Robust singular value decomposition with application to multi-location plant breeding trials

Paulo Canas Rodrigues

Department of Statistics, Federal University of Bahia, Brazil
CMA/FCT/UNL, Nova University of Lisbon, Portugal

(joint work with Andreia Monteiro and Vanda Lourenço)

FCT
Fundação para a Ciência e a Tecnologia

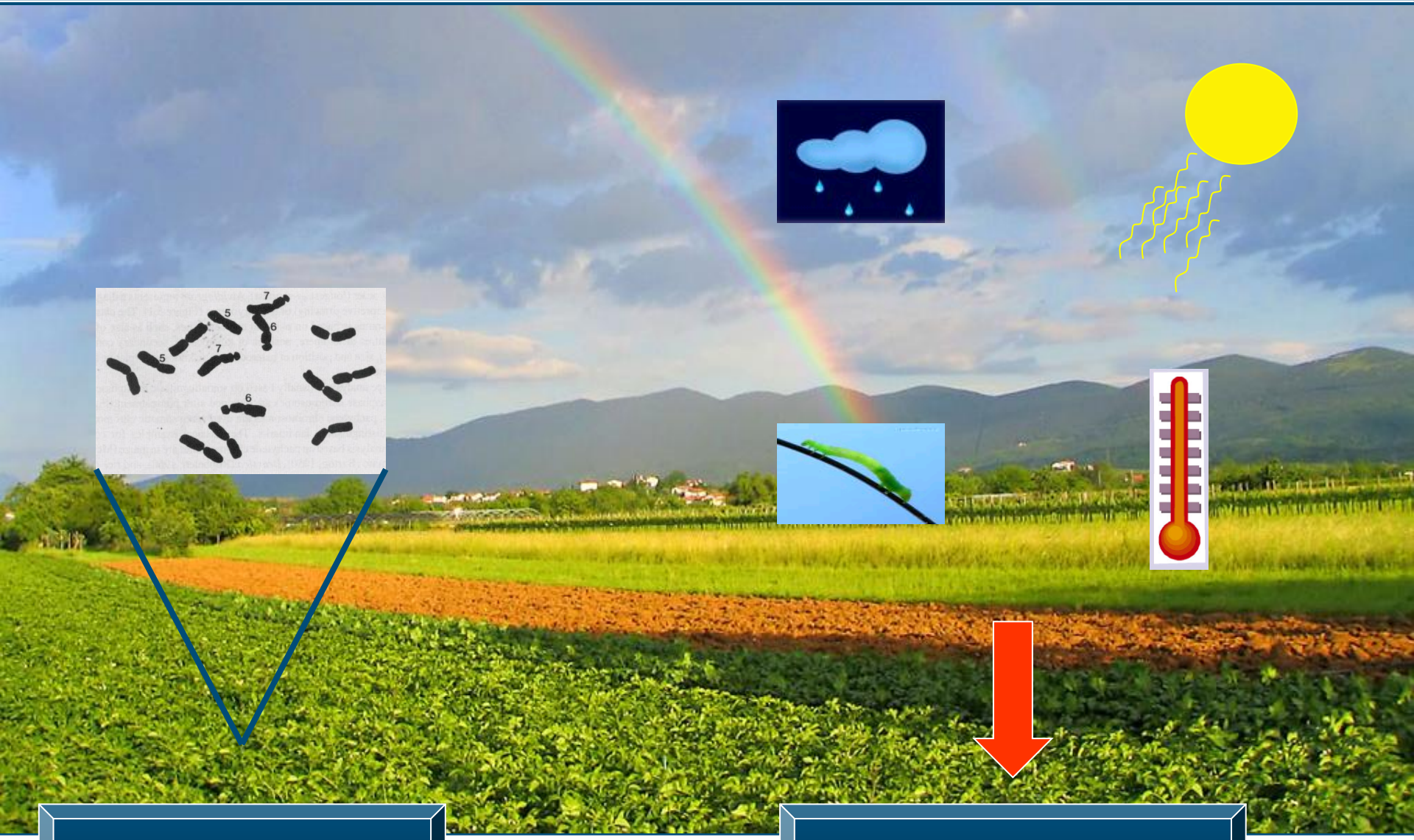
Outline of the presentation

1. Introduction and background
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Robust statistics
2. Materials and methods
   - Additive main effects and multiplicative interaction (AMMI) model
   - Robust AMMI model
   - Simulation study
3. Results
   - Simulation study
   - Real data example
4. Concluding Remarks

The ultimate aim in plant breeding…

- To produce “better” cultivars:
  - Quantity
  - Quality
- Successful breeding companies: sell better phenotypes
- How to improve phenotypes?
  - By improving the environment (agronomy)
  - By providing better genotypes (breeding)
  - Combination of the two: finding a good matching between genotypes and environments

