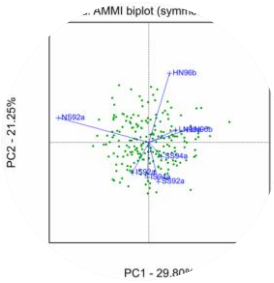
Extensions of the Finlay-Wilkinson model: AMMI models and factorial regression

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



The issue

- ❑ What environmental information to use to explain the GxE?

- ❑ Implicit information present in y
  - Finlay Wilkinson regression
    - E = environmental main effect
  - AMMI-model
    - Optimal environmental info explaining the GxE

- ❑ Explicit information not present in y
    - Temperature, rainfall, soil parameters, ...
  - Factorial regression models

Implicit environmental information

The AMMI-model

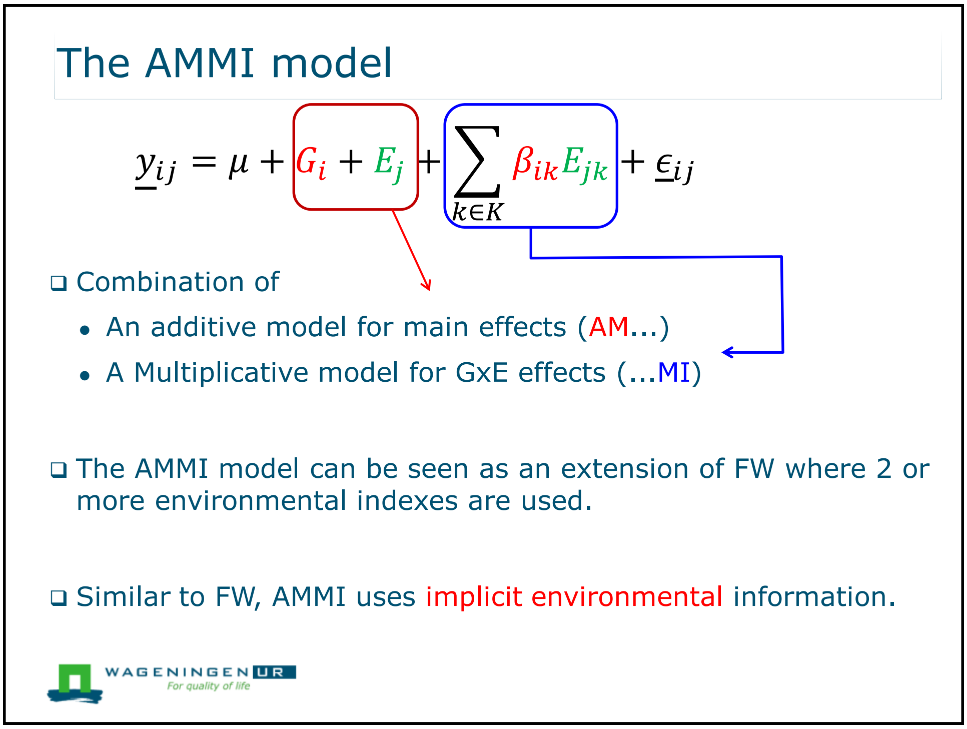
Extension of the Finlay Wilkinson model

□ FW is a multiplicative model with one dimension:

- $\underline{y}_{ij} = \mu + G_i + E_j + \beta_i E_j + \underline{\epsilon}_{ij}$
  - ○ β_i genotypic sensitivity
  - ○ E_j environmental characterization (index)

□ Goal: use more/better environmental indexes in the model.

