G x E: Genotype-environment interaction

Bruce Walsh lecture notes
Uppsala EQG 2012 course
version 8 Feb 2012

Detailed reading: Chapters 38, 39
**G x E**

- Introduction to G x E
  - Basics of G x E
  - Some suggested rules
  - Treating as a correlated-trait problem
- Estimation of G x E terms
  - Finlay-Wilkinson regressions
- SVD-based methods
  - The singular value decomposition (SVD)
  - AMMI models
- Factorial regressions
- Mixed-Model approaches
  - BLUP
  - Structured covariance models
Genotypes vs. individuals

• Much of the G x E theory is developed for plant breeders who are using pure (= fully inbred) lines, so that every individual has the same genotype.

• The same basic approaches can be used by taking family members as the replicates for outbred species. Here the “genotype” over the family members is some composite value (the mean breeding value of the family).
Yield in Environment 1

Genotype 1

\[ G_{11} \quad E_1 \quad G_{21} \]

Genotype 2

\[ E_i = \text{mean value in environment } i \]

Yield in Environment 2

\[ G_{22} \quad E_2 \quad G_{12} \]

Overall means

\[ E_1 \quad G_1 G_2 \quad E_2 \]
$G_{ij} = \text{mean of genotype i in environment j}$

Under base model of Quantitative Genetics,

$G_{ij} = \mu + G_i + E_j$

When $G \times E$ present, there is an interaction between a particular genotype and a particular environment so that $G_{ij}$ is no longer additive, $G_{ij} = \mu + G_i + E_i + GE_{ij}$
Components measured as deviations from the mean $\mu$

$$GE_{ij} = g_{ij} - g_i - e_j$$
Which genotype is the best?

Depends:
If the genotypes are grown in both environments, \( G_2 \) has a higher mean.

If the genotypes are only grown in environment 1, \( G_2 \) has a higher mean.

If the genotypes are only grown in environment 2, \( G_1 \) has a higher mean.
\[ G \times E: \text{ Both a problem and an opportunity} \]

- A line with little \( G \times E \) has stability across environments.
- However, a line with high \( G \times E \) may outperform all others in specific environments.
- \( G \times E \) implies the opportunity to fine-tune specific lines to specific environments.
- \( G \times E \) implies high \( G \times E \) in at least some lines in the sample.

\[ \sigma^2(GE) \] implies high \( G \times E \) in at least some lines in the sample.
<table>
<thead>
<tr>
<th>Amount of $G \times E$</th>
<th>Mean Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Potential for</td>
</tr>
<tr>
<td></td>
<td>locally-adapted</td>
</tr>
<tr>
<td></td>
<td>lines</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Potential for</td>
</tr>
<tr>
<td></td>
<td>locally-adapted</td>
</tr>
<tr>
<td></td>
<td>lines</td>
</tr>
<tr>
<td></td>
<td>Ideal. Potential for</td>
</tr>
<tr>
<td></td>
<td>widely adaptive lines</td>
</tr>
<tr>
<td></td>
<td>Undesirable</td>
</tr>
</tbody>
</table>

Ideal: high mean performance, low $G \times E$

Low $G \times E = \text{widely adaptive lines/genotypes}$

High $G \times E = \text{locally adaptive lines/genotypes}$
Major vs. minor environments

• An identical genotype will display slightly different traits values even over apparently identical environments due to low micro-environmental variation and developmental noise.

• However, macro-environments (such as different locations or different years <such as a wet vs. a dry year>) can show substantial variation, and genotypes (pure lines) may differentially perform over such macro-environments ($G \times E$).

• Problem: The mean environment of a location may be somewhat predictable (e.g., corn in the tropics vs. temperate North American), but year-to-year variation at the same location is essentially unpredictable.

• Decompose $G \times E$ into components
  - $G \times E_{\text{locations}} + G \times E_{\text{years}} + G \times E_{\text{years} \times \text{locations}}$
  - Ideal: strong $G \times E$ over locations, high stability over years.
Components of $\sigma^2_{G \times E}$: Variance Heterogeneity and Lack of Correlations

It is useful to remind the reader that there are two different sources for $G \times E$ — differences in the genetic variances across environments (genetic heterogeneity, often referred to as scale effects) and lack of perfect correlation among breeding values across environments (LW Chapter 22). For two environments, Robertson (1959) showed that the $G \times E$ interaction variance can be partitioned into these two sources,

$$\sigma^2_{G \times E} = \frac{(\sigma_{A_1} - \sigma_{A_2})^2}{2} + \sigma_{A_1} \sigma_{A_2} (1 - r_A)$$  \hspace{1cm} (38.1a)

where $\sigma^2_{A_i}$ is the additive variance in environment $i$ and $r_A$ is the additive genetic correlation across environments. Cockerham (1963) and Itoh and Yamada (1990) extended Robertson’s decomposition to $n_e$ environments,

$$\sigma^2_{G \times E} = \frac{1}{n_e - 1} \sum_{j}^{n_e} (\bar{\sigma}_{A_j} - \bar{\sigma}_A)^2 + \frac{2}{n_e(n_e - 1)} \sum_{i<j}^{n_e} \sigma_{A_i} \sigma_{A_j} [1 - r_A(i,j)]$$  \hspace{1cm} (38.1b)

Key: differences in scale and lack of perfect correlation over environments both generate $G \times E$
Falconer: $G \times E$

- The modern treatment of $G \times E$ starts with Falconer (1952)
  - Measures of the same trait in different environments are correlated traits
  - Hence, if measured in $k$ environments, it's a $k$-dimensional trait
  - Hence, results from direct and correlated responses apply to selection on $G \times E$

- If selection in environment $i$, expected change in environment $j$ is
  $$CR_j = i_i h_i h_j r_A \sigma_p(j)$$
Hammond’s Conjecture

• Hammond (1947) suggested that selection be undertaken in a more favorable environment to maximize progress in a less favorable one.