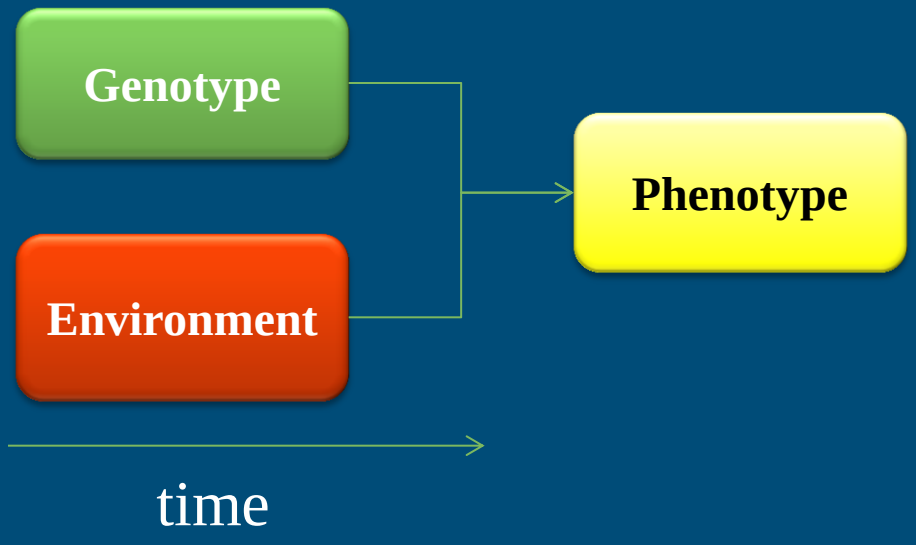
What's the phenotype?

