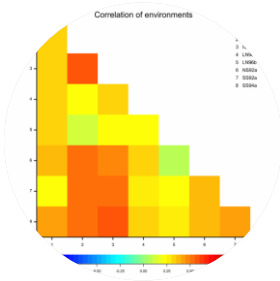
Mixed models for GxE data (Part I) Modelling genetic variances and covariances

Marcos Malosetti & Piter Bijma
Wageningen University & Research
Armidale, January 2017



WAGENINGEN UR
For quality of life

Contents

- GxE and heterogeneity of variance and correlation
- Compound symmetry as base model
- A series of models
- Model comparison, an example in maize

WAGENINGEN UR
For quality of life

So far discussed fixed-effects models

- Usually we only have a sample of the genotypes
- Random genotypic effects seems more natural for this type of data...

WAGENINGEN UR
For quality of life

For the animal breeder used to P = G + E:

- The data y_{ij} are averages of a large number of individuals per genotype-environment combination.
  - E.g. many plots of each genotype in each environment
  - Genotypes are discrete, e.g. clones or lines
- So y_{ij} is the genotypic value of genotype i in environment j,
  - not an individual phenotypic value
- In other words, for the k^{th} individual of genotype i in environment j:
  - $P_{ijk} = \mu + G_i + E_j + GE_{ij} + e_{ijk}$
  - But y_{ij} = average(P_{ijk}) = $\mu + G_i + E_j + GE_{ij}$
  - There is no individual environmental effect as in "P = G + E"
    - E_j is like HYS
    - GE_{ij} is like HYS*breed

Modelling mean and VCOV for MET data in LMM

$$\underline{y}_{ij} = \mu_{ij} + \underline{\epsilon}_{ij} \qquad \underline{\epsilon}_{ij} \sim MVN(0, \Sigma)$$

- Aim of statistical modelling for MET/GxE data
  - μ_{ij}: predictable
    - **Goal:** Describe μ_{ij} as much as possible in terms of single indexed parameters
    - = Separate G_i and E_j terms
  - $VCOV(\underline{\epsilon}_{ij}) = \Sigma$: unpredictable
    - Find an appropriate structure for $\underline{\epsilon}_{ij}$
      - Genetic (co)variances.
    - Model dependencies between:
      - Genotypes (kinship)
      - Environments (genetic correlation)

For animal breeders:
Interest is in the fixed effects here

Modelling GxE in segregating populations

$$\underline{y}_{ij} = \mu + E_j + \underline{\epsilon}_{ij} \qquad \underline{\epsilon}_{ij} \sim MVN(0, \Sigma)$$

- μ is a fixed intercept, and E_j fixed environment effect.
- $\underline{\epsilon}_{ij}$ is the total random variation (including environment-specific genotypic effects).
- $\Sigma = \Sigma^G \otimes \Sigma^E$: variance-covariance matrix.
  - Σ^G: takes here a simple form (identity).
    - We assume a set of unrelated genotypes ("lines")
  - Σ^E: Can take different forms (genetic correlations between environments).
- Both Σ^G and Σ^E define similarities among genetic effects!

WAGENINGEN UR
For quality of life

1. Basic compound symmetry model (CS)

$$\underline{y}_{ij} = \mu + E_j + \underline{G}_i + \underline{\epsilon}_{ij} \qquad \begin{matrix} \underline{G}_i \sim N(0, \sigma_G^2) \\ \underline{\epsilon}_{ij} \sim N(0, \sigma^2) \end{matrix}$$

- We take genotypes as random effects ($\underline{G}_i$).

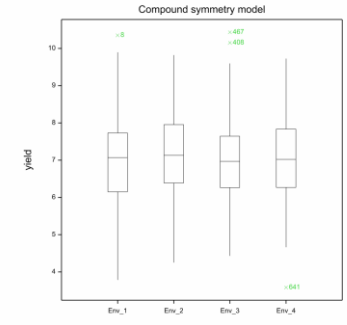
$$\Sigma = \begin{bmatrix} \sigma_G^2 + \sigma^2 & & & \\ \sigma_G^2 & \sigma_G^2 + \sigma^2 & & \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma^2 & \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma^2 \end{bmatrix}$$

$$corr(E_j; E_{j^*}) = \frac{\sigma_G^2}{\sqrt{\sigma_G^2 + \sigma^2}\sqrt{\sigma_G^2 + \sigma^2}} = \frac{\sigma_G^2}{\sigma_G^2 + \sigma^2}$$

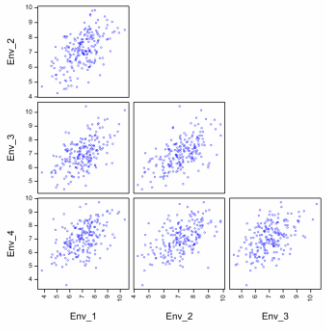
WAGENINGEN UR For quality of life

What is curious in this model?

