3.1 Variance structures

Arthur Gilmour
Overview

- Traditional variance models assume independent effects: $\sigma^2 I$

- General variance structures
  - Unstructured - every variance and covariance is a separate parameter
  - Structured - variances and covariances are functions of parameters

- Spatial models
  - correlation based on distance
  - parameterized in terms of correlation and variance
Overview

- Traditional variance models
- General variance structures
  - Unstructured - Structured
- Spatial models
  - correlation based on distance
  - parameterized in terms of correlation and variance
- Compound variance structures
  - formed as a direct product
General Variance structures

- Unstructured (US) is parameterised directly as variances and covariances
- Symmetric Lower triangle rowwise

\[
\begin{bmatrix}
V_{11} & & \\
V_{21} & V_{22} & \\
V_{31} & V_{32} & V_{33}
\end{bmatrix}
\]
Reduced parameterization

- Diagonal (DIAG) has zero covariances
- Factor Analytic (FACV, XFA): \( \Sigma = \Lambda \Lambda' + \Psi \)
- Cholesky (CHOLn, CHOLnC): \( \Sigma = LDL' \)
  where \( L \) is unit lower triangle
- Antedependence (ANTEn): \( \Sigma^{-1} = UDU' \)
  where \( U \) is unit lower triangle
Aim in using alternate forms is
–to accommodate the variance heterogeneity adequately while minimising the number of parameters
–force a positive definite structure.

ANTE (a generalization of AR) is suited to ordered levels (e.g. times)

CHOL, XFA, FACV are suited to unordered levels (e.g. sites, traits)
General variance structures

- **DIAG** - off diagonal is zero
- **CHOL**\(_i\) - \( \Sigma = LDL' \)
  - \( L \) is lower triangle unit matrix with \( i \) off-diagonal bands
  - \( D \) is diagonal matrix of conditional variances.
e.g. in CHOL1 $L = \begin{pmatrix}
1 & 0 & 0 & 0 \\
a & 1 & 0 & 0 \\
0 & b & 1 & 0 \\
0 & 0 & c & 1
\end{pmatrix}$

$D = \text{diag}(A, B, C, D)$ so that

$$\Sigma = \begin{pmatrix}
A & aA & 0 & 0 \\
aA & aAa + B & bB & 0 \\
0 & bB & bBb + C & cC \\
0 & 0 & cC & cCc + D
\end{pmatrix}$$
**CHOL1C of order 4**

- e.g. in CHOL1C $L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ a & 1 & 0 & 0 \\ b & 0 & 1 & 0 \\ c & 0 & 0 & 1 \end{pmatrix}$

$D = \text{diag}(A, B, C, D)$ so that

$$\Sigma = \begin{pmatrix} A & aA & bA & cA \\ aA & aAa + B & bAa & cAa \\ bA & bAa & bAb + C & bAc \\ cA & aAc & cAb & cAc + D \end{pmatrix}$$
Antedependence

- is a generalized form of Autoregressive
- ANTE\(_i\) - \(\Sigma^{-1} = UD\)\(D\)'\n  - \(U\) is upper triangle unit matrix with \(i\) off-diagonal bands
  - \(D\) is diagonal matrix of conditional inverse variances.

- Since parameterization is obtuse for CHOL and ANTE, you may supply an unstructured matrix as starting values and ASReml will factorize it.
Factor Analytic

- Correlation Form: $FA_i$
  \[ \Sigma = D(LL' + E)D' \]
  Parameters are elements of $p \times i$ matrix $L$ and $\text{diag}(\Sigma) = DD$; $E$ is defined such that $\text{diag}(LL' + E)$ is Identity.