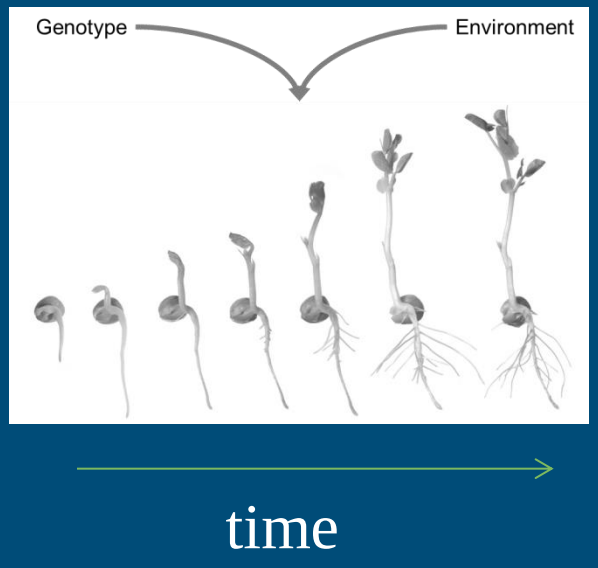


genotype

environment

Genotype & Environments *vs.* time



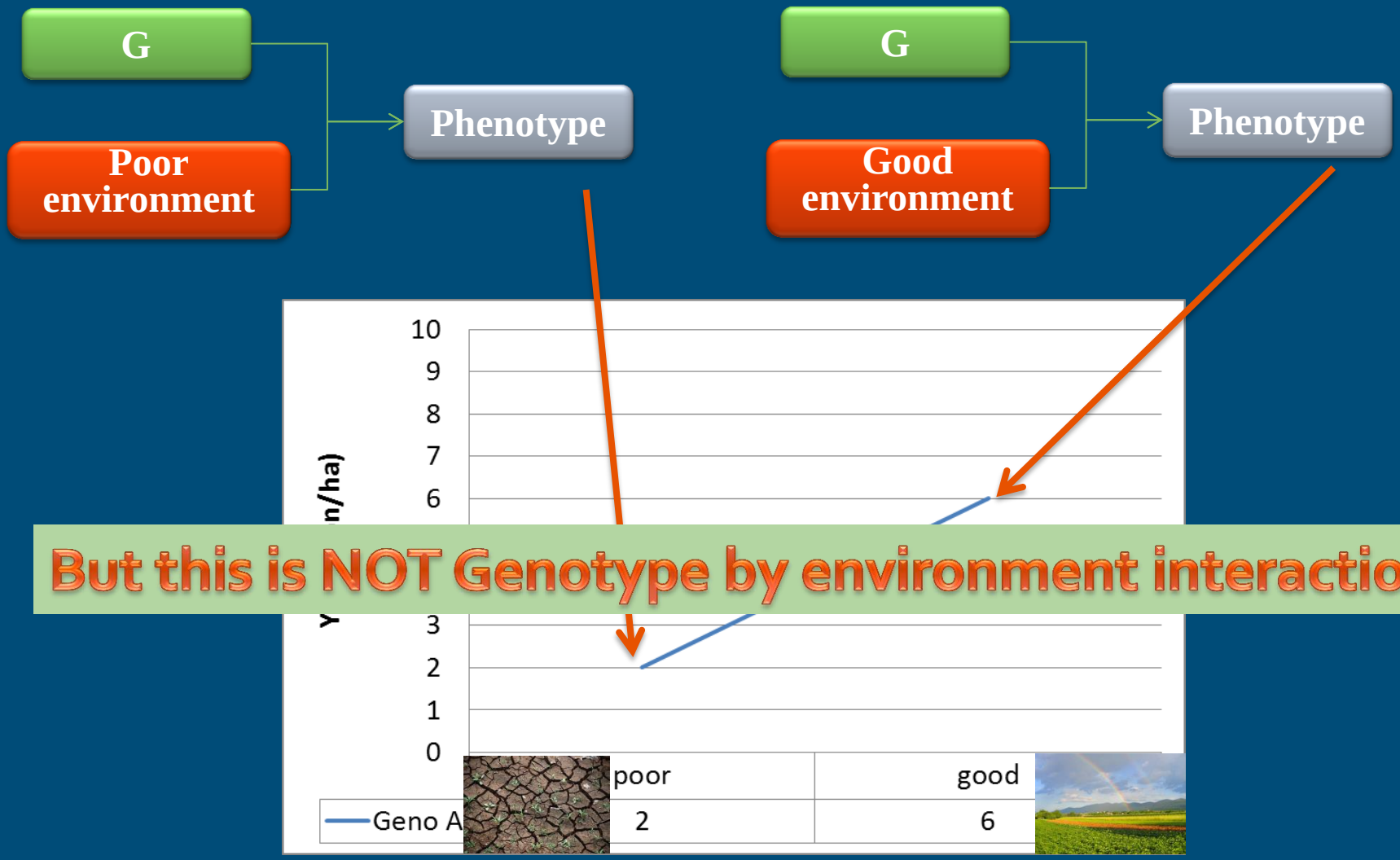


- Phenotype is the cumulative result of:
  - genetic make-up of plant
  - environment
  - over time and changing over the development of the plant

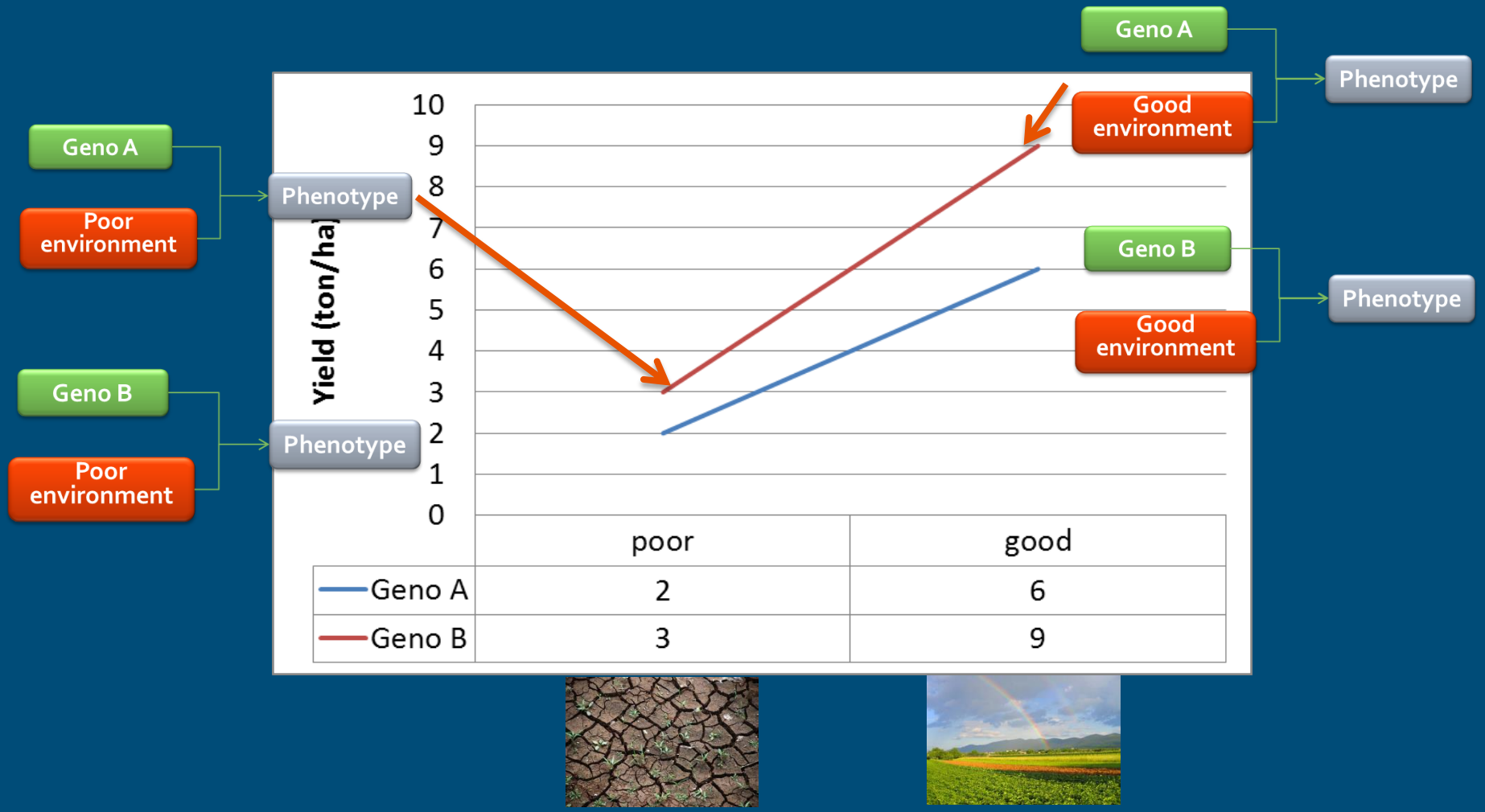
Environments can differ…

- Moisture (drought)
- Temperature (heat, chilling, freezing)
- Soil-Chemical (nutrients, minerals)
- Radiation

