.g. $AR1 \otimes AR1 + \eta$

$$Var\ (e_{ij}) = \sigma_s^2 + \sigma_{ms}^2$$

$$Cov\ (e_{ij}\ ,\ e_{i'j'}) = \sigma_s^2\ \rho_x^{hx}\ \rho_y^{hy}$$



SPATIAL ANALYSIS

Strategy in ASReml (regular grid)

- Begin with an separable autorregressive error structure: AR1⊗AR1. This is a first order autorregressive model that assumes separate correlations ρ_x and ρ_y for columns and rows, respectively (i.e. `ar1`).
- Evaluate if a nugget effect is required (i.e. `units`).
- Check variogram and incorporate additional random or fixed effects for trends.
- Use a likelihood ratio test (LRT), BIC or AIC to compare models.

Strategy in ASReml (irregular grid)

- Begin with an isotropic exponential (i.e. `iexp`) and then move to more complex models (e.g. `aexp`) .
- As before, evaluate if a nugget effect is required (i.e. `units`), check variogram and incorporate additional random or fixed effects.

VARIANCE STRUCTURES

Correlation/Spatial structures

`ar1()`	First order autoregressive	**1**
`ar2()`	Second order autoregressive	**2**
`arma()`	Autoregressive and moving average	**2**
`iexp()`	Isotropic Exponential	**1**
`aexp()`	Anisotropic Exponential	**2**

Relevant functions in ASReml

`units`	includes nugget (microsite) random error
`pol(y,n)`	forms a set of orthogonal polynomials of order `n`
`lin(f)`	transform the factor `f` into a covariate
`spl(v,k)`	defines a spline model term for the variable `v` with `k` knots

SPATIAL ANALYSIS

Heritability in spatial models

- *Traditional* expression is only valid when distance between individuals is assumed to be zero.
- Generic expression for spatial analyses:

$$h^2 = \frac{4\sigma_{\mathrm{g}}^2}{\sigma_{\mathrm{g}}^2 + (\rho_{\mathrm{x}}^{|\mathrm{dx}|} \times \rho_{\mathrm{y}}^{|\mathrm{dy}|}) \times \sigma_{\mathrm{e}}^2 + \sigma_0^2}$$

- An alternative is to use the PEVs to approximate the *mean parental heritability:*

$$h_{\mathrm{PEV}}^2 = 1 - \frac{mean\{PEV(\mathbf{g})\}}{\sigma_g^2}$$

SPATIAL ANALYSIS

Comparing spatial models

- Use LRT when models are nested and have the same fixed effect terms.
- Compare AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) to select among non-nested models (but with same fixed effect terms).
- Use a h^2_{PEV} to compare among different models.
- Calculate one of the proposed R^2 expressions for mixed models.

$$\mathrm{AIC} = -\,2 \times \log\mathrm{L} + 2 \times t$$
$$\mathrm{BIC} = -\,2 \times \log\mathrm{L} + 2 \times t \times \log(v)$$

t number of variance parameters in the model

v residual degrees of freedom, $v = n - p$

n number of observations

p number of parameters in fixed effect factors

SPATIAL TRIAL

Example: /Day2/Spatial/ROWCOL.txt

An experiment was established to evaluate a group of open-pollinated families. The experiment consisted in row-column design with 4 replicates. The plants within the experiment where arranged in a 16x16 grid and is of interest to rank female parents based on the response yield (YA) by fitting an spatial model.

ID	REP	ROW	COL	PLOT	TREE	FEMALE	X	Y	YA
1	2	4	1	14	2	4	1	1	8.628352
2	2	4	1	14	1	4	1	2	7.718902
3	2	3	1	26	2	7	1	3	8.041164
4	2	3	1	26	1	7	1	4	9.593278
5	2	2	1	62	2	16	1	5	8.739841
6	2	2	1	62	1	16	1	6	8.456119
7	2	1	1	50	2	13	1	7	9.557565
8	2	1	1	50	1	13	1	8	10.639179
9	1	4	1	1	2	1	1	9	9.938713
10	1	4	1	1	1	1	1	10	8.332414
11	1	3	1	53	2	14	1	11	10.495654
12	1	3	1	53	1	14	1	12	10.130853
13	1	2	1	37	2	10	1	13	11.983712
14	1	2	1	37	1	10	1	14	12.080121
15	1	1	1	33	2	9	1	15	11.203263
16	1	1	1	33	1	9	1	16	10.757546
17	2	4	1	14	4	4	2	1	9.797591
18	2	4	1	14	3	4	2	2	9.206996
19	2	3	1	26	4	7	2	3	8.786462
...									

