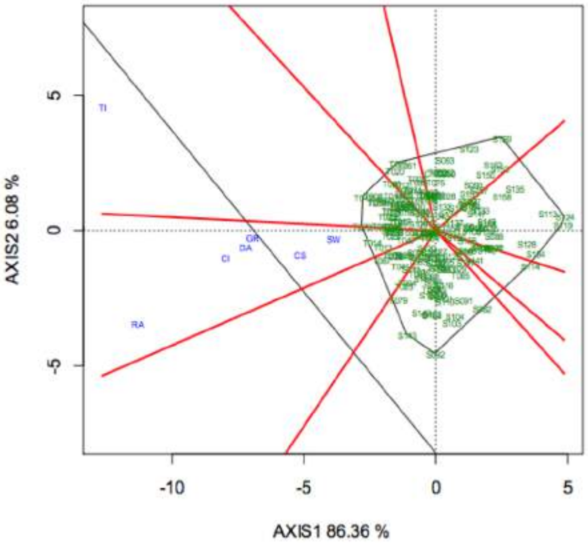
AGR 6322 Advanced Plant Breeding

Fall 2018

Genotype by Environment and Multi-Site Analysis



Goals for today

Multi-environment and Multi-variate analysis

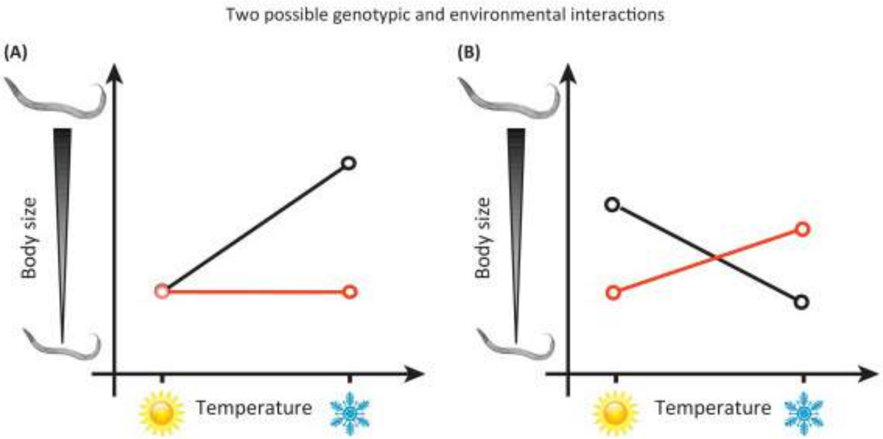
Linear Mixed Model

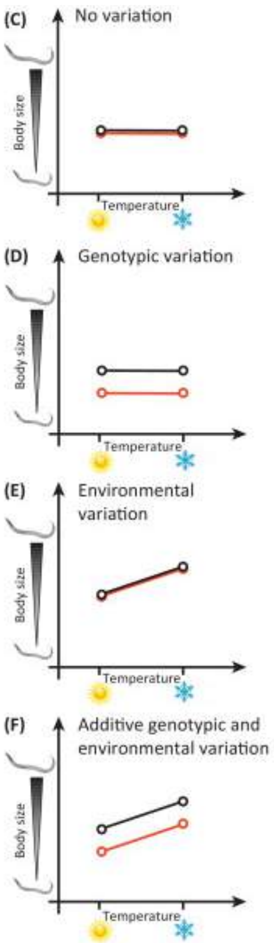
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where:

- **y**: response vector; observations
- **β**: vector of fixed effects
- **u**: vector of random effects; $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{G})$
- **X** and **Z**: (known) incidence matrices
- **e**: residual vector; $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$

Genotype by Environment and Multi-Site Analysis





$$P = G + E + GxE$$

$$V_P = V_G + V_E + V_{GxE}$$

GxE = when the difference in performance of two genotypes depends on the environment in which the performance is measured.

GxE can refer to a change in size of the difference in performance, or to a change in ranking in different environments

Genotype by Environment and Multi-Site Analysis

Strategies for fitting multi-environment models

- 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).
- 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).
- Considering favoring the simplest model that suits your requirements.
- Ideal objective: to fit a US structure to the GxE matrix to understand the genetic structure and evaluate stability of genotypes and breeding zones.
- Progress *slowly* to more complex variance