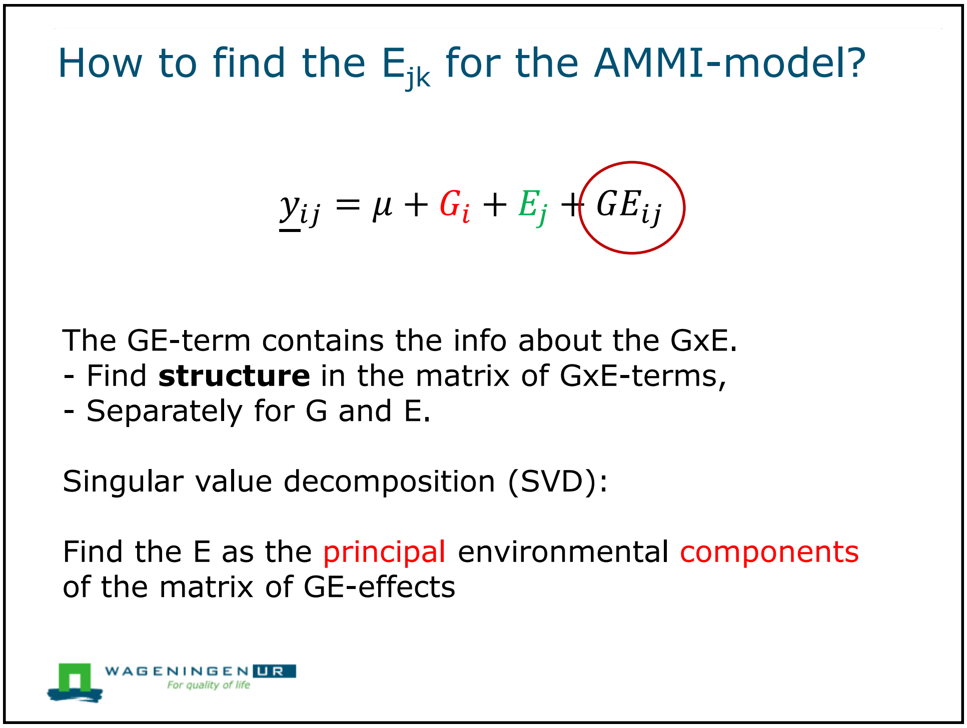
□ Issue: how to define those environmental indexes?



The AMMI model

$$\underline{y}_{ij} = \mu + \boxed{G_i + E_j} + \boxed{\sum_{k \in K} \beta_{ik} E_{jk}} + \underline{\epsilon}_{ij}$$

- ❑ Combination of
  - An additive model for main effects (AM…)
  - A Multiplicative model for GxE effects (…MI)
- ❑ The AMMI model can be seen as an extension of FW where 2 or more environmental indexes are used.
- ❑ Similar to FW, AMMI uses implicit environmental information.



How to find the E_{jk} for the AMMI-model?

$$\underline{y}_{ij} = \mu + G_i + E_j + \boxed{GE_{ij}}$$

The GE-term contains the info about the GxE.
- Find **structure** in the matrix of GxE-terms,
- Separately for G and E.

Singular value decomposition (SVD):

Find the E as the principal environmental components of the matrix of GE-effects

Intermezzo: singular value decomposition

❑ Full rank singular value decomposition (SVD) of a matrix:

$$A_{n\times m} = U_{n\times n}\,\Lambda_{n\times m}\,V^T_{m\times m}$$

- $U = (u_1, u_2 \dots u_n)$: matrix of left singular vectors
- Λ: rectangular diagonal matrix of singular values $\lambda_{\min(n,m)}$
- $V = (v_1, v_2 \dots v_n)$: matrix of right singular vectors

❑ A low rank approximation of matrix $A_{n\times m}$

$$A^*_{n\times m} = U_{n\times p}\,\Lambda_{p\times p}\,V^T_{p\times m}$$

- $A^*_{n\times m}$ is an approximation of $A_{n\times m}$ obtained by taking the p largest singular values.

Fitting an AMMI model: 2 Steps

❑ Step 1: fit $\underline{y}_{ij} = \mu + G_i + E_j + GE_{ij}$

- Obtain residuals as $\widehat{GE}_{ij} = y_{ij} - (\hat{\mu} + \hat{G}_i + \hat{E}_j)$

❑ Step 2: SVD of the residuals matrix $\widehat{GE}_{ij}$

- Left singular vectors: genotypic sensitivities.
- Right singular vectors: environmental indexes.

❑ Choose the *n* (e.g. 2) largest principal components

- These are the implicit “environments”

❑ We get a low-rank approximation of the $\widehat{GE}_{ij}$ matrix.

