

Lecture 08:  
 $G \times E$ : Genotype-by-  
environment interactions:  
Standard methods

Bruce Walsh lecture notes  
Tucson Winter Institute  
9 - 11 Jan 2013

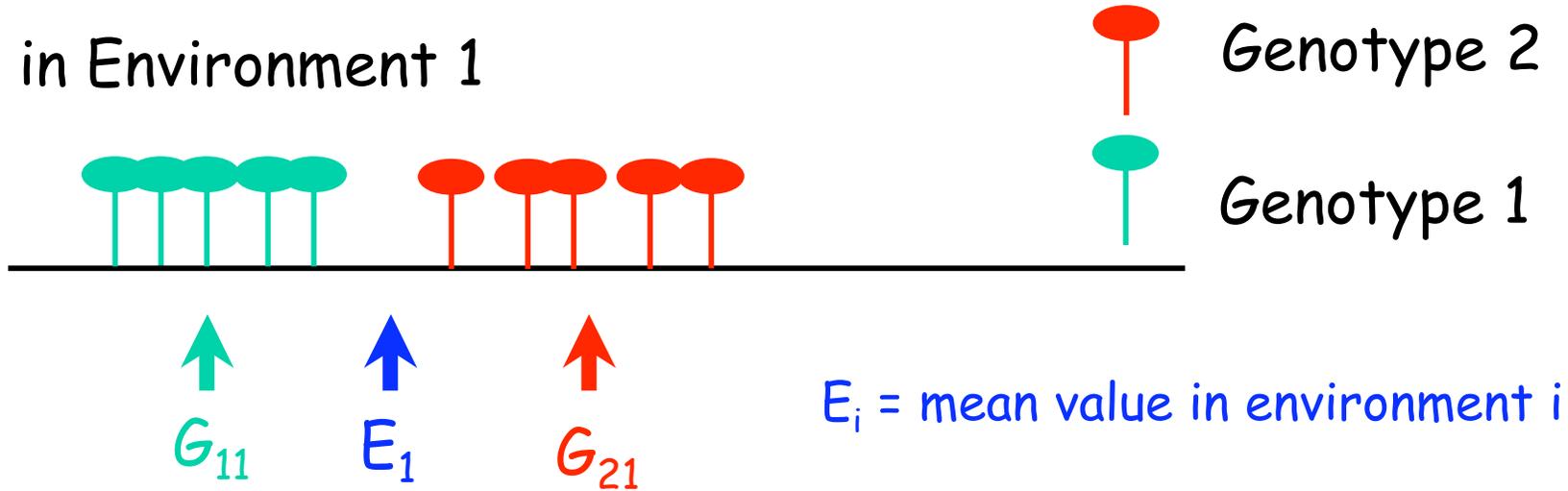
# $G \times E$

- Introduction to  $G \times E$ 
  - Basics of  $G \times E$
  - $G \times E$  is a correlated characters problem
  - Finlay-Wilkinson regressions
- SVD-based methods
  - The singular value decomposition (SVD)
  - AMMI models
- Factorial regressions

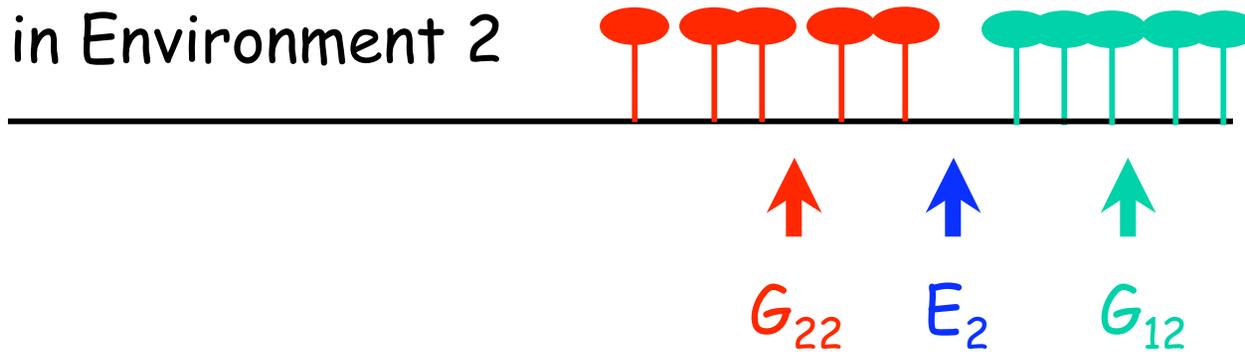
# Genotypes vs. individuals

- Much of the  $G \times E$  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.