Variance Structures

id/idv: identity

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

diag: diagonal

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

corv: uniform correlation

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

ar1v: autocorrelation 1st order

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

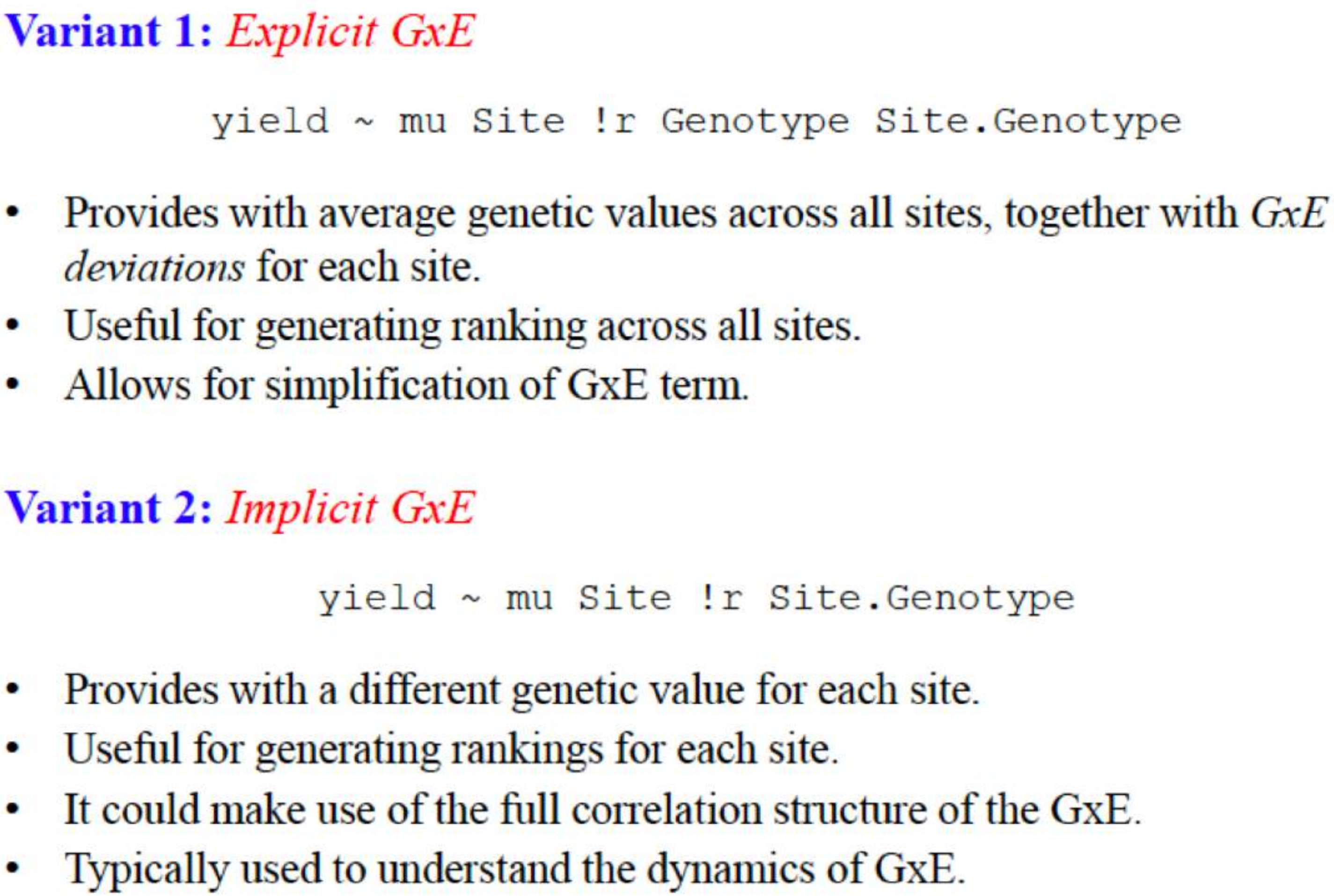
corh: uniform heterogeneous

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

us: unstructured

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

Genotype by Environment and Multi-Site Analysis



Variant 1: *Explicit GxE*

```
yield ~ mu Site !r Genotype Site.Genotype
```

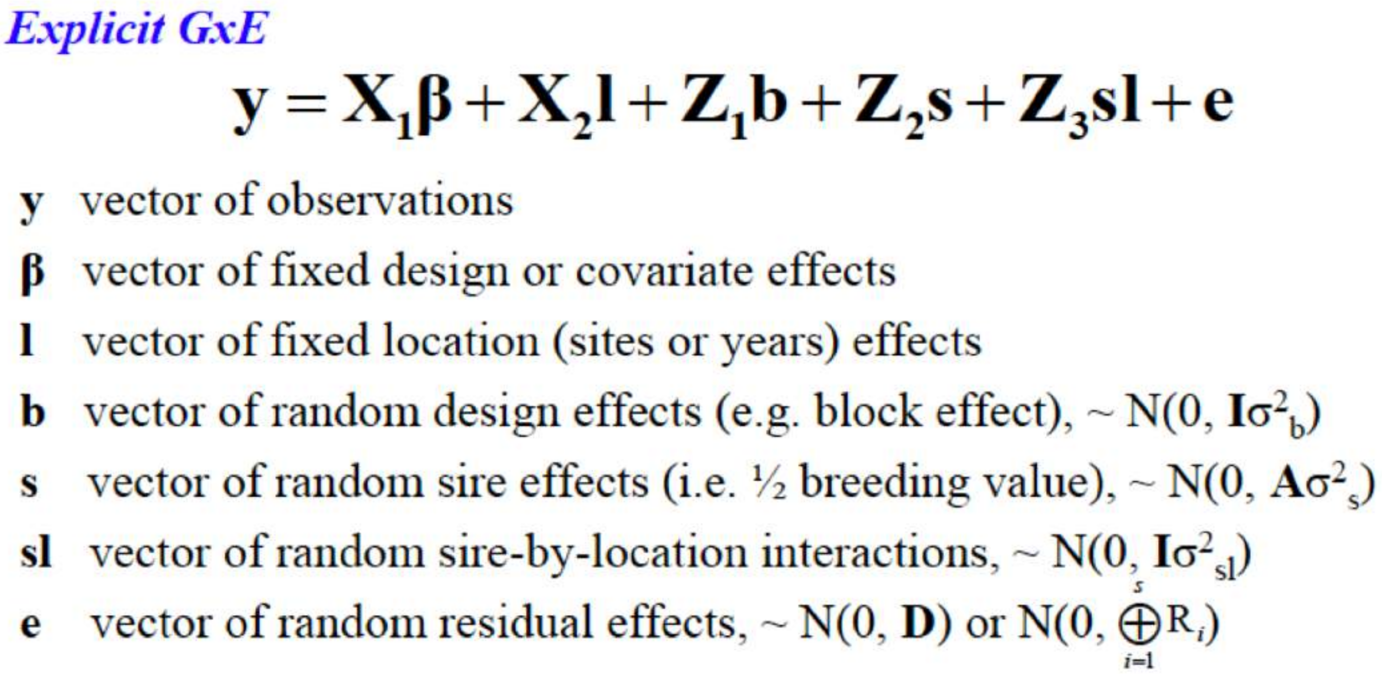
- Provides with average genetic values across all sites, together with *GxE deviations* for each site.
- Useful for generating ranking across all sites.
- Allows for simplification of GxE term.

Variant 2: *Implicit GxE*

```
yield ~ mu Site !r Site.Genotype
```

- 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.

Genotype by Environment and Multi-Site Analysis



Explicit GxE

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

y vector of observations

β vector of fixed design or covariate effects

l vector of fixed location (sites or years) effects

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

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

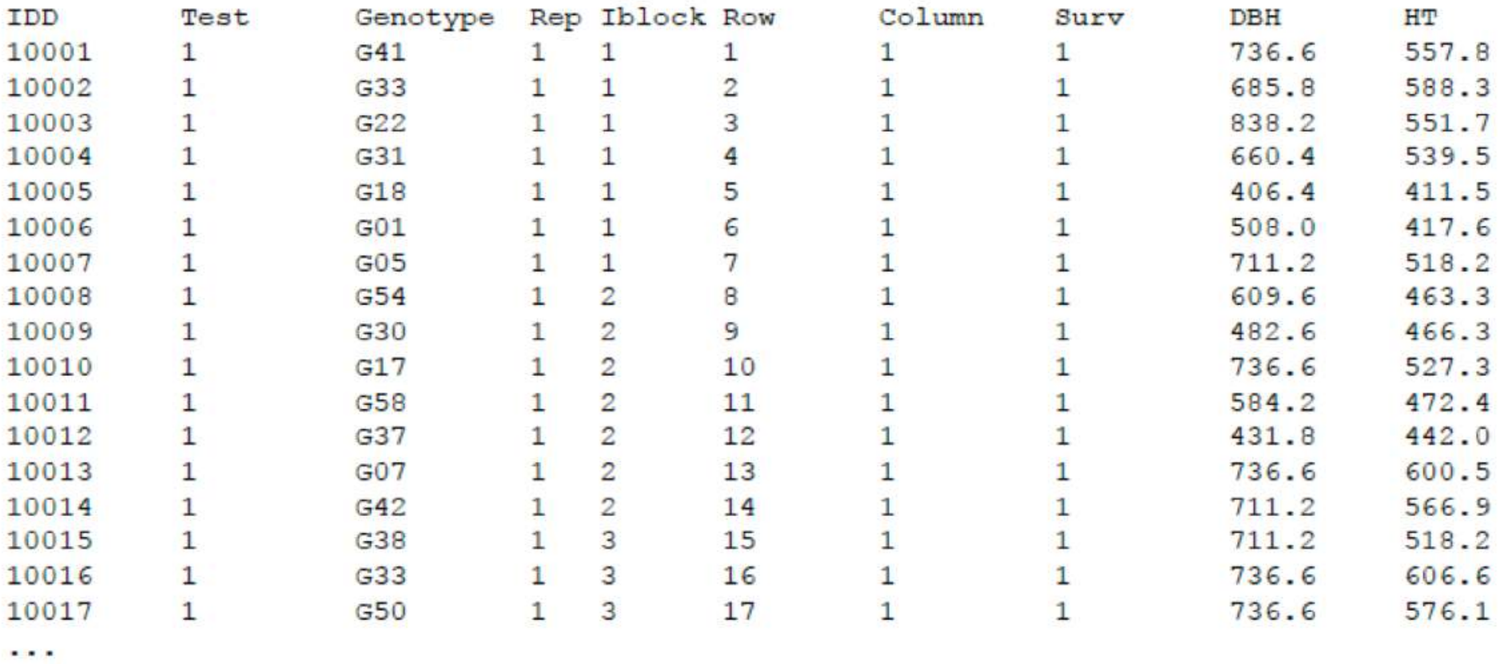
sl vector of random sire-by-location interactions, ~ N(0, $\mathbf{I}\sigma^2_{\mathrm{sl}}$)

e vector of random residual effects, ~ N(0, **D**) or N(0, $\bigoplus_{i=1}^{s} \mathrm{R}_i$)