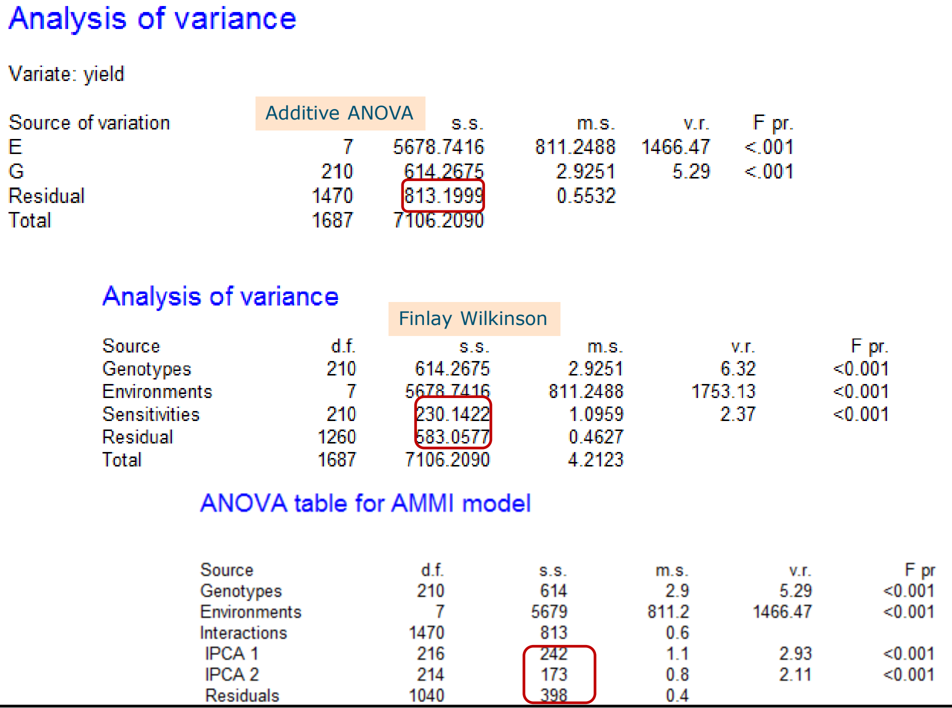
- AMMI-2: $\widehat{GE}^*_{ij} \cong \lambda_1 u_{i1} v_{j1} + \lambda_2 u_{i2} v_{j2}$
- Each environment has two environmental variables, v_{j1}, v_{j2}
- Each genotype has two sensitivities, u_{i1}, u_{i2}
  - One for each environmental variable

“Fitting” an AMMI model

$$\underline{y}_{ij} = \mu + G_i + E_j + \sum_{k\in K} \beta_{ik} E_{jk} + \underline{\epsilon}_{ij}$$

- ❑ We don’t really need to “fit” the AMMI-model
  - The SVD immediately produces
    - The environmental indexes (v)
      - These are the E’s
    - The genotypic sensitivities to these indexes (u)
      - These are the beta’s

- ❑ But we may still fit the PCA’s from the SDV for:
  - Testing of significance
  - Deciding how many PCA’s to include



Analysis of variance

Additive ANOVA

```
Variate: yield

Source of variation       d.f.       s.s.        m.s.      v.r.    F pr.
E                            7    5678.7416    811.2488   1466.47   <.001
G                          210     614.2675      2.9251      5.29   <.001
Residual                  1470     813.1999      0.5532
Total                     1687    7106.2090
```

Analysis of variance

Finlay Wilkinson

```
Source              d.f.        s.s.        m.s.        v.r.      F pr.
Genotypes            210     614.2675     2.9251        6.32    <0.001
Environments           7    5678.7416   811.2488     1753.13    <0.001
Sensitivities        210     230.1422     1.0959        2.37    <0.001
Residual            1260     583.0577     0.4627
Total               1687    7106.2090     4.2123
```

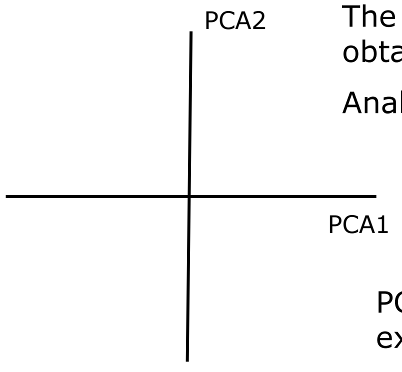
ANOVA table for AMMI model

```
Source          d.f.     s.s.     m.s.      v.r.      F pr
Genotypes        210      614      2.9      5.29    <0.001
Environments       7     5679    811.2   1466.47    <0.001
Interactions    1470      813      0.6
 IPCA 1          216      242      1.1      2.93    <0.001
 IPCA 2          214      173      0.8      2.11    <0.001
 Residuals      1040      398      0.4
```