UNREPLICATED TRIALS (UR)

- Field experiments that allows testing several hundreds of genotypes with little or no replication.
- Useful for initial stages of genotype screening.
- Most treatments (with the exception of controls or checks) have a **single** replication.
- Checks are used for estimation of local control and to detect trends, and they allow estimation of the residual variance.
- Typically augmented designs are the base for unreplicated trials.
- Using too many check plots could be expensive.
- Checks should have a similar response than test genotypes.
- Statistical analysis can be based in simple (e.g. RCBD) or spatial models (e.g. $AR1 \otimes AR1$).

UNREPLICATED TRIALS (UR)

General recommendations

- More control plots improve the efficiency of UR experiments.
- Important gains in efficiency are achieved by using spatial analyses.

11	C2	24	112	23	69	C1	96	22	6	34	C1
85	101	48	C1	28	7	89	60	C2	108	74	56
47	C1	10	43	C2	16	52	5	38	33	C2	93
65	111	64	100	81	104	C2	78	C1	113	21	106
12	C2	44	68	42	C1	97	17	32	73	C1	35
25	C1	27	C2	15	88	29	4	53	C2	55	75
102	84	1	49	C1	61	70	C2	18	95	37	C1
46	86	C2	63	2	51	79	39	59	92	C2	57
66	13	C1	82	41	98	C2	90	C1	77	20	36
C1	45	83	87	C2	62	3	30	72	54	105	76
26	C2	9	14	50	8	40	C1	31	19	C2	C1
110	103	67	C1	99	80	C2	71	91	58	109	94

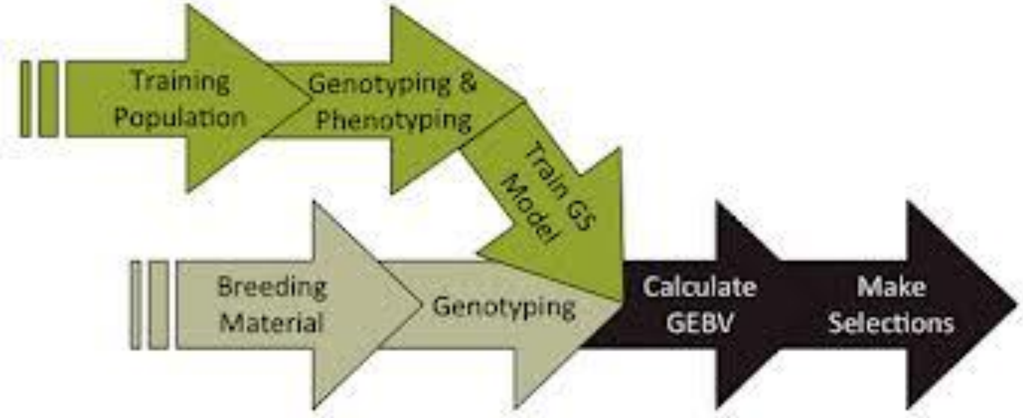
UNREPLICATED TRIALS (UR)

Example: `/Day2/Unreplicated/PEPPER.txt`

An unreplicated pepper trial was established to evaluate a total of 824 pepper genotypes planted in single plots and arranged as a RCBD with 4 blocks. In addition, a total of 10 control genotypes were planted with 20 replications each (i.e. 5 replications per block). All these individuals were arranged in a 32x32 grid, and the response variable yield, YD, was obtained. It is of interest to rank all the single replicated genotypes.

```
Gens      Control  Rep       X         Y         YD
6         0        1         1         25        7.91
16        0        1         7         17        9.04
18        0        1         11        26        9.53
19        0        1         16        20        10.08
22        0        1         2         27        9.78
35        0        1         10        26        9.21
39        0        1         4         30        8.86
40        0        1         8         24        9.15
42        0        1         11        25        9.38
45        0        1         15        22        10.64
48        0        1         10        32        10.32
50        0        1         10        31        11.22
51        0        1         8         26        11.45
...
```