Genotype by Environment and Multi-Site Analysis

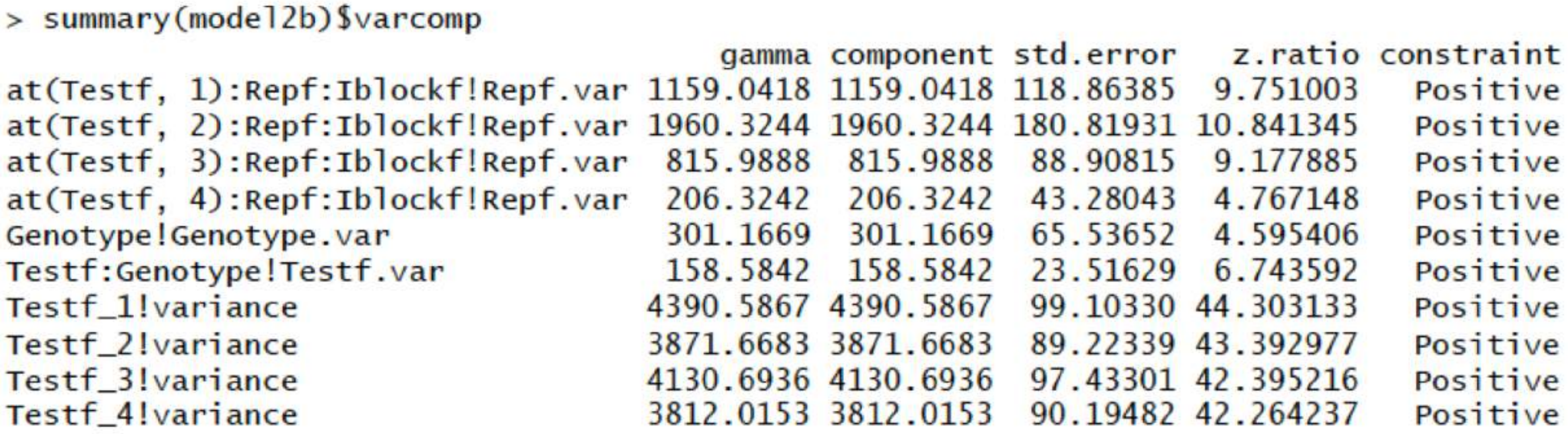
EXAMPLE

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	Rep	Iblock	Row	Column	Surv	DBH	HT
10001	1	G41	1	1	1	1	1	736.6	557.8
10002	1	G33	1	1	2	1	1	685.8	588.3
10003	1	G22	1	1	3	1	1	838.2	551.7
10004	1	G31	1	1	4	1	1	660.4	539.5
10005	1	G18	1	1	5	1	1	406.4	411.5
10006	1	G01	1	1	6	1	1	508.0	417.6
10007	1	G05	1	1	7	1	1	711.2	518.2
10008	1	G54	1	2	8	1	1	609.6	463.3
10009	1	G30	1	2	9	1	1	482.6	466.3
10010	1	G17	1	2	10	1	1	736.6	527.3
10011	1	G58	1	2	11	1	1	584.2	472.4
10012	1	G37	1	2	12	1	1	431.8	442.0
10013	1	G07	1	2	13	1	1	736.6	600.5
10014	1	G42	1	2	14	1	1	711.2	566.9
10015	1	G38	1	3	15	1	1	711.2	518.2
10016	1	G33	1	3	16	1	1	736.6	606.6
10017	1	G50	1	3	17	1	1	736.6	576.1
...									

Genotype by Environment and Multi-Site Analysis



```
> summary(model2b)$varcomp
                                         gamma  component  std.error   z.ratio constraint
at(Testf, 1):Repf:Iblockf!Repf.var 1159.0418 1159.0418 118.86385  9.751003   Positive
at(Testf, 2):Repf:Iblockf!Repf.var 1960.3244 1960.3244 180.81931 10.841345   Positive
at(Testf, 3):Repf:Iblockf!Repf.var  815.9888  815.9888  88.90815  9.177885   Positive
at(Testf, 4):Repf:Iblockf!Repf.var  206.3242  206.3242  43.28043  4.767148   Positive
Genotype!Genotype.var               301.1669  301.1669  65.53652  4.595406   Positive
Testf:Genotype!Testf.var            158.5842  158.5842  23.51629  6.743592   Positive
Testf_1!variance                   4390.5867 4390.5867  99.10330 44.303133   Positive
Testf_2!variance                   3871.6683 3871.6683  89.22339 43.392977   Positive
Testf_3!variance                   4130.6936 4130.6936  97.43301 42.395216   Positive
Testf_4!variance                   3812.0153 3812.0153  90.19482 42.264237   Positive
```

$\mathbf{V_a} = 4\, s^2_g = 4 \text{ x } 301.2 = 1204.7$

$\mathbf{V_{axs}} = 4\, s^2_{gs} = 4 \text{ x } 158.6 = 634.3$

$\mathbf{V_p} = 301.2 + 158.6 + (4141.7)/4 + (16235.0)/4 = 5553.9$

$h^2 = \mathbf{V_a} / \mathbf{V_p} = 1204.7 / 5553.9 = 0.217$

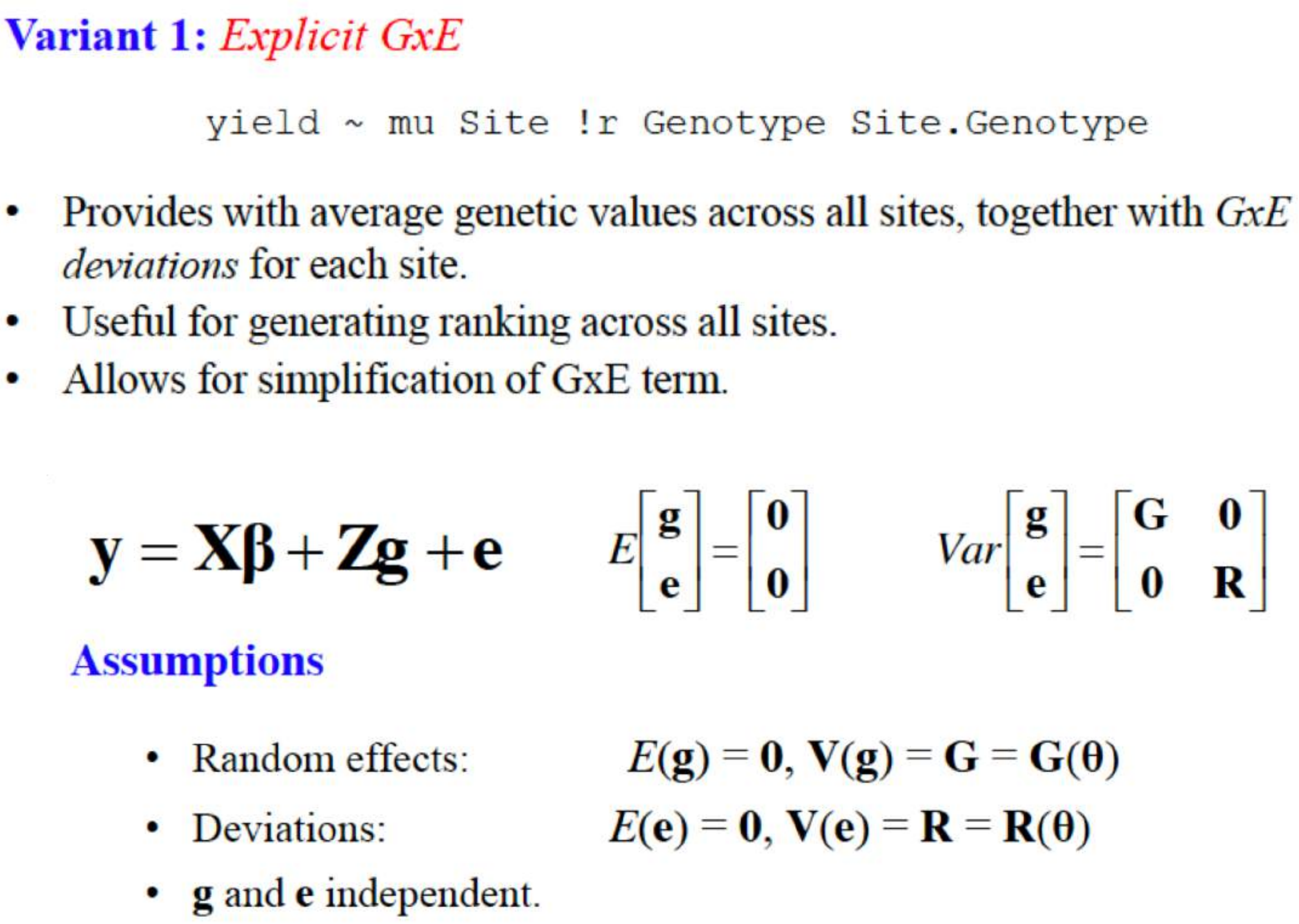
$rg_{\mathrm{B(a)}} = \mathbf{V_a} / [\mathbf{V_a} + \mathbf{V_{axs}}] = 1204.7 / [1204.7 + 634.3] = 0.655$

Note: individual site heritabilites can also be calculated.