AMMI biplots: visualizing the structure of the GxE

- ❑ Use multiplicative terms of SVD of $\widehat{GE}_{ij}$ as coordinates for genotypes and environments
  - u_{ik}define vectors for genotypes
  - v_{jk}define vectors for environments
- ❑ Origin of vectors = zero interaction
- ❑ The length of the vectors is proportional to the amount of GE for a genotype/environment
- ❑ The angle between genotypic vectors is proportional to the correlation
  - 0 degrees, r= 1; 90 degrees, r= 0; 180 degrees, r= -1
- ❑ Projecting genotypic vectors on environmental vectors approximates their GE

AMMI biplots: visualizing the structure of the GxE



Biplot

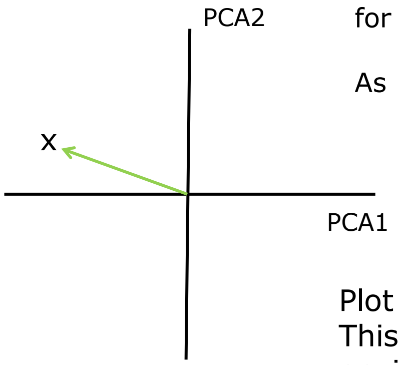
The PCA are implicit environments obtained from the data.

Analogous to the E_j in FW-regression

PCA1 = “environment” that explains most of the GxE,

PCA2 = “environment” that explains the 2nd most of the GxE

AMMI biplots: visualizing the structure of the GxE



Biplot

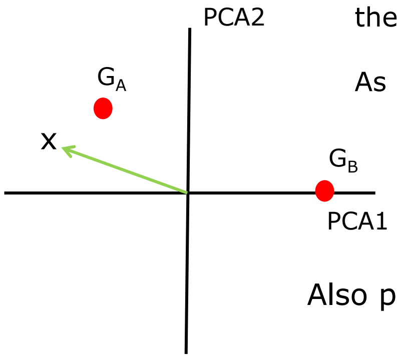
Each environment has a value for these "environments" (v_{1j} and v_{2j})

As if it were temperature, rain fall, etc.

Plot the environments in the biplot,
This shows the values of each environment for both PCA

Environment x has a negative value for v_1 and a positive for v_2

AMMI biplots: visualizing the structure of the GxE



Each genotype has a sensitivity to these "environments" (u_{1i} and u_{2i})

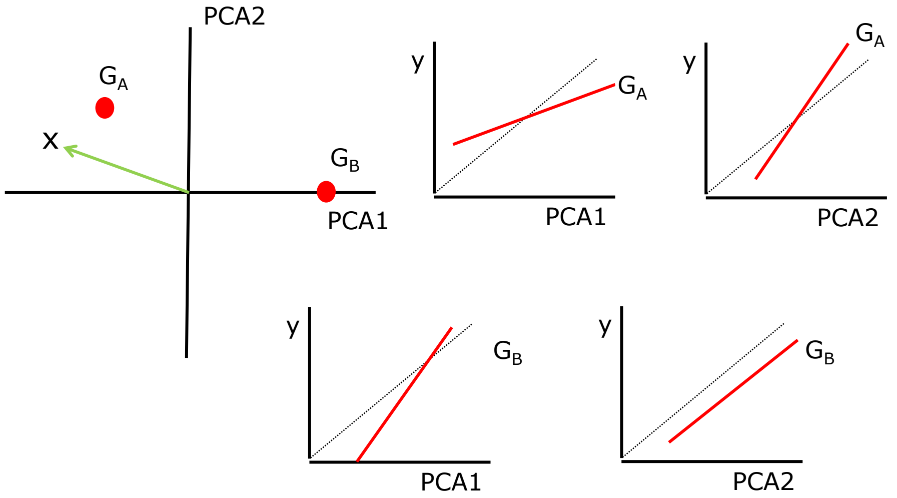
As if it were temperature tolerance

Also plot the genotypes in the biplot,

G_A performs relatively better if PCA1 is low and PCA2 is high
- $u_{1A} < 0$, $u_{2A} > 0$.

