ASReml-R: an R package for mixed models using residual maximum likelihood

David Butler[1] Brian Cullis[2] Arthur Gilmour[3]

[1]Queensland Department of Primary Industries
Toowoomba

[2]NSW Department of Primary Industries
Wagga Wagga Agricultural Institute

[3]NSW Department of Primary Industries
Orange Agricultural Institute

Outline

Introduction

- ASReml: standalone program (Gilmour *et al.*, 1999)
- Designed to fit complex mixed models to large problems.
- Efficient computing strategies
  - Average Information algorithm (Gilmour *et al.*, 1995)
    - avoids forming expensive trace terms
  - Sparse matrix methods
    - avoid forming and storing zero cells
    - exploit variance structures with sparse inverses
    - optimize solution order
  - Direct product structures exploited
    - $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$

Asreml-R

- ASReml-R is the R interface to the ASReml fitting routines
- model specified as formula objects
- initial values specified as list objects
- ASReml object
  - BLUPs of random effects
  - GLS estimates of fixed effects
  - REML estimates of variance components
  - predictions from the linear model (if requested)

The linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

$\boldsymbol{y}$	denotes the $n \times 1$ vector of observations
$\boldsymbol{\tau}$	is a $p \times 1$ vector of fixed treatment effects
$\boldsymbol{X}$	is a $n \times p$ design matrix
$\boldsymbol{u}$	is a $q \times 1$ vector of random effects
$\boldsymbol{Z}$	is a $n \times q$ design matrix
$\boldsymbol{e}$	is a $n \times 1$ vector of residual errors

The linear model

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \theta \begin{bmatrix} \boldsymbol{G}(\gamma) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}(\phi) \end{bmatrix} \right)$$

- Where:
  - $\boldsymbol{G}, \boldsymbol{R}$ parameterized variance matrices
  - γ a vector of variance parameters relating to $\boldsymbol{u}$
  - ϕ a vector of variance parameters relating to $\boldsymbol{e}$
  - θ is a scale parameter
- $\boldsymbol{y} \sim N(\boldsymbol{X}\tau, \boldsymbol{H})$
- $\boldsymbol{H} = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$.

Variance structures for the errors
R structures

- $\boldsymbol{R}$ may comprise t independent sections

$$\boldsymbol{R} = \oplus_{j=1}^{t} \boldsymbol{R}_j = \begin{bmatrix} \boldsymbol{R}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{R}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{R}_t \end{bmatrix}$$

- Each section may be the direct product of two or more dimensions

$$\boldsymbol{R}_i = \boldsymbol{R}_{i_1} \otimes \boldsymbol{R}_{i_2} \otimes \dots$$

Variance of the random effects
G structures

- The vector of random effects is often composed of b subvectors
$$\boldsymbol{u} = [\boldsymbol{u}_1' \, \boldsymbol{u}_2' \, \ldots \, \boldsymbol{u}_b']'$$
- The $\boldsymbol{u}_i$ are assumed $N(\boldsymbol{0}, \theta \boldsymbol{G}_i)$.
- As for $\boldsymbol{R}$

$$\boldsymbol{G} = \oplus_{i=1}^{b} \boldsymbol{G}_i = \begin{bmatrix} \boldsymbol{G}_1 & 0 & \ldots & 0 \\ 0 & \boldsymbol{G}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{G}_b \end{bmatrix}$$

- Assuming separability $\boldsymbol{G}_i = \boldsymbol{G}_{i1} \otimes \boldsymbol{G}_{i2} \otimes \ldots \otimes \boldsymbol{G}_{if}$

Specifying the linear model

fit.asr $<-$ *asreml (fixed=, random=, rcov=, data=)*

- Fixed effects
  fixed = *y* $\sim$ *model formula*
- Random effects (G structures)
  random = $\sim$ *model formula*
- Error model (R structures)
  rcov = $\sim$ *model formula*
- Sparse fixed
  sparse = $\sim$ *model formula*

  Variance matrix for solutions not available
- Factors crossed or nested - determined by coding.
- *y* may be a matrix

Specifying variance models

G structures

- The default variance model is (scaled) identity.
- Variance models for random terms are specified using *special functions*.
- For example

  random = ∼ diag(A):B

  specifies a diagonal variance structure of order length(levels(A)) for A and a (default) identity for B.

