Doctoral Course (Doctoral Schools ABIES and GAIA)
Environmental Genetics

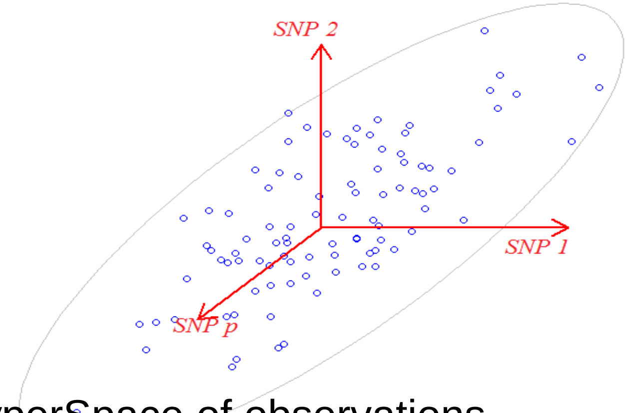
Multivariate Data Analysis:

Geometrical aspects - The Inertia

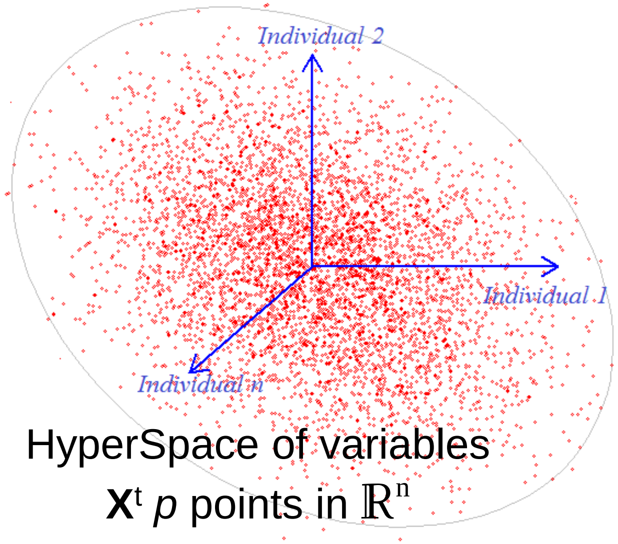
Denis Laloë, Tatiana Zerjal, Xavier Rognon

UMR GABI – INRA/AgroParisTech

Data analysis and duality diagram
The story so far



HyperSpace of observations
X *n* points in $\mathbb{R}^p$



HyperSpace of variables
$\mathbf{X}^t$ *p* points in $\mathbb{R}^n$

$$\mathbf{X} = \begin{bmatrix} & SNP_1 & SNP_2 & \dots & SNP_p \\ Ind_1 & 0 & 2 & \dots & 2 \\ Ind_2 & 2 & 1 & \dots & 0 \\ & \dots & \dots & \dots & \dots \\ Ind_n & 1 & 2 & \dots & 2 \end{bmatrix}$$

$$\mathbf{X}^t = \begin{bmatrix} & Ind_1 & Ind_2 & \dots & Ind_n \\ SNP_1 & 0 & 2 & \dots & 1 \\ SNP_2 & 2 & 1 & \dots & 2 \\ & \dots & \dots & \dots & \dots \\ SNP_p & 2 & 0 & \dots & 2 \end{bmatrix}$$