G_B is not sensitive to PCA2
- u_{2B} is very small

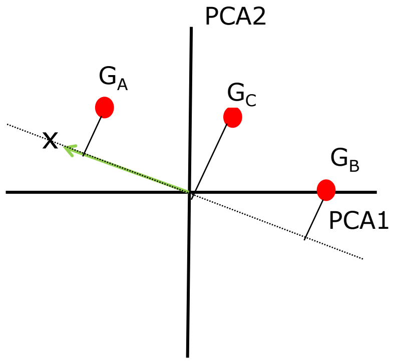
AMMI biplots: visualizing the structure of the GxE



Note: only look at the slope (relative to the average slope)

AMMI biplots: visualizing the structure of the GxE

However, the PCA are not environments that occur in reality



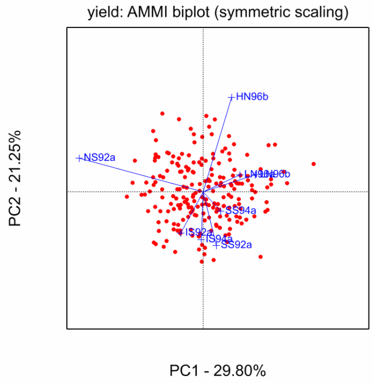
Projection on environmental axis
for real environments (such as x)
reveals the GxE

G_A has positive interaction with environment x

G_B has negative interaction with in environment x

G_C has little interaction with environment x

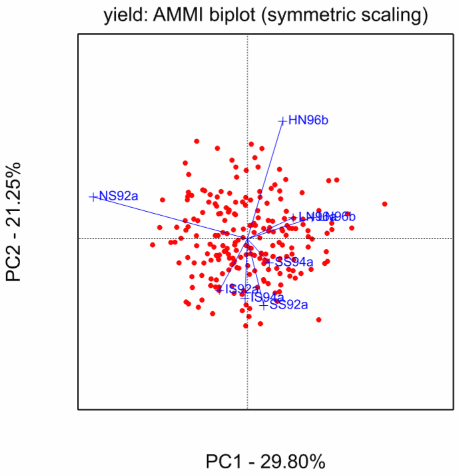
Graphical visualization of AMMI: biplots

