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

General Points

Factorial analyses *a la francaise* : an empirical approach

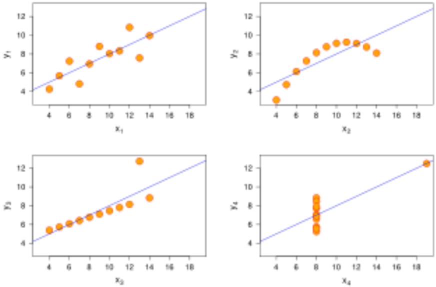
- *The model must follow the data, not the reverse, J P Benzécri*
- Observation *vs* Experimentation
  - Pre-existing (Social sciences / Ecology)

Graphics

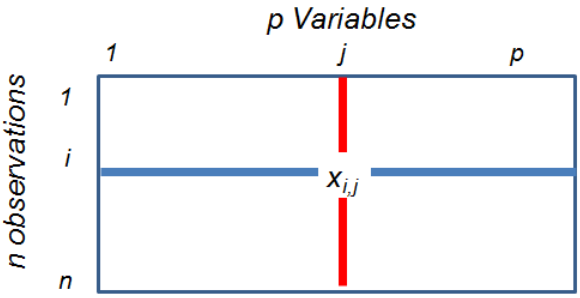
Data : Emphasis on graphical representation

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

Geometrical approach : Data $\longmapsto$ cloud of points

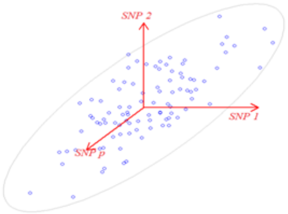


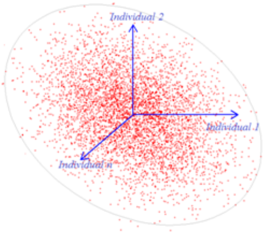
A data table



Two geometric representations

- Observations : cloud of *n* points in a p-dimensional space
- Variables : cloud of *p* points in a n-dimensional space

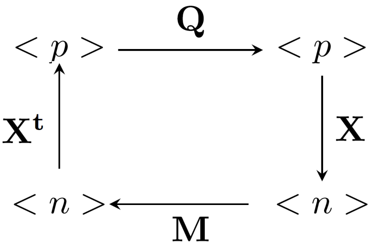




The duality diagram

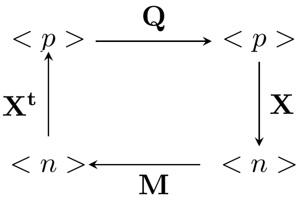
Dray and Dufour, 2007

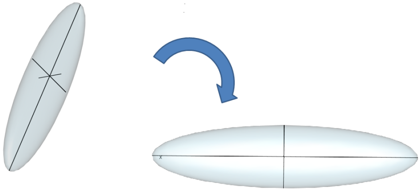
- **X** (resp. $\mathbf{X^t}$) switches from a cloud to another
- **M** diagonal matrix of weights of observations
- **Q** diagonal matrix of weights of variables



Data transformation

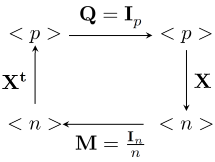
- to condense data into some representative features
- Internal Criteria : Inertia
- finding directions with the maximal projected inertia
- Canonical decomposition of
  - Observations : $\mathbf{XQX^tM}$
  - Variables : $\mathbf{X^tMXQ}$





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 $\frac{\mathbf{I_n}}{n}$
  - weights of observations
  - *Metric (distance) on variables*
- **Q** diagonal matrix of weights of variables $\mathbf{I_p}$
  - *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*

$$
\begin{array}{ccc}
<p> & \xrightarrow{\mathbf{Q}=\mathbf{I}_p} & <p> \\
\uparrow \mathbf{X}^t & & \downarrow \mathbf{X} \\
<n> & \xleftarrow[\mathbf{M}=\frac{\mathbf{I}_n}{n}]{} & <n>
\end{array}
$$