The data matrix X: a linear operator between spaces

Let us consider the vector

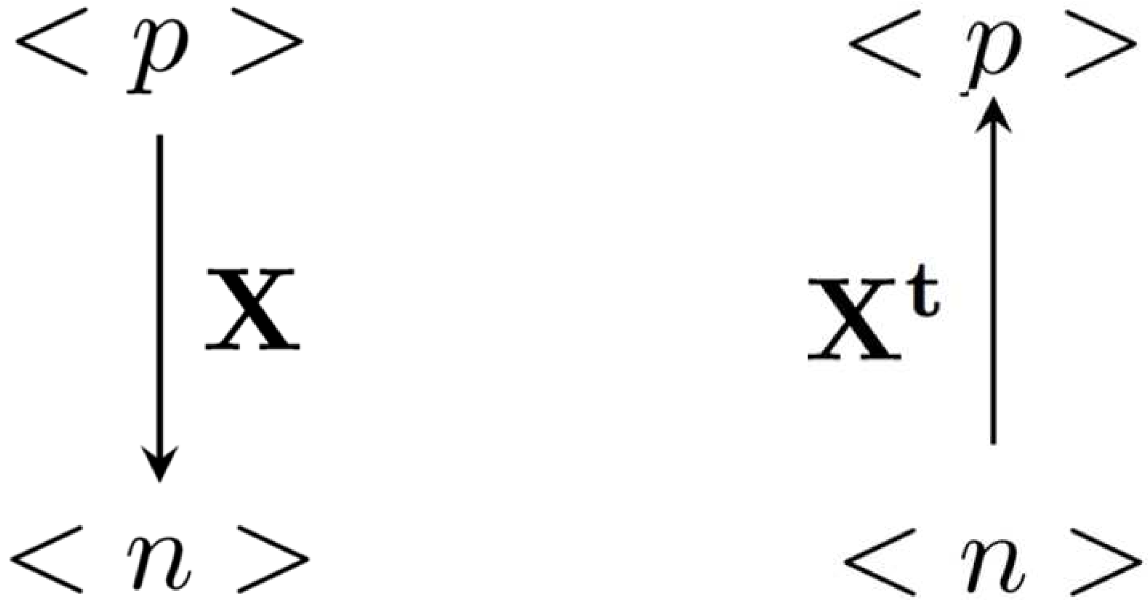
$$\mathbf{u} = \frac{\mathbf{1_n}}{2n} = \begin{bmatrix} \frac{1}{2n} \\ \frac{1}{2n} \\ \cdots \\ \frac{1}{2n} \\ \frac{1}{2n} \end{bmatrix}$$

u is in $\mathbb{R}^n$

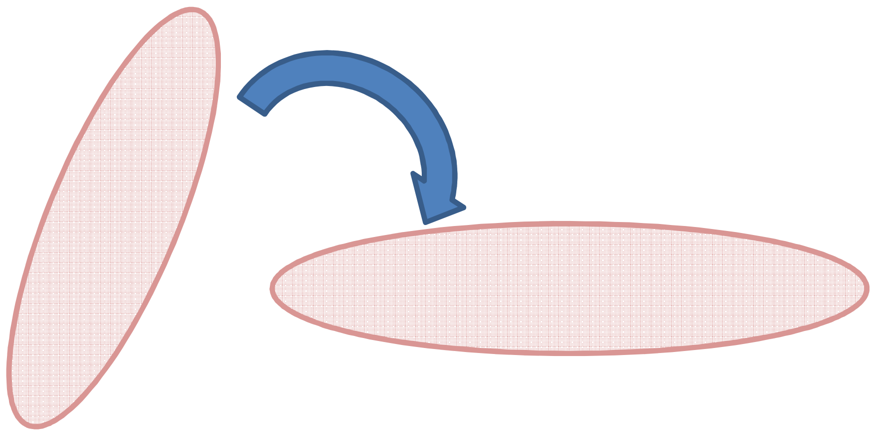
$$\mathbf{X^t u} = \mathbf{X^t} \begin{bmatrix} \frac{1}{2n} \\ \frac{1}{2n} \\ \cdots \\ \frac{1}{2n} \\ \frac{1}{2n} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} x_{i1}}{2n} \\ \frac{\sum_{i=1}^{n} x_{i2}}{2n} \\ \cdots \\ \frac{\sum_{i=1}^{n} x_{ip}}{2n} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \cdots \\ f_p \end{bmatrix}$$