• Idea: perhaps more genetic variation, and hence greater discrimination between genotypes.

• Downside: don’t know if \( Var(\mathcal{G}) \) greater in “better” environments, even if it is, between-environment correlation can be small.
Example 38.2. Falconer and Latyszewski (1952) and Falconer (1960) selected for growth rate in mice in two nutritional environments (this work was also discussed in Example 30.7). In one environment, mice were housed individually and food was restricted to around 75% of normal intake, while in the other, mice were housed in groups of four to six and given unlimited food. Selection for increased weight gain was effective in both environments, although heritability was higher (0.29 to 0.20) in the restricted diet environment (although this difference was not significant). The higher heritability value arose because while the additive genetic variance was reduced in the poorer environment (by around 45%), the environmental variance was reduced even more (around 66%). Falconer suggested that this reduction in $\sigma^2_e$ may be, in part, due to rearing single versus multiple individuals.

When the restricted-diet selected individuals were grown in the unrestricted environment, they showed a significant weight gain, but when the unrestricted-selected individuals were reared in the restricted diet environment, they did not. These results are a direct contradiction to Hammond’s conjecture, in that selection in the poorer environment gave the larger response in the target population. Further, there were other significant differences. The high-feed selected lines contained around 24% more body fat than the restricted-diet lines when both were grown in the high-feed environment. Thus, selection in the restricted diet also resulted in leaner mice, which (in many cases) would also be economically favored in a selection program.
Jinks-Connolly rule

- Stability of the genotypic value over environments is a measure of G x E sensitivity.
  - High stability = low sensitivity
- Antagonistic G x E selection
  - Up-selecting in the bad environment
- Synergistic G x E selection
  - Up-selecting in the good environment
- Jinks-Connolly rule:
  - Antagonistic selection improves stability (decreases environmental sensitivity), while synergistic selection decreases stability
Antagonistic

Synergistic
While Jinks-Connolly suggests a general trend and is expected to hold more often than not, Falconer (1990) noted that a modification of this rule held in all 24 experimental cases he examined, namely that the sensitivity is less after antagonistic selection than after synergistic selection. Since the sensitivity is a slope, this means that the change in the numerator of Equation 38.5 is greater under antagonistic selection than under synergistic selection. When selecting to decrease a trait, this requires

\[(R_H - CR_L) - (CR_H - R_L) > 0\]  \hspace{1cm} (38.6a)

which rearranges to recover

\[R_H + R_L > CR_H + CR_L\]  \hspace{1cm} (38.6b)

with this same condition holding for selection to increase a trait. Hence, for Falconer’s modification to hold, the less restrictive assumption that the sum of the direct responses is greater than the sum of correlated responses must hold.
Falconer’s suggestion

• Suggested mean performance over the two environments is best improved by antagonistic selection
What about Falconer's (1989) suggestion that mean performance over the two environments is best improved by antagonistic selection? If the mean change is equally weighted in both environments, then when selecting to increase a trait, under antagonistic selection direct response occurs in the low environment, while under synergistic selection direct response occurs in the high environment. Thus, Falconer’s (1989) suggestion holds when the average of the direct response in low and the correlated response in high exceeds the direct response in high and the correlated response in low,

$$R_L + CR_H > R_H + CR_L$$  \hspace{1cm} (38.7a)

Assuming equal selection in both environments, then from Equation 38.3a, this reduces to

$$h_L (\sigma_{A_L} + r_A \sigma_{A_H}) > h_H (\sigma_{A_H} + r_A \sigma_{A_L})$$  \hspace{1cm} (38.7b)

Conversely, when selecting to decrease trait value, this condition becomes

$$R_H + CR_L > R_L + CR_H$$  \hspace{1cm} (38.7c)

Note that Equations 38.7a and 38.7c are mutually exclusive, so that if antagonistic selection is better in one direction, it will be worse in the opposite direction. Thus, as Falconer (1990) pointed out, there is little theoretical justification for his earlier (1989) suggestion.
The Cost to Response from $G \times E$

As a benchmark for selection when $G \times E$ is present, if environmental structure is ignored and simple mass selection used (choosing the best performing individuals based solely on their phenotypic values), then the expected response becomes

$$R = \bar{t} \sigma_z h^2 = \bar{t} \frac{\sigma^2_A}{\sigma_z} = \bar{t} \frac{\sigma^2_A}{\sqrt{\sigma^2_G + \sigma^2_{G \times E} + \sigma^2_E}} \quad (38.9a)$$

where $\sigma^2_G$ and $\sigma^2_E$ are the genetic and environmental variances. When $\sigma^2_{G \times E}$ is large relative to $\sigma^2_A$, the heritability is low and selection very inefficient, as an individual’s phenotypic value in one environment is a poor predictor of their average breeding value over all environments. If we are selecting among clones (or pure lines) then $\sigma^2_G$ replaces $\sigma^2_A$. Setting $\sigma^2_{G \times E}$ to zero, Matheson and Cotterill (1990) note that the “cost” (loss of potential gain) of genotype-environment interaction when using standard mass selection is

$$1 - \sqrt{\frac{\sigma^2_G + \sigma^2_E}{\sigma^2_G + \sigma^2_{G \times E} + \sigma^2_E}} \quad (38.9b)$$
Replication over environments can reduce effect of $G \times E$ in selection response

If members of the same genotype/line are replicated over $n_e$ random environments, response to selection based on line (or family) means is

$$R = \bar{t} \frac{\sigma^2_G}{\sigma^2_z} = \bar{t} \frac{\sigma^2_G}{\sqrt{\sigma^2_G + (\sigma^2_E + \sigma^2_{G \times E})/n_e + \sigma^2_e/(n_T n_e)}}$$
Estimating the GE term

• While GE can be estimated directly from the mean in a cell (i.e., $G_i$ in $E_j$) we can usually get more information (and a better estimate) by considering the entire design and exploiting structure in the GE terms.