Example

Genotype-by-environment Analysis in Bermudagrass

Genotype-by-environment Analysis in Bermudagrass



Variant 1: *Explicit GxE*

```
yield ~ mu Site !r Genotype Site.Genotype
```

- 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.

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

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.

Variance Structures

id/idv: 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

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

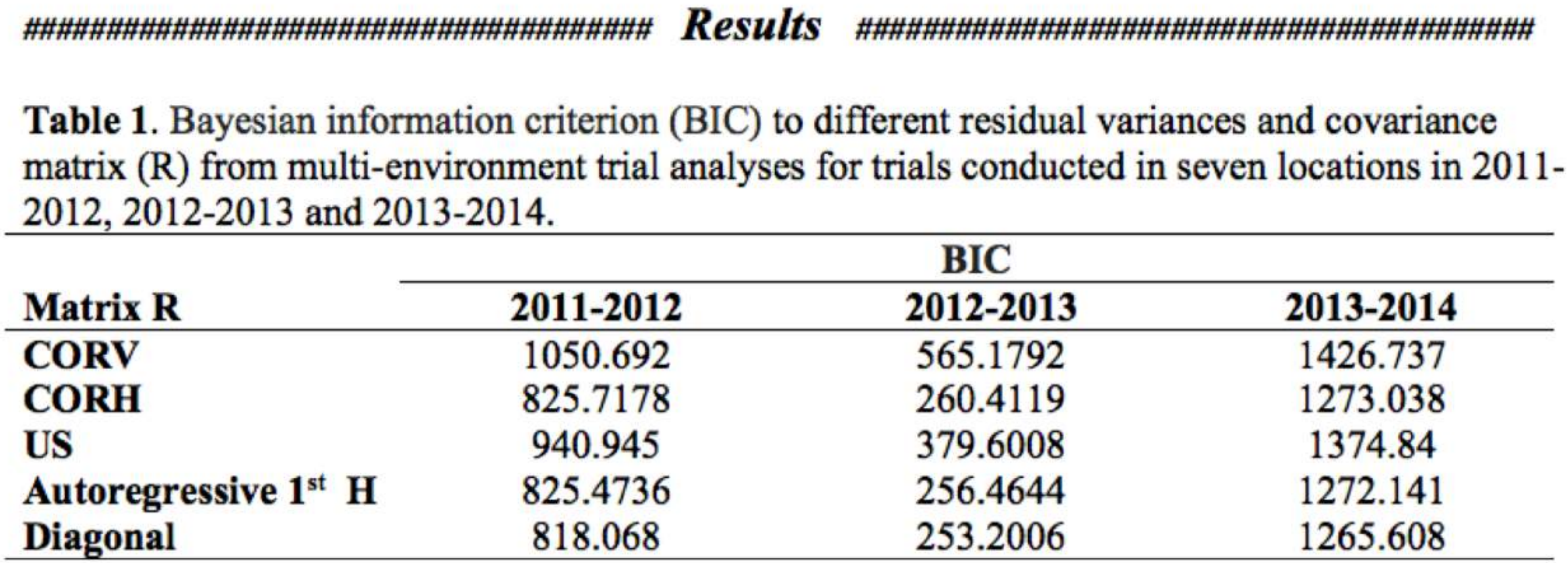
corv: uniform correlation

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

us: unstructured

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

Genotype-by-environment Analysis in Bermudagrass



Results

Table 1. Bayesian information criterion (BIC) to different residual variances and covariance matrix (R) from multi-environment trial analyses for trials conducted in seven locations in 2011-2012, 2012-2013 and 2013-2014.

	BIC		
Matrix R	**2011-2012**	**2012-2013**	**2013-2014**
CORV	1050.692	565.1792	1426.737
CORH	825.7178	260.4119	1273.038
US	940.945	379.6008	1374.84
Autoregressive 1st H	825.4736	256.4644	1272.141
Diagonal	818.068	253.2006	1265.608

Genotype-by-environment Analysis in Bermudagrass

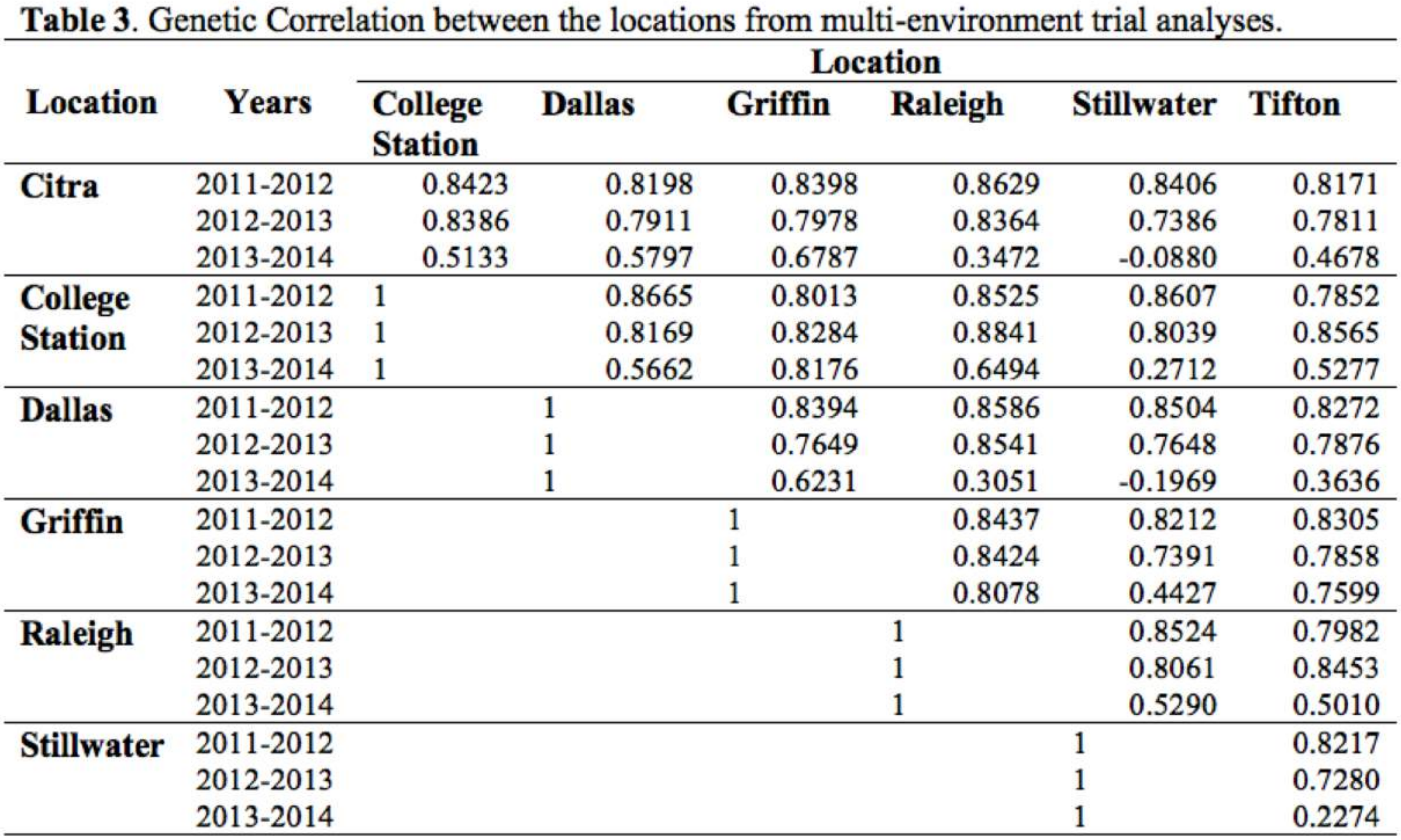


Table 3. Genetic Correlation between the locations from multi-environment trial analyses.