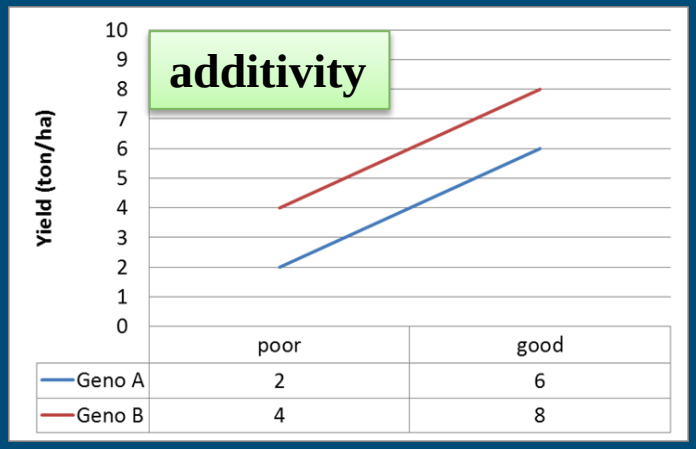
Environments are different → phenotype is different

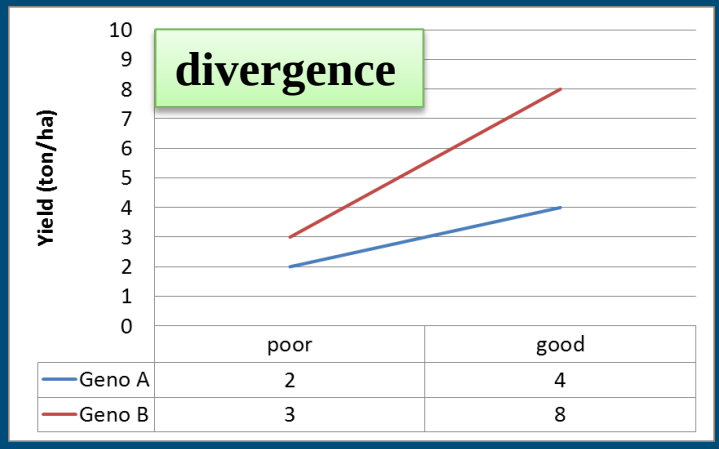


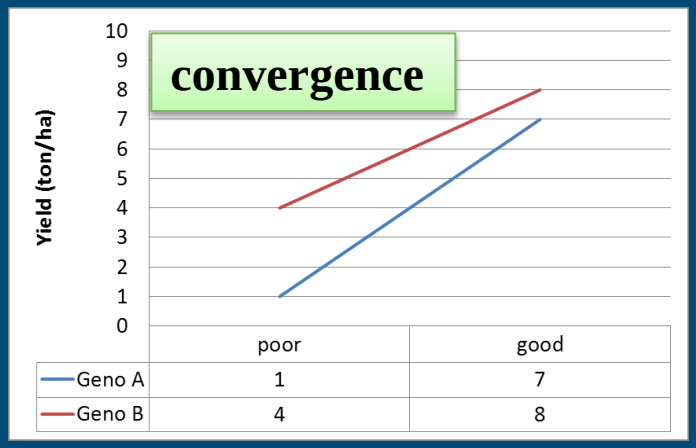
This is Genotype by Environment Interaction