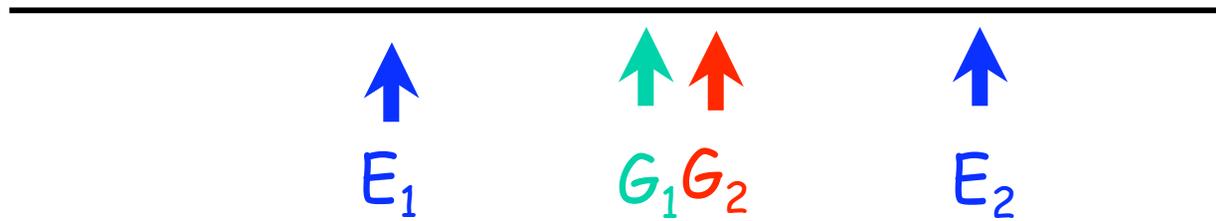
Yield in Environment 1



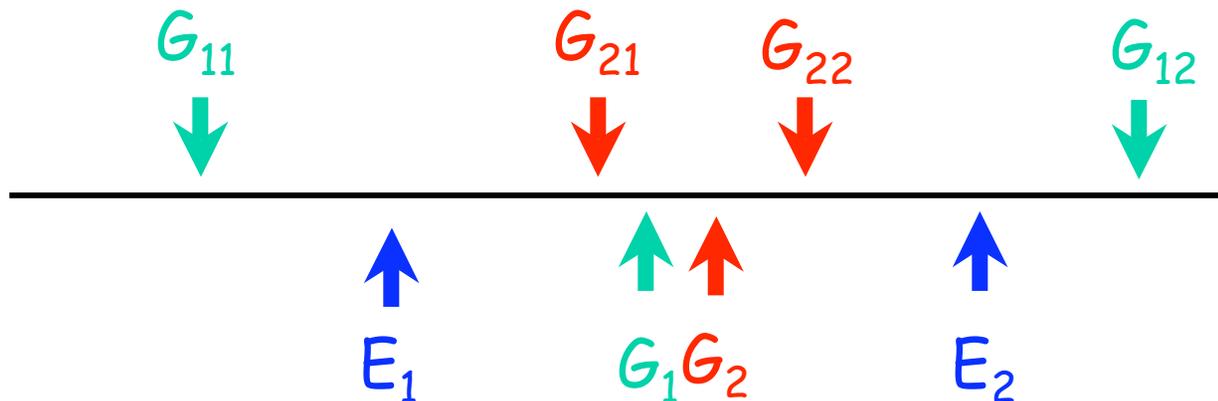
Yield in Environment 2



Overall means



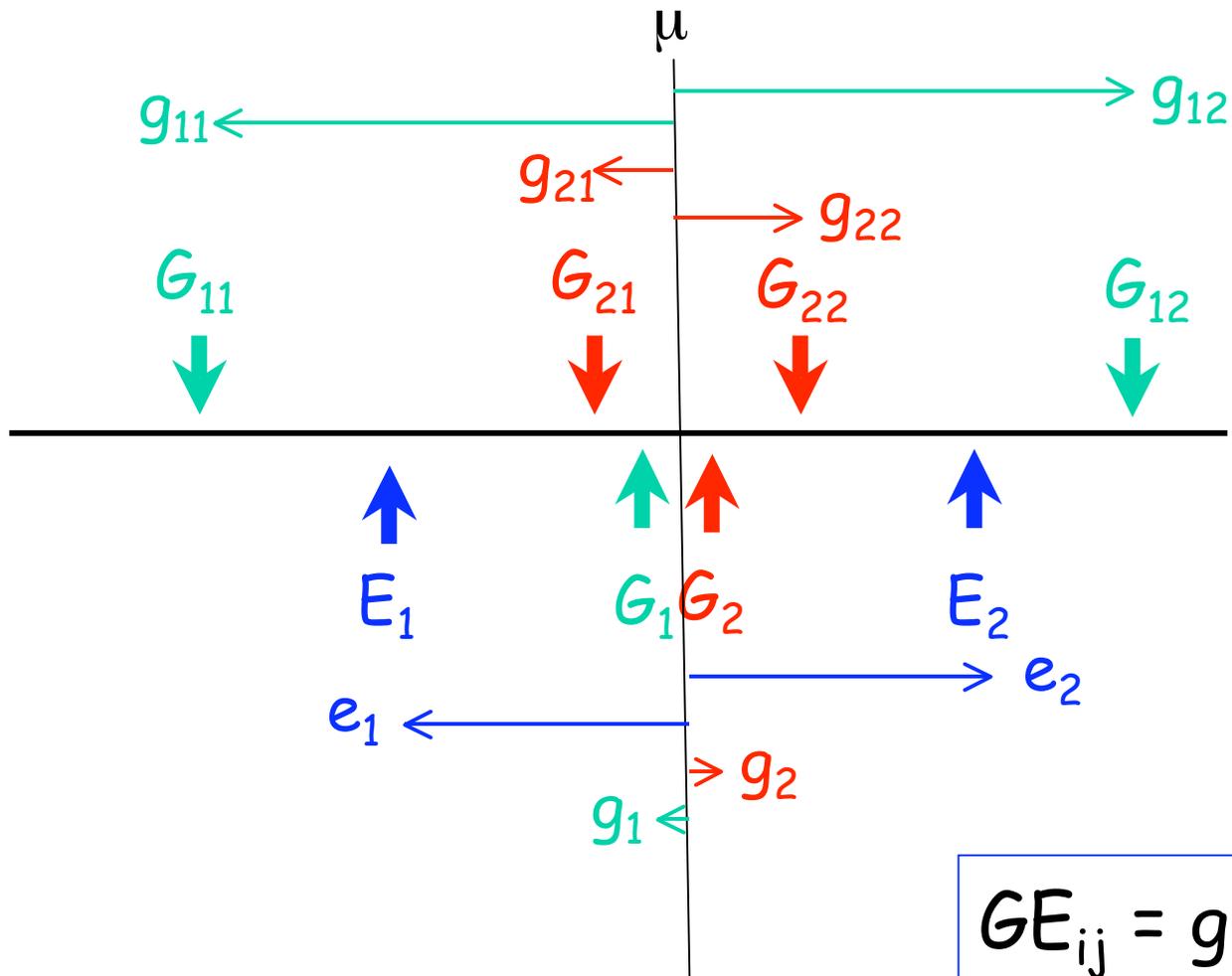
$G_{ij}$  = mean of genotype  $i$  in environment  $j$



Under base model of Quantitative Genetics,

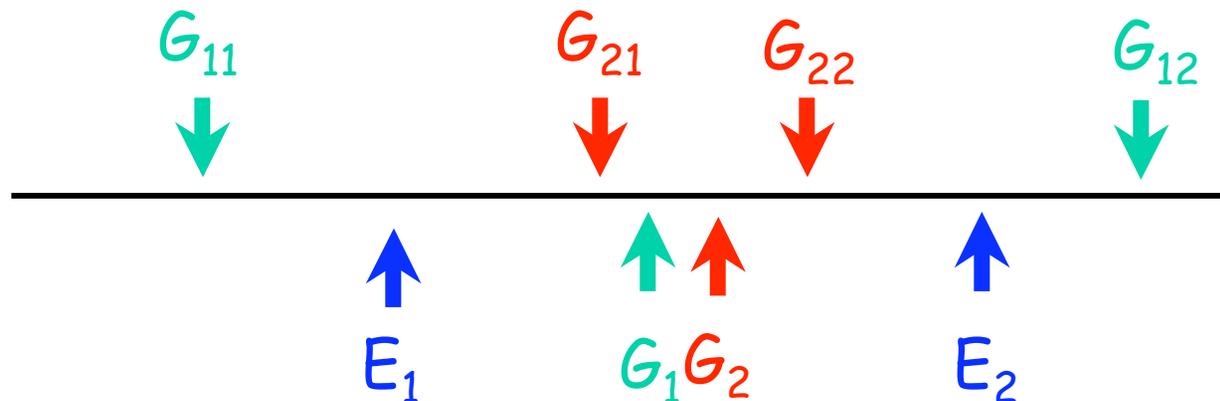
$$G_{ij} = \mu + G_i + E_j$$

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



Components measured as deviations  
from the mean  $\mu$

# 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

# $G \times E$ : 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**
- High  $\sigma^2(GE)$  implies high  $G \times E$  in at least some lines in the sample.