- Variance Form: $FACV_i$
  \[ \Sigma = \Lambda \Lambda' + \Psi \]
  Parameters are $\Lambda = DL$ and $\Psi = DED$
Extended Factor Analytic

- Same parameterization as FACV but in order $(\Psi) \ vec(\Lambda)$
- Elements of $\Psi$ may be zero (making $\Sigma$ singular)
- Requires use of $xfa(T, i)$ model term which inserts $i$ columns of zeros into the design matrix corresponding to the $i$ factors.
- Much faster than $FA_i$ and $FACV_i$ when more than 10 levels in term.
Extended Factor Analytic

... xfa(Trait,1).dam ...
  xfa(Trait,1).dam 2
  xfa(Trait,1) 0 XFA1
  2*0
  1.1 0.9

dam

Covariance/Variance/Correlation Matrix

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.550</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1.437</td>
<td>1.332</td>
<td>1.000</td>
</tr>
<tr>
<td>1.245</td>
<td>1.154</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Other structures

- US - unstructured
- OWN\(i\) - user supplies program to calculate G and the derivatives of G
- AINV - Use fixed relationship matrix
- GIV\(i\) - Use user defined fixed relationship matrix (see .giv, .grm)
Spatial structures

- ID - Identity
- CORU - uniform correlation
- AR1: $\rho \rho^2 \rho^3 \rho^4 \rho^5 \ldots$
- AR2, MA1, MA2, ARMA, SAR1, SAR2, CORU, CORB, CORH
- EXP, GAU
- IEXP, AEXP, IGAU, AGAU, IEUC, LVR, ISP, SPH, MAT

one or two dimensional distance
Variances

- Equal variance correlation
  append V to code e.g. AR1V, CORUV

- Unequal (Heterogeneous) variance correlation
  append H to code e.g. AR1H, CORUH

- If $D$ is the diagonal matrix of variances, and $C$ is a correlation matrix, $\Sigma = D^{0.5} CD^{0.5}$
3.2 Spatial Analysis

Arthur Gilmour
Two basic kinds

- Regular grid e.g. field trial
  - interest is in adjusting for other effects
Two basic kinds

- Regular grid e.g. field trial
  - interest is in adjusting for other effects
- Irregular grid e.g. survey
  - interest is in modelling the spatial pattern
  - kriging
Two basic kinds

- Regular grid e.g. field trial
  - interest is in adjusting for other effects

- Irregular grid e.g. survey
  - interest is in modelling the spatial pattern
  - kriging

- ASReml is regularly used for former
  - developing capability for latter
Single field trial

- Slate Hall Farm - Barley 1976
  - Balanced Incomplete block design
  - 25 varieties, 6 replicates
  - layout 10 rows by 15 columns

- BIB Model
  fixed: treatments
  random: rep block

- Spatial Model
  Autoregressive error model $R = \Sigma_R \otimes \Sigma_C$
Slate Hall base

- Slate Hall 1976 Cereal trial
  - rep 6 latrow 30 latcol 30 fldrow 10 fldcol 15
  - variety 25
  - yield !/100
  - shf.dat !DOPART $1
  - !DISPLAY 15 !SPATIAL !TWOWAY
Slate Hall - Design based

- PART 1 RCB Analysis
  \[ \text{yield} \sim \mu, \text{var} !r, \text{rep} \]

- PART 2 # BIB analysis
  \[ \text{yield} \sim \mu, \text{var} !r, \text{rep}, \text{latrow}, \text{latcol} \]
!PART 3 #  Fitting AR1.AR1
yield ~ mu var
predict var
 1 2
fldrow fldrow AR .1
fldcol fldcol AR .1
PART 4 # Fitting AR1.AR1
yield ~ mu var !r rep latrow latcol
predict var
  1 2
fldrow fldrow AR .1
fldcol fldcol AR .1
### Slate Hall - summary

<table>
<thead>
<tr>
<th>Model</th>
<th>LogL(l)</th>
<th>$-2\Delta(l)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCB</td>
<td>-167.694</td>
<td>2</td>
</tr>
<tr>
<td>BIB design</td>
<td>-132.134</td>
<td>4</td>
</tr>
<tr>
<td>Spatial model</td>
<td>-124.676</td>
<td>3</td>
</tr>
<tr>
<td>BIB+Spatial</td>
<td>-124.312</td>
<td>6</td>
</tr>
</tbody>
</table>