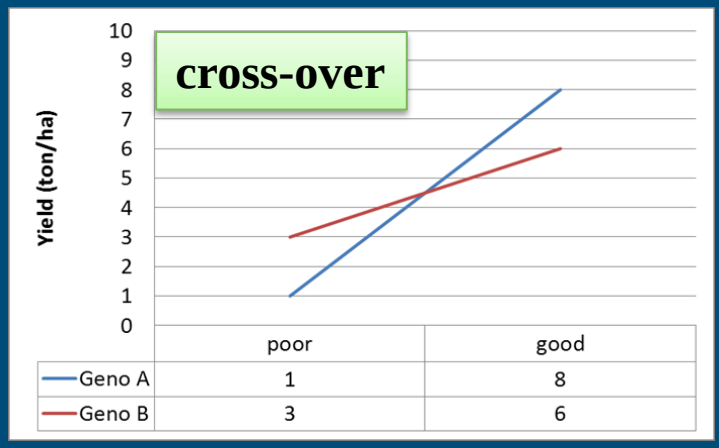


GxE as changing mean performance across environments









Statistical models for GxE

$y_{i,j} = \mu + G_i + E_j + \varepsilon_{i,j}$	additive model
$y_{i,j} = \mu + G_i + E_j + (G.E)_{i,j} + \varepsilon_{i,j}$	full interaction model
$y_{i,j} = \mu + G_i + E_j + \beta_i E_{j.} + \varepsilon_{i,j}$	Finlay and Wilkinson regression model
$y_{i,j} = \mu + G_i + E_j + \sum_{n=1}^{N} \lambda_n \gamma_{i,n} \delta_{j,n} + \varepsilon_{i,j}$	additive main effects and multiplicative interaction (AMMI) model
$y_{i,j} = \mu + E_j + \sum_{n=1}^{N} a_{i,n} b_{j,n} + \varepsilon_{i,j}$	genotype main effect plus genotype-by-environment interaction (GGE) model

The challenge

- to extract information from huge data sets, usually in the form of matrices;
- to analyze these big data matrices efficiently and to get results within a reasonable amount of time;
- to develop robust methodologies which can deal with problems such as missing values, and noisy data;
- to develop time-wise efficient computational algorithms;
- to develop data visualization tools for a quick analysis of the results and outputs.

A big challenge for Statisticians!!

Robust statistics

- **The problem**: For many observed data sets, the distribution of the quantitative traits is not normal and often shows heavy tails and outlying observations.

- **Some possible solutions**:
  - Transforming the data to normality
    - frequently comes together with interpretation issues
  - Removing outlying observations from the data
    - true outliers are not always visible, due to masking and swamping effects, among other reasons;
    - removing outliers from the data reduces sample size, may effect the distribution theory and variances may be underestimated from the cleaned data;
  - Use robust statistical methods

Outline of the presentation

1. Introduction and background
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Robust statistics
2. Materials and methods
   - Additive main effects and multiplicative interaction (AMMI) model
   - Robust AMMI model
   - Simulation study
3. Results
   - Simulation study
   - Real data example
4. Concluding Remarks

Why to use low-rank approximations?

- Compression of the data (mostly to reduce processing time);
- Prediction;
- Reconstructing latent signal, i.e. separation between signal and noise;
- Easier interpretation of results.

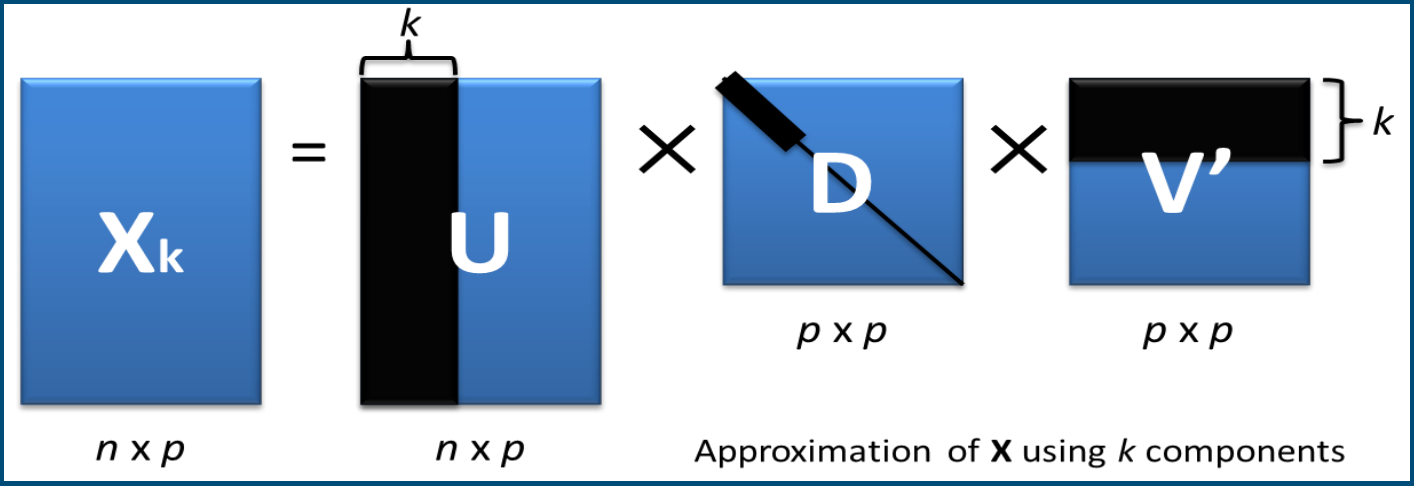
Singular Value Decomposition

We decompose the matrix $\boldsymbol{X}, rank(X) = r$ as

$$\boldsymbol{X} = \boldsymbol{UDV^T}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns and $\boldsymbol{D}$ is diagonal; $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$. Columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right singular vectors. The diagonal of $\boldsymbol{D}$ contains the p singular values.

A low-rank approximation of $\boldsymbol{X}$ using only k factors can be written as:



The approximation minimizes the error $||\boldsymbol{X} - \boldsymbol{X}_k||_F$ (Frobenius norm).