Maximisation of the dispersion of individuals

Observations

W=XX' /*n*

Diagonalisation

X'X **XX'**

same positive eigenvalues

$\lambda 1 > \lambda 2 > ... > \lambda r > 0$

$\Lambda = \mathrm{diag}(\lambda 1, \ldots, \lambda r)$

VA=AΔ
A'A=I
Principal axes

WB=BΔ
B'B=I
Principal components

Variable scores
C=X'B

Transition formulae

$XA\Delta^{-0,5} = B$

$X'B\Delta^{-0,5} = A$

Observation scores
L=XA

Singular value decomposition

Best approximation (rank *l*)
Eckart and Young

$$\hat{\mathbf{X}}_l = \sum_{i=1,l} \sqrt{\lambda_i}\,\mathbf{b}_i \mathbf{a}_i'$$

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

R package ade4

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

PCA with ade4: dudi.pca

The R package ade4 is the most complete software for exploratory data methods displayed in the duality diagram scheme.

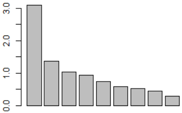
dudi

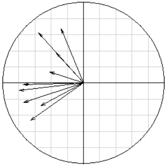
Object	Meaning	Duality Diagram
tab	Transformed data table	$\mathbf{X}$ (centered/normed)
cw	Column weights	$\mathbf{Q}$
lw	Row weights	$\mathbf{M}$
eig	Non-null eigenvalues	$\mathbf{\Delta}$
rank	number of non-null eigenvalues	$n-1$ $(n<p)$
c1	Variable loadings	$\mathbf{A}$
l1	Observation loadings	$\mathbf{B}$
co	Variable scores	$\mathbf{C}=\mathbf{X}^t\mathbf{B}$
li	Observation scores	$\mathbf{L}=\mathbf{XA}$

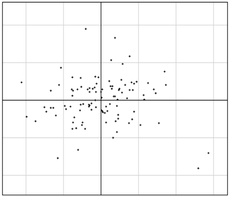
PCA with ade4: dudi.pca

Some graphics

- screeplot (eigenvalues barplot)
- Correlation circle (Variable coordinates: $\mathbf{C} = \mathbf{X}^t\mathbf{B}$)
- 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

$$\mathbf{X} = \begin{bmatrix} & SNP_1 & SNP_2 & ... & SNP_p \\ Ind_1 & 0.5 & 1 & ... & 0.5 \\ Ind_2 & 0.5 & 1 & ... & 0.5 \\ & ... & ... & ... & ... \\ Ind_n & 0.5 & 1 & ... & 1 \end{bmatrix}$$

Transformation of the data

$$\mathbf{X} = \begin{bmatrix} & SNP_1 & SNP_2 & \dots & SNP_p \\ Ind_1 & 0.5 & 1 & \dots & 0.5 \\ Ind_2 & 0.5 & 1 & \dots & 0.5 \\ & \dots & \dots & \dots & \dots \\ Ind_n & 0.5 & 1 & \dots & 1 \end{bmatrix}$$

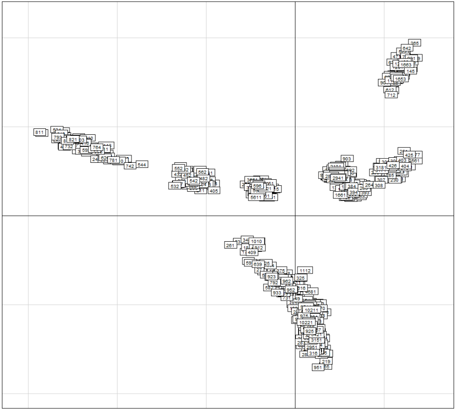
- Centering by column
- Normalization by $\sqrt{f_j(1-f_j)}$, where f_j is the allelic frequency of SNP_j
  - Links between inertia and F_{st}