• This approach also allows us to potentially predict the GE terms in specific environments.

• Basic idea: replace $GE_{ij}$ by $\alpha_i \gamma_j$ or more generally by $\Sigma_k \alpha_{ki} \gamma_{kj}$. These are called biadditive or bilinear models. This (at first sight) seems more complicated. Why do this?

• With $n_G$ genotypes and $n_E$ environments, we have
  - $n_G n_E$ GE terms (assuming no missing values)
  - $n_G + n_E$ $\alpha_i$ and $\gamma_i$ unique terms
  - $k(n_G + n_E)$ unique terms in $\Sigma_k \alpha_{ki} \gamma_{kj}$.

• Suppose 50 genotypes in 10 environments
  - 500 $GE_{ij}$ terms, 60 unique $\alpha_i$ and $\gamma_i$ terms, and (for $k=3$), 180 unique $\alpha_{ki}$ and $\gamma_{ki}$ terms.
Finlay-Wilkinson Regression

Also called a joint regression or regression on an environmental index.

Let \( \mu + G_i \) be the mean of the \( i \)th genotype over all the environments, and \( \mu + E_j \) be the average yield of all genotypes in environment \( j \)

\[
\mu_{ij} = \mu + G_i + E_j (1 + \beta_i) + \delta_{ij}
\]

The FW regression estimates \( GE_{ij} \) by the regression \( GE_{ij} = \beta_i E_j + \delta_{ij} \).

The regression coefficient is obtained for each genotype from the slope of the regression of the \( G_{ij} \) over the \( E_j \). \( \delta_{ij} \) is the residual (lack of fit). If \( \sigma^2(GE) \gg \sigma^2(\delta) \), then the regression accounted for most of the variation in \( GE \).
Application

• Yield in lines of wheat over different environments was examined by Calderini and Slafer (1999). The lines they examined were lines from different eras of breeding (for four different countries).

• Newer lines had larger values, but also had higher slopes (large $\beta_i$ values), indicating less stability over mean environmental conditions than seen in older lines.
Regression slope for each genotype is $\beta_i$. 
SVD approaches

- In Finlay-Wilkinson, the $GE_{ij}$ term was estimated by $\beta_i E_j$, where $E_j$ was observed. We could also have used $\gamma_j G_i$, where $\gamma_j$ is the regression of genotype values over the j-th environment. Again $G_i$ is observable.

- **Singular-value decomposition (SVD) approaches** consider a more general approach, approximating $GE_{ij}$ by $\sum_k \alpha_{ki} \gamma_{kj}$ where the $\alpha_{ki}$ and $\gamma_{kj}$ are determined by the first $k$ terms in the SVD of the matrix of GE terms.

- The SVD is a way to obtain the best approximation of a full matrix by some matrix of lower dimension.
The Singular-Value Decomposition (SVD)

An \( n \times p \) matrix \( \mathbf{A} \) can always be decomposed as the product of three matrices: an \( n \times p \) diagonal matrix \( \Lambda \) and two unitary matrices, \( \mathbf{U} \) which is \( n \times n \) and \( \mathbf{V} \) which is \( p \times p \). The resulting **singular value decomposition** (SVD) of \( \mathbf{A} \) is given by

\[
\mathbf{A}_{n \times p} = \mathbf{U}_{n \times n} \Lambda_{n \times p} \mathbf{V}^T_{p \times p}
\]  

(39.16a)

We have indicated the dimensionality of each matrix to allow the reader to verify that each matrix multiplication conforms. The diagonal elements \( \lambda_1, \ldots, \lambda_s \) of \( \Lambda \) correspond to the **singular values** of \( \mathbf{A} \) and are ordered by decreasing magnitude. Returning to the unitary matrices \( \mathbf{U} \) and \( \mathbf{V} \), we can write each as a row vector of column vectors,

\[
\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots \mathbf{u}_n), \quad \mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_i, \ldots \mathbf{v}_p)
\]  

(39.16b)

where \( \mathbf{u}_i \) and \( \mathbf{v}_i \) are \( n \) and \( p \)-dimensional column vectors (often called the **left** and **right singular vectors**, respectively). Since both \( \mathbf{U} \) and \( \mathbf{V} \) are unitary, by definition (Appendix 4) each column vector has length one and are mutually orthogonal (i.e., if \( i \neq j \), \( \mathbf{u}_i \mathbf{u}_j^T = \mathbf{v}_i \mathbf{v}_j^T = 0 \)). Since \( \Lambda \) is diagonal, it immediately follows from matrix multiplication that we can write any element in \( \mathbf{A} \) as

\[
A_{ij} = \sum_{k=1}^{s} \lambda_k u_{ik} v_{kj}
\]  

(39.16c)

where \( \lambda_k \) is the \( k \)-th singular value and \( s \leq \min(p, n) \) is the number of non-zero singular values.
The importance of the singular value decomposition in the analysis of G × E arises from the **Eckart-Young theorem** (1938), which relates the best approximation of a matrix by some lower-rank (say \( k \)) matrix with the SVD. Define as our measure of goodness of fit between a matrix \( \mathbf{A} \) and a lower rank approximation \( \hat{\mathbf{A}} \) as the sum of squared differences over all elements,

\[
\sum_{ij} (A_{ij} - \hat{A}_{ij})^2
\]