- Spatial correlation model fits better than the BIB model
Spatial components

<table>
<thead>
<tr>
<th>Source</th>
<th>terms</th>
<th>Gamma</th>
<th>Component</th>
<th>Comp/SE</th>
<th>% C</th>
</tr>
</thead>
<tbody>
<tr>
<td>rep</td>
<td>6</td>
<td>6</td>
<td>2.003E-05</td>
<td>7.24166E-05</td>
<td>0.00</td>
</tr>
<tr>
<td>latrow</td>
<td>30</td>
<td>30</td>
<td>6.327E-01</td>
<td>2.28684</td>
<td>0.71</td>
</tr>
<tr>
<td>latcol</td>
<td>30</td>
<td>30</td>
<td>1.608E-03</td>
<td>5.81362E-03</td>
<td>0.00</td>
</tr>
<tr>
<td>Variance</td>
<td>150</td>
<td>125</td>
<td>1.000</td>
<td>3.61464</td>
<td>4.28</td>
</tr>
<tr>
<td>Residual</td>
<td>AutoR</td>
<td>10</td>
<td>0.4652</td>
<td>0.465209</td>
<td>4.85</td>
</tr>
<tr>
<td>Residual</td>
<td>AutoR</td>
<td>15</td>
<td>0.6741</td>
<td>0.674095</td>
<td>8.76</td>
</tr>
</tbody>
</table>
Variogram

Slate Hall 1976 Cereal trial
Variogram of residuals 31 Jan 2005 16:15:30

Outer displacement  Inner displacement
Residual to plan

Slate Hall 1976 Cereal trial
Field plot of residuals 31 Jan 2005 16:15:30
Range: -4.80  5.37
Slate Hall 1976 Cereal trial
Residuals V Row and Column position: 31 Jan 2005 16:15:30
Range: -4.80 5.37
Multi environment trial

- In early generational cereal breeding, run several trials with 1 or two replicates of test lines, 20 percent check lines for error estimation.

- More power from fitting as correlated effects across sites.
MET in ASReml

- Three Multi Environment Trial
  seq
  col 15  # Actually 12 12 and 15 respectively
  row 34  # Actually 34 34 and 28 respectively
  chks 7  # Check 7 is the test lines
  test 336  # coded 0 for check lines
  geno 337
  yld  !*.01
  site 3
  met.dat  !section site
Spatial models

\[ \text{yld} \sim \text{site} \ \text{chk.site} \ !r \ \text{at(site,3).row} \ .02, \]
\[ \ \text{at(site).col} \ .90 \ .40 \ .036 \ \text{site.test} \]

\[
\begin{array}{ccc}
\text{site} & 2 & 1 \\
12 \ \text{col} & \text{AR} & .1271 \\
34 \ \text{row} & \text{AR} & .751 \\
12 \ \text{col} & \text{AR} & .25 \\
34 \ \text{row} & \text{AR} & .56 \\
15 \ \text{col} & \text{ID} & \\
28 \ \text{row} & \text{AR} & .38
\end{array}
\]

\!S2=2.19
\!S2=0.84
\!S2=0.19
Model genetic variation

- site.test 2
  - site 0 FA1
    - .5 .5 .5
    - .1 .1 .1
  - test
## Components

<table>
<thead>
<tr>
<th>Source</th>
<th>Model terms</th>
<th>Component</th>
<th>Comp/SE % C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual</td>
<td>1236</td>
<td>1213</td>
<td></td>
</tr>
<tr>
<td>at(site,01).col</td>
<td>15</td>
<td>15</td>
<td>0.323302E-05 0.00  0 B</td>
</tr>
<tr>
<td>at(site,02).col</td>
<td>15</td>
<td>15</td>
<td>0.142114 1.32  0 P</td>
</tr>
<tr>
<td>at(site,03).col</td>
<td>15</td>
<td>15</td>
<td>0.446791E-01 1.77 0 P</td>
</tr>
<tr>
<td>at(site,3).row</td>
<td>34</td>
<td>34</td>
<td>0.241380E-01 2.80  0 P</td>
</tr>
<tr>
<td>Variance[ 1]</td>
<td>408</td>
<td>0</td>
<td>2.60271 5.18  0 P</td>
</tr>
<tr>
<td>Residual</td>
<td>AR=AutoR</td>
<td>12</td>
<td>0.407051 4.45  0 U</td>
</tr>
<tr>
<td>Residual</td>
<td>AR=AutoR</td>
<td>34</td>
<td>0.882580 33.50 0 U</td>
</tr>
<tr>
<td>Variance[ 2]</td>
<td>408</td>
<td>0</td>
<td>1.00339 8.29  0 P</td>
</tr>
<tr>
<td>Residual</td>
<td>AR=AutoR</td>
<td>12</td>
<td>0.282407 4.84  0 U</td>
</tr>
<tr>
<td>Residual</td>
<td>AR=AutoR</td>
<td>34</td>
<td>0.580701 11.37 0 U</td>
</tr>
<tr>
<td>Variance[ 3]</td>
<td>420</td>
<td>0</td>
<td>0.105411 5.59  0 P</td>
</tr>
<tr>
<td>Residual</td>
<td>AR=AutoR</td>
<td>28</td>
<td>0.687455 10.14 0 U</td>
</tr>
</tbody>
</table>
Factor Analytic