Session 12

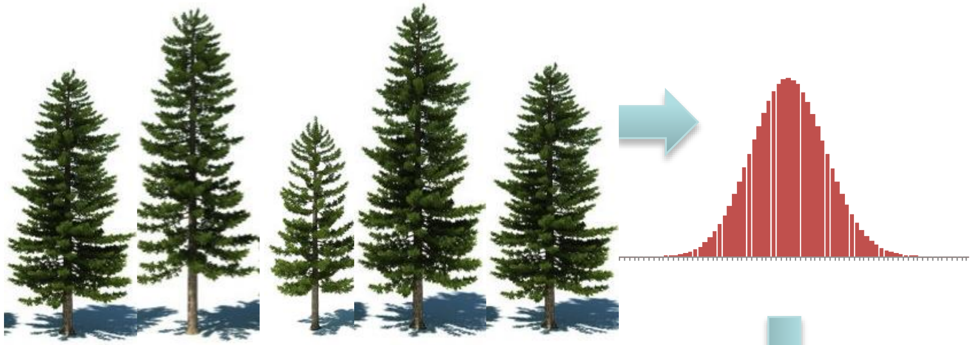


Introduction to Genomic Selection

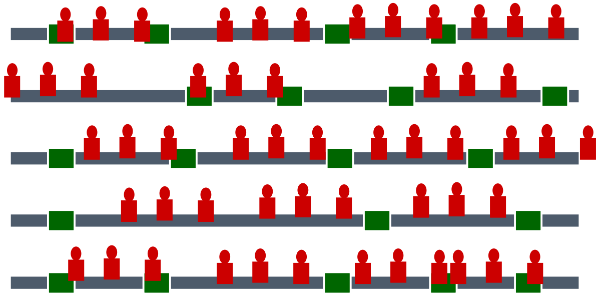
GENOMIC SELECTION

- Construct prediction models using the current breeding population phenotype and molecular markers capturing most of the quantitative variation

Quantitative phenotypic information



Genotypic information



Breeding Value (BV) + Molecular Markers

Prediction model construction:

$$BV_j = \mu + \sum_{j=1}^{p} M_j m_j + e_j$$

GENOMIC SELECTION

- Future individuals are genotyped to be use as input on prediction models to select superior genotypes in next cycles

$$\mathbf{M} = \begin{array}{c|cccc|} & m_1 & m_2 & m_3 & m_4 \\ g_1 & 1 & 0 & 1 & 2 \\ g_2 & 2 & 2 & 0 & 2 \\ g_3 & 2 & 1 & 1 & 0 \\ g_4 & 0 & 2 & 2 & 1 \\ g_5 & 0 & 0 & 2 & 1 \end{array} \qquad \hat{m} = \begin{bmatrix} 0.24 \\ 0.02 \\ -0.08 \\ 0.14 \end{bmatrix} \qquad \hat{a} = \mathbf{M}\hat{m} \qquad \hat{a} = \begin{bmatrix} 0.44 \\ 0.80 \\ 0.42 \\ 0.02 \\ -0.02 \end{bmatrix}$$

- If the markers are capturing all genetic variation, then we can assume that:
- If we also assume: $V(m) = \mathbf{I}\,\sigma_m^2$
- Then we get: $V(a) = \mathbf{M}\,\mathbf{M}'\sigma_m^2$
- An by scaling: $V(a) = \mathbf{MM}'\dfrac{\sigma_a^2}{\sum_{\mathrm{i}} 2p_i q_i} = \mathbf{G}_A\,\sigma_a^2$

BENEFITS OF GS

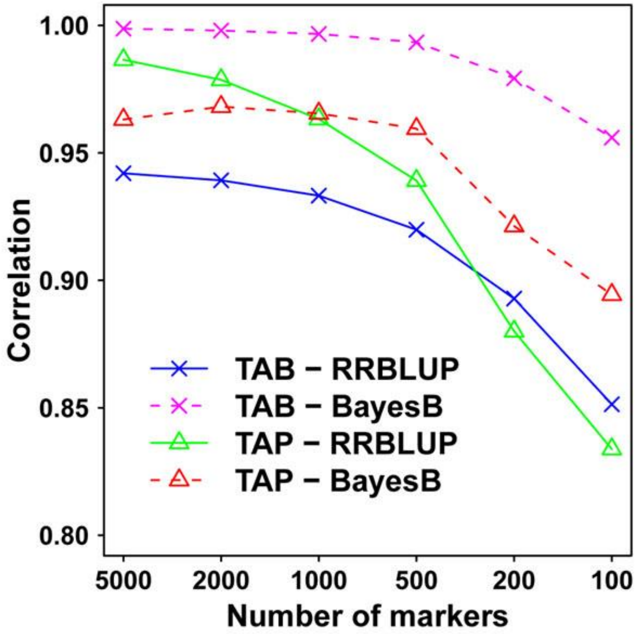
- Decrease the generation cycle of breeding (e.g. Perennials, Cattle).
- Decrease the cost of testing (e.g. Cattle, Maize).
- Screening a larger number of genotypes without field testing, thus increasing the selection pressure (e.g. Maize, other cereals).
- Predict performance for difficult and/or expensive traits (e.g. Cattle, Salmon).
- Predict performance for diseases avoiding challenging and losing the germplasm (all species).
- Can be used regardless the genetic architecture of the trait.