Specifying variance models

R structures

- Default $\sigma^2 \boldsymbol{I}_n$, where n $<-$ nrow(data)
- Specified using *special functions*.
- Example: a series of t independent experiments indexed by the factor *Trial*,

  rcov = $\sim$ at(Trial):ar1(A):ar1(B)

  specifies separable autoregressive processes across *A* and *B* at each level of *Trial*

Special Functions

Model functions

lin(obj=x)	Includes the named factor as a variate.
spl(obj=x)	Spline **random** factor.
pol(obj=x, t)	Orthogonal polynomials of order $\|t\|$.

Time series type models

ar1(), ar2()	Autoregressive
ma1(), ma2()	Moving average

Metric based models in $\Re$ or $\Re^2$

exp(), gau()	One dimensional
aexp(), agau()	Anisotropic 2D
mtrn()	Matérn class

General structure models

cor(), corb(), corg()	Correlation
diag(), us(), ante(), chol()	Variance
fa(obj=x, q)	Factor Analytic with q factors

Known structures

ped(), giv()	Use known inverse matrices.

The asreml class

Component	Description
loglik	log likelihood at termination
gammas	vector of variance parameter estimates
coefficients	list of fixed, random and sparse coefficients
vcoeff	variance of the coefficients
fitted.values	fitted values
residuals	residuals
sigma2	residual variance
predictions	list of predictions if specified
G.param	list object of variance models for random terms
R.param	list object of variance models for error term

asreml methods

coef()	List with components *fixed*, *random* and *sparse*.
resid()	Vector of residuals.
fitted()	Vector of fitted values.
summary()	List including the asreml() call, REML log-likelihood, variance parameters, coefficients, residuals and components of $\boldsymbol{C}^{-1}$ if requested.
wald()	A table of Wald tests for each fixed term.
plot()	Residual plots including the sample variogram, distribution, fitted values and trend plots.
predict()	Predictictions from the linear model (eg, tables of adjusted means). See Gilmour *et al.* (2004) and Welham *et al.* (2004).

Example: Multi-environment trials

In the context of a plant genetic improvement program,

- It is important to know how genotype performance varies with a change in environment, that is, to investigate (G$\times$E) interaction.
- Identify genotypes with broad or specific adaptation.
- G$\times$E is assessed in a series of designed experiments in a range of environments (METs)
- Environments may be geographic locations and/or years
- Smith *et al.* (2005) present a useful review.

Example: Mixed model for MET data

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_o\boldsymbol{u}_o + \boldsymbol{e}$$

- Assume $\text{var}\left(\boldsymbol{u}_g\right) = \boldsymbol{G}_g = \boldsymbol{G}_e \otimes \boldsymbol{I}_g$
- $\boldsymbol{G}_e$ is the genetic variance matrix:

$$\boldsymbol{G}_e = \begin{bmatrix} \sigma^2_{g_1} & \sigma_{g_{12}} & \sigma_{g_{13}} & \cdots & \sigma_{g_{1t}} \\ & \sigma^2_{g_2} & \sigma_{g_{23}} & \cdots & \sigma_{g_{2t}} \\ & & \sigma^2_{g_3} & \cdots & \sigma_{g_{3t}} \\ & & & \ddots & \\ & & & & \sigma^2_{g_t} \end{bmatrix}$$

- Allow separate spatial covariance structures for the errors for each trial

$$\boldsymbol{R}_j = \sigma^2_j \boldsymbol{R}_{c_j}(\phi_{c_j}) \otimes \boldsymbol{R}_{r_j}(\phi_{r_j})$$

Example: METs in ASReml-R

asreml(yield $\sim$ trial + . . . ,
random = $\sim$ us(trial):genotype + . . . ,
rcov = $\sim$ at(trial):ar1(column):ar1(row), . . .)

- *trial* is a (fixed) factor with t levels
- *genotype* is a (random) factor with g levels
- us(trial):genotype models genotype effects in each trial with variance $\boldsymbol{G}_e \otimes \boldsymbol{I}_g$ where $\boldsymbol{G}_e$ is an unstructured form
- at(trial):ar1(column):ar1(row) models the residual effects for each trial with an AR1$\times$AR1 correlation structure.