Compound symmetry assumes this…



Constant variance within environments



Constant covariance (and correlation)

WAGENINGEN UR For quality of life

Is this realistic ?

2. Heterogeneous compound symmetry model (HCS)

$$\underline{y}_{ij} = \mu + E_j + \underline{G}_i + \underline{\epsilon}_{ij}$$

$$\underline{G}_i \sim N(0, \sigma_G^2)$$
$$\underline{\epsilon}_{ij} \sim N(0, \sigma_j^2)$$

$$\Sigma = \begin{bmatrix} \sigma_G^2 + \sigma_1^2 & & & \\ \sigma_G^2 & \sigma_G^2 + \sigma_2^2 & & \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma_3^2 & \\ \sigma_G^2 & \sigma_G^2 & \sigma_G^2 & \sigma_G^2 + \sigma_4^2 \end{bmatrix}$$

$$corr(E_j; E_{j^*}) = \frac{\sigma_G^2}{\sqrt{\sigma_G^2 + \sigma_j^2}\sqrt{\sigma_G^2 + \sigma_{j^*}^2}}$$

- Residual variances are allowed to change (σ_j^2)
- Co-variance still constant, but correlations do change!
  - A little bit, only due to scaling of the variance

3. Diagonal model (heterogeneous variance, zero correlation)

$$\underline{y}_{ij} = \mu + E_j + \underline{\epsilon}_{ij}$$

$$\underline{\epsilon}_{ij} \sim N\left(0, \sigma_{G_j}^2\right)$$

$$\Sigma = \begin{bmatrix} \sigma_{G_1}^2 & & & \\ & \sigma_{G_2}^2 & & \\ & & \sigma_{G_3}^2 & \\ & & & \sigma_{G_4}^2 \end{bmatrix}$$

$$corr(E_j; E_{j^*}) = \frac{0}{\sigma_j\, \sigma_{j^*}}$$

- Each environment has its own (residual) genetic variance
- There is no genetic correlation between environments
  - Probably unlikely: in two similar environments, the same genotype will perform similarly

4: Unstructured model: FULL model ("multitrait model")

$$\underline{y}_{ij} = \mu + E_j + \underline{\epsilon}_{ij} \qquad \underline{\epsilon}_{ij} \sim N(0, \Sigma)$$

E.g., for one genotype in 4 environments:

$$\Sigma = \begin{bmatrix} \sigma^2_{G_1} & & & \\ \sigma_{12}\sigma^2_{G_2} & & & \\ \sigma_{13}\sigma_{23}\sigma^2_{G_3} & & & \\ \sigma_{14}\sigma_{24}\sigma_{34}\sigma^2_{G_4} & & & \end{bmatrix}$$

$$genetic\ corr(y_{ij}; y_{ij^*}) = \frac{\sigma_{jj^*}}{\sigma_j\ \sigma_{j^*}}$$

- One genetic variance per environment.
- One covariance per pair of environments.
  - Like a full multi-trait model in animal breeding
- Fits the full GxE, but provides no structure
  - Many VC to estimate (10 here)

5. Structure: Factor analytic model on the environments

Interpretation of a factor-analytical model:

- There exists an overall genotypic effect of each genotype (G_i)
- Each environment j captures an amount λ_j of this effect
- An amount ε_{ij} remains unexplained

$$\underline{y}_{ij} = \mu + E_j + \underline{\epsilon}_{ij} \quad \text{where} \quad \underline{\epsilon}_{ij} = \lambda_j G_i + \varepsilon_{ij}$$

The scale of G and λ cannot be separated $\rightarrow$ Var(G) = 1

$$var(\underline{\epsilon}_{ij}) = \lambda_j^2 + \psi_j \qquad cov(\underline{\epsilon}_{ij}, \underline{\epsilon}_{ij*}) = \lambda_j \lambda_{j*}$$

