A weighted AMMI algorithm to study genotype-by-environment interaction and QTL-by-environment interaction

Paulo Canas Rodrigues
Department of Statistics, Federal University of Bahia, Brazil
CMA-FCT/UNL, Nova University of Lisbon, Portugal
(joint work with Malosetti, M., Gauch, H.G. and van Eeuwijk, F.A.)

Outline of the presentation

1. Introduction and background
   - Examples of big data
   - Genotype by environment interaction (GxE)
   - Statistical models for GxE
   - Quantitative trait loci (QTL) by environment interaction (QEI)

2. Low-rank approximations
   - Principal component analysis
   - Singular value decomposition
   - Weighted low-rank approximations

3. Application
4. Other applications
5. Concluding Remarks

High dimensional data and its challenges

- Massive data sets – BIG data
  - rapid advances in technology (storage, computation, remote sensing, etc.) leads to enormous amounts of data being collected.
  - “Big data” is a data set whose size is beyond the ability of standard software to manage, process and analyze, within a adequate amount of time. “Big data” depends on the degree of complexity and the size of the data set.
  - Examples: financial data: stocks, currencies, derivatives, etc.; genetics, genomics, proteomics, biogeochimical and biological sciences; data visualization and image analysis; astronomy, high energy physics and atmospheric science; technological Era: web logs, radio-frequency identification, sensor networks, social networks, Internet search indexing, cell detail records, etc. and many more...

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!!

GxE as changing mean performance across environments
Strategies used to study the GxE and QTLxE

- Regression and SVD based techniques:
  - the Linear Regression model;
  - the GGE biplot;
  - the AMMI model.

- Heterogeneity of variance based technique:
  - the Linear Mixed Models framework.

Quantitative Trait Locus (QTL)

- The word locus (plural loci) is Latin for “place”. In genetics “Locus” is the position of a gene (or other significant sequence) on a chromosome.

- QTLs are stretches of DNA that are closely linked to the genes that underlie the trait in question.

- The QTLs that affect the variation of the phenotypic trait constitute its genetic architecture. It may tell us that plant height is controlled by many genes of small effect, or by a few genes of large effect.

Statistical models for GxE

\[ y_{ij} = \beta + G_i + E_j + G_i \cdot E_j \]

Where \( y_{ij} \) is the phenotype for the \( i \)-th genotype and \( j \)-th environment.

Statistical models for QTLxE

\[ Y_{ij} = \mu + \gamma_{ij} = \mu + G_i + E_j + (G_i \cdot E_j) + e_{ij} \]

Where \( Y_{ij} \) is the phenotype for the \( i \)-th genotype and \( j \)-th environment.

The data

Marker and Map data

Outline of the presentation

1. Introduction and background
   - Examples of leg data
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Quantitative trait loci (QTL) by environment interaction (QEI)

2. Low-rank approximations
   - Principal component analysis
   - Singular value decomposition
   - Weighted low-rank approximations

3. Application

4. Other applications

5. 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.

Principal Component Analysis

- When dealing with large data sets, one of the main problems is how to extract the information.
- Principal Component Analysis (PCA) is the most used technique for reducing the data set while preserving significant features.
- PCA transforms the n original (correlated) variables in n principal components (uncorrelated linear combinations of the original variables). One of the main features of PCs is that the first (few) explain a big amount of the original variance and can be used to “replace” the initial data set.

Singular Value Decomposition

We decompose the matrix $X : \text{rank}(X) = r$ as

$$X = U \Sigma V^T$$

where $U$ and $V$ have orthonormal columns and $\Sigma$ is diagonal: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$. Columns of $U$ and $V$ are the left and right singular vectors. The diagonal of $\Sigma$ contains the $r$ singular values.

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

$$X_k = \hat{X} = U_k \Sigma_k V_k^T$$

The approximation minimizes the error $||X - \hat{X}||_F$ (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 some observations, individuals (i.e. rows) or variables (i.e. columns), are known to have more importance to the analysis, is likely to lead to erroneous and non-significant results.

Alternative: Weighted Low-rank Approximations!

Why to use weighted low-rank approximations?

- External information about noise variance (error variance) for each measurement (e.g. statistical genetics and gene expression);
- Missing data (0/1 weights);
- Different number of samples;
- Different importance of different entries (columns, rows or cells).

Low-rank approximation

$$X = X_k + Z$$

$$\text{log } L(X_k, X) = \sum_{i=1}^{n} \log P(A_{ij} | X_k) = -\frac{1}{2} \sum_{i=1}^{n} (X_{ij} - X_{kj})^2 + \text{const}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \left( X_{ij} - \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right)^2$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} \left( X_{ij} - \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left( X_{ij} - \frac{1}{n} \sum_{i=1}^{n} X_{ij} \right)^2$$
Weighted low-rank approximation

- Given:
  - A target matrix $X \in \mathbb{R}^{m \times n}$;
  - A corresponding non-negative matrix $W \in \mathbb{R}^{m \times n}$;
  - A desired integer rank $k$;
- We would like to find a matrix $X_k \in \mathbb{R}^{m \times n}$ of rank (at most) $k$, that minimizes the weighted Frobenius distance $f(X_k) = \sum_{i,j} W_{ij} (X_{ij} - X_{kij})^2$ (Snirbo and Jaakkola, 2003).

$$X = X_k + Z$$

(target data) (rank k) (Gaussian noise)

How to do the minimization?

- Minimization of $f(X_k) = \sum_{i,j} W_{ij} (X_{ij} - X_{kij})^2$:
  - Use an Expectation-Maximization approach;
  - While the sum of squares of two consecutive iterates $X_k^{(t+1)}$ and $X_k^{(t)}$ is greater than a small value (e.g., $10^{-5}$);
  - $X_k^{(t+1)} = \text{SD} (W \odot X + (1 - W) \odot X_k^{(t)})$;
  - Where $W$ is the $m \times m$ matrix with weights, $0 \leq W_{ij} \leq 1$, $I$ is a $n \times n$ matrix with ones, $\odot$ is the Hadamard (or entrywise) product of matrices, and $t$ is the iteration number (Snirbo and Jaakkola, 2003);
  - $X_k$ should be initialized at $X_k^{(0)} = X$ or $X_k^{(0)} = 0$;
  - The outputs of this EM algorithm are the matrices $U_D$, $D_D$, and $V_D$ such that $X_k = U_D D_D V_D^T$, being $k$ the rank of approximation.

Outline of the presentation

1. Introduction and background
   - Examples of big data
   - Genotype by environment interaction (GxE)
   - Statistical models for GxE
   - Quantitative trait loci (QTL) by environment interaction (QEI)
2. Low-rank approximations
   - Principal component analysis
   - Singular value decomposition
   - Weighted low-rank approximations
3. Application
4. Other applications
5. Concluding Remarks

WAMMI model

- **AMMI model:** $y = 1_i \ 1_j^T \mu + \alpha_j^T + \beta_j^T + U^T D V + \varepsilon$
  - It gives good results when the error variance is constant across environments. However it does not account for heterogeneity of error variance across environments!
- **WAMMI model:** replace the SVD in AMMI by the weighted low-rank SVD. It is based on the expectation-maximization algorithm
  $$X^{(t+1)} = \text{SD} (W \odot Y + (1 - W) \odot X^{(t)}), \quad X^{(0)} = U^0 D^0 V^0$$
  - After convergence, the matrices $U_D$, $D_D$, and $V_D$ such that $Y = U_D D_D V_D^T$ are obtained.

AMMI2 (left) and WAMMI2 (right) biplots

WAQ analysis for QTL detection - 100 simulations

Summary of the total number of detected QTL in 100 simulations for the WAMMI model.
Outline of the presentation

1. Introduction and background
   - Examples of big data
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Quantitative trait loci (QTL) by environment interaction (QEI)

2. Low-rank approximations
   - Principal component analysis
   - Singular value decomposition
   - Weighted low-rank approximations

3. Application

4. Other applications

5. Concluding Remarks

Other applications in statistical genetics

- Analysis and interpretation when there is a clear NMAR pattern of missing values in the data matrix:

- Robust methods in statistical genetics:
  - Standard PCA fits a low-rank matrix to a data set
  - A small amount of data corruption can have a large influence.

Outline of the presentation

1. Introduction and background
   - Examples of big data
   - Genotype by environment interaction (GEI)
   - Statistical models for GEI
   - Quantitative trait loci (QTL) by environment interaction (QEI)

2. Low-rank approximations
   - Principal component analysis
   - Singular value decomposition
   - Weighted low-rank approximations

3. Application

4. Other applications in statistical genetics

5. Concluding Remarks
Concluding Remarks

- Low-rank approximation and weighted low-rank approximations represent a big challenge and opportunity to deal with big data sets.

- Weighted low-rank approximations can be used in two-way data matrices where some observations, individuals (i.e. rows) or variables (i.e. columns), are known to have more importance to the final model.

- Weighted low-rank approximations can easily deal with missing values and provide a powerful tool to deal with complex data matrices.

- Challenges for statisticians:
  - To model low-dimensional structures;
  - Deal with high-dimensional scaling;
  - Interact closely with other scientists in multidisciplinary teams.

Thank you for your attention

Questions/Remarks/Suggestions?
(paulocanas@gmail.com)

Some references