MET data set

- Stage 2 trials taken from the Qld barley program (Kelly *et al.*, 2007)
- 14 environments over 2 years of trialling: 2003/4
- 1255 unique genotypes tested
  - 698 in 2003
  - 720 in 2004
  - 163 genotypes common across years
- Partially replicated designs (Cullis *et al.*, 2006)
- Response variate is grain yield
- Pedigrees traced back four generations

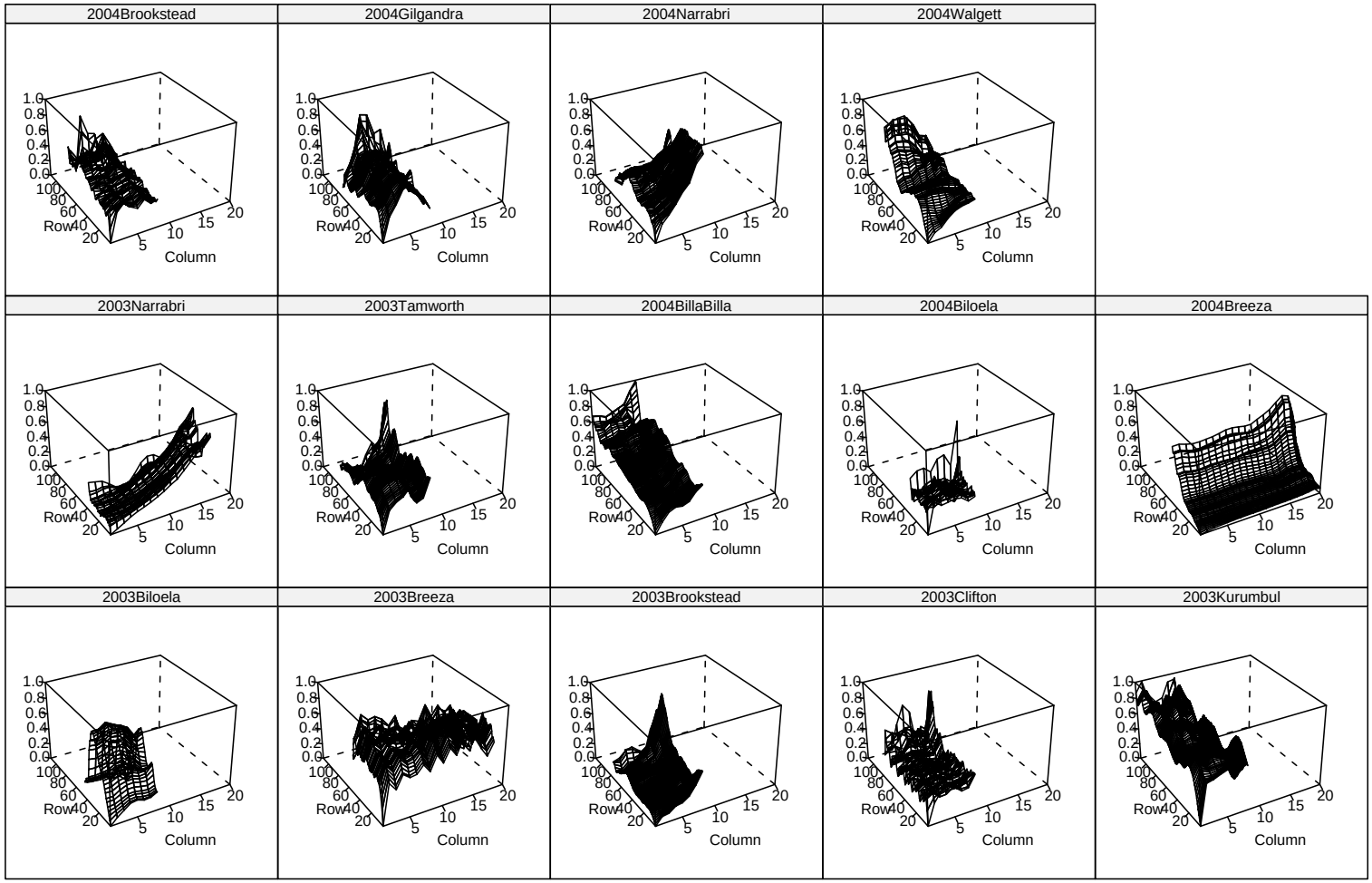
Analysis strategy

1. Initial spatial model for each experiment
   - analyse each trial separately, or
   - joint analysis with a diagonal variance model

```
qb.asr1 <- asreml(yield ~ Site,
                  random = ~ diag(Site):Genotype,
                  rcov = ~ at(Site):ar1(Column):ar1(Row),
                  data = qb)
```

   17,663 equations
   56 variance parameters
2. Model global trend and extraneous effects.
3. Model G$\times$E.
4. Predict genotype effects

plot(qb.asr1,option='v')



Models for G×E

- Diagonal variance structure analagous to individual analyses.
- Assumes that the genetic effects in different environments are un-correlated. Unlikely to be sensible.
- The us() model is the most general form for $\boldsymbol{G}_e$. Difficulties:
  - With many environments, the number of parameters is large
  - Difficult to fit REML estimate of matrix can be singular - not full rank
- Factor Analytic (FA) variance model a good approximation to US and handles not full rank