Eckart and Young show that the best fitting approximation \( \hat{\mathbf{A}} \) of rank \( m < s \) is given from the first \( m \) terms of the singular value decomposition (the **rank-m SVD**),

\[
\hat{A}_{ij} = \sum_{k=1}^{m} \lambda_k u_{ik} v_{kj}
\]

(39.17a)

For example, the best rank-2 approximation for the G × E interaction is given by

\[
GE_{ij} \simeq \lambda_1 u_{i1} v_{j1} + \lambda_2 u_{i2} v_{j2}
\]

(39.17b)

where \( \lambda_i \) is the \( i \)th singular value of the \( \mathbf{GE} \) matrix, \( \mathbf{u} \) and \( \mathbf{v} \) are the associated singular vectors (see Example 39.3). The fraction of total variation of a matrix accounted for by taking the first \( m \) terms in its SVD is

\[
\sum_{k=1}^{m} \frac{\lambda_k^2}{\sum_{ij} A_{ij}^2} = \frac{\lambda_1^2 + \cdots + \lambda_m^2}{\lambda_1^2 + \cdots + \lambda_s^2}
\]
A data set for soybeans grown in New York (Gauch 1992) gives the GE matrix as

\[
GE = \begin{pmatrix}
57 & 176 & -233 \\
-36 & -196 & 233 \\
-45 & -324 & 369 \\
-66 & 178 & -112 \\
89 & 165 & -254 \\
\end{pmatrix}
\]

Where \( GE_{ij} = \text{value for Genotype } i \text{ in envir. } j \)

In R, the compact SVD (Equation 39.16d) of a matrix \( X \) is given by \texttt{svd(x)}, returning the SVD of \( GE \) as

\[
\begin{pmatrix}
0.40 & 0.21 & 0.18 \\
-0.41 & 0.00 & 0.91 \\
-0.66 & 0.12 & -0.30 \\
0.26 & -0.83 & 0.11 \\
0.41 & 0.50 & 0.19 \\
\end{pmatrix}
\begin{pmatrix}
746.10 & 0 & 0 \\
0 & 131.36 & 0 \\
0 & 0 & 0.53 \\
\end{pmatrix}
\begin{pmatrix}
0.12 & 0.64 & -0.76 \\
0.81 & -0.51 & -0.30 \\
0.58 & 0.58 & 0.58 \\
\end{pmatrix}
\]

The first singular value accounts for \( 746.10^2 / (743.26^2 + 131.36^2 + 0.53^2) = 97.0\% \) of the total variation of \( GE \), while the second singular value accounts for 3.0%, so that together they account for essentially all of the total variation. The rank-1 SVD approximation of \( GE \) is given by setting all of the diagonal elements of \( A \) except the first entry to zero,

\[
GE_1 = \begin{pmatrix}
0.40 & 0.21 & 0.18 \\
-0.41 & 0.00 & 0.91 \\
-0.66 & 0.12 & -0.30 \\
0.26 & -0.83 & 0.11 \\
0.41 & 0.50 & 0.19 \\
\end{pmatrix}
\begin{pmatrix}
746.10 & 0 & 0 \\
0 & 131.36 & 0 \\
0 & 0 & 0.53 \\
\end{pmatrix}
\begin{pmatrix}
0.12 & 0.64 & -0.76 \\
0.81 & -0.51 & -0.30 \\
0.58 & 0.58 & 0.58 \\
\end{pmatrix}
\]
Similarly, the rank-2 SVD is given by setting all but the first two singular values to zero,

\[
GE_2 = \begin{pmatrix}
0.40 & 0.21 & 0.18 \\
-0.41 & 0.00 & 0.91 \\
-0.66 & 0.12 & -0.30 \\
0.26 & -0.83 & 0.11 \\
0.41 & 0.50 & 0.19
\end{pmatrix}
\begin{pmatrix}
746.10 & 0 & 0 \\
0 & 131.36 & 0 \\
0 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
0.12 & 0.64 & -0.76 \\
0.81 & -0.51 & -0.30 \\
0.58 & 0.58 & 0.58
\end{pmatrix}
\]

For example, the rank-1 SVD approximation for GE_{32} is

\[g_{31}\lambda_1 e_{12} = 746.10*(-0.66)*0.64 = -315\]

While the rank-2 SVD approximation is\[g_{31}\lambda_2 e_{12} + g_{32}\lambda_2 e_{22} = 746.10*(-0.66)*0.64 + 131.36* 0.12*(-0.51) = -323\]

Actual value is -324

Generally, the rank-2 SVD approximation for GE_{ij} is\[g_{i1}\lambda_1 e_{1j} + g_{i2}\lambda_2 e_{2j}\]
**AMMI models**

*Additive main effects, multiplicative interaction (AMMI) models use the first m terms in the SVD of GE:*

\[ GE_{ij} = \sum_{k=1}^{m} \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij} \]

*Giving*

\[ \mu_{ij} = \mu + G_i + E_j + \sum_{k=1}^{m} \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij} \]

*AMMI is actually a *family* of models, with AMMI_m denoting AMMI with the first m SVD terms*
AMMI models

\[ \mu_{ij} = \mu + G_i + E_j + \sum_{k=1}^{m} \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij} \]

Fit main effects

Fit principal components to the interaction term (SVD is a generalization of PC methods)
Why do AMMI?

- One can plot the SVD terms ($\gamma_{ki}$, $\eta_{kj}$) to visualize interactions
  - Called biplots (see Chapter 39 for details)
- AMMI can better predict mean values of $GE_{ij}$ than just using the cell value (the observed mean of Genotype i in Environment j)
- A huge amount more on AMMI in Chapter 33!
Factorial Regressions

- While AMMI models attempt to extract information about how G x E interactions are related across sets of genotypes and environments, **factorial regressions** incorporate **direct measures of environmental factors** in an attempt to account for the observed pattern of G x E.