		Location					
Location	**Years**	**College Station**	**Dallas**	**Griffin**	**Raleigh**	**Stillwater**	**Tifton**
Citra	2011-2012	0.8423	0.8198	0.8398	0.8629	0.8406	0.8171
	2012-2013	0.8386	0.7911	0.7978	0.8364	0.7386	0.7811
	2013-2014	0.5133	0.5797	0.6787	0.3472	-0.0880	0.4678
College Station	2011-2012	1	0.8665	0.8013	0.8525	0.8607	0.7852
	2012-2013	1	0.8169	0.8284	0.8841	0.8039	0.8565
	2013-2014	1	0.5662	0.8176	0.6494	0.2712	0.5277
Dallas	2011-2012		1	0.8394	0.8586	0.8504	0.8272
	2012-2013		1	0.7649	0.8541	0.7648	0.7876
	2013-2014		1	0.6231	0.3051	-0.1969	0.3636
Griffin	2011-2012			1	0.8437	0.8212	0.8305
	2012-2013			1	0.8424	0.7391	0.7858
	2013-2014			1	0.8078	0.4427	0.7599
Raleigh	2011-2012				1	0.8524	0.7982
	2012-2013				1	0.8061	0.8453
	2013-2014				1	0.5290	0.5010
Stillwater	2011-2012					1	0.8217
	2012-2013					1	0.7280
	2013-2014					1	0.2274

Genotype-by-environment Analysis in Bermudagrass

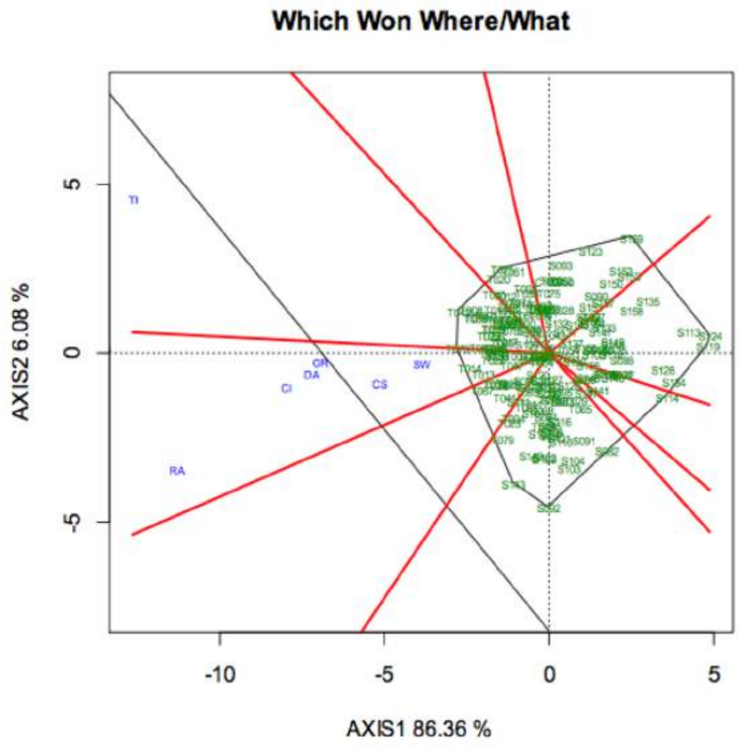


FIGURE 1. GGE 2011-2012.

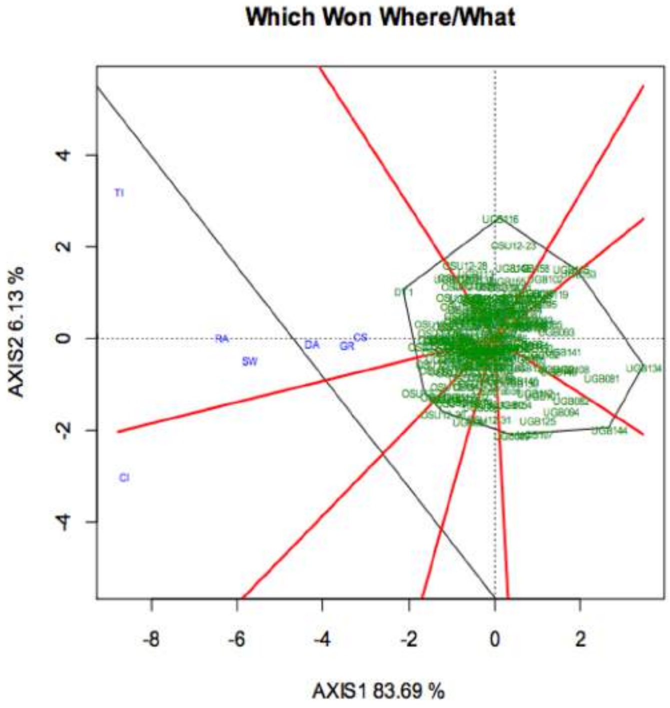
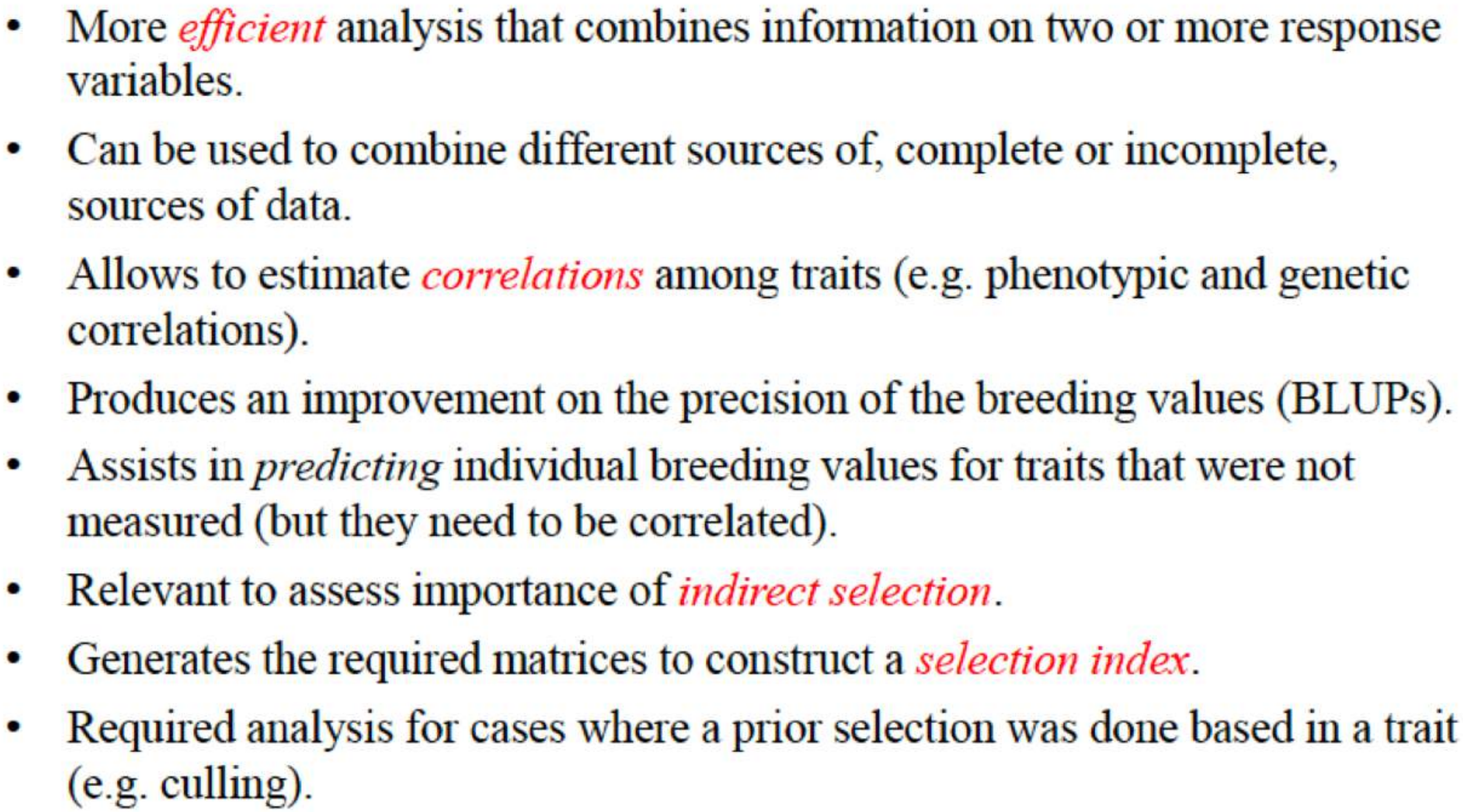