The limitations of LRA

- Classical PCA and LRA assume that all cells (row and column combinations) contribute equally to the final model.
- Performing a PCA or LRA or determining the solution of a low-rank approximation of a two-way data matrix where influent factors as outlying observations are present, is likely to led to erroneous and non-significant results.

Alternative: Robust Low-rank Approximations!

Robust AMMI model

- AMMI model: $$\boldsymbol{y} = \mathbf{1}_I \, \mathbf{1}_J^T \, \mu + \boldsymbol{\alpha}_I \, \mathbf{1}_J^T + \mathbf{1}_I \, \boldsymbol{\beta}_J^T + \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T + \boldsymbol{\varepsilon}$$

It gives good results when the data is not contaminated by outlying observations. However, field data such as data resulting from MET is prone to contamination and thus outlying observations are often found.

- Robust AMMI model: the linear fit is replaced by a robust fit (M-regression) and the use of the standard SVD by a robust SVD approach.

The choice of M-regression was based on the fact that in this kind of analysis contamination is only seen at the response variable level and not also at the explanatory variables level, in which case high breakdown and efficient MM-regression should be considered.

Simulation study

- Simulation of 1000 two-way data tables with 100 genotypes/rows and 8 environments/columns (the interaction is explained by two multiplicative terms);

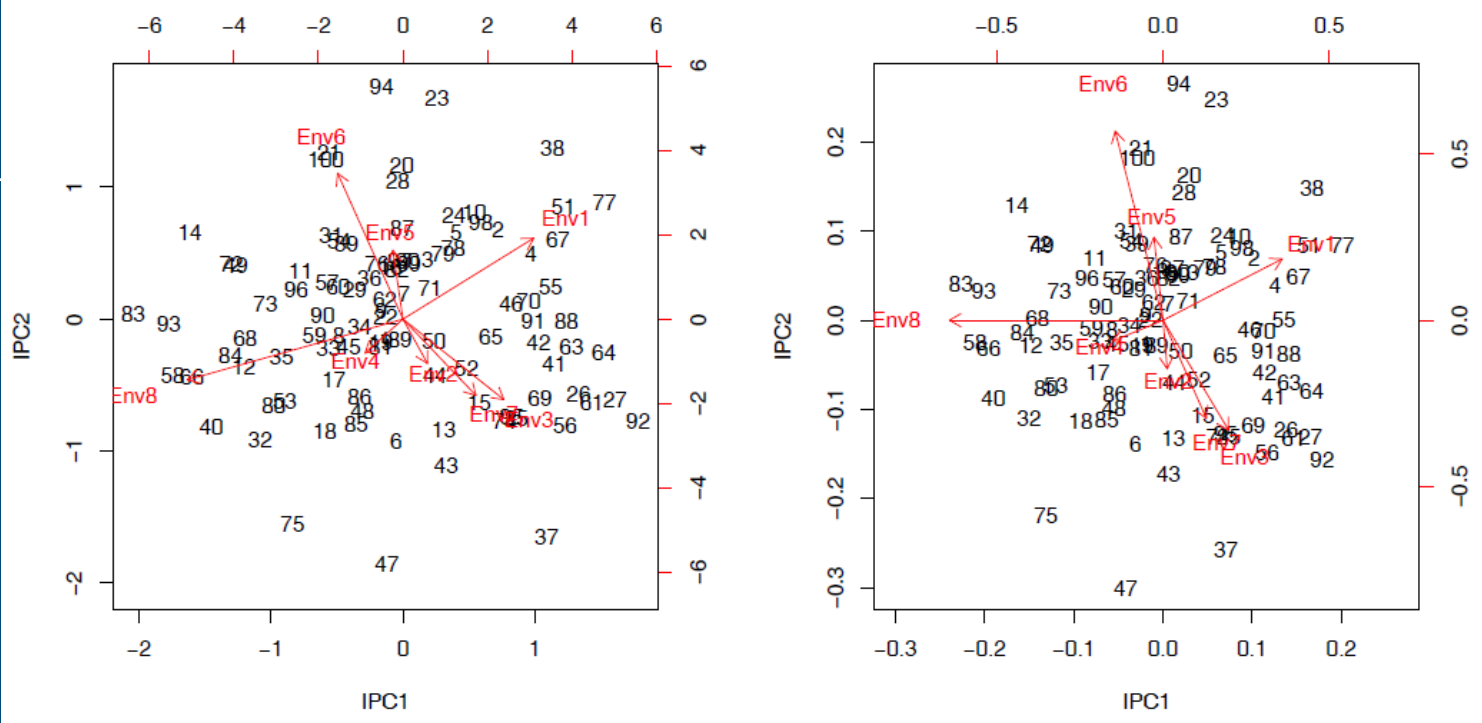
- In each run, the AMMI and robust AMMI models were used to fit/analyze the data. Biplots were obtained and the MSE was obtained for the singular values;