An example : 11 french local freeds

- 30 animals by breed
- 770 K SNP chip

The PCA on individuals

Individuals : the factorial map

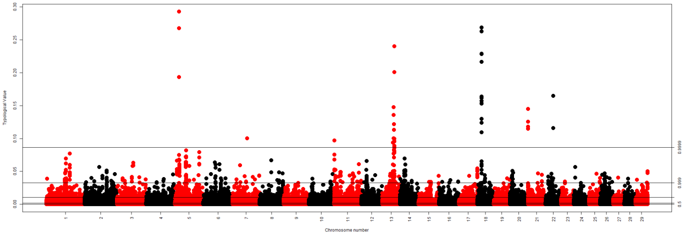


The PCA on individuals

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

$\longleftrightarrow$

Linked to
Explained by

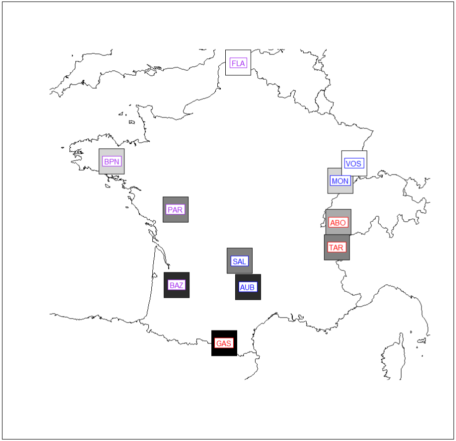
a priori
External

Set of variables
Space
Geography
Breeds
Environment

Indirect / A posteriori methods

- Factor: Factor map with classes of points (*s.class*) .
- Quantitative variables
  - Supplementary variables (Projection of variables on the factorial map)
  - Correlation of supplementary variables with axes

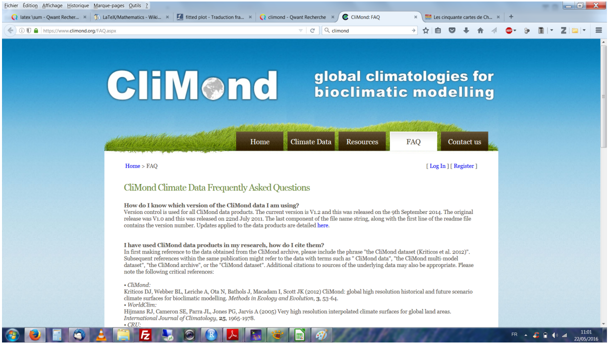
An example : 11 french local freeds



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

Bioclimatic variables

Climond data https://www.climond.org *Kriticos et al, 2012*



Bioclimatic variables

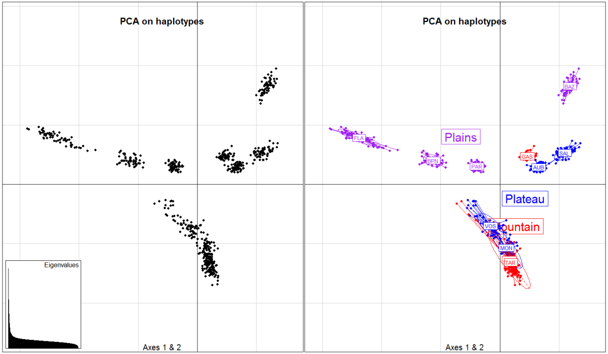
Climond data https://www.climond.org

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

Bio01	Annual mean temperature (°C)
Bio04	Temperature seasonality (C of V)
Bio05	Max temperature of warmest week (°C)
Bio12	Annual precipitation (mm)
Bio14	Precipitation of driest week (mm)
Bio15	Precipitation seasonality (C of V)
Bio20	Annual mean radiation (W m-2)
Bio21	Highest weekly radiation (W m-2)
Bio23	Radiation seasonality (C of V)
Bio28	Annual mean moisture index
Bio29	Highest weekly moisture index
Bio31	Moisture index seasonality (C of V)
Bio35	Mean moisture index of coldest quarter

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

- Symetrical methods.
  - Comparison of structures : (multiple) Co-Inertia Analysis / Multiple Factor Analysis
- Asymetrical methods
  - Modelling by instrumental variables (Redundancy Analysis)
  - Model $\mathbf{X=Y+e}$