- The power of this approach is that if we can determine which genotypes are more (or less) sensitive to which environmental features, the breeder may be able to more finely tailor a line to a particular environment without necessarily requiring trials in the target environment.
Suppose we have a series of $m$ measured values from the environments of interest (such as average rainfall, maximum temperature, etc.) Let $x_{kj}$ denote the value of the $k$-th environmental variable in environment $j$.

Factorial regressions then model the GE term as the sensitivity $\zeta_{ki}$ of environmental value $k$ to genotype $i$, (this is a regression slope to be estimated from the data):

$$GE_{ij} = \sum_{k=1}^{m} \zeta_{ki} x_{kj} + \delta_{ij}$$

Note that the Finlay-Wilkinson regression is a special case where $m=1$ and $x_j$ is the mean trait value (over all genotypes) in that environment.
<table>
<thead>
<tr>
<th>Model</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finlay-Wilkinson</td>
<td>$\beta_i = \text{sensitivity of genotype } i \text{ to the average effect } E_j \text{ of the environment.}$</td>
</tr>
<tr>
<td>$GE_{ij} = \beta_i (E_j - \mu) + \delta_{ij}$</td>
<td>First $m$ terms of the SVD of the $GE$ matrix</td>
</tr>
<tr>
<td></td>
<td>$\lambda_k^2$ is the amount of variation explained by axis $k$</td>
</tr>
<tr>
<td></td>
<td>$\gamma_{ki} = \text{sensitivity of genotype } i \text{ to environmental axis } k$</td>
</tr>
<tr>
<td></td>
<td>$\eta_{kj} = \text{value of environment } j \text{ on the } k\text{th environmental axis}$</td>
</tr>
<tr>
<td>AMMI</td>
<td>Model $G \times E$ using $m$ measured environmental factors</td>
</tr>
<tr>
<td>$GE_{ij} = \sum_{k=1}^{m} \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij}$</td>
<td>$x_{kj} = \text{value of } k\text{th environmental factor in environment } j$</td>
</tr>
<tr>
<td></td>
<td>$\zeta_{ki} = \text{sensitivity of genotype } i \text{ to } k\text{th environmental factor}$</td>
</tr>
<tr>
<td>Factorial Regression</td>
<td>Model $G \times E$ based on a reduced dimensional set of the observed</td>
</tr>
<tr>
<td></td>
<td>environmental factors by constructing $m$ combinations (axes) of these effects.</td>
</tr>
<tr>
<td>Reduced rank Factorial Regression</td>
<td>$c_{kp} = \text{loading of } p\text{th environmental factor on axis } k.$</td>
</tr>
<tr>
<td></td>
<td>$\zeta_{ki} = \text{sensitivity of genotype } i \text{ to } k\text{th environmental combination (axis)}$</td>
</tr>
<tr>
<td>AMMI using Reduced rank Factorial Regression</td>
<td>The environmental axes $\eta_{kj}$ under AMMI are replaced by the environmental axes generated by linear combinations of measured environmental factors generated by a reduced rank factorial regression, with $\eta_{kj} = \sum_p c_{kp} x_{pj}$.</td>
</tr>
</tbody>
</table>
Mixed model analysis of $G \times E$

- Thus far, our discussion of estimating $GE$ has be set in terms of fixed effects.
- Mixed models are a powerful alternative, as they easily handle missing data (i.e., not all combinations of $G$ and $E$ explored).
- As with all mixed models, key is the assumed covariance structure
  - Structured covariance models
    - Compound symmetry
    - Finlay–Wilkinson
    - Factor-analytic models (closely related to AMMI)
- Much more complete treatment and discussion in Chapter 39
Basic GxE Mixed model

• Typically, we assume either G or E is fixed, and the other random (making GE random)
• Taking E as fixed, basic model becomes
• $z = X\beta + Z_1g + Z_2ge + e$
  - The vector $\beta$ of fixed effects includes estimates of the $E_j$.
    The vector $g$ contains estimates of the $G_i$ values, while the vector $ge$ contains estimates of all the $GE_{ij}$.
  - Typically we assume $e \sim 0, \sigma_e^2I$, and independent of $g$ and $ge$.
  - Models significantly differ on the variance/covariance structure of $g$ and $ge$. 
Example

We have two genotypes and three environments. Let $z_{ijk}$ denote the k-th replicate of genotype i in environment j. Suppose we have single replicates of genotype 1 in all three environments, two replicates of genotype 2 in environment 1, and one in environment 3.

\[
\begin{pmatrix}
  z_{111} \\
  z_{121} \\
  z_{131} \\
  z_{211} \\
  z_{212} \\
  z_{231}
\end{pmatrix}, \quad \begin{pmatrix}
  E_1^* \\
  E_2^* \\
  E_3^*
\end{pmatrix}, \quad \begin{pmatrix}
  G_1 \\
  G_2
\end{pmatrix}, \quad \begin{pmatrix}
  GE_{11} \\
  GE_{12} \\
  GE_{13} \\
  GE_{21} \\
  GE_{22} \\
  GE_{23}
\end{pmatrix}, \quad \begin{pmatrix}
  \epsilon_{111} \\
  \epsilon_{121} \\
  \epsilon_{131} \\
  \epsilon_{211} \\
  \epsilon_{212} \\
  \epsilon_{231}
\end{pmatrix}
\]