FIGURE. GGE 2012-2013.

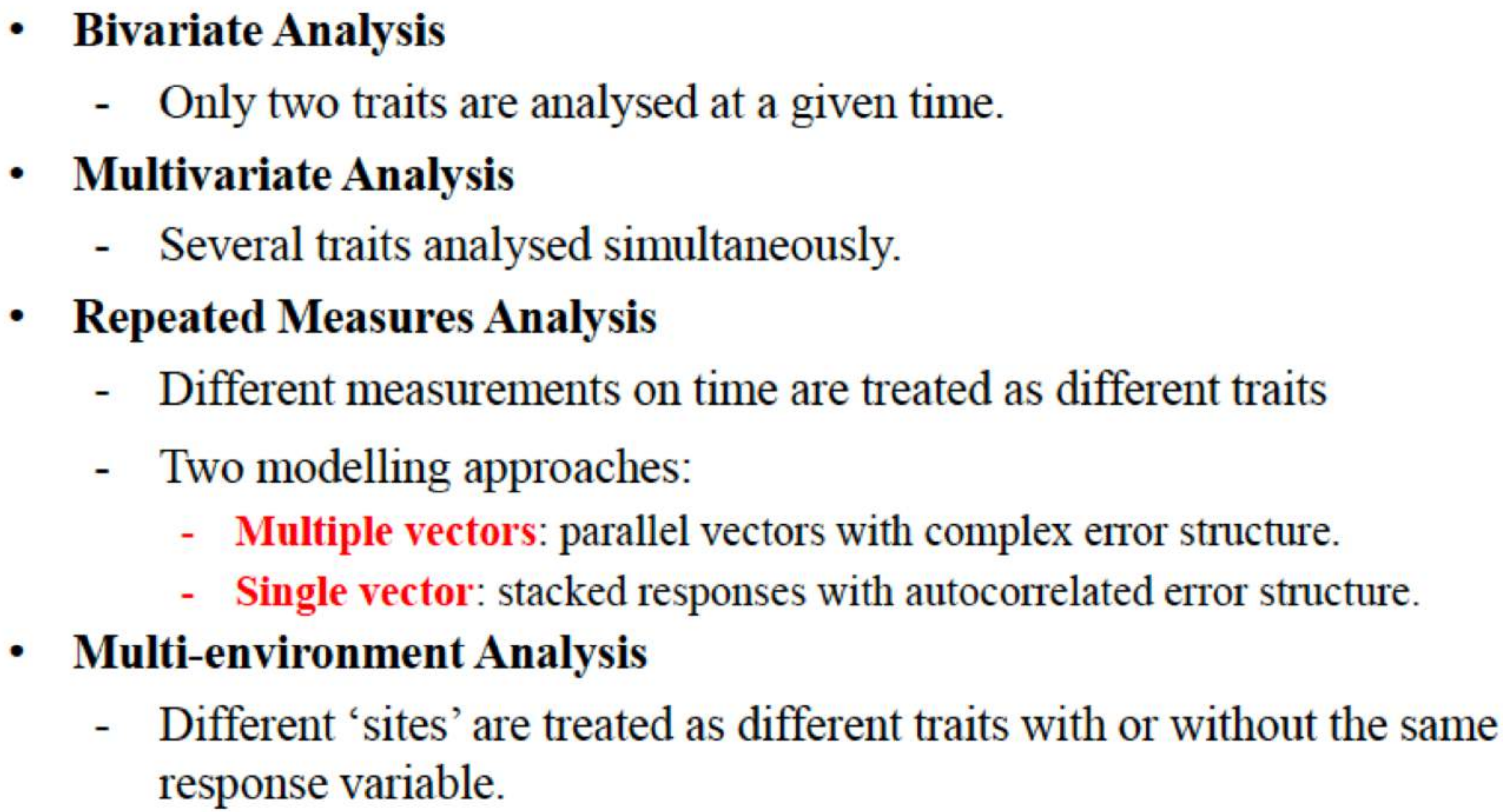
GGEBiplotGUI

Multivariate Analysis



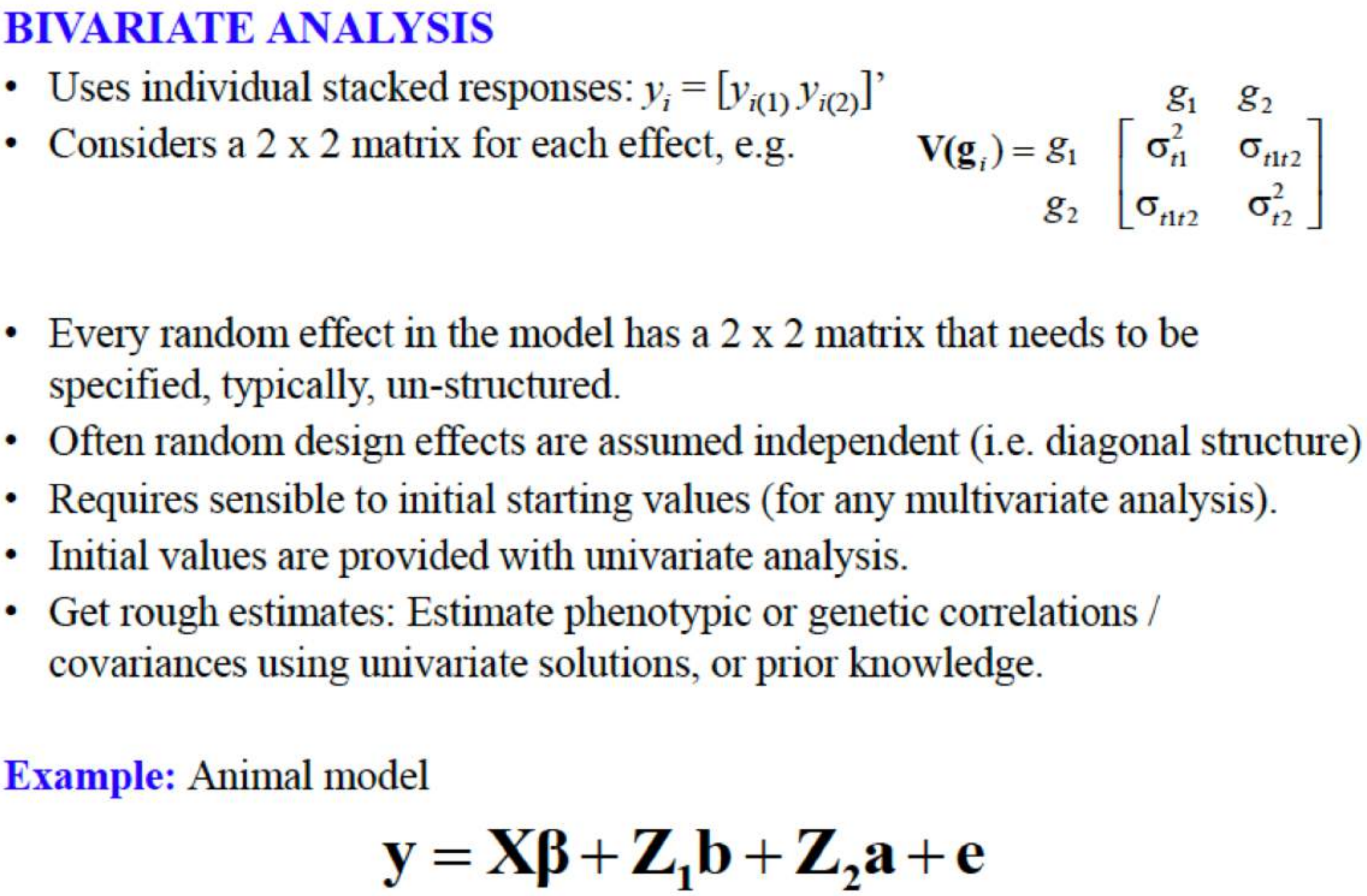
- More *efficient* analysis that combines information on two or more response variables.
- Can be used to combine different sources of, complete or incomplete, sources of data.
- Allows to estimate *correlations* among traits (e.g. phenotypic and genetic correlations).
- Produces an improvement on the precision of the breeding values (BLUPs).
- 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*.
- Generates the required matrices to construct a *selection index*.
- Required analysis for cases where a prior selection was done based in a trait (e.g. culling).