A simple modelling. The breed

Model

$$\delta_{ik}^{j} = Breed_{ik}^{j} + \epsilon_{ik}^{j}$$
$$= f_{k}^{j} + \epsilon_{ik}^{j}$$
$$\mathbf{X} = \mathbf{F} + \mathbf{E}$$

PCA

$$\mathrm{ACP(X)} = \mathrm{ACP(F)} + \mathrm{ACP(E)}$$

ACP = ACP between breeds + PCA within breed

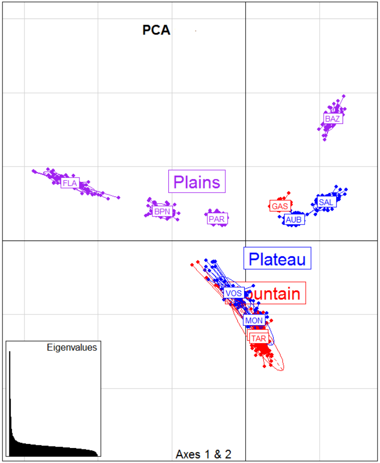
Genetic interpretation

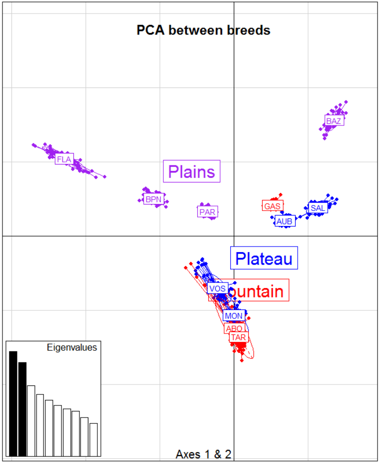
Chessel et Laloë, 2001; Laloë et Gautier,2011

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

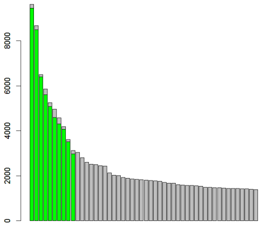
The PCA between breeds

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 : ≥ 0.98

General Modelling. The PCA on Instrumental Variables

Model : Breed

$$\begin{aligned}\delta^j_{ik} &= Breed^j_{ik} + \epsilon^j_{ik}\\ &= f^j_k + \epsilon^j_{ik}\\ \mathbf{X} &= \mathbf{F} + \mathbf{E}\end{aligned}$$

PCA

$$\mathrm{PCA(X)} = \mathrm{PCA(F)} + \mathrm{PCA(E)}$$

PCA = PCA between breeds + PCA within breed

Genetic interpretation

$$\begin{aligned}1 &= \sum_F (c_j^2) + \sum_E (c_j^2)\\ 1 &= F_{st} + [1 - F_{st}]\end{aligned}$$

General Model : Instrumental Variables

$$\begin{aligned}\delta^j_{ik} &= \hat{\delta}^j_i + \epsilon^j_i\\ \mathbf{X} &= \hat{\mathbf{X}} + \mathbf{E}\end{aligned}$$

PCA

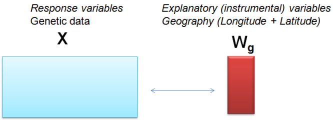
$$\mathrm{PCA(X)} = \mathrm{PCA(\hat{X})} + \mathrm{PCA(E)}$$

PCA = PCAIV + orthogonal PCAIV

Genetic interpretation

$$\begin{aligned}1 &= \sum_{\hat{X}} (c_j^2) + \sum_E (c_j^2)\\ 1 &= F_{st} + [1 - F_{st}]\end{aligned}$$