Here $E_i^* = \mu + E_i$, with the $E_i$ constrained to sum to zero. The resulting design matrices are

\[
X = \begin{pmatrix}
  1 & 0 & 0 \\
  0 & 1 & 0 \\
  0 & 0 & 1 \\
  1 & 0 & 0 \\
  1 & 0 & 0 \\
  0 & 0 & 1
\end{pmatrix}, \quad Z_1 = \begin{pmatrix}
  1 & 0 \\
  1 & 0 \\
  1 & 0 \\
  1 & 0 \\
  1 & 0 \\
  0 & 1
\end{pmatrix}, \quad Z_2 = \begin{pmatrix}
  1 & 0 & 0 & 0 & 0 & 0 \\
  0 & 1 & 0 & 0 & 0 & 0 \\
  0 & 0 & 1 & 0 & 0 & 0 \\
  0 & 0 & 0 & 1 & 0 & 0 \\
  0 & 0 & 0 & 1 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}
\]
Compound Symmetry assumption

- To proceed further on the analysis of the mixed model, we need covariance assumptions on $g$ and $ge$.
- The **compound symmetry** assumption is
  - $\sigma^2(G_i) = \sigma_g^2$, $\sigma^2(GE_{ij}) = \sigma_{GE}^2$
  - Under this assumption, the covariance of any genotype across any two (different) environments is the same.
  - Likewise, the genetic variance within any environment is constant across environments
  - Net result, the genetic covariance is the same between any two environments
Expected genetic variance within a given environment

\[ \sigma(z_{ijk}, z_{ijl}) = \sigma(G_i + GE_{ij}, G_i + GE_{ij}) \]
\[ = \sigma(G_i, G_i) + \sigma(GE_{ij}, GE_{ij}) \]
\[ = \sigma_G^2 + \sigma_{GE}^2 \]

Genetic covariance of the same genotype across environments

\[ \sigma(z_{ij}, z_{ik}) = \sigma(G_i + GE_{ij}, G_i + GE_{ik}) = \sigma(G_i, G_i) \]

Genetic correlation across environments is constant

\[ \rho_G = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2} \]

For our example, the resulting covariance matrix becomes

\[ V_Z = Z_1 V_g Z_1^T + Z_2 V_{ge} Z_2^T + V_e \]
\[ V_g = \sigma_G^2 I_{2 \times 2}, \quad V_{ge} = \sigma_{G\times E}^2 I_{6 \times 6}, \quad V_e = \sigma_e^2 I_{6 \times 6} \]
Mixed-model allows for missing values.

Under fixed-model model, estimate of $\mu_{ij} = \bar{z}_{ij}$.

**BLUP estimates under mixed-model (E fixed, G random)**

Assuming equal number of replicates for each $ij$ combination, the predicted yield $\mu_{ij}$ given an observed mean of $z_{ij}$ is given by

$$\text{BLUP}(\mu_{ij}) = \bar{z}_{.j} + h_G^2 (\bar{z}_{i.} - \bar{z}_{..}) + h_{GE}^2 (z_{ij} - \bar{z}_{.j} - \bar{z}_{i.} + \bar{z}_{..})$$

$$= \hat{\mu} + \hat{E}_j + h_G^2 \hat{G}_i + h_{GE}^2 \hat{GE}_{ij}$$

where for $n_e$ environments, the repeatability of genetic effects and interactions, are respectively,

$$h_G^2 = \frac{\sigma_{GE}^2 + n_e \sigma_G^2}{\sigma_{GE}^2 + n_e \sigma_G^2 + \sigma_e^2}, \quad h_{GE}^2 = \frac{\sigma_{GE}^2}{\sigma_{GE}^2 + \sigma_e^2}$$

Contrasting the estimated cell mean under LS (given by the observed sample mean $z_{ij}$) with the predicted cell means under BLUP shows how BLUP shrinks the contributions from the two random effects ($G_i, GE_{ij}$).
BLUP shrinks (regresses) the BLUE estimate back towards zero

\[ h_G^2 = \frac{\sigma_{GE}^2 + n_e \sigma_G^2}{\sigma_{GE}^2 + n_e \sigma_G^2 + \sigma_e^2}, \quad h_{GE}^2 = \frac{\sigma_{GE}^2}{\sigma_{GE}^2 + \sigma_e^2} \]

In particular, the BLUP contribution of the genotypic effect is \( h_G^2 \hat{G}_i \), which is a shrinkage of the BLUE estimate \( \hat{G}_i \) back to its mean (zero). The same is true for the \( G \times E \) effect. The amount of shrinkage is proportional to the lack of repeatability of these two contributions. If \( h^2 \) is near one, there is very little shrinkage, while if \( h^2 \) is near zero, its contribution is shrunk back towards nearly zero. An informal (but helpful) way of thinking about shrinkage is that the coefficient of shrinkage is the ratio of signal over signal plus noise, and is a measure of the “borrowing strength” from correlated observations. If there is little such information, there is much more noise than signal, and the resulting shrinkage is considerable, while if there is a strong signal, there is little shrinkage.
Modification of the residual covariance

Under compound symmetry, all the covariance matrices are a variance component times an identity matrix. More realistic models replace these simple matrices with more complex ones. We could allow residual variances to vary over lines \( [\sigma^2(e_{ijk}) = \sigma^2_{e_i}] \) or environments \( [\sigma^2(e_{ijk}) = \sigma^2_{e_j}] \), in which case \( \mathbf{V}_e \) becomes a diagonal matrix with the diagonal the appropriate residual variance component (e.g., Cullis et al. 1996). For our hypothetical design, if residual variances are genotype-dependent,

\[
\mathbf{V}_e = \text{diagonal}(\sigma^2_{e_1}, \sigma^2_{e_1}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma^2_{e_2})
\]

while if they are environment-dependent

\[
\mathbf{V}_e = \text{diagonal}(\sigma^2_{e_1}, \sigma^2_{e_2}, \sigma^2_{e_3}, \sigma^2_{e_1}, \sigma^2_{e_2})
\]