Multivariate Analysis



- **Bivariate Analysis**
  - Only two traits are analysed at a given time.
- **Multivariate Analysis**
  - Several traits analysed simultaneously.
- **Repeated Measures Analysis**
  - Different measurements on time are treated as different traits
  - Two modelling approaches:
    - **Multiple vectors**: parallel vectors with complex error structure.
    - **Single vector**: stacked responses with autocorrelated error structure.
- **Multi-environment Analysis**
  - Different 'sites' are treated as different traits with or without the same response variable.

Multivariate Analysis



BIVARIATE ANALYSIS

- Uses individual stacked responses: $y_i = [y_{i(1)}\, y_{i(2)}]'$
- Considers a 2 x 2 matrix for each effect, e.g.

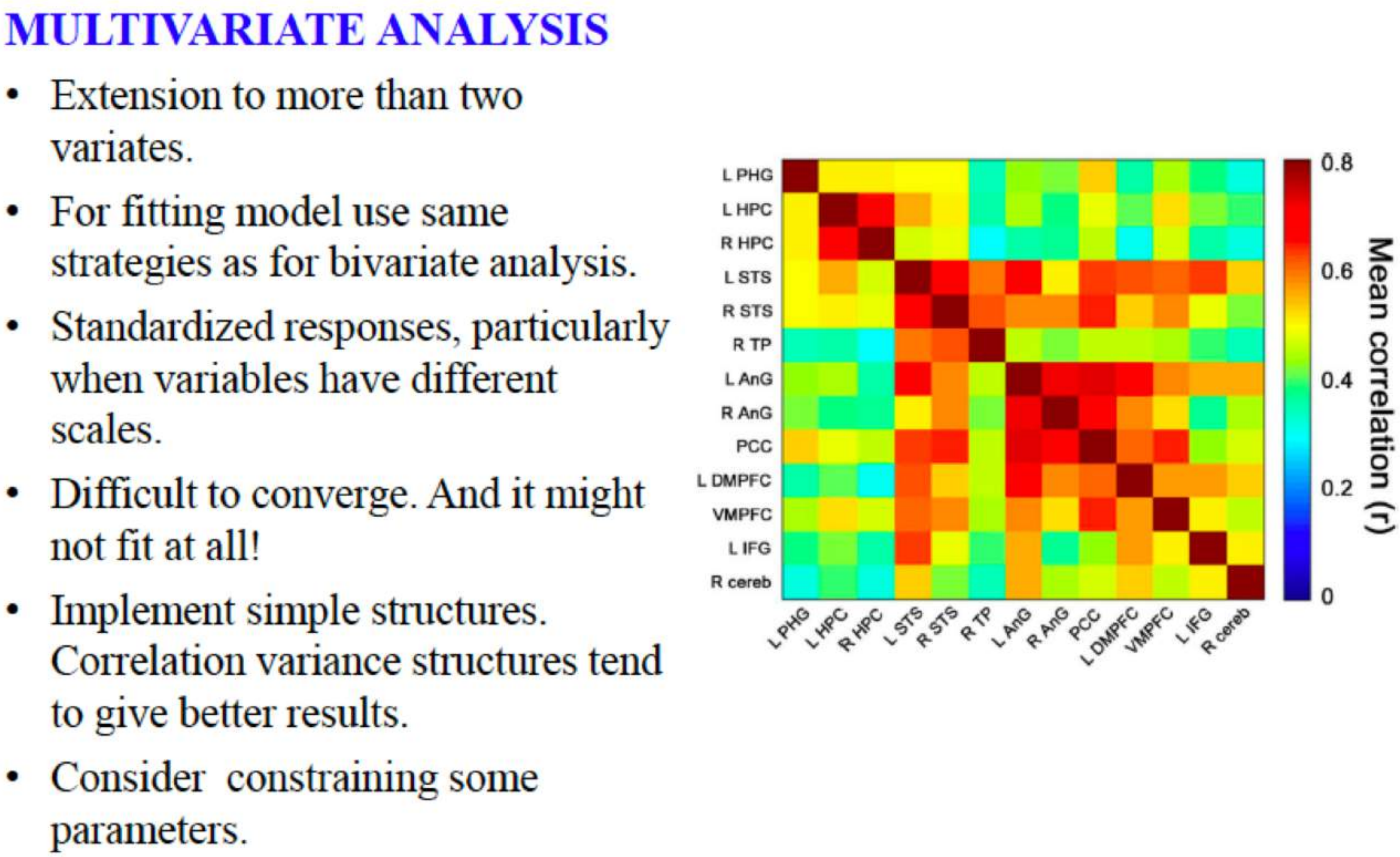
$$\mathbf{V}(\mathbf{g}_i) = \begin{array}{cc} & \begin{array}{cc} g_1 & g_2 \end{array} \\ \begin{array}{c} g_1 \\ g_2 \end{array} & \begin{bmatrix} \sigma^2_{t1} & \sigma_{t1t2} \\ \sigma_{t1t2} & \sigma^2_{t2} \end{bmatrix} \end{array}$$

- Every random effect in the model has a 2 x 2 matrix that needs to be specified, typically, un-structured.
- Often random design effects are assumed independent (i.e. diagonal structure)
- Requires sensible to initial starting values (for any multivariate analysis).
- Initial values are provided with univariate analysis.
- Get rough estimates: Estimate phenotypic or genetic correlations / covariances using univariate solutions, or prior knowledge.