<table>
<thead>
<tr>
<th>site.test</th>
<th>FA</th>
<th>D(L)</th>
<th>1</th>
<th>1</th>
<th>0.518516</th>
<th>5.35</th>
<th>0</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>site.test</td>
<td>FA</td>
<td>D(L)</td>
<td>1</td>
<td>2</td>
<td>1.13028</td>
<td>2.18</td>
<td>0</td>
<td>U</td>
</tr>
<tr>
<td>site.test</td>
<td>FA</td>
<td>D(L)</td>
<td>1</td>
<td>3</td>
<td>0.735010</td>
<td>6.04</td>
<td>0</td>
<td>U</td>
</tr>
<tr>
<td>site.test</td>
<td>FA</td>
<td>D(L)</td>
<td>0</td>
<td>1</td>
<td>0.991585</td>
<td>7.99</td>
<td>0</td>
<td>U</td>
</tr>
<tr>
<td>site.test</td>
<td>FA</td>
<td>D(L)</td>
<td>0</td>
<td>2</td>
<td>0.731805E-01</td>
<td>1.07</td>
<td>0</td>
<td>U</td>
</tr>
<tr>
<td>site.test</td>
<td>FA</td>
<td>D(L)</td>
<td>0</td>
<td>3</td>
<td>0.121810</td>
<td>7.17</td>
<td>0</td>
<td>U</td>
</tr>
</tbody>
</table>

Covariance/Variance/Correlation

FA D(LL’+E)D

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9916</td>
<td>0.5865</td>
<td>0.3811</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1579</td>
<td>0.7308E-01</td>
<td>0.8313</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1325</td>
<td>0.7844E-01</td>
<td>0.1218</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Spatial analysis in Forest Genetic trials.

- Typically not a complete rectangle
  - add missing values to complete the pattern
  - use map points (if < 5000 trees)
- With Tree model, must include Nugget variance
  - either Nugget is residual, spatial is in G
    or spatial is residual and Nugget is G,
- spatial model typically superior to 'design' model for growth/production traits
  - less so for disease and conformation traits
MicroArray

- spatial pattern
3.3 Repeated Measures

Arthur Gilmour
Main approaches

- General variance structure
  (Multivariate approach)
  UnStructured, Autoregressive, EXPponential

- Repeated measures
  Longitudinal model
  Repeated measures.
Multivariate approach

- Suited when most animals have most measures
- Repeats are at significant standard times
  Say WWT, 200dayWT, 400dayWT, 600dayWT
- Discuss
Multivariate

\[ \text{WWT WT200 WT400 WT600} \sim \text{Trait Tr.sex,} \]
\[ !r \text{Tr.animal} !f \text{Tr.cohort} \]

\[
\begin{array}{ccc}
1 & 2 & 1 \\
0 & & \\
\text{Trait 0 US} & & \\
10*0 & & \\
\text{Tr.animal 2} & & \\
\text{Tr 0 US} & & \\
10*0 & & \\
\text{animal 0 AINV} & & \\
\end{array}
\]
Multivariate

\textbf{WWT} WT200 WT400 WT600 \sim \text{Trait} Tr.\text{sex},
\text{!r Tr.animal} \text{!f Tr.cohort}