WAGENINGEN UR
For quality of life

5. Structure: Factor analytic model on the environments

$$\underline{y}_{ij} = \mu + E_j + \underline{\epsilon}_{ij} \qquad \underline{\epsilon}_{ij} \sim N(0, \Sigma)$$

$$\Sigma = \begin{bmatrix} \lambda_1^2 + \psi_1 & & & \\ \lambda_1\lambda_2 & \lambda_2^2 + \psi_2 & & \\ \lambda_1\lambda_3 & \lambda_2\lambda_3 & \lambda_3^2 + \psi_3 & \\ \lambda_1\lambda_4 & \lambda_2\lambda_4 & \lambda_3\lambda_4 & \lambda_4^2 + \psi_4 \end{bmatrix}$$

$$corr(E_j; E_{j^*}) = \frac{\lambda_j\lambda_{j^*}}{\sqrt{\lambda_j^2 + \psi_j}\sqrt{\lambda_{j^*}^2 + \psi_{j^*}}}$$

- Heterogeneity of variances and covariances allowed.
- Relatively few parameters ($2 \times n_j$, n_j = number of environments; 8 here)
  - Specially useful when number of environments increases
    - Identical to unstructured when n = 3

Summary of (some) models for VCOV

Variance	Covar	N par	Description
$\sigma_G^2 + \sigma_{GE}^2$	σ_G^2	2	Homog var, uniform covar
$\sigma_{G_j}^2$	0	n_j	Heterog var, no covar
$\sigma_G^2 + \sigma_{GE_j}^2$	σ_G^2	$n_j + 1$	Heterog var, uniform covar
$\sigma_{G_j}^2$	$\theta\sigma_{G_j}^2\sigma_{G_{j^*}}^2$	$n_j + 1$	Heterog var, uniform corr
$\lambda_j^2 + \psi_j$	$\lambda_j\lambda_{j^*}$	$2 \times n_j$	Factor analytic
$\sigma_{G_j}^2$	$\sigma_{G_{jj^*}}$	$n_j \times (n_j + 1)/2$	Unstructured

- Best model?
- Trade-off between parsimony (few parameters) and goodness of fit (depends on the data set).

WAGENINGEN UR
For quality of life

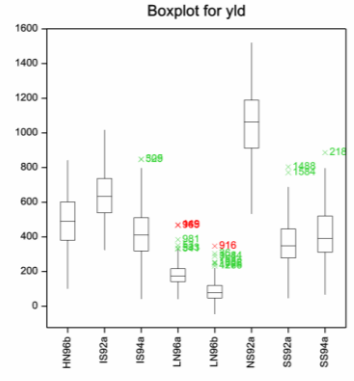
The maize example again (CIMMYT, Mexico)

- **Managed stress trials**
- **8 environments**
- 1992 (Tlaltizapán, México)
  - Well watered (WW)
  - Intermediate stress (IS)
  - Severe stress (SS)
- 1994 (Tlaltizapán, México)
  - Intermediate stress (IS)
  - Severe stress (SS)
- 1996 (Poza Rica, México)
  - Low Nitrogen (2 seasons)
  - High Nitrogen

WAGENINGEN UR
For quality of life

Indications for GxE: heterogeneity of var

- Make e.g. boxplots
- Heterogeneity of genetic variance suggests GxE…

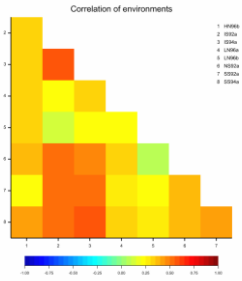


WAGENINGEN UR
For quality of life

Indications for GxE: Corr. between environments

Correlations

_e['HN96b']	1	-						
_e['IS92a']	2	0.3303	-					
_e['IS94a']	3	0.3481	0.5879	-				
_e['LN96a']	4	0.3192	0.2367	0.3177	-			
_e['LN96b']	5	0.3304	0.1471	0.2277	0.2302	-		
_e['NS92a']	6	0.3699	0.5313	0.4590	0.3165	0.0921	-	
_e['SS92a']	7	0.2005	0.5155	0.5281	0.2522	0.2474	0.3816	-
_e['SS94a']	8	0.4298	0.5155	0.5776	0.3487	0.2538	0.3901	0.4316
		1	2	3	4	5	6	7



- Low genetic correlation between environments
  - Implies GxE!