Again, these can be estimated via REML. Another modification is when pedigree information exists on the genotypes, in which case \( \mathbf{V}_g \) may have off-diagonal elements reflecting relationships among genotypes (Crossa et al. 2006, Oakey et al. 2007, Piepho et al. 2008). Finally, one can allow for differential correlations among genotype-environment interactions by suitably modifying \( \mathbf{V}_{ge} \), a point we develop in detail shortly.
Extending genetic covariances

- **Shukla’s model**: starts with the compound symmetry model, but allows for different $G \times E$ variances over genotypes,
  - $GE_{ij} \sim N(0, \sigma^2_{GiE})$
  - $G_i \sim N(0, \sigma_G^2)$
  - $\text{Cov}(ge) = \text{Diagonal} (\sigma^2_{G1E}, \ldots, \sigma^2_{GnE})$
  - The covariance of a genotype across environments is still $\sigma_G^2$

- **Structured covariance models** allow more complicated (and more general) covariance matrices
Covariances based on Finlay-Wilkinson

\[ z_{ijk} = \mu + G_i + (1 + \beta_i)E_j + \delta_{ij} + \epsilon_{ijk} \]

Previously we analyzed this model assuming a fixed-effects framework. Digby (1979) showed that one could use an iterative least-squares approach to accommodate missing data (certain genotype-environment combinations are missing). This is reasonable, as one can borrow information from other observations in an attempt to predict the missing observation. Suppose that line five was not measured in environment three. Data from the other genotypes can be used to estimate \( E_3 \) while observations on genotype five from other environments can be used to estimate \( \beta_5 \), with \( GE_{53} \) being estimated by \( (1 + \beta_5)E_3 \). This shows how information can be borrowed from other observations under this model by using correlations between observations. In a mixed-model framework, such information borrowing occurs through the covariance matrix associated with the vector of random effects.

We treat \( G_i \) and \( \beta_i \) as fixed effects, \( \delta_{ij} \) and \( \epsilon_{ijk} \) as random (fixed genetic effects, random environmental effects)
Assume that the environmental effect, regression deviation, and residual error are all independent random effects and have constant variances,

\[ E_j \sim N(0, \sigma_E^2), \quad \delta_{ij} \sim N(0, \sigma_\delta^2), \quad \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \]

Hence,

\[ \sigma(E_j, E_\ell) = \begin{cases} 0 & j \neq \ell \\ \frac{\sigma_E^2}{\sigma_\delta^2} & j = \ell \end{cases}, \quad \sigma(\delta_{ij}, \delta_{k\ell}) = \begin{cases} 0 & ij \neq k\ell \\ \frac{\sigma_\delta^2}{\sigma_\delta^2} & ij = k\ell \end{cases} \]

The variance of the trait value from an individual from line \( i \) randomly drawn across environments is

\[ \sigma^2(z_{ijk}) = (1 + \beta_i)^2 \sigma_E^2 + \sigma_\delta^2 + \sigma_\varepsilon^2 \]

Peipho (1997a) notes that the regression residual and normal residual variances (if both homoscedastic) cannot be separately estimated and hence can be combined into a single general residual variance. The covariance between two different genotypes \( (i \) and \( k \) \) in the same environment \( (j) \) similarly becomes

\[ \sigma^2(z_{ij}, z_{kj}) = \sigma [(1 + \beta_i)E_j + \delta_{ij}, (1 + \beta_k)E_j + \delta_{kj}] \\
= (1 + \beta_i)(1 + \beta_k)\sigma_E^2 \]
Let $z_j$ be a vector of observations of the line means within environment $j$ (for simplification we assume a single observation, but multiple, and unequal, replication is easily accommodated by modification of $V_e$). In matrix form, the covariance matrix for $z_j$ is

$$V_{z_j} = \sigma^2_E \lambda \lambda^T + (\sigma^2_\delta + \sigma^2_e) I,$$

where

$$\lambda = \begin{pmatrix}
1 + \beta_1 \\
\vdots \\
1 + \beta_{ng}
\end{pmatrix}$$

Observe that the assumed structure of the Finlay-Wilkinson model translates underlying independent random effects $(E, \delta)$ into correlated effects across the vector $z$ of observations. This is a simple example of a factor-analytic covariance structure where the covariance structure is determined by a small number of interacting factors.

**Factor-analytic covariance structures** allow one to consider more general covariance structures informed by the data, rather than assumed by the investigator.

Piepho (1997a) and Denis et al. (1997) showed how this general framework can be extended to cases (e.g., Shulka 1972) where the lines have different variances,

$$V_{z_j} = \sigma^2_E \lambda \lambda^T + \text{diag}(\sigma^2_{\delta_1}, \ldots, \sigma^2_{\delta_{ng}}) + \sigma^2_e I$$

Likewise much more general covariance structures for the residuals can be incorporated,

$$V_{z_j} = \sigma^2_E \lambda \lambda^T + \text{diag}(\sigma^2_{\delta_1}, \ldots, \sigma^2_{\delta_{ng}}) + V_e$$
AMMI-based structured covariance models

The same logic used to generated a mixed-model Finlay-Wilkinson regression easily extends to other biadditive models, such as AMMI. The AMMIₙ model is given by

\[
z_{ijl} = \mu + G_i + E_j + \sum_{k=1}^{m} \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij} + \epsilon_{ijl}
\]

\[
= \mu + G_i + E_j + \sum_{k=1}^{m} w_{ki} v_{kj} + \epsilon_{ijl}^*
\]