Known genetic effects

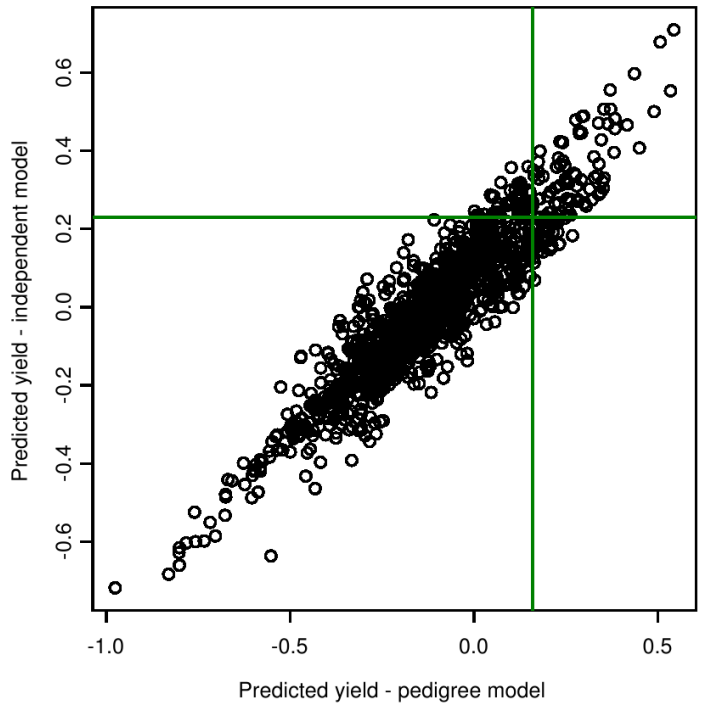
- A better genetic variance model most likely achieved by partitioning genetic effects into additive and non-additive.
- If $\boldsymbol{u}_g = \boldsymbol{a}_g + \boldsymbol{i}_g$, then
  - Assume $\boldsymbol{a}_g \sim N(\boldsymbol{0}, \sigma_a^2 \boldsymbol{A})$
  - Assume $\boldsymbol{i}_g \sim N(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I})$
  - $\text{var}\left(\boldsymbol{u}_g\right) = \boldsymbol{G}_{ae} \otimes \boldsymbol{A} + \boldsymbol{G}_{ie} \otimes \boldsymbol{I}$
- Asreml-R
  1. ainv $<-$ asreml.Ainverse(pedigree)$ginv
  2. asreml(. . . , ped(genotype), . . . + . . . , ide(genotype), . . . , ginverse=list(genotype=ainv), . . .)

The final model

```
asreml(yield ~ Site+at(Site,c(3,6,8,13)):lincol + at(Site,c(3,8,10,11)):linrow +
                at(Site,3):lincol:linrow + at(Site,4):fx4 + at(Site,6):fx6,
       random = ~ fa(Site,3):ped(Genotype) + fa(Site):ide(Genotype) +
                at(Site,c(2,4,5,7,9,11,12)):Column + at(Site,c(2)):Row,
       rcov = ~ at(Site):ar1(Column):ar1(Row),
       ginverse = list(Genotype=ainv), data=qb)
```

50,115 equations

134 parameters



References

Cullis, B., Smith, A., and Coombes, N. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological and Environmental Statistics*, **(in press)**.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**, 1440–1450.

Gilmour, A. R., Cullis, B. R., Welham, S. J., and Thompson, R. (1999). ASREML, reference manual. Biometric bulletin, no 3, NSW Agriculture, Orange Agricultural Institute, Forest Road, Orange 2800 NSW Australia.

Gilmour, A. R., Cullis, B. R., Welham, S. J., Gogel, B. J., and Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis*, **44**, 571–586.

Kelly, A., Cullis, B. R., Gilmour, A., Smith, A. B., Eccleston, J. A., and Thompson, R. (2007). Estimation in a multiplicative mixed model involving a genetic relationship matrix. *In preparation*.

Smith, A., Cullis, B., and Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *Journal of Agricultural Science, Cambridge*, **143**, 1–14.

Welham, S. J., Cullis, B. R., Gogel, B. J., Gilmour, A. R., and Thompson, R. (2004). Prediction in linear mixed models. *Australian and New Zealand Journal of Statistics*, **46**, 325–347.

More about ASReml-R and ASReml

- Visit: www.vsni.co.uk

VSN International Ltd.
2 Amberside House, Wood Lane
Hemel Hempstead
Herts HP2 4TP

United Kingdom