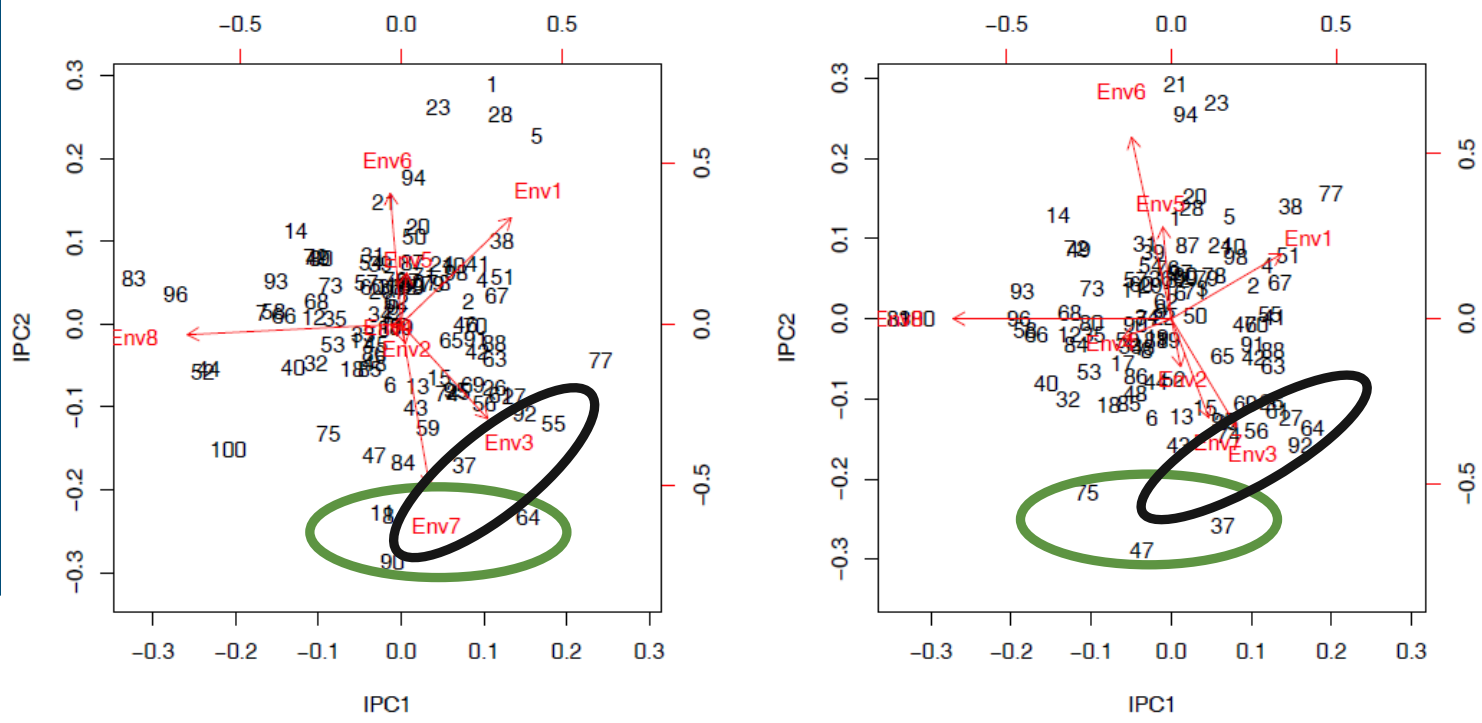
- Different contamination schemes (shift outliers and point-mass outliers) and several contamination rate were considered;

- Example (5% of shift-outliers):
  1. 5% positions are randomly selected in the two-way table thus assigning contamination positions in different environments for distinct genotypes;
  2. the 5% bad data is generated from a $N(\mu_j + k, \sigma^2{}_j)$ (pure shift outliers; $k = 4\sigma_j$ units) where μ_j and $\sigma^2{}_j$ are taken as the sample phenotypic mean and sample phenotypic variance according to the correspondent environment j, $j = 1, \ldots, 8$;
  3. the bad data replaces the 5% of the good data from the two-way table at the positions assigned in 1.

Outline of the presentation

1. Introduction and background
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Robust statistics
2. Materials and methods
   - Additive main effects and multiplicative interaction (AMMI) model
   - Robust AMMI model
   - Simulation study
3. Results
   - Simulation study
   - Real data example
4. Concluding Remarks