## $G \times E$ is Both a Challenge and an Opportunity

		Mean Performance	
		High	Low
Amount of $G \times E$	High	Potential for locally-adapted lines	Potential for locally-adapted lines
	Low	Ideal. Potential for widely adaptive lines	Undesirable

High  $G \times E$  = potential for locally-adapted lines

High  $G \times E$  = poor stability across environments

$G \times E$  can be generated by either differences in the additive variance over environments or by lack of perfect genetic correlation among environments

For two environments, Robertson (1959) showed that the  $G \times E$  interaction variance can be partitioned into these two sources,

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

where  $\sigma_{A_i}^2$  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,

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

Here  $\bar{\sigma}_A$  is the average of the square root of the genetic variances over all environments, and  $r_A(i, j)$  is the correlation between environments  $i$  and  $j$ .

# Major vs. minor environments

- An identical genotype will display slightly different traits values even over apparently identical environments due to 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.**

# Where to select?

- Suppose can only select in one environment when  $G \times E$  is present
  - Selection with  $G \times E$  is a correlated traits problem
  - Direct response = change when selected in that environment
  - Correlated response = change when selected in another environment
- Is it better to select in the better, or in the poorer, environment?
  - **Hammond's conjecture**: Best to select in the poor environment. Support mixed

# Jinks-Connolly Rule

- **Antagonistic selection**
  - Select in opposite direction from environmental effect
    - Up-select in an environment with reduced trait value
- **Synergistic selection**
  - Select in same direction as environmental effect
    - Up-select in an environment with increased trait value
- Jink and Connolly (1973) suggested that
  - **Antagonistic selection reduces environmental sensitivity** (i.e., improves stability)
  - **Synergistic selection increases sensitivity** (decreases stability)
  - While not always true, this is often true, and hence is a **trend** (rather than a rule)
- Falconer's (1990) generalization is that
  - **sensitivity is less after antagonistic selection than after synergistic selection**

# 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  $\sum_k \alpha_{ki}\gamma_{kj}$ . 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  $\sum_k \alpha_{ki}\gamma_{kj}$ .
- 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**.

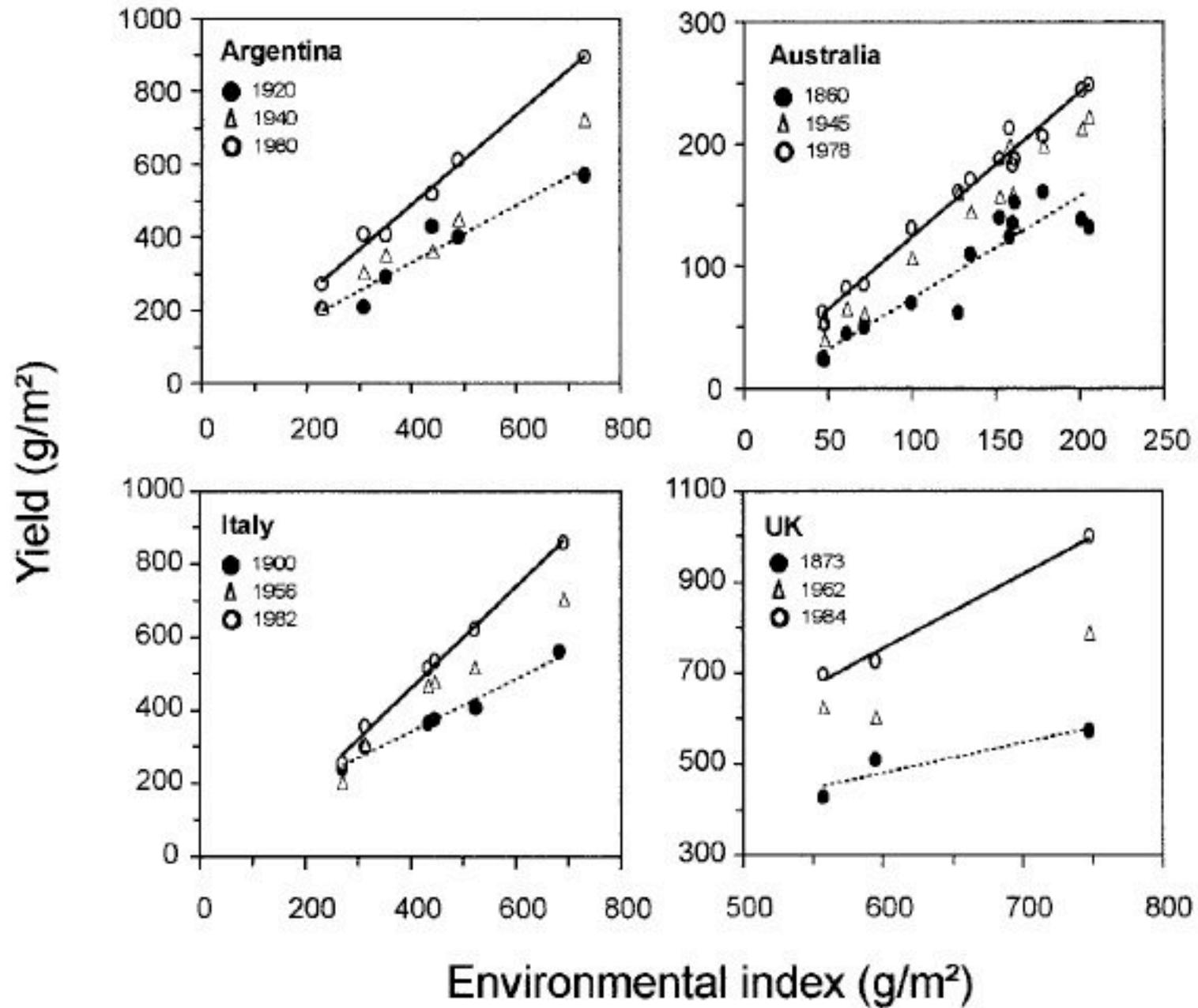
Let  $\mu + G_i$  be the mean of the  $i$ th genotype over all the environments, and  $\mu + E_j$  be the average yield of all genotypes in environment  $j$