$\mathbf{X^t u}$ is in $\mathbb{R}^p$

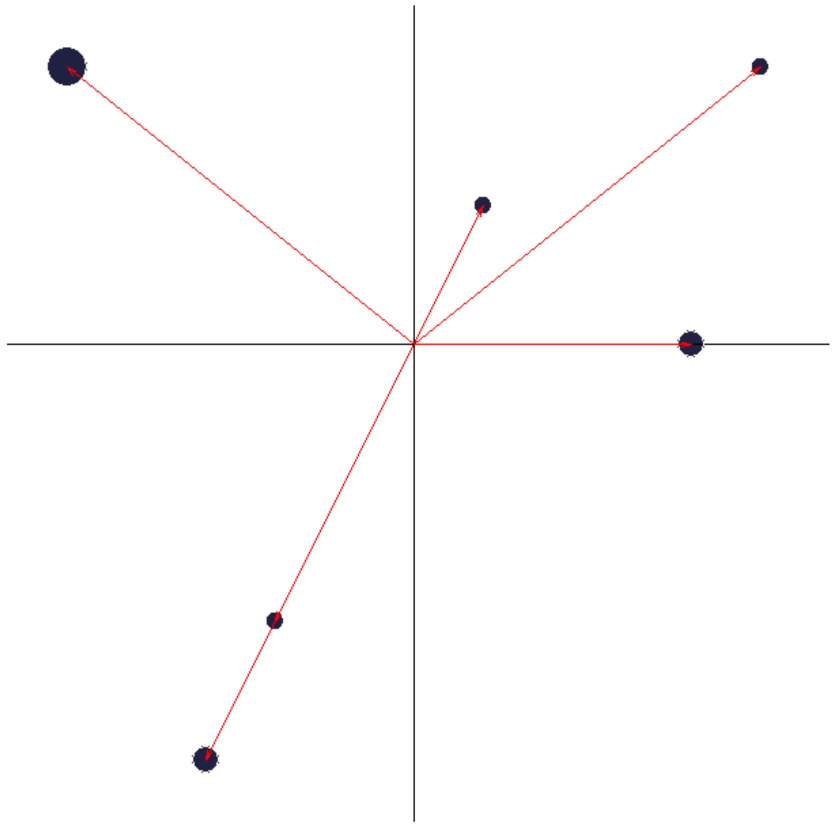
The data matrix X: a linear operator between spaces



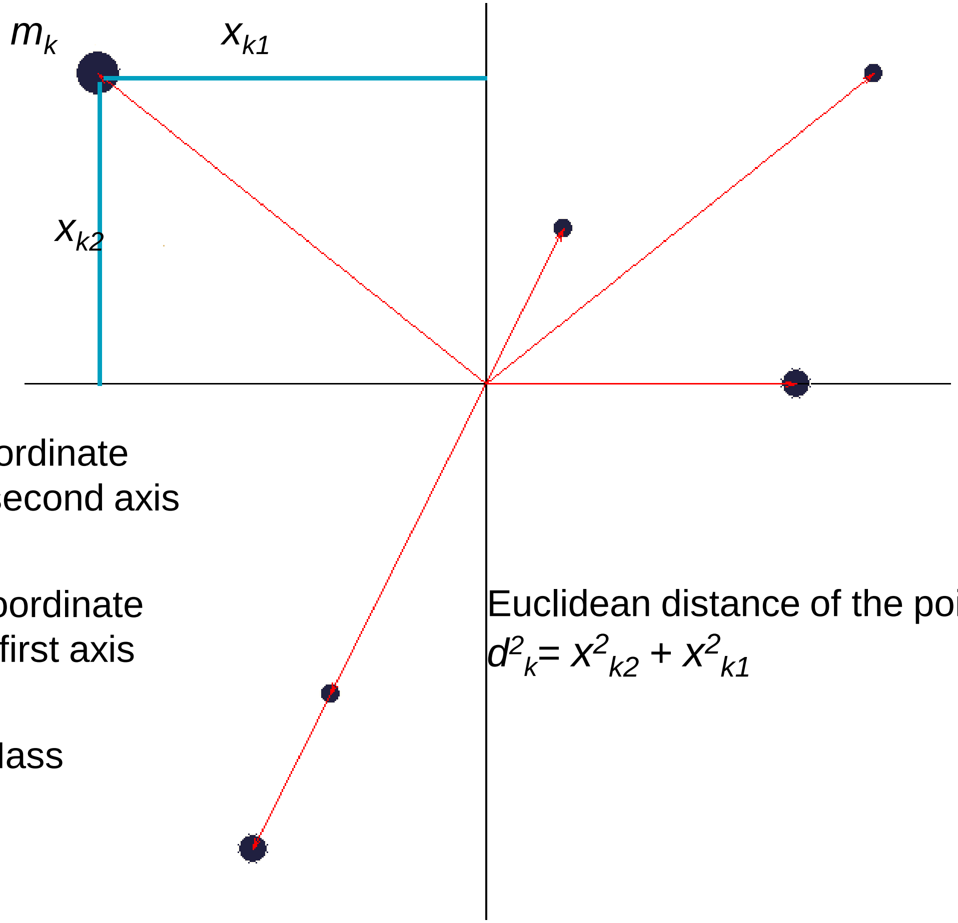
Data transformation
Rotation and Dimension Reduction



The inertia



The euclidean distance



x_{k2} : Coordinate on the second axis