Note

- To apply GS successfully the constructed models need to accurately predict the genetic performance.

ANALYTIC METHODS FOR GS

- BLUP-Based: G-BLUP, RR-BLUP
- Bayes-Based: Bayes A, Bayes B, Bayes Cπ, Bayes RR
- LASSO-Based: Bayesian Lasso, Improved Lasso
- Semi-Parametric Regression: RKHS
- Non-Parametrics: Suport Vector Machine, Neural-Networks
- Others...



GBLUP

- Genomic BLUP (GBLUP) is a Genomic Selection method that uses the same framework than BLUP analysis, but replaces:
  - The **numerator relationship matrix (A)** derived from the pedigree by,
  - The **realized relationship matrix (G_A)** derived from molecular markers.

- G_A is also known as **observed relationship matrix** or **genomic matrix.**
- GBLUP is equivalent to RR_BLUP but it is simpler to implement.

$$\mathbf{A}=\begin{bmatrix} 1 & 0.5 & 0.25 & 0 \\ 0.5 & 1 & 0.25 & 0 \\ 0.25 & 0.25 & 1 & 0.25 \\ 0 & 0 & 0.25 & 1 \end{bmatrix} \qquad \mathbf{G}_A=\begin{bmatrix} 0.98 & 0.42 & 0.23 & -0.02 \\ 0.42 & 0.99 & 0.26 & 0.01 \\ 0.23 & 0.26 & 1.03 & 0.20 \\ -0.02 & 0.01 & 0.20 & 0.99 \end{bmatrix}$$

GBLUP

$$\mathbf{A} = \begin{bmatrix} 1 & 0.5 & 0.25 & 0 \\ 0.5 & 1 & 0.25 & 0 \\ 0.25 & 0.25 & 1 & 0.25 \\ 0 & 0 & 0.25 & 1 \end{bmatrix} \qquad \mathbf{G}_A = \begin{bmatrix} 0.98 & 0.42 & 0.23 & -0.02 \\ 0.42 & 0.99 & 0.26 & 0.01 \\ 0.23 & 0.26 & 1.03 & 0.20 \\ -0.02 & 0.01 & 0.20 & 0.99 \end{bmatrix}$$

Advantages and Considerations

- The use of GBLUP instead of the *pedigree-based* BLUP was shown to partition better the genetic from environmental variation.
- The **A** matrix is derived based on the infinitesimal model and represents and average relationship.
- The relationship matrix derived from the markers is more informative because the relationships estimates include the *Mendelian sampling.*
- Finally, GBLUP is unbiased: E($\mathbf{G}_A$) = **A**

GBLUP

- GBLUP uses the same framework that BLUP (Linear Mixed Models).
- Fewer normal equations need to be solved in the fitting of the model.
- Allows the direct estimation of individual's accuracies.
- Permits the simultaneous analysis of genotyped an non-genotyped individuals.

Animal Model - GBLUP

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2\mathbf{a} + \mathbf{e}$$

β vector of fixed effects

b vector of random design effects (e.g. block effect), ~ N(0, $\mathbf{I}\sigma^2{}_b$)

a vector of random additive effects (i.e. BV), ~ N(0, $\mathbf{G}_A\sigma^2{}_a$)

e vector of random residual effects, ~ N(0, $\mathbf{I}\sigma^2$)

Note:

- The variance-covariance matrix (**G**$_A$) of the additive effects is now derived from molecular markers, and it replaces the old **A** matrix.
- All marker manipulation and matrices can be obtained with **GenoMatrix**

GBLUP

Computing the Realized Relationship Matrix

- There are several different algorithms to compute the $\mathbf{G}_A$ matrix from SNP data:
  - Hayes and Goddard (2008)
  - Van Raden (2008) – 2 methods
  - Yang *et al.* (2010) – Human genetics
- Relationship matrices work well to model the variance-covariance of additive effects assuming a large number of markers is used.
- Overall, the different algorithms to calculate $\mathbf{G}_A$ do not differ considerably in their predictive ability.

Problem:

  - $\mathbf{G}_A$ matrix is usually not positive definite

Solution:

  - Bending the matrix (e.g. diag($\mathbf{G}_A$) + 0.00001).
  - Blending the matrix (e.g. $\mathbf{G}_A{}^{*} = 0.99\ \mathbf{G}_A + 0.01\ \mathbf{A}$).