Data without contamination

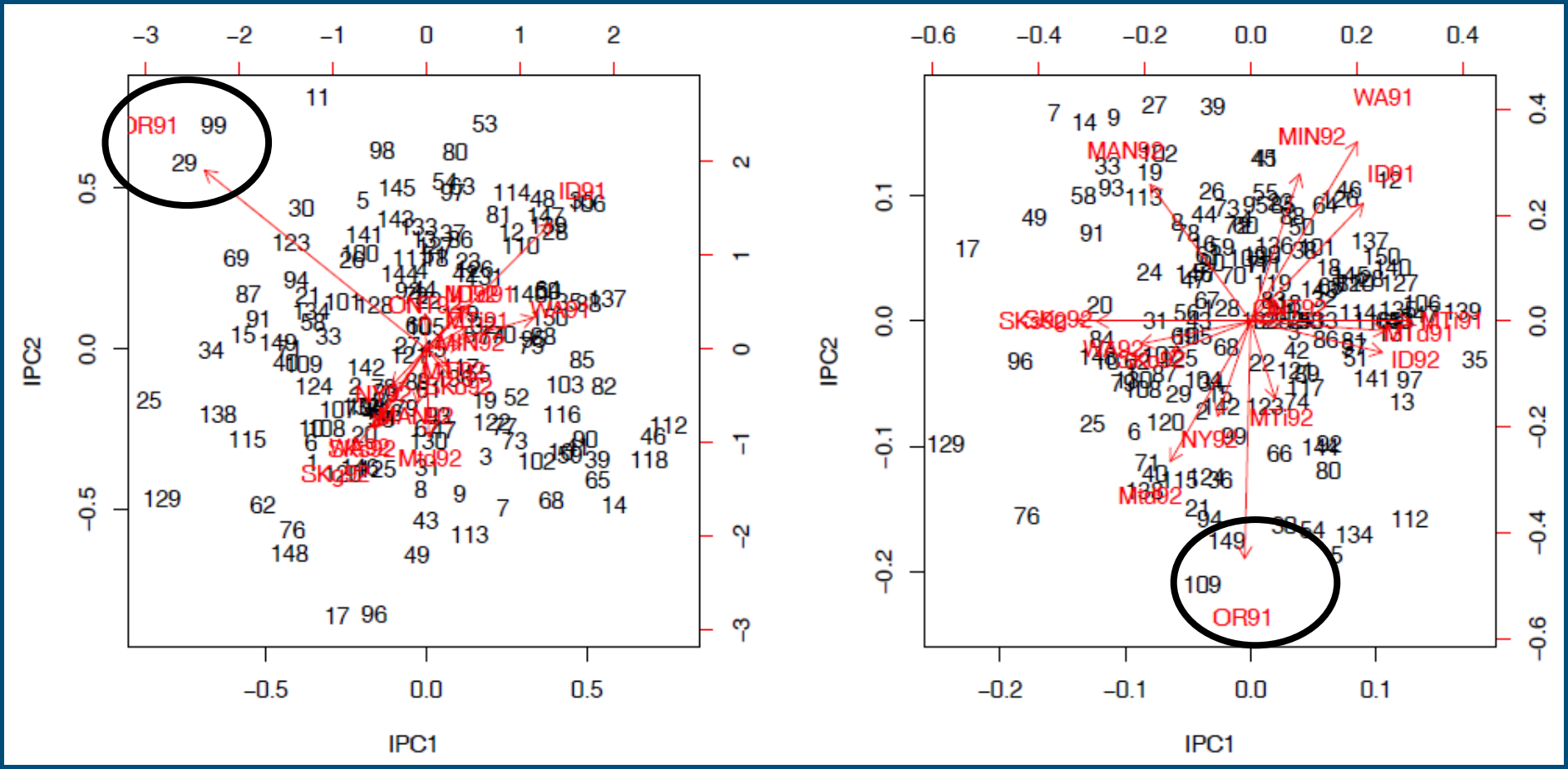
Data with 5% contamination

Results – Simulation study

$$MSE(\hat{\lambda}_j) = \frac{1}{1000} \sum_{l=1}^{1000} \left(\hat{\lambda}_j^{(l)} - \lambda_j \right)^2$$

Model	IPC1	IPC2
Robust AMMI	26.00	34.03
AMMI (5% contaminated data)	1159.22	1268.81
Robust AMMI (5% contaminated data)	186.57	275.33

Results – Real data example



Steptoe x Morex (SxM) barley mapping population: 150 genotypes and 15 environments

Outline of the presentation

1. Introduction and background
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Robust statistics
2. Materials and methods
   - Additive main effects and multiplicative interaction (AMMI) model
   - Robust AMMI model
   - Simulation study
3. Results
   - Simulation study
   - Real data example
4. Concluding Remarks

Concluding Remarks

- The use of the robust methodologies proposed, not only provided results similar to the classical ones when there was no contamination but also proved to provide better results when the data was in fact contaminated;

- Further simulation schemes are being studied in this AMMI model as well as other models widely used in quantitative genetics to model genotype by environment interaction;

- Challenges for statisticians:
  - To model low-dimensional structures while accounting for contaminated data and other model misspecifications;
  - To interact closely with other scientists in multidisciplinary teams, and to understand and to bridge the needs in different fields.

Thank you for your attention

Questions/Remarks/Suggestions?

(paulocanas@gmail.com)

References

- Gauch, H.G., Rodrigues, P.C., Munkvold, J.D., Heffner, E.L. and Sorrells, M. (2011). Two New Strategies for Detecting and Understanding QTL by Environment Interactions. Crop Science 51: 96–113.
- Rodrigues, P.C., Malosetti, M., Gauch, H.G. and van Eeuwijk, F.A. (2014). Weighted AMMI to study genotype-by-environment interaction and QTL-by-environment interaction. Crop Science 54: 1555–1570.

Founding

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

Projects:

- PTDC/MATSTA/ 0568/2012
- PEst-OE/MAT/UI0297/2014 (CMA)