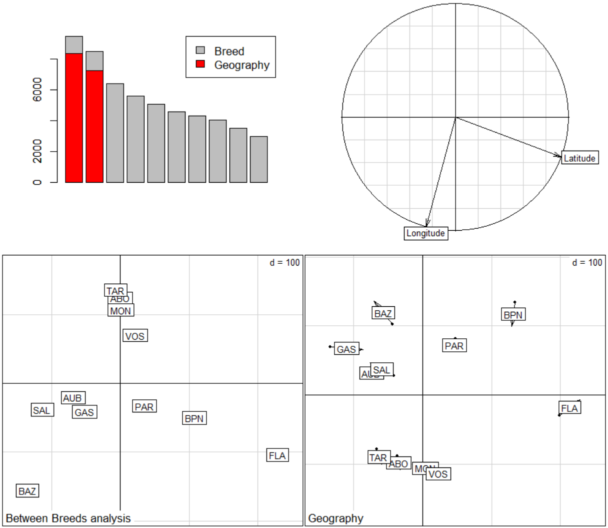
PCA on Instrumental Variables Geography



PCAIV of **X,W**:

- Predicted**X** by **W** : $\hat{\mathbf{X}} = (\mathbf{W}^t\mathbf{W})^{-}\mathbf{W}^t\mathbf{X}$
- PCA of : $\hat{\mathbf{X}}$
- Comparison between PCAIV and PCA

PCA on Instrumental Variables Geography (Latitude + Longitude)

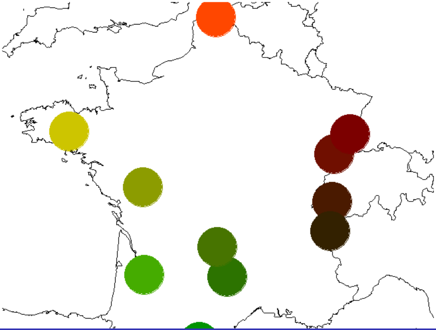


PCA on Instrumental Variables Geography (Latitude + Longitude)

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

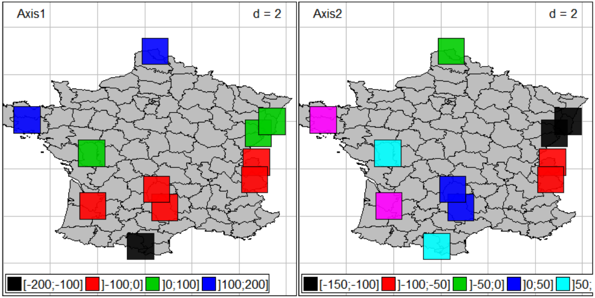
Up to three dimensions

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



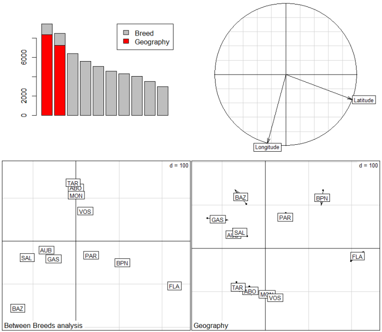
PCA on Instrumental Variables Geography (Latitude + Longitude)

To visualize geography on genetic diversity with a bubble plot (ade4/adegraphics package)
One dimension



PCA on Instrumental Variables Geography (Latitude + Longitude)

Inertia	Cum. inertia	Constr. inertia	Cum. Constr. inertia	ratio	R^2	λ
9467	9467	8893	8893	0.939	0.942	8377
8493	17960	7695	16587	0.924	0.945	7271



PCA on Instrumental Variables Geography (Latitude + Longitude)

Variability

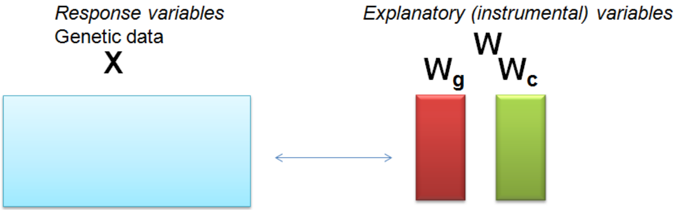
Inertia	Cum. inertia	Constr. inertia	Cum. Constr. inertia	**ratio**	R^2	λ
9467	9467	8893	8893	**0.939**	0.942	8377
8493	17960	7695	16587	**0.924**	0.945	7271