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

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 accounts 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  $\beta_i$

# $\beta$ and Stability

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

- Since predicted  $GE_{ij}$  term is  $\beta_i E_j$ ,  $\beta_i$  measures the sensitivity of genotype  $i$  over the sampled environments
  - Positive  $\beta$  implies sign of  $GE =$  sign of  $E$ 
    - Good environment = extra gain from positive  $GE$
    - Poorer environment = extra loss from negative  $GE$
  - Negative  $\beta$  implies sign of  $GE$  opposite of sign of  $E$ 
    - Performs better in poorer environments
  - Large  $|\beta|$  implies a higher sensitivity over environments
  - $\beta_i = -1$  implies  $\mu_{ij} = \mu + G_i + \delta_{ij}$  (no dependence on  $E_j$ )

# Types of Stability

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

- **Type I stability** ( $\beta_i = -1$ ):
  - Genotypic value is constant over environments
- **Type II stability** ( $\beta_i = 0$ ):
  - No  $G \times E$ , but this also implies that genotypic value changes over environments
- **Type III stability** ( $\sigma^2(\delta)$  small):
  - The FW regression accounts for most of  $G \times E$

# 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  $\mathbf{\Lambda}$  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

$$\mathbf{A}_{n \times p} = \mathbf{U}_{n \times n} \mathbf{\Lambda}_{n \times p} \mathbf{V}_{p \times p}^T \quad (39.16a)$$

We have indicated the dimensionality of each matrix to allow the reader to verify that each matrix multiplication conforms. The diagonal elements  $\lambda_1, \dots, \lambda_s$  of  $\mathbf{\Lambda}$  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,

$$\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_i, \dots, \mathbf{u}_n), \quad \mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_i, \dots, \mathbf{v}_p) \quad (39.16b)$$

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  $\mathbf{\Lambda}$  is diagonal, it immediately follows from matrix multiplication that we can write any element in  $\mathbf{A}$  as

$$A_{ij} = \sum_{k=1}^s \lambda_k u_{ik} v_{kj} \quad (39.16c)$$

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  $G \times E$  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**),

$$\hat{A}_{ij} = \sum_{k=1}^m \lambda_k u_{ik} v_{kj} \quad (39.17a)$$

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

$$GE_{ij} \simeq \lambda_1 u_{i1} v_{j1} + \lambda_2 u_{i2} v_{j2} \quad (39.17b)$$

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

$$\sum_{k=1}^m \lambda_k^2 / \sum_{ij} A_{ij}^2 = \frac{\lambda_1^2 + \dots + \lambda_m^2}{\lambda_1^2 + \dots + \lambda_s^2}$$

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  $GE_{ij}$  = value for Genotype  $i$  in enviro.  $j$