GBLUP in ASReml

User supplied special variance structures

- The relationship matrix ($\mathbf{G}_A$) that is previously computed using a given algorithm from other software (R, Fortran, etc.) based on molecular markers, is read in R and then supplied to ASReml-R.

- Inverse of $\mathbf{G}_A$ matrix is an independent file in ASCII format that is supplied in SPARSE form (lower diagonal).
- SPARSE format (and column names): `Row`, `Column`, `Value` (lower triangular row-wise sorted column within rows).
- Need to specify `attr(gmatrix,"rowNames")` with the same number of levels than the factor (from data).
- All diagonal elements of the matrix must be included in the file (even 1s).
- In some versions of ASReml the $\mathbf{G}_A$ matrix can be read in DENSE form.

Warning

- The **number** and **order** of levels have to *match* perfectly the ones used for the associated factor, e.g. `animalID`, read in the data.

GBLUP in ASReml

Example: `/Day2/GBLUPTest/`

An experiment consisting in evaluating a total of 10 individuals originating from full-sib families of 4 sires and 4 dams. The objective is to fit a parental model (i.e. select sires) that considers the molecular pedigree information.

DATA.txt

Indiv	Sire	Dam	Resp
1001	10	50	155
1002	10	60	121
1003	10	70	130
1004	20	50	141
1005	20	60	130
1006	20	70	162
1007	30	50	118
1008	30	60	108
1009	30	70	119
1010	40	80	143

PEDSIRE.txt

Indiv	Sire	Dam
10	1	0
20	2	0
30	2	0
40	1	0

GBLUP in ASReml

$$\mathbf{A} = \begin{array}{c} \begin{array}{cccc} 10 & 20 & 30 & 40 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0.25 \\ 0 & 1 & 0.25 & 0 \\ 0 & 0.25 & 1 & 0 \\ 0.25 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$\mathbf{G}_A = \begin{bmatrix} 1.023 & 0.012 & -0.036 & 0.364 \\ 0.012 & 0.992 & 0.226 & 0.023 \\ -0.036 & 0.226 & 1.016 & 0.068 \\ 0.364 & 0.023 & 0.068 & 0.987 \end{bmatrix}$$

$$\mathbf{G}^{-1}{}_A = \begin{bmatrix} 1.130 & -0.020 & 0.073 & -0.421 \\ -0.020 & 1.062 & -0.237 & -0.001 \\ 0.073 & -0.237 & 1.046 & -0.093 \\ -0.421 & -0.001 & -0.093 & 1.175 \end{bmatrix}$$

```
GINVM.giv

Row Column GINV
1 1  1.1302492
2 1 -0.0204900
2 2  1.0623199
3 1  0.0728078
3 2 -0.2369711
3 3  1.0457936
4 1 -0.4213681
4 2 -0.0008723
4 3 -0.0933796
4 4  1.1750231
```

GBLUP in ASReml

Predictions for ‘new’ individuals

$$\mathbf{G}_A = \begin{array}{c} \begin{array}{cccccc} 10 & 20 & 30 & 40 & 50 & 60 \end{array} \\ \begin{bmatrix} 1.023 & 0.012 & -0.036 & 0.364 & 0.083 & 0.176 \\ 0.012 & 0.992 & 0.226 & 0.023 & 0.023 & 0.508 \\ -0.036 & 0.226 & 1.016 & 0.068 & -0.011 & 0.136 \\ 0.364 & 0.023 & 0.068 & 0.987 & 0.123 & 0.495 \\ 0.083 & 0.023 & -0.011 & 0.083 & 0.996 & 0.077 \\ 0.176 & 0.508 & 0.136 & 0.495 & 0.077 & 1.010 \end{bmatrix} \end{array}$$

GBLUP in ASReml

Final comments

- Modifications can be done that incorporate observed relationships of parents and all offspring.
- Individuals with measurements correspond to training population and 'new' individuals in $\mathbf{G_A}$ matrix are treated as prediction population.
- It is possible to combine pedigree data (**A**) with observed relationships ($\mathbf{G_A}$) into a single matrix. This will allows to consider individuals without molecular data.
- Observed dominance ($\mathbf{G_D}$) relationship matrix can also be incorporated to model these interactions or higher order interactions, e.g. **A#D**.
- Further understanding of the construction (and properties) of the $\mathbf{G_A}$ matrix are required.
- Special care must be considered with the number of decimal places of the inverses of these matrices.
- Non-definitive matrices are automatically handled by ASReml routines.