1 2 1
0
\text{Trait 0 US}
10*0

\text{Tr.animal 2}
\text{Tr 0 US}
10*0
\text{animal 0 AINV}
Multivariate

\[
\begin{align*}
\text{WT} & \quad \text{WT200} \quad \text{WT400} \quad \text{WT600} \\
\text{~ Trait } & \quad \text{Tr.sex,} \\
\text{!r } & \quad \text{Tr.animal} \\
\text{!f } & \quad \text{Tr.cohort} \\
1 & \quad 2 \quad 1 \\
0 & \\
\text{Trait} & \quad 0 \quad \text{US} \\
10*0 & \\
\text{Tr.animal} & \quad 2 \\
\text{Tr} & \quad 0 \quad \text{US} \\
10*0 & \\
\text{animal} & \quad 0 \quad \text{AINV}
\end{align*}
\]
Random Regression

- Appropriate when
  - there is considerable unbalance in times of measurement
  - there are varying numbers of measurements
  - all animals have multiple measures

- Concept: Regression for each individual consisting of an overall response pattern (fixed) plus an individual (random) adjustment.
RR principles

- This is a reduced parameterization model which must be well formulated
  - mean profile of higher order than random profile - random profile generally low order

- Usually formulated as polynomial but could be low order spline
RR Example

!WORK 150

This is random regression analysis of
animal !P sire 89 !I dam 1052 !I
year 2 !I !V21=V4 !==2 !*-365
flock 5 sex 2 !A aod
tobr 3 !I dob !-14800 !+V21
age wt fat emd

sdf01a.ped !SKIP 1
sdfwfm1.csv !SKIP 1 !MVremove !DOPART
!DDF !TYPEIIISS !MAXIT 20
PART 1  # Linear RR

emd ~ mu age year wt sex sex.wt flock,
tobr aod dob year.dob year.age,
year.sex year.flock year.tobr,
sex.dob tobr.dob,
!r animal animal.age,
  ide(animal) ide(animal).age,
  at(year,1,2).spl(age,20)
RR G structure

0 0 2
animal 2
2 0 US !GP  # Intercept and slope
1.3 0.01 0.01
animal 0 AINV
ide(animal) 2  # Intercept and slope
2 0 US !GP
1.6 0.01 0.03
ide(animal)
Fitting PART 1

- Fixed terms year.age, year.sex year.tobr are NS
- Variance of ide(animal).age is at boundary
- LogL after dropping 3 interactions was -726.867
PART 2  # Quadratic RR using pol
emd ~ mu age year wt sex sex.wt flock
dob year.dob year.flock sex.dob tob
!r pol(age,2).animal pol(age,1).ide
at(year,1,2).spl(age,20)
0 0 2
pol(age,2).animal 2
3 0 US
1.6 .6 .6 .3 .3 .3
animal 0 AINV
pol(age,1).ide(animal)
2 0 US
PART 2 G structures

0 0 2
pol(age,2).animal 2
3 0 US
1.6 .6 .6 .3 .3 .3
animal 0 AINV
pol(age,1).ide(animal)
2 0 US
2.1 .6 1.3
ide(animal)
PART 2

- LogL -643.67 so significant quadratic curvature
- Obtained initial values by ignoring G structure in initial run.
Spline curvature

!PART 3

!SPLINE spl(age,3) 4 0 6
emdmu age year wt sex sex.wt flock
year.dob year.age year.sex year.fl
!r animal animal.age animal.spl(age,3)
ide(animal) ide(animal).age,
ide(animal).spl(age,3),
at(year,1,2).spl(age,20)
0 0 2
animal 2
3 0 US !GU # Icept,slope,spl
1.3 0.1 0.01
Simpler

!PART 4

e md ~ mu age year wt sex sex.wt flock tobr aod

year.dob year.age -year.sex year.flock year.tobr

!r pol(age,2).animal ide(animal),

at(year,1).spl(age,20) at(year,2).spl(age,20)

0 0 1
pol(age,2).animal 2
3 0 US
1.6
.6 .6
3 3 3
Interpretation

.res file has pol() coefficients. say T
Form TGT’ to get full matrix of variances (all
times).