In  $\mathbf{R}$ , the compact SVD (Equation 39.16d) of a matrix  $X$  is given by  $\mathbf{svd}(X)$ , 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  $746.10^2 / (746.10^2 + 131.36^2 + 0.53^2) = 97.0\%$  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  $\Lambda$  except the first entry to zero,

$$GE_1 = \begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

Similarly, the rank-2 SVD is given by setting all but the first two singular values to zero,

$$\mathbf{GE}_2 = \begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 131.36 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

For example, the rank-1 SVD approximation for  $\mathbf{GE}_{32}$  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  $\mathbf{GE}_{ij}$  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  $GE$ :

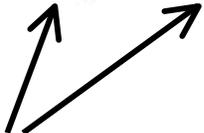
$$GE_{ij} = \sum_{k=1}^m \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij}$$

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)

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

# Why do AMMI?

- One can plot the SVD terms ( $\gamma_{ki}, \eta_{kj}$ ) to visualize interactions
  - Called **biplots** (see online notes Chapter 33 for details)
- AMMI can better estimate mean values of  $GE_{ij}$  than just using the cell value (the observed mean of Genotype  $i$  in Environment  $j$ )
- AMMI can predict GE values for genotype-environment combination not measured
- A huge amount more on AMMI in the online notes (Chapter 33)!

## Modifications of the Basic AMMI Family of Models

A variety of modifications of Equation 39.19 appear in the literature. Two common variations are the **sites regression model**, or **SREG** (Crossa and Cornelius 1997) wherein the genetic main effect ( $G_i$ ) is absorbed into the interaction terms and hence the regression main effects are only over sites ( $E_j$ ),

$$\mu_{ij} = \mu + E_j + \sum_{k=1}^m \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij} \quad (39.20a)$$

and the **shifted multiplicative model**, or **SHMM**, Crossa and Cornelius 1997) where *both* the environment and genetic main effects are absorbed into the interaction terms,

$$\mu_{ij} = \mu + \sum_{k=1}^m \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij} \quad (39.20b)$$

As with AMMI, these models are usually subscripted to indicate the number of multiplicative terms included, so that SREG<sub>3</sub> is Equation 39.20a with  $m = 3$  terms. As we will shortly see, these variants can prove useful for joint visualization of both genetic main effects plus GE interactions. Other variants have also been proposed, again based on which terms are kept as main effects versus being absorbed into a general interaction term, see Cornelius and Crossa (1999) for details.

# Factorial Regressions

- While AMMI models attempt to extract information about how  $G \times E$  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  $G \times E$ .
- 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.

Model	Interpretation
<b>Finlay-Wilkinson</b> $GE_{ij} = \beta_i(E_j - \mu) + \delta_{ij}$	$\beta_i$ = sensitivity of genotype $i$ to the average effect $E_j$ of the environment.
<b>AMMI</b> $GE_{ij} = \sum_{k=1}^m \lambda_k \gamma_{ki} \eta_{kj} + \delta_{ij}$	First $m$ terms of the SVD of the <b>GE</b> matrix $\lambda_k^2$ is the amount of variation explained by axis $k$ $\gamma_{ki}$ = sensitivity of genotype $i$ to environmental axis $k$ $\eta_{kj}$ = value of environment $j$ on the $k$ th environmental axis
<b>Factorial Regression</b> $GE_{ij} = \sum_{k=1}^m \zeta_{ki} x_{kj} + \delta_{ij}$	Modeling $G \times E$ using $m$ measured environmental factors $x_{kj}$ = value of $k$ th environmental factor in environment $j$ $\zeta_{ki}$ = sensitivity of genotype $i$ to $k$ th environmental factor
<b>Reduced rank Factorial Regression</b> $GE_{ij} = \sum_{k=1}^m \zeta_{ki} (\sum_p c_{kp} x_{pj}) + \delta_{ij}$	Modeling $G \times E$ based on a reduced dimensional set of the observed environmental factors by constructing $m$ combinations (axes) of these effects. $c_{kp}$ = loading of $p$ th environmental factor on axis $k$ . $\zeta_{ki}$ = sensitivity of genotype $i$ to $k$ th environmental combination (axis)
<b>AMMI using Reduced rank Factorial Regression</b> $GE_{ij} = \sum_k \lambda_k \gamma_{ki} (\sum_p c_{kp} x_{pj}) + \delta_{ij}$	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 $\eta_{kj} = \sum_p c_{kp} x_{pj}$ .