The second line simplifies the AMMI model (to allow for estimability) in two ways. First, the singular value \(\lambda_k\) is absorbed into the genotype sensitivity \((w_{ki})\) and environmental \((v_{kj})\) coefficients. Second, the error in predicting GE from the AMMI approximation \((\delta_{ij})\) and the model residual \((\epsilon_{ijl})\) are combined into a single residual \(\epsilon_{ijl}^*\). Assume genotypes are random and environments are fixed, so that \(E_j\) and \(v_{kj}\) are fixed, while \(G_i\) and \(w_{ki}\) are random (as is, of course, \(\epsilon_{ijl}^*\)). Assume these underlying components are independent and homoscedastic,

\[
G_i \sim N(0, \sigma_G^2), \quad w_{ki} \sim N(0, \sigma_k^2), \quad \epsilon_{ijl}^* \sim N(0, \sigma_e^2 + \sigma_8^2)
\]

The resulting variance for the trait value of a random individual drawn from environment \(j\) becomes

\[
\sigma^2(z_{ij}) = \sigma_G^2 + \sum_{k=1}^{m} \sigma_k^2 v_{kj}^2 + \sigma_e^2 + \sigma_8^2
\]

Hence, the resulting genetic variance in environment \(j\) is just

\[
\sigma^2(z_{ij}) = \sigma_G^2 + \sum_{k=1}^{m} \sigma_k^2 v_{kj}^2 + \sigma_8^2
\]
Likewise, the covariance between the same random genotype over different environments \((j\) and \(\ell\)) becomes

\[
\sigma^2(z_{ij}, z_{i\ell}) = \sigma \left( G_i + \sum_{k=1}^{m} w_{ki} v_{kj}, G_i + \sum_{k=1}^{m} w_{ki} v_{k\ell} \right)
\]

\[
= \sigma_G^2 + \sum_{k=1}^{m} \sigma(w_{ki} v_{kj}, w_{ki} v_{k\ell}) = \sigma_G^2 + \sum_{k=1}^{m} v_{kj} v_{k\ell} \sigma(w_{ki}, w_{ki})
\]

\[
= \sigma_G^2 + \sum_{k=1}^{m} \sigma_k^2 v_{kj} v_{k\ell}
\]

The resulting covariance matrix for the vector \(z_i\) of observations of genotypes over the \(n_e\) environments can be written in the form

\[
\mathbf{V}_{z_i} = \sigma_G^2 \mathbf{J} + \lambda \lambda^T + \sigma_e^2 \mathbf{I}
\]

where \(\mathbf{J}\) is a matrix of ones, and \(\lambda\) is the \(n_e \times m\) matrix,

\[
\lambda = \begin{pmatrix} \lambda_1 & \cdots & \lambda_m \end{pmatrix}, \quad \text{where} \quad \lambda_k = \sigma_k^2 \begin{pmatrix} v_{k1} \\ \vdots \\ v_{kn_e} \end{pmatrix} = \begin{pmatrix} \lambda_{k1} \\ \vdots \\ \lambda_{k n_e} \end{pmatrix}
\]

Chapter 39 goes into detail on further analysis of this model, and its connection to stability analysis and the response to selection.
### Summary: Structured Covariance models

<table>
<thead>
<tr>
<th></th>
<th>$G$ random, $E$ Fixed</th>
<th>$E$ random, $G$ Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2(z_{ij})$</td>
<td>$\sigma(z_{ij}, z_{ik})$</td>
<td>$\sigma^2(z_{ij})$</td>
</tr>
<tr>
<td>Index range</td>
<td>$1 \leq j \leq n_e$</td>
<td>$1 \leq i \leq n_g$</td>
</tr>
<tr>
<td>$C$</td>
<td>$\sigma_G^2 + \sigma_{GE}^2$</td>
<td>$\alpha_i \sigma_E^2 + \sigma_{\delta_i}^2$</td>
</tr>
<tr>
<td>$S$</td>
<td>$\sigma_G^2 + \sigma_{GE_j}^2$</td>
<td>$\sigma_E^2 + \sigma_{C_i E}^2$</td>
</tr>
<tr>
<td>FW</td>
<td>$\sigma_G^2 + E_j \sigma_\beta^2 + \sigma_{\delta_j}^2$</td>
<td>$\alpha_i \sigma_E^2 + \sigma_{\delta_i}^2$</td>
</tr>
<tr>
<td>FA(m)</td>
<td>$\sigma_G^2 + \sum_{\ell}^m \lambda_{\ell j}^2$</td>
<td>$\sigma_E^2 + \sum_{\ell}^m \lambda_{\ell i} \lambda_{\ell k}$</td>
</tr>
<tr>
<td>U</td>
<td>$\sigma_j^2$</td>
<td>$\sigma_i^2$</td>
</tr>
</tbody>
</table>

Summary the covariance structures for various mixed-models for $G \times E$. When $G$ is taken as random with $E$ fixed, $\sigma^2(z_{ij})$ is genetic variance in environment $j$, while $\sigma(z_{ij}, z_{ik})$ is the covariance between a random genotype ($i$) measured in environments $j$ and $k$. When $E$ taken as random, $\sigma^2(z_{ij})$ corresponds to the variance for an individual from genotype $i$ drawn from a random environment, while $\sigma(z_{ij}, z_{kj})$ is the covariance between genotypes $i$ and $k$ when measured in across a random environment ($j$). $C$ corresponds to the Compound Symmetry model, $S$ is Shukla’s extension, $FW$ is Finlay-Wilkinson where $\alpha_i = 1 + \beta_i$, $FA(m)$ is factor-analytic model (i.e., a mixed AMMI-type model) with $m$ factors, $U$ is the completely Unstructured model.