Predictability

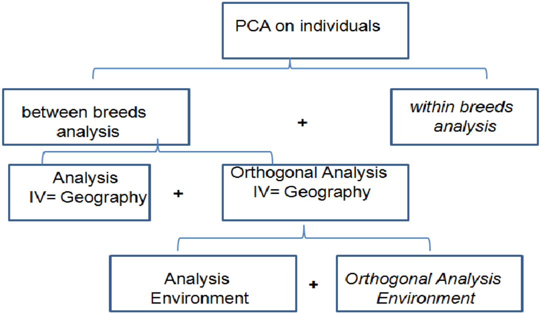
Inertia	Cum. inertia	Constr. inertia	Cum. Constr. inertia	ratio	$\mathbf{R^2}$	λ
9467	9467	8893	8893	0.939	**0.942**	8377
8493	17960	7695	16587	0.924	**0.945**	7271

PCA on Instrumental Variables Partial analyses

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



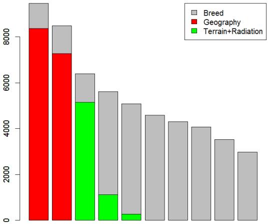
Sequence of analyses



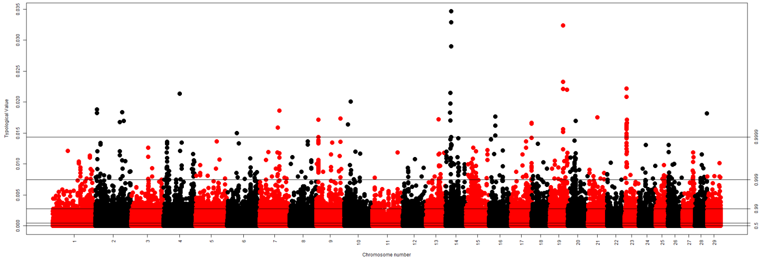
Decomposition of inertia

Output

Factor	Fst (%)	% of between-breed inertia
Breed	8.4	100
Geo (Lat.+Lon.)	2.4	29
Radiation+Terrain / Geo	1.0	12



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.

PCA on Instrumental Variables = Redudancy Analysis

- Another package of interest : vegan (J Oksanen et al)

A warning

- Avoid the overparametrization

References / Packages R

- Chessel, D. et Laloë, D. Les tableaux de fréquences alléliques. (2001). Consultable à *ftp://pbil.univ-lyon1.fr/pub/mac/ADE/ADE4/DocThemPDF/Thema2D.pdf*
- Kriticos, D.J., Webber, B.L., Leriche, A., Ota, N., Macadam, I., Bathols, J. and Scott, J.K. (2012) CliMond: global high resolution historical and future scenario climate surfaces for bioclimatic modelling. *Methods in Ecology and Evolution 3: 53-64*
- Dray, S. and Dufour, A-B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22(4).
- Laloë, D. and M. Gautier (2011). On the genetic interpretation of between-group PCA on SNP data. HAL hal-00661214
- Laloë, D., Chiquet, J., Jaffrezic, F., Gautier, M. (2014). FLPCA: A Fused-Lasso based approach to identify footprints of selection in differentiated populations from dense SNP data: applications to human and cattle data. *In International Biometry Conference, Firenze, Italy, july 2014*
- Lebart, L., Piron, M, Morineau, A. (2006). Statistique exploratoire multidimensionnelle. *Dunod*.
- Legendre P., Legendre L. (2012). Numerical ecology. *Elsevier*.
- Le Roux B., Rouanet, H, 2004. Geometric Data Analysis. *Kluwer*
- Sork V.L., Aitken S.N., Dyer R.J., Eckert A.J., Legendre P., Neale D.B.,2013. Putting the landscape into the genomics of trees: approaches for understanding local adaptation and population responses to changing climate. *Tree Genetics and Genomes, 9:901-911.*

- ade4. *http://pbil.univ-lyon1.fr/ade4/*
- vegan *http://cran.r-project.org/web/packages/vegan/*