Example: Animal model

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

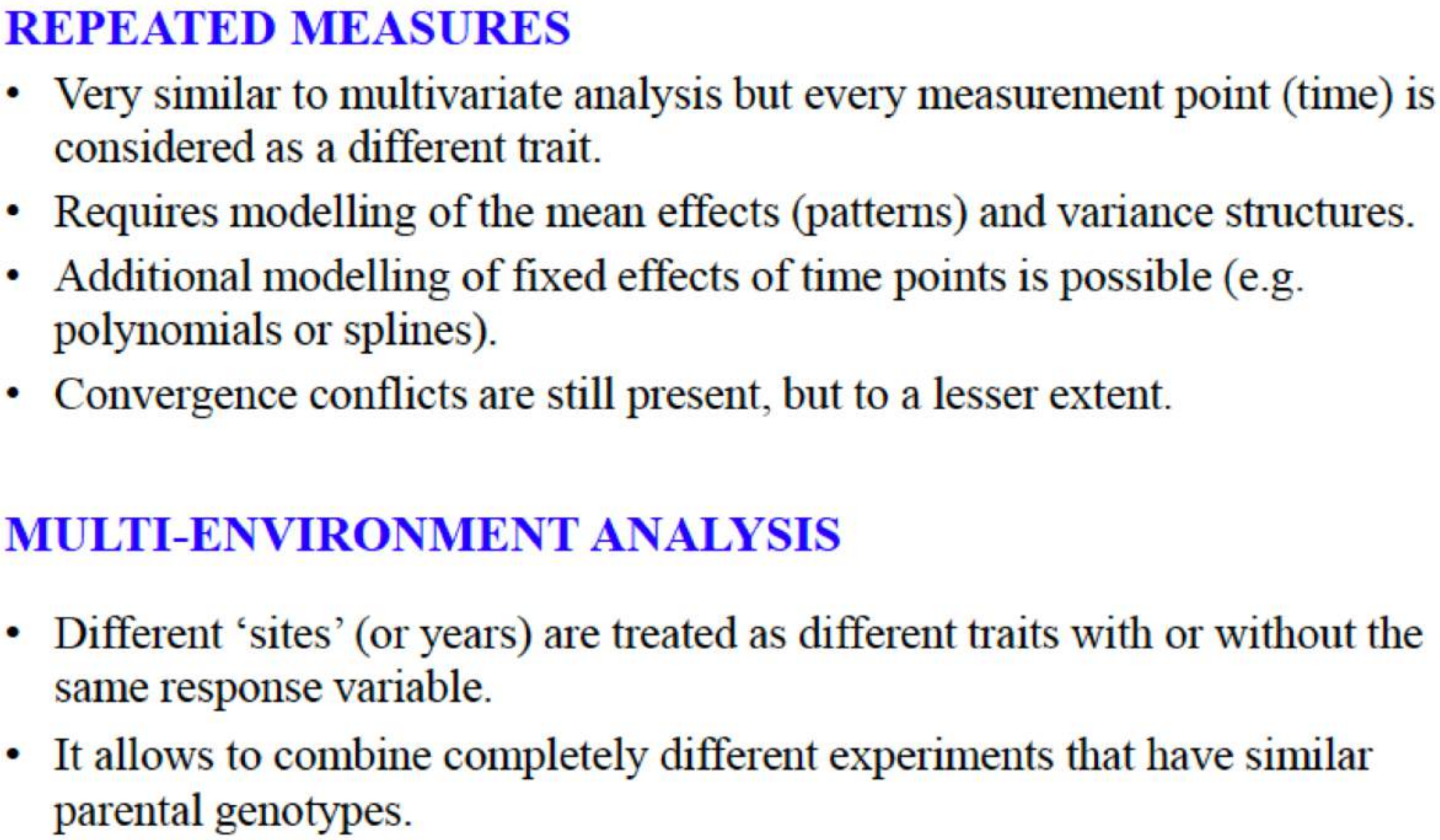
Multivariate Analysis



MULTIVARIATE ANALYSIS

- Extension to more than two variates.
- For fitting model use same strategies as for bivariate analysis.
- Standardized responses, particularly when variables have different scales.
- Difficult to converge. And it might not fit at all!
- Implement simple structures. Correlation variance structures tend to give better results.
- Consider constraining some parameters.

Multivariate Analysis



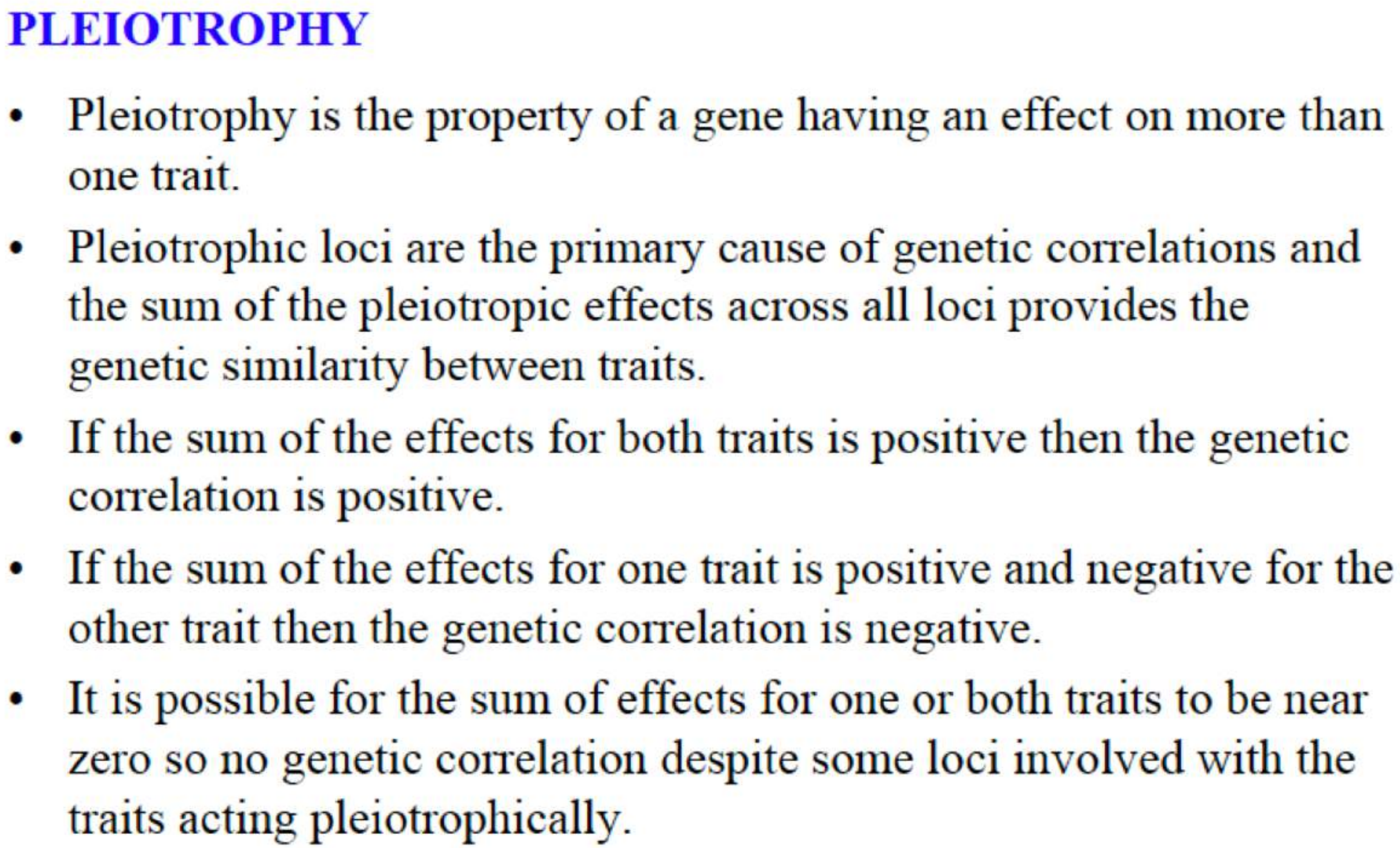
REPEATED MEASURES

- 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.

MULTI-ENVIRONMENT ANALYSIS

- Different ‘sites’ (or years) are treated as different traits with or without the same response variable.
- It allows to combine completely different experiments that have similar parental genotypes.

Multivariate Analysis



PLEIOTROPHY

- Pleiotrophy is the property of a gene having an effect on more than one trait.
- Pleiotrophic loci are the primary cause of genetic correlations and the sum of the pleiotropic effects across all loci provides the genetic similarity between traits.
- If the sum of the effects for both traits is positive then the genetic correlation is positive.
- If the sum of the effects for one trait is positive and negative for the other trait then the genetic correlation is negative.
- It is possible for the sum of effects for one or both traits to be near zero so no genetic correlation despite some loci involved with the traits acting pleiotrophically.

Multivariate Analysis

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

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

Multivariate Analysis

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).
- $\mathbf{V}_{\mathbf{axs}}$ 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).

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