Model selection based on AIC or BIC

Summary

Trait: yield

Model	AIC	BIC	Deviance	NParameters
FA	634.9	688.6	602.9	16
FA2	619.6	696.7	573.6	23
OUTSIDE	687.7	717.9	669.7	9
UNSTRUCTURED	620.7	741.4	548.7	36
HCS	856.3	886.5	838.3	9
CS	1081.9	1088.6	1077.9	2
DIAGONAL	1070.6	1097.4	1054.6	8
IDENTITY	1451.2	1454.5	1449.2	1

Best model: FA (on basis of criterion BIC)

- Best model is FA when using BIC as criterion
  - BIC puts a higher penalty on the number of parameters
  - Unstructured has too many parameters

Fit of the best model (FA1 model)

Residual variance model

Term	Factor	Model(order)	Parameter	Estimate	s.e.
G.E			Sigma2	1.000	fixed
	G	Identity	-	-	-
	E	FA(1) (covariance form)			
			g_11	0.5001	0.0706
			g_21	0.7251	0.0618
			g_31	0.7698	0.0612
			g_41	0.1902	0.0314
			g_51	0.1291	0.0307
			g_61	0.8174	0.0884
			g_71	0.5574	0.0581
			g_81	0.7203	0.0644
			psi_1	0.7515	0.0779
			psi_2	0.4242	0.0529
			psi_3	0.3790	0.0509
			psi_4	0.1573	0.0160
			psi_5	0.1571	0.0156
			psi_6	1.068	0.117
			psi_7	0.4523	0.0501
			psi_8	0.4895	0.0587

λ_j

ψ_j

- Genetic variance in environment 1
  - $\lambda_1^2 + \psi_1 = 0.5001^2 + 0.7515 = 1.002$
- Genetic covariance between environment 1 and 2:
  - $\lambda_1\lambda_2 = 0.5001 \times 0.7251 = 0.362$

Estimated VCOV (and correlation)

Covariance matrix:

$\lambda_1^2 + \psi_1 = 0.5001^2 + 0.7515 = 1.002$

$\lambda_1\lambda_2 = 0.5001 \times 0.7251 = 0.363$

1	1.002							
2	0.363	0.950						
3	0.385	0.558	0.972					
4	0.095	0.138	0.146	0.194				
5	0.065	0.094	0.099	0.025	0.174			
6	0.409	0.593	0.629	0.155	0.106	1.736		
7	0.279	0.404	0.429	0.106	0.072	0.456	0.763	
8	0.360	0.522	0.554	0.137	0.093	0.589	0.401	1.008
	1	2	3	4	5	6	7	8

Observed vs estimated correlations

Correlations

Correlations from the data

_e['HN96b']	1	-						
_e['IS92a']	2	0.3303	-					
_e['IS94a']	3	0.3481	0.5879	-				
_e['LN96a']	4	0.3192	0.2367	0.3177	-			
_e['LN96b']	5	0.3304	0.1471	0.2277	0.2302	-		
_e['NS92a']	6	0.3699	0.5313	0.4590	0.3165	0.0921	-	
_e['SS92a']	7	0.2005	0.5155	0.5281	0.2522	0.2474	0.3816	-
_e['SS94a']	8	0.4298	0.5155	0.5776	0.3487	0.2538	0.3901	0.4316
		1	2	3	4	5	6	7

Correlation matrix:

Correlations from the FA1 model

HN96b	1.0000							
IS92a	0.3717	1.0000						
IS94a	0.3902	0.5810	1.0000					
LN96a	0.2161	0.3217	0.3377	1.0000				
LN96b	0.1548	0.2305	0.2419	0.1340	1.0000			
NS92a	0.3100	0.4615	0.4845	0.2683	0.1922	1.0000		
SS92a	0.3189	0.4747	0.4983	0.2759	0.1977	0.3959	1.0000	
SS94a	0.3584	0.5336	0.5602	0.3102	0.2222	0.4450	0.4577	1.0000
	HN96b	IS92a	IS94a	LN96a	LN96b	NS92a	SS92a	SS94a

WAGENINGEN UR
For quality of life

Summary

- Modelling of GxE from variance-covariance perspective.
- Simple diagnostic plots for GxE
  - heterogeneous variances
  - Correlations.
- Models of increasing complexity and model selection

WAGENINGEN UR
For quality of life