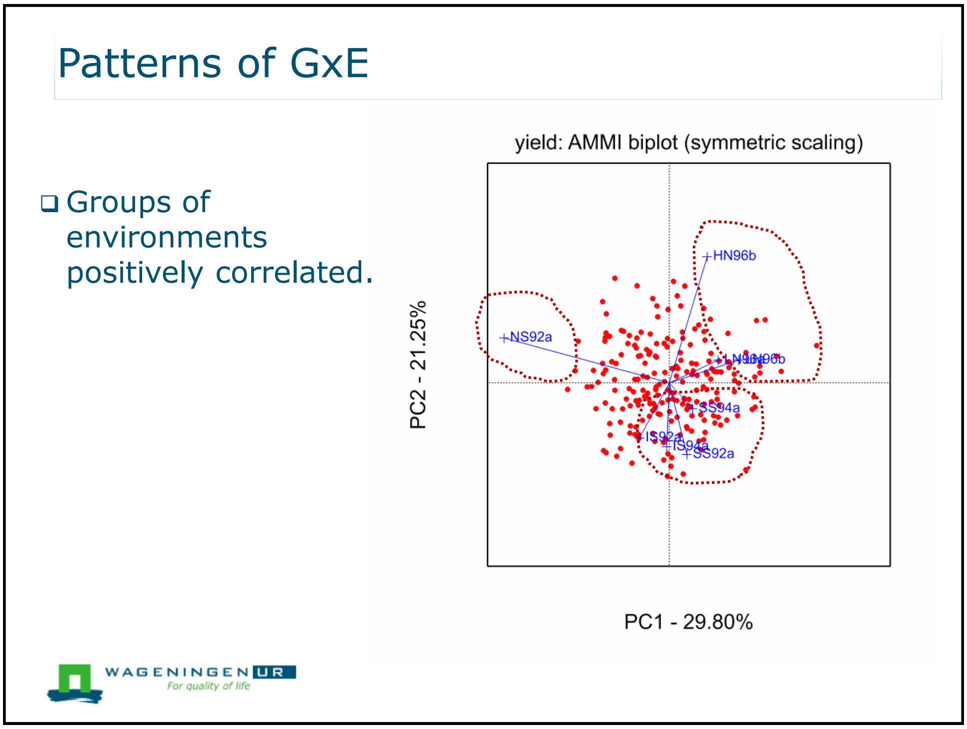


- □ Genotypes AND environments in the same plot.
- □ The coordinates define vectors for genotypes (u_{ik}) and environments (v_{jk})
    - Environmental vectors depart from the centre.

Graphical visualization of AMMI: biplots

- □ Length of vector reflect contribution to the total GxE:
    - NS92a and HN96b larger GxE than IS92a or SS94a
- □ Angle between vectors reflect correlation between environments:
    - 90° $r = 0$, high GxE, eg: NS92a and HN96b.
    - <90° $r > 0$ low GxE, eg: IS94a and SS92a.
    - >90° $r < 0$ high GxE, eg: HN96b and IS92a.

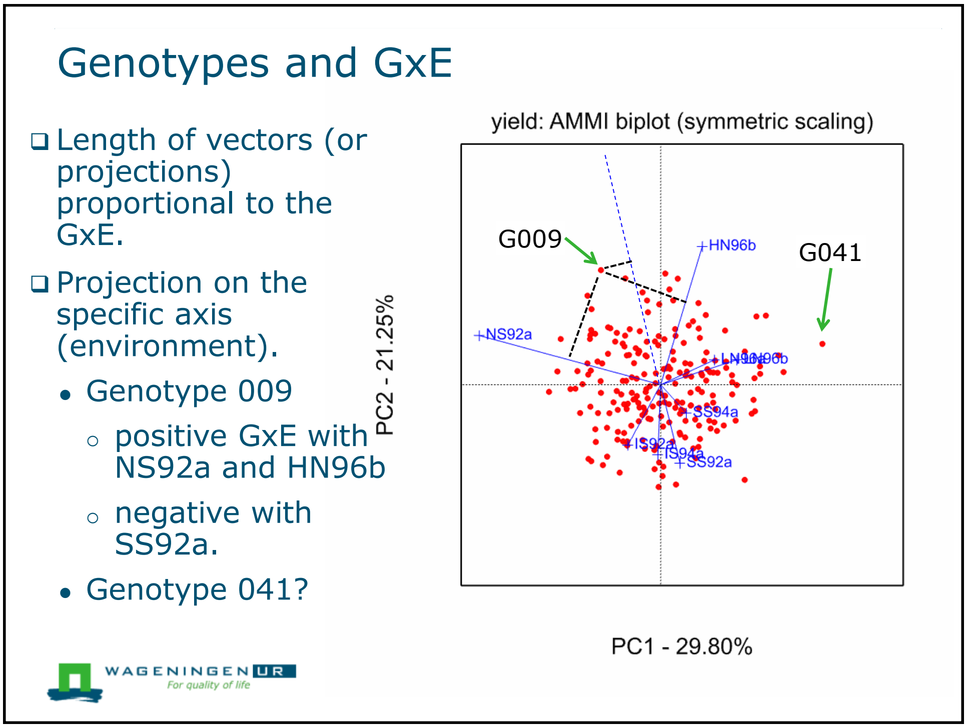




Patterns of GxE

- ❑ Groups of environments positively correlated.

WAGENINGEN UR
For quality of life



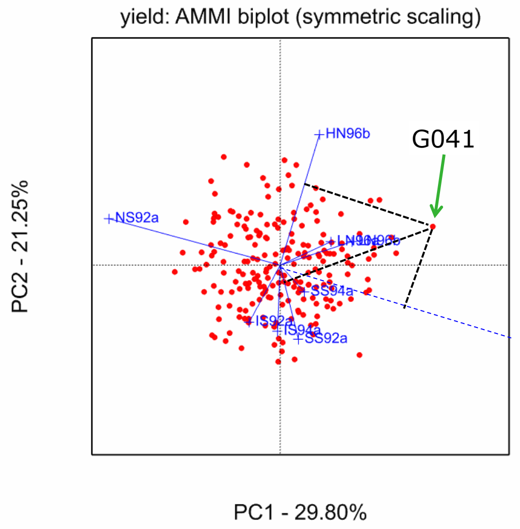
Genotypes and GxE

- ❑ Length of vectors (or projections) proportional to the GxE.
- ❑ Projection on the specific axis (environment).
  - Genotype 009
    - ○ positive GxE with NS92a and HN96b
    - ○ negative with SS92a.
  - Genotype 041?

