Data Analysis (Factorial Analyses and Environmental Genomics)

Principal Components on Instrumental Variables Analysis / Redundancy Analysis

Denis Laloë
Populations, Statistique et Génome
GABI, INRA

25 May 2016
Factorial analyses *a la francaise* : an empirical approach

- *The model must follow the data, not the reverse, J P Benzécri*
- Observation vs Experimentation
  - Pre-existing (Social sciences / Ecology)
Data: Emphasis on graphical representation

- Graphs are essential to good statistical analysis, F J Anscombe

Geometrical approach: Data $\longrightarrow$ cloud of points
## A data table

A data table typically represents a matrix of observations. Each row corresponds to an observation and each column to a variable. The table is usually denoted as $n \times p$, where $n$ is the number of observations and $p$ is the number of variables. The entries in the table, denoted as $x_{i,j}$, represent the value of the $j$th variable for the $i$th observation.}

```
<table>
<thead>
<tr>
<th>1</th>
<th>j</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

**Variables**: $p$ variables

**Observations**: $n$ observations

**Table Entry**: $x_{i,j}$

In the context of genomics, this table could represent a dataset where rows are samples and columns are gene expressions. In environmental genomics, it might represent environmental factors affecting gene expression.
Two geometric representations

- Observations: cloud of $n$ points in a $p$-dimensional space
- Variables: cloud of $p$ points in a $n$-dimensional space
The duality diagram

*Dray and Dufour, 2007*

- $X$ (resp. $X^t$) switches from a cloud to another
- $M$ diagonal matrix of weights of observations
- $Q$ diagonal matrix of weights of variables

\[
\begin{array}{c}
\begin{tikzpicture}
\node (A) at (0,0) {$< p >$};
\node (B) at (2,0) {$< p >$};
\node (C) at (0,-2) {$< n >$};
\node (D) at (2,-2) {$< n >$};
\node (E) at (1,-1) {$\mathbf{Q}$};
\node (F) at (1,-3) {$\mathbf{M}$};
\node (G) at (1,1) {$\mathbf{X}$};
\node (H) at (-1,1) {$\mathbf{X}^t$};
\draw[->] (A) -- (B);\draw[->] (A) -- (C);\draw[->] (B) -- (D);\draw[->] (C) -- (D);
\end{tikzpicture}
\end{array}
\]
Data transformation

- to condense data into some representative features
- Internal Criteria: Inertia
- finding directions with the maximal projected inertia
- Canonical decomposition of
  - Observations: $XQX^tM$
  - Variables: $X^tMXQ$

\[ <p> \xrightarrow{Q} <p> \]
\[ X^t \]
\[ <n> \xleftarrow{M} <n> \]
The duality diagram for a PCA

- $X$ a data matrix of centered, and possibly normed $p$ variables measured on $n$ observations
  - normed variables: normed PCA (PCA on correlations)
  - non-normed variables: non-normed PCA (PCA on covariances)

- $M$ diagonal matrix $\frac{I_n}{n}$
  - weights of observations
  - Metric (distance) on variables

- $Q$ diagonal matrix of weights of variables $I_p$
  - weights of variables
  - Metric (distance) on observations
The duality diagram for a PCA

Maximisation of the correlation between variables and components

Variables

\[ V = X'X/n \]

VA = AΔ

A'A = I

Principal axes

Variable scores

C = X'B

Maximisation of the dispersion of individuals

Observations

\[ W = XX'/n \]

Diagonalisation

\[ X'X \quad XX' \]

same positive eigenvalues

\[ \lambda_1 > \lambda_2 > ... > \lambda_r > 0 \]

\[ \Lambda = \text{diag}(\lambda_1, ..., \lambda_r) \]

Transition formulae

\[ XA^{-0.5} = B \]

\[ X'B\Delta^{-0.5} = A \]

Singular value decomposition

Best approximation (rank l)

Eckart and Young

\[ \hat{X}_i = \sum_{i=1,l} \sqrt{\lambda_i} b_i a_i \]
In short

- Two points of view (Variable vs Observation)
- Maximisation of a criteria : Inertia
  - Observations : maximal variance/dispersion
  - Variables : maximal correlation with axes
- Same computations: a canonical decomposition
  - $I = \sum_{i=1}^{r} \lambda_i$
- Transition formulae between space of variables and space of observations
- Singular Value Decomposition
  - Reconstitution of the data matrix
Many softwares

- **R**
  - `princomp, pcadapt,...`
  - FactoMineR, vegan, **ade4**

- **non R**
  - SAS, xlstat,...
  - smartpca
The R package ade4 is the most complete software for exploratory data methods displayed in the duality diagram scheme. Dray, S. et Dufour, A-B. (2007)

- dudi.pca: Principal Component Analysis
- dudi.coa: Correspondence Analysis
- dudi.acm: Multiple Correspondence Analysis
- dudi.fca: Fuzzy Correspondence Analysis
- dudi.mix: mixed analysis (numeric and factors)
- dudi.nsc: Non Symetric Correspondence Analysis
The R package ade4 is the most complete software for exploratory data methods displayed in the duality diagram scheme.

<table>
<thead>
<tr>
<th>dudi Object</th>
<th>Meaning</th>
<th>Duality Diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>tab</td>
<td>Transformed data table</td>
<td>X (centered/normed)</td>
</tr>
<tr>
<td>cw</td>
<td>Column weights</td>
<td>Q</td>
</tr>
<tr>
<td>lw</td>
<td>Row weights</td>
<td>M</td>
</tr>
<tr>
<td>eig</td>
<td>Non-null eigenvalues</td>
<td>(\Delta)</td>
</tr>
<tr>
<td>rank</td>
<td>number of non-null eigenvalues</td>
<td>(n - 1) ((n &lt; p))</td>
</tr>
<tr>
<td>c1</td>
<td>Variable loadings</td>
<td>A</td>
</tr>
<tr>
<td>l1</td>
<td>Observation loadings</td>
<td>B</td>
</tr>
<tr>
<td>co</td>
<td>Variable scores</td>
<td>(C = X^tB)</td>
</tr>
<tr>
<td>li</td>
<td>Observation scores</td>
<td>(L = XA)</td>
</tr>
</tbody>
</table>
PCA with ade4: dudi.pca

Some graphics

- screeplot (eigenvalues barplot)
- Correlation circle (Variable coordinates: $C = X^tB$)
- Scatterplot (Observations)
## SNP data: the raw data table

### Individuals
- $n$ Individuals
- Number of copies of an allelic form: 0, 1, 2
- Allelic frequency at an individual level: 0, 0.5, 1

### Bi-allelic SNPs

$$
\mathbf{X} = \begin{bmatrix}
SNP_1 & SNP_2 & \ldots & SNP_p \\
Ind_1 & 0.5 & 1 & \ldots & 0.5 \\
Ind_2 & 0.5 & 1 & \ldots & 0.5 \\
& \dotsc & \dotsc & \dotsc & \dotsc \\
Ind_n & 0.5 & 1 & \ldots & 1 
\end{bmatrix}
$$
Transformation of the data

\[
\mathbf{X} = \begin{bmatrix}
    SNP_1 & SNP_2 & \ldots & SNP_p \\
    Ind_1 & 0.5 & 1 & \ldots & 0.5 \\
    Ind_2 & 0.5 & 1 & \ldots & 0.5 \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    Ind_n & 0.5 & 1 & \ldots & 1
\end{bmatrix}
\]

- Centering by column
- Normalization by \( \sqrt{f_j(1-f_j)} \), where \( f_j \) is the allelic frequency of \( SNP_j \)
  - Links between inertia and \( F_{st} \)
An example: 11 French local freeds

- 30 animals by breed
- 770 K SNP chip
The PCA on individuals

Individuals: the factorial map
The PCA on individuals

SNPs: a Manhattan Plot

**SNPs**
- Squared coordinates (correlation) of SNPs:
- Contributions to inertia
- Typological Values, Fst
- May be summed over axes
How to take into account environmental data into such analyses

Phenomenon
Set of variables
Genomics

a priori
External
Set of variables
Space
Geography
Breeds
Environment

Linked to
Explained by
Indirect / A posteriori methods

- Factor: Factor map with classes of points \((s.\text{class})\).
- Quantitative variables
  - Supplementary variables (Projection of variables on the factorial map)
  - Correlation of supplementary variables with axes
An example: 11 french local freeds

- 30 animals by breed
- 770 K SNP chip
- Terrain
- Mean Radiation

30 animals by breed
770 K SNP chip
Terrain
Mean Radiation
Bioclimatic variables

Bioclimatic variables

Climond data https://www.climond.org

- Bioclim gridded data layers at 10’ and 30’ for recent historical (‘current’) climate
- Temperature, precipitation, moisture, radiation
- annual and seasonal mean values, seasonality,

<table>
<thead>
<tr>
<th>Bio</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bio1</td>
<td>Annual mean temperature (°C)</td>
</tr>
<tr>
<td>Bio4</td>
<td>Temperature seasonality (C of V)</td>
</tr>
<tr>
<td>Bio5</td>
<td>Max temperature of warmest week (°C)</td>
</tr>
<tr>
<td>Bio12</td>
<td>Annual precipitation (mm)</td>
</tr>
<tr>
<td>Bio14</td>
<td>Precipitation of driest week (mm)</td>
</tr>
<tr>
<td>Bio15</td>
<td>Precipitation seasonality (C of V)</td>
</tr>
<tr>
<td>Bio20</td>
<td>Annual mean radiation (W m(^{-2}))</td>
</tr>
<tr>
<td>Bio21</td>
<td>Highest weekly radiation (W m(^{-2}))</td>
</tr>
<tr>
<td>Bio23</td>
<td>Radiation seasonality (C of V)</td>
</tr>
<tr>
<td>Bio28</td>
<td>Annual mean moisture index</td>
</tr>
<tr>
<td>Bio29</td>
<td>Highest weekly moisture index</td>
</tr>
<tr>
<td>Bio31</td>
<td>Moisture index seasonality (C of V)</td>
</tr>
<tr>
<td>Bio35</td>
<td>Mean moisture index of coldest quarter</td>
</tr>
</tbody>
</table>
The PCA on individuals

Stratification according to breeds with the function `s.class`

- A clear stratification according to breeds
- Why not directly account for this factor
Direct methods / A priori methods

- Symmetrical methods.
  - Comparison of structures: (multiple) Co-Inertia Analysis / Multiple Factor Analysis

- Asymmetrical methods
  - Modelling by instrumental variables (Redundancy Analysis)
  - Model $X = Y + e$
A simple modelling. The breed

Model

\[ \delta_{ik}^j = \text{Breed}_{ik}^j + \epsilon_{ik}^j \]
\[ = f_k^i + \epsilon_{ik}^j \]
\[ X = F + E \]

PCA

\[ \text{ACP}(X) = \text{ACP}(F) + \text{ACP}(E) \]

ACP = ACP between breeds + PCA within breed

Genetic interpretation

Chessel et Laloë, 2001; Laloë et Gautier, 2011

\[ 1 = \sum_F (c_j^2) + \sum_E (c_j^2) \]
\[ 1 = F_{st} + [1 - F_{st}] \]
The PCA between breeds

Instrumental variable: breed
The PCA between breeds

- Inertia between breeds: 97% of the inertia of the 10 first axes of the PCA
- Correlations between markers scores in both analyses: $\geq 0.98$
**General Modelling. The PCA on Instrumental Variables**

**Model: Breed**

\[
\delta_{ik}^{j} = \text{Breed}_{ik}^{j} + \epsilon_{ik}^{j}
\]
\[
= f_{k}^{j} + \epsilon_{ik}^{j}
\]
\[
X = F + E
\]

**PCA**

\[
\text{PCA}(X) = \text{PCA}(F) + \text{PCA}(E)
\]
PCA = PCA between breeds + PCA within breed

**Genetic interpretation**

\[
1 = \sum_{F} (c_{j}^{2}) + \sum_{E} (c_{j}^{2})
\]
\[
1 = F_{st} + [1 - F_{st}]
\]

**General Model: Instrumental Variables**

\[
\delta_{ik}^{j} = \hat{\delta}_{ik}^{j} + \epsilon_{ik}^{j}
\]
\[
X = \hat{X} + E
\]

**PCA**

\[
\text{PCA}(X) = \text{PCA}(\hat{X}) + \text{PCA}(E)
\]
PCA = PCAIV + orthogonal PCAIV

**Genetic interpretation**

\[
1 = \sum_{\hat{X}} (c_{j}^{2}) + \sum_{E} (c_{j}^{2})
\]
\[
1 = F_{st} + [1 - F_{st}]
\]
PCA on Instrumental Variables

**PCAIV of** \( \mathbf{X}, \mathbf{W} \):

- Predicted \( \mathbf{X} \) by \( \mathbf{W} \): \( \hat{\mathbf{X}} = (\mathbf{W}^t \mathbf{W})^{-1} \mathbf{W}^t \mathbf{X} \)
- PCA of: \( \hat{\mathbf{X}} \)
- Comparison between PCAIV and PCA
PCA on Instrumental Variables

Geography (Latitude + Longitude)

Between Breeds analysis

Geography
PCA on Instrumental Variables  Geography (Latitude + Longitude)

To visualize geography on genetic diversity with a colorplot (adegenet package)

Up to three dimensions

Each dot component is represented as intensity of a given color channel. The first PC is shown in red, the second PC in green, and the third PC in blue.
To visualize geography on genetic diversity with a bubble plot (ade4/adegraphics package)

One dimension
PCA on Instrumental Variables  Geography (Latitude + Longitude)

<table>
<thead>
<tr>
<th>Inertia</th>
<th>Cum. inertia</th>
<th>Constr. inertia</th>
<th>Cum. Constr. inertia</th>
<th>ratio</th>
<th>$R^2$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9467</td>
<td>9467</td>
<td>8893</td>
<td>8893</td>
<td>0.939</td>
<td>0.942</td>
<td>8377</td>
</tr>
<tr>
<td>8493</td>
<td>17960</td>
<td>7695</td>
<td>16587</td>
<td>0.924</td>
<td>0.945</td>
<td>7271</td>
</tr>
</tbody>
</table>
## PCA on Instrumental Variables

### Geography (Latitude + Longitude)

**Variability**

<table>
<thead>
<tr>
<th>Inertia</th>
<th>Cum. inertia</th>
<th>Constr. inertia</th>
<th>Cum. Constr. inertia</th>
<th>ratio</th>
<th>$R^2$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9467</td>
<td>9467</td>
<td>8893</td>
<td>8893</td>
<td><strong>0.939</strong></td>
<td>0.942</td>
<td>8377</td>
</tr>
<tr>
<td>8493</td>
<td>17960</td>
<td>7695</td>
<td>16587</td>
<td><strong>0.924</strong></td>
<td>0.945</td>
<td>7271</td>
</tr>
</tbody>
</table>

**Predictability**

<table>
<thead>
<tr>
<th>Inertia</th>
<th>Cum. inertia</th>
<th>Constr. inertia</th>
<th>Cum. Constr. inertia</th>
<th>ratio</th>
<th>$R^2$</th>
<th>$\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9467</td>
<td>9467</td>
<td>8893</td>
<td>8893</td>
<td>0.939</td>
<td><strong>0.942</strong></td>
<td>8377</td>
</tr>
<tr>
<td>8493</td>
<td>17960</td>
<td>7695</td>
<td>16587</td>
<td>0.924</td>
<td><strong>0.945</strong></td>
<td>7271</td>
</tr>
</tbody>
</table>
PCA on Instrumental Variables  Partial analyses

- Joint effect of the geography (Noise) and the environment (Interest)
- Partial analysis
- Geography
- Environment | Geography

Response variables
Genetic data

Explanatory (instrumental) variables

\[ X, \quad W, \quad W_g, \quad W_c \]
Sequence of analyses

1. **PCA on individuals**
   - between breeds analysis
     - Analysis IV = Geography
   - within breeds analysis
     - Orthogonal Analysis IV = Geography
   - Analysis Environment
   - Orthogonal Analysis Environment
# Decomposition of inertia

<table>
<thead>
<tr>
<th>Factor</th>
<th>Fst (%)</th>
<th>% of between-breed inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breed</td>
<td>8.4</td>
<td>100</td>
</tr>
<tr>
<td>Geo (Lat.+Lon.)</td>
<td>2.4</td>
<td>29</td>
</tr>
<tr>
<td>Radiation+Terrain / Geo</td>
<td>1.0</td>
<td>12</td>
</tr>
</tbody>
</table>

The output shows the decomposition of inertia, with Breed having the highest contribution at 8.4% of between-breed inertia, followed by Geo (Lat.+Lon.) at 2.4% and Radiation+Terrain / Geo at 1.0%.
Manhattan plot
PCA on Instrumental Variables

In short

- PCA on predicted values
- An other optimisation criteria : Variability * Predictability
- Contribution of instrumental variables to axes
- Permutation tests
- Partition of inertia according to instrumental variables
  - Loss in variability
  - Gain in Predictability/Interpretability
- Contribution of SNPs to axes.
Another package of interest: vegan (J Oksanen et al)
A warning

- Avoid the overparametrization
References / Packages R


- Laloë, D. and M. Gautier (2011). On the genetic interpretation of between-group PCA on SNP data. HAL hal-00661214


- vegan [http://cran.r-project.org/web/packages/vegan/](http://cran.r-project.org/web/packages/vegan/)