

Modelling variance structures of multi-term effects in ASReml

Brian Cullis

`bcullis@uow.edu.au`

University of Wollongong

Modelling variance structures of multi-term effects

Collaborations and Acknowledgements

- This presentation is joint work with Alison Smith (UOW) and David Butler (Crop and Food Science, Agri-Science Queensland DEEDI)
- Thanks to Colleen Hunt for the motivation to work on sorghum(!)
- Thanks to Yusuf Genc and David Jordan for use of their data-sets
- Grains Research and Development Corporation for financial support.



Linear mixed model

The model

Let $\mathbf{y}^{(n \times 1)}$ be vector of observations. The general linear mixed model can be written as

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

- $\boldsymbol{\tau}^{(p \times 1)}$ vector of fixed effects with design matrix \mathbf{X} (assumed full rank)
- $\mathbf{u}^{(b \times 1)}$ vector of random effects with design matrix \mathbf{Z}
- $\mathbf{e}^{(n \times 1)}$ vector of residuals.

Linear mixed model

Assumptions

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

- $\mathbf{G} = \mathbf{G}(\boldsymbol{\gamma})$
- $\mathbf{R} = \mathbf{R}(\boldsymbol{\phi})$
- $\text{var}(\mathbf{y}) = \mathbf{H}$ where $\mathbf{H} = \mathbf{ZGZ}' + \mathbf{R}$
- Overall scale parameters have been omitted for simplicity of presentation: inclusion leads to use of variance components (rather than variance component ratios)

- Assume an indexing factor which delineates the **sections** of the data and partition e conformably with this indexing.
- Thus $e = [e'_1 e'_2 \dots, e'_s]'$.
- The variance matrix for each **section** may differ, but generally we assume that the errors from different sections are independent. Thus

$$\mathbf{R} = \oplus_{j=1}^s \mathbf{R}_j$$

G-structures

General Framework

- Assume $\mathbf{u} = [\mathbf{u}'_1 \ \mathbf{u}'_2 \ \dots \ \mathbf{u}'_q]'$
- Correspondingly $\mathbf{Z} = [\mathbf{Z}_1 \ \mathbf{Z}_2 \ \dots \ \mathbf{Z}_q]$
- \mathbf{u}_i relate to separate terms and **mostly** assumed **mutually independent**
- This leads to

$$\mathbf{G} = \bigoplus_{i=1}^q \mathbf{G}_i$$

Separability

Direct products in R -structures

Consider one section, data ordered by two factors, namely `COLUMNS` with c levels, termed the **outer** factor, and `ROWS` with r levels, termed the **inner** factor. The residuals may be presented in a $c \times r$ matrix \mathbf{E} , say, where

$$\mathbf{e} = \text{vec}(\mathbf{E})$$

The direct product form for the variance structure of \mathbf{e} is then given by

$$\mathbf{R} = \mathbf{R}_c \otimes \mathbf{R}_r$$

Typical examples include the two-dimensional separable auto-regressive model ($\text{AR1} \otimes \text{AR1}$) used for modelling the variance structure of the residuals in field experiments.

Separability

Direct products in G -structures

Simple illustration:

- Analysis of a multi-environment trial with $p = 2$ sites and $m = 20$ varieties
- If u represents (only) the effects for a first-order interaction between Site and Variety then $q = 1$ and,
- if the effects are ordered varieties within sites, then Site is the **outer** factor and Variety is the **inner** factor.

Separability

Direct products in G -structures

Simple illustration (continued):

- If $\mathbf{u} = \text{vec}(\mathbf{U}^{20 \times 2}) = [\mathbf{u}'_1, \mathbf{u}'_2]'$ then a reasonable and commonly used variance model is

$$\text{var}(\mathbf{u}_1) = g_{11} \mathbf{I}_{20}$$

$$\text{var}(\mathbf{u}_2) = g_{22} \mathbf{I}_{20}$$

$$\text{cov}(\mathbf{u}_1, \mathbf{u}_2) = g_{12} \mathbf{I}_{20}$$

- Collectively we have

$$\text{var}(\mathbf{u}) = \mathbf{G}_e \otimes \mathbf{I}_{20}$$

say where $\mathbf{G}_e^{2 \times 2} = \{g_{ij}\}$ is a symmetric positive definite matrix.

Modelling variance structures for multi-term effects

- Until now G and R variance structures have been applied to a single term.
- There are applications which require variance models to be applied to more than one term
- Three (perhaps four - time permitting) examples will be considered to illustrate the (relatively) new syntax
 - the previous toy example,
 - random regressions with a twist,
 - competition modelling in field trials using the random treatment interference (R-TIM) model and
 - linear mixed models for partial compositing.

Syntax for multi-term variance models

ASReml-R syntax

random= ~ **str(form, vmodel)**

form model formula specifying
a set of terms that have
an associated variance
model

vmodel a formula object
containing ASReml
variance functions
separated by the “:”
operators presenting
each component of the
direct product variance
model

Syntax for multi-term variance models

ASReml-R syntax

`random= ~ str(form, vmodel)`

form model formula specifying a set of terms that have an associated variance model

vmodel a formula object containing ASReml variance functions separated by the “:” operators presenting each component of the direct product variance model

ASReml syntax

`!r ![form !]`

:

:

`0 0 1`

`vmodel1`

`vmodel2`

form list of terms to assign variance model to

vmodel1 specifies *model-term* and *d* being the term (or first term if there is more than one term) which the variance model is to be applied and the number of components of the direct product variance model

vmodel2 *d* lines each line specifies *order*, *key*, *model*, typically the name of the factor or number of levels, 0 and the variance model (eg US)

Multi-site toy example

Recall: multi-environment trial with $p = 2$ sites and $m = 20$ varieties

ASReml-R **single term variance**

syntax

random= ~ **us(Site):id(Variety)**

- (i) applies `us()` to `Site` and `ID` to `Variety`
- (ii) gives $\mathbf{G} = \mathbf{G}_e \otimes \mathbf{I}_{20}$

Multi-site toy example

Recall: multi-environment trial with $p = 2$ sites and $m = 20$ varieties

ASReml-R single term variance syntax

`random = ~ us(Site):id(Variety)`

- (i) applies `us()` to `Site` and `ID` to `Variety`
- (ii) gives $G = G_e \otimes I_{20}$

ASReml-R multi-term syntax

`random = ~ str(~ at(Site,1):id(Variety) + at(Site,2):Variety, ~ us(2):id(Variety))`

- (i) `vmodel` has been applied to two separate terms
- (ii) the overall size (ie product of the number of terms) of the variance model must match the overall number of effects in `form`

Salinity tolerance in bread wheat

An example of random regressions

- Dryland salinity is a major limitation to agriculture and food production
- Two complementary approaches to alleviate the problem:
 - soil management - sometimes difficult and expensive especially where water is a limiting factor
 - exploit genetic diversity in salinity tolerance (ST) - despite apparent genetic diversity so far limited success in determining genetic architecture of ST in bread wheat

Salinity tolerance in bread wheat

An example of random regressions

- A previous study (Genc *et al.*, 2010) reported QTL affecting Na^+ exclusion, K^+ accumulation and seedling biomass in a bread wheat mapping population (Berkut/Krichuaff)
- Here we consider data from a set of six field trials grown in 2007 (2 locations) and 2008 (4 locations)
- A total of 151 doubled-haploid (DH) lines were sown at six sites (location by year combinations) according to a resolvable complete block design with two replicates at 5 sites and three at the other site
- Additional named varieties were sown at some or most sites with varying replication. These additional varieties included the parents.

Salinity tolerance in bread wheat

Phenotyping and soils data

- Grain yield and tissue concentrations of Na^+ and K^+ were measured on all plots - our focus will be on grain yield
- Apparent soil electrical conductivity (EC_a) was measured on all plots using an EM38 meter in the vertical dipole position at the same time that plants were sampled for nutrient analysis
- A total of 60 readings were taken for each plot
- The EM38 reading will be used as a covariate (ie a measure of soil salinity in our analysis)

Salinity tolerance in bread wheat

Aims of analysis

- Identify QTL associated with so-called ST, as measured through grain yield
- As a measure of ST we consider the joint effects of the regression coefficients of site-mean adjusted yield on EM
- EM varies both within and across sites and so this “random regression” has to be embedded within the framework of a complex multi-environment trial analysis
- Subsequent mapping (not presented here) involves bivariate QTL analysis using an approach which is an extension of the univariate spatial smoothing model presented earlier today (Smith *et al.*)

Salinity tolerance in bread wheat

Multi-environment (MET) ST analysis

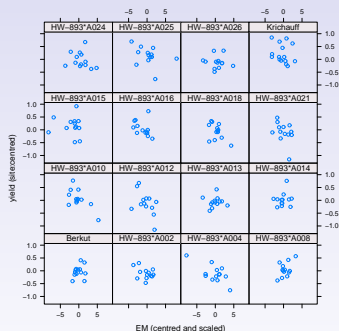
$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_{st}\mathbf{u}_{st} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

- \mathbf{y} : data for six sites, ordered as rows within columns within sites
- $\boldsymbol{\tau}$: fixed effects, including site by D_{type} effects
- \mathbf{u}_{st} : random DH ST effects, ie intercepts and slopes of the regression of yield on EM
- \mathbf{u}_g : random site by genetic (ie DH) effects
- \mathbf{u}_p : random non-genetic (or peripheral) effects, including replicate effects
- \mathbf{e} : residuals

Salinity tolerance in bread wheat

Motivation for ST model

Scatter plot of the site mean adjusted yield against EM for the parents and the first 14 DH's



Salinity tolerance in bread wheat

MET-ST model and syntax

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_{st}\mathbf{u}_{st} + \mathbf{Z}_g\mathbf{u}_g + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

```
st.asr <- asreml(yield ~ Site + EM2 + Site:Dtype,  
random = ~ str(~ DHGeno + DHGeno:EM2, ~ corh(2):id(151)) +  
diag(Site):DHGeno + at(Site):Repl + . . . ,  
rcov = ~ at(Site):ar1(Column):ar1(Row), data=em.dat,  
R.param=gam.table,G.param=gam.table,workspace=40e6,  
na.method.X='include')
```


Competition modelling in sorghum plant breeding trials

Background

- Stringer *et al.* (2011) considered joint modeling of spatial variability and within-row interplot competition in field trials grown by the BSES sugar breeding programme
- Earlier empirical evidence suggested that $AR1 \times AR1$ model provided an inferior fit to these trials (Stringer and Cullis, 2002)
- They showed that a random effects treatment interference model (R-TIM) provided an improved fit to data from sugar cane breeding trials which exhibited substantial inter-plot competition

Competition modelling in sorghum plant breeding trials

Background

- In this example we consider data from the DEEDI sorghum breeding programme which extends their approach to incorporate information on pedigrees
- This programme runs two separate pedigree breeding programmes, one for males and one for females
- All field evaluation of lines within each sub-programme is undertaken using F_1 hybrids
- The aim of each programme is to provide elite fully in-bred parental lines for commercial use within a hybrid breeding programme

Competition modelling in sorghum plant breeding trials

Field trial design

- Focus on a preliminary yield trial for males (PYTM) grown in 2008 at the Hermitage Research Station in Warwick Queensland
- The trial design was a resolvable p -rep design (Cullis *et al.* 2006) involving 791 F_1 hybrids
- The number of plots sown for each type of F_1 hybrid is presented below. Test lines (the lines of main interest) were either sown in 1 or 2 plots.

| Plots | Test: F_1 | Check: F_1 | Commercial: F_1 | Total: F_1 |
|-------|-------------|--------------|-------------------|--------------|
| 1 | 512 | 0 | 0 | 512 |
| 2 | 271 | 0 | 2 | 273 |
| 3 | 0 | 0 | 2 | 2 |
| 4 | 0 | 2 | 2 | 4 |
| Total | 783 | 2 | 6 | 791 |

Competition modelling in sorghum plant breeding trials

Trial characteristics and agronomy

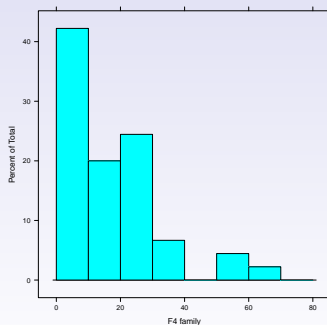
- Trials are sown as a rectangular array
- Plots are $1.5 \times 10\text{m}$ and contain two plot-rows of plants
- *Midge* necessitates spray-out rows as shown below (denoted by “x”, others given by hybrid code)

| Row | Column | | | | | | |
|-----|--------|-----|-----|-----|-----|-----|-----|
| | 1 | 2 | 3 | ... | 18 | 19 | 20 |
| 1 | 729 | 103 | 175 | ... | 234 | 669 | 493 |
| 2 | 184 | 511 | 18 | ... | 22 | 465 | 786 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 10 | 733 | 47 | 379 | ... | 769 | 80 | 485 |
| 11 | x | x | x | ... | x | x | x |
| 12 | x | x | x | ... | x | x | x |
| 13 | 361 | 179 | 524 | ... | 8 | 221 | 189 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 22 | 371 | 151 | 747 | ... | 788 | 660 | 326 |
| 23 | x | x | x | ... | x | x | x |
| 24 | x | x | x | ... | x | x | x |
| 25 | 255 | 621 | 196 | ... | 194 | 69 | 247 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 63 | 474 | 635 | 38 | ... | 697 | 734 | 11 |
| 64 | 326 | 786 | 309 | ... | 547 | 70 | 598 |

Competition modelling in sorghum plant breeding trials

Genetic design

- Aim of the PYTM trial is to promote about 10% of the F_4 males to the next level of testing
- 783 F_4 males were crossed with one female
- The 783 males came from 48 full-sib families



Competition modelling in sorghum plant breeding trials

R-TIM model including pedigrees

- The R-TIM assumes that as well as having a direct effect that each entry (ie all F_1 hybrids with data and all ancestors) also has a so-called neighbour effect
- Furthermore, the total genetic effect is partitioned into an additive effect and a residual genetic effect (we ignore dominance effects given the genetic design)

Hence our full model is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_a + \mathbf{Z}_g\mathbf{u}_l + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

where $\mathbf{u}'_a = (\mathbf{u}'_{a_d} \mathbf{u}'_{a_n})$ and $\mathbf{u}'_l = (\mathbf{u}'_{l_d} \mathbf{u}'_{l_n})$. The associated genetic design matrices are given by $\mathbf{Z}_g = [\mathbf{Z}_{g_d} \mathbf{N}_g \mathbf{Z}_{g_d}]$ where $\mathbf{N}_g = \mathbf{I}_c \otimes \mathbf{N}_r$ and \mathbf{N}_r is the within row first order neighbour incidence matrix.

Competition modelling in sorghum plant breeding trials

R-TIM model including pedigrees-distributional assumptions

The most general variance matrix for the vectors of genetic effects is given by

$$\begin{aligned}\text{var}(\mathbf{u}_a) &= \begin{pmatrix} \sigma_{add}^2 & \sigma_{adn} \\ \sigma_{adn} & \sigma_{ann}^2 \end{pmatrix} \otimes \mathbf{A} = \mathbf{G}_a \otimes \mathbf{A} \\ \text{var}(\mathbf{u}_l) &= \begin{pmatrix} \sigma_{l dd}^2 & \sigma_{l dn} \\ \sigma_{l dn} & \sigma_{l nn}^2 \end{pmatrix} \otimes \mathbf{I}_m = \mathbf{G}_l \otimes \mathbf{I}_m\end{aligned}$$

where m is the total number of entries ($m = 1778$).

The usual variance model is applied to \mathbf{u}_p while Stringer *et al.* (2011) present an extended class of residual variance models which jointly model competition and trend. This class includes the standard AR1 \times AR1 variance model.

Competition modelling in sorghum plant breeding trials

R-TIM model including pedigrees-distributional assumptions

Often the so-called reduced rank form of a factor analytic model of order 1 often provides a more parsimonious fit. This model is equivalent to the random effects Draper and Guttman model, in which the neighbour and direct effects are given by, say for

$s = a,$

$$\mathbf{u}_{a_n} = \rho_a \mathbf{u}_{a_d}$$

This model can be fitted using the extended factor analytic algorithm and these lead to rank one forms for G_s given by $G_s = \lambda_s \lambda_s'$ where $\lambda_s' = (\lambda_{s_1}, \lambda_{s_2})$ for $s = a, l$.

Competition modelling in sorghum plant breeding trials

R-TIM model including pedigrees-fitting the preferred model

Recall the vector-matrix representation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_g\mathbf{u}_a + \mathbf{Z}_g\mathbf{u}_\iota + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}$$

Take a deep breath

Competition modelling in sorghum plant breeding trials

R-TIM model including pedigrees-fitting the preferred model

Recall the vector-matrix representation as

$$y = X\tau + Z_g u_a + Z_g u_l + Z_p u_p + e$$

Take a deep breath

```
pytm.asr5 <- asreml(yield~ stand,  
random=~ str(~ ped(Geno)+ped(Gleft)+and(ped(Gright)), ~  
fa(2):id(ped(Geno)))+  
str(~ ide(Geno)+ide(Gleft)+and(ide(Gright)),~ fa(2):id(ide(Geno))) +  
Replicate + Column + Row,rcov=~ ar1(Column):ar1(Row),  
data=pytm,na.method.X='include',  
ginverse=list(Geno=pytm.ginv,Gleft=pytm.ginv,Gright=pytm.ginv),  
G.param=temp,R.param=temp)
```

Competition modelling in sorghum plant breeding trials

R-TIM model including pedigrees-model summary

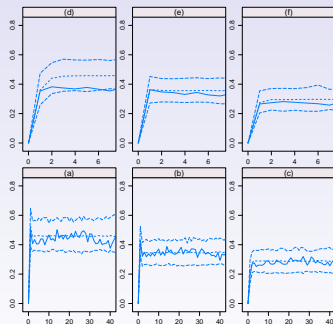
Summary of models fitted. The notation RR() denotes the Draper and Guttman variance model for the bracketed term; D - direct effects, N - neighbour effects.

| Model | Add | Nonadd | Other | REMLLL | Test | P-val |
|-------|---------|---------|---------|--------|----------|-------|
| 1 | D | D | | -32.54 | | |
| 2 | D | D | Row,Col | -10.32 | | |
| 2a | D | D | Col | -30.35 | M2a v M2 | 0.000 |
| 2b | D | D | Row | -14.46 | M2b v M2 | 0.002 |
| 3 | RR(D,N) | RR(D,N) | Row,Col | 0.00 | | |
| 3a | RR(D,N) | D | Row,Col | -1.94 | M3a v M3 | 0.049 |
| 3b | D | RR(D,N) | Row,Col | -6.20 | M3b v M3 | 0.000 |

Competition modelling in sorghum plant breeding trials

Variogram slicing

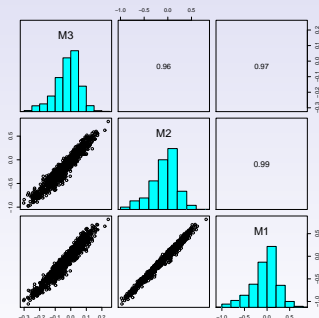
Plots of row and column faces of the empirical semi-variogram for the residuals from models 1,2,3 (solid line). Plots are augmented with the mean and 95% point-wise coverage intervals of the row and column faces of the empirical semi-variogram from a parametric bootstrap sample of size 100.



Competition modelling in sorghum plant breeding trials

Additive E-BLUPS

Pairwise scatter plots (lower triangle), simple correlation coefficients (upper triangle) and histograms (diagonals) of the E-BLUPS of the pure-stand effects from model 3, and the E-BLUPS of the direct effects from models 1 and 2 for the additive effects of the F_4 male parents



Partial compositing example from Smith *et al.*

Single trial analysis

$$Dy = DX\tau + DZ_g u_g + DZ_p u_p + De$$

- Data has been “averaged” commensurate with compositing process, ie. started with n plots and have reduced to s samples (a mixture of composite and individual plot samples)
- D is $s \times n$ averaging matrix
- Our example: $n = 180$, $s = 120$ and we have 60 samples that are composites of 2 plots and 60 that are individual plot samples
- Model involves non-standard design matrices: use “grp” facility in ASReml-R

Partial compositing example from Smith *et al.*

Single trial analysis

- Model fitted to example data (individual reps only):

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

```
kel.asr1 <- asreml(oil ~ 1 + linrow,  
random = ~ Entry + Block +  
Column, rcov = ~ ar1(Column):ar1(Row), data=kel.df)
```

- Now assume mixture of composite and individual plot samples and fit same model

$$D\mathbf{y} = D\mathbf{X}\boldsymbol{\tau} + D\mathbf{Z}_g\mathbf{u}_g + D\mathbf{Z}_p\mathbf{u}_p + D\mathbf{e}$$

```
kelpc.asr <- asreml(oil ~ 1 + linrow,  
random = ~ Entry + grp('Block') + grp('Column') +  
str( ~('Plot'), ~ar1v(6):ar1(30)),  
family = asreml.gaussian(dispersion=0.0001), data=kelpc.df,  
control=asreml.control(group=list(Block=184:185,Column=186:191,Plot=4:183))
```

Modelling variance structures for multi-term effects

Conclusions and Further Work

- We have illustrated some of the applications of modelling variance structures for multi-term effects
- There are numerous other examples (see for example, identification of QTL using spatial smoothing (Smith *et al.* (2011)))
- We hope that this talk has provided the stimulus to many other exciting and challenging applications

THANK YOU