WAGENINGEN UR
For quality of life

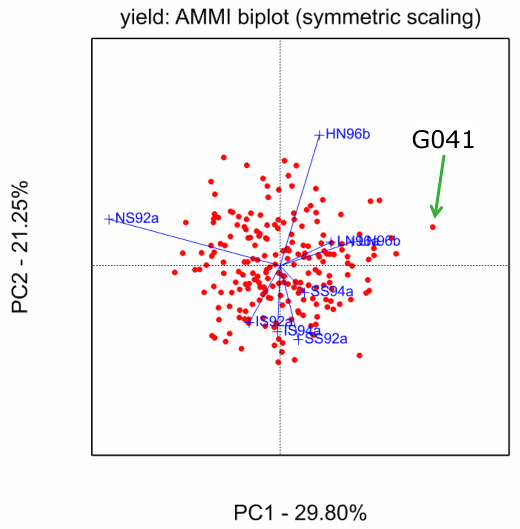
Genotype 041…

- ❑ Positive interaction with HN96b and SS92a, but negative with NS92a!
- ❑ Larger positive interaction with HN96b than with SS92a.



Biplots do not show the main effect!

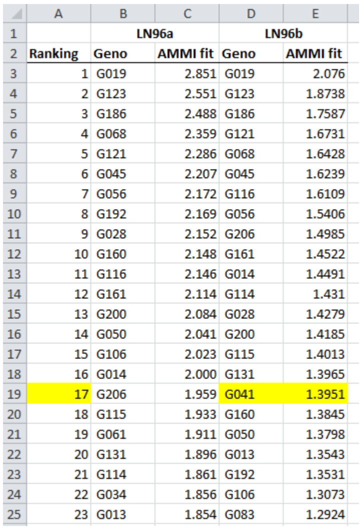
- ❑ G041 has by far the largest positive interaction with LN96a (and LN96b)
- ❑ Is G041 also the best performer in those two environments?
- ❑ We can't see overall performance in the AMMI biplot → it is only about GxE!



Fitted values from AMMI model

$$\underline{\hat{y}}_{ij} = \mu + G_i + E_j + \sum_{k\in K} \lambda_k u_{ik} v_{jk}$$

- Fitted values combine:
  - Main effects
  - GxE interaction effects
- G041 only appears 17th in the ranking in LN96b and not within the first 25 in LN96a (46th).
- For ranking genotypes, fitted values from the AMMI model should be used (and not only the multiplicative component)!



	A	B	C	D	E
1		LN96a		LN96b	
2	Ranking	Geno	AMMI fit	Geno	AMMI fit
3	1	G019	2.851	G019	2.076
4	2	G123	2.551	G123	1.8738
5	3	G186	2.488	G186	1.7587
6	4	G068	2.359	G121	1.6731
7	5	G121	2.286	G068	1.6428
8	6	G045	2.207	G045	1.6239
9	7	G056	2.172	G116	1.6109
10	8	G192	2.169	G056	1.5406
11	9	G028	2.152	G206	1.4985
12	10	G160	2.148	G161	1.4522
13	11	G116	2.146	G014	1.4491
14	12	G161	2.114	G114	1.431
15	13	G200	2.084	G028	1.4279
16	14	G050	2.041	G200	1.4185
17	15	G106	2.023	G115	1.4013
18	16	G014	2.000	G131	1.3965
19	17	G206	1.959	G041	1.3951
20	18	G115	1.933	G160	1.3845
21	19	G061	1.911	G050	1.3798
22	20	G131	1.896	G013	1.3543
23	21	G114	1.861	G192	1.3531
24	22	G034	1.856	G106	1.3073
25	23	G013	1.854	G083	1.2924

WAGENINGEN UR
For quality of life

AMMI summarized

- Similarly to FW model: uses implicit environmental indexes to characterize environments.
- Environmental indexes defined from a PCA analysis of residuals from an additive model.
  - Capture as much GxE as possible
- Higher flexibility than FW model.
- Graphical displays (biplots) to investigate patterns of GxE.

WAGENINGEN UR
For quality of life

Explicit environmental information

Factorial regression

Implicit environmental information

- Advantage:
  - Follows from y, no additional information needed.
- Disadvantage:
  - Interpretation?
  - Cannot predict unobserved environments.

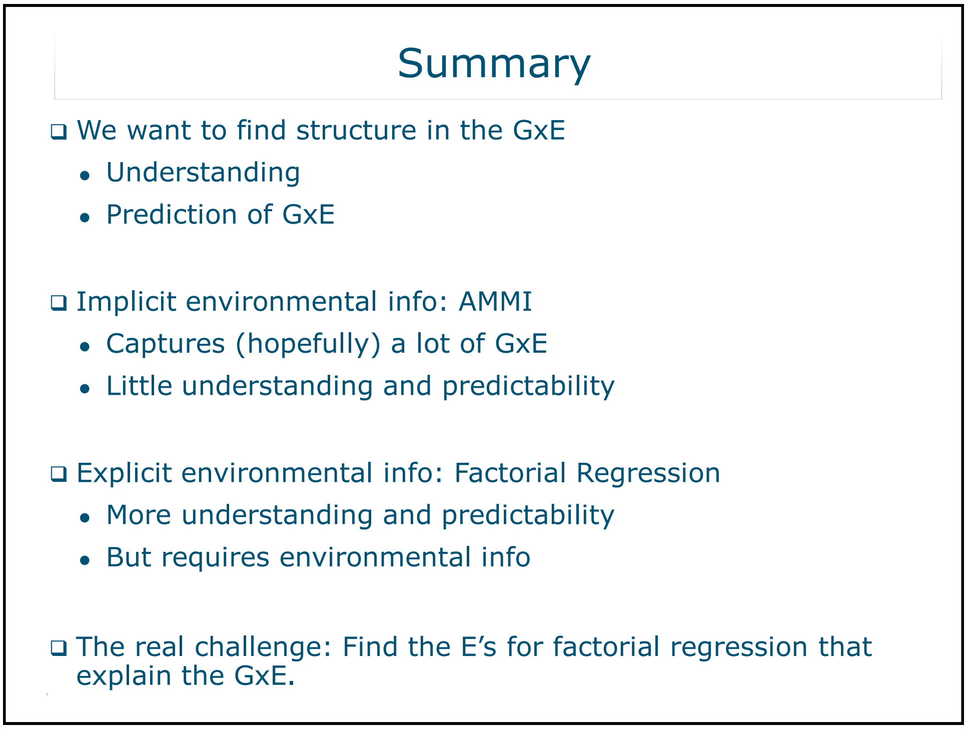
Explicit environmental information

- ❑ Environmental parameters
  - Temperature / Water availability / Radiation / Latitute – longitud / amount of Nitrogen, etc
- ❑ Advantages:
  - Better interpretation of GxE
  - Predict unobserved environments
  - Understanding mechanisms of GxE
    - ○ Connect crop physiology and statistics.

Factorial regression

$$\underline{y}_{ij} = \mu + G_i + E_j + \sum_{k \in K} \beta_{ik} Z_{jk} + \underline{\epsilon}_{ij}$$

- ❑ The Z_{jk} are observed environmental covariables
  - Temperature / %humidity / radiation / water / etc.
- ❑ Higher order relationships (eg: quadratic) possible
- ❑ Slope (β_{ik}) is sensitivity of genotype i to environmental covariable k
- ❑ The β_{ik} is interpretable!
  - E.g. sensitivity to temperature



Summary

- We want to find structure in the GxE
  - Understanding
  - Prediction of GxE

- Implicit environmental info: AMMI
  - Captures (hopefully) a lot of GxE
  - Little understanding and predictability

- Explicit environmental info: Factorial Regression
  - More understanding and predictability
  - But requires environmental info

- The real challenge: Find the E's for factorial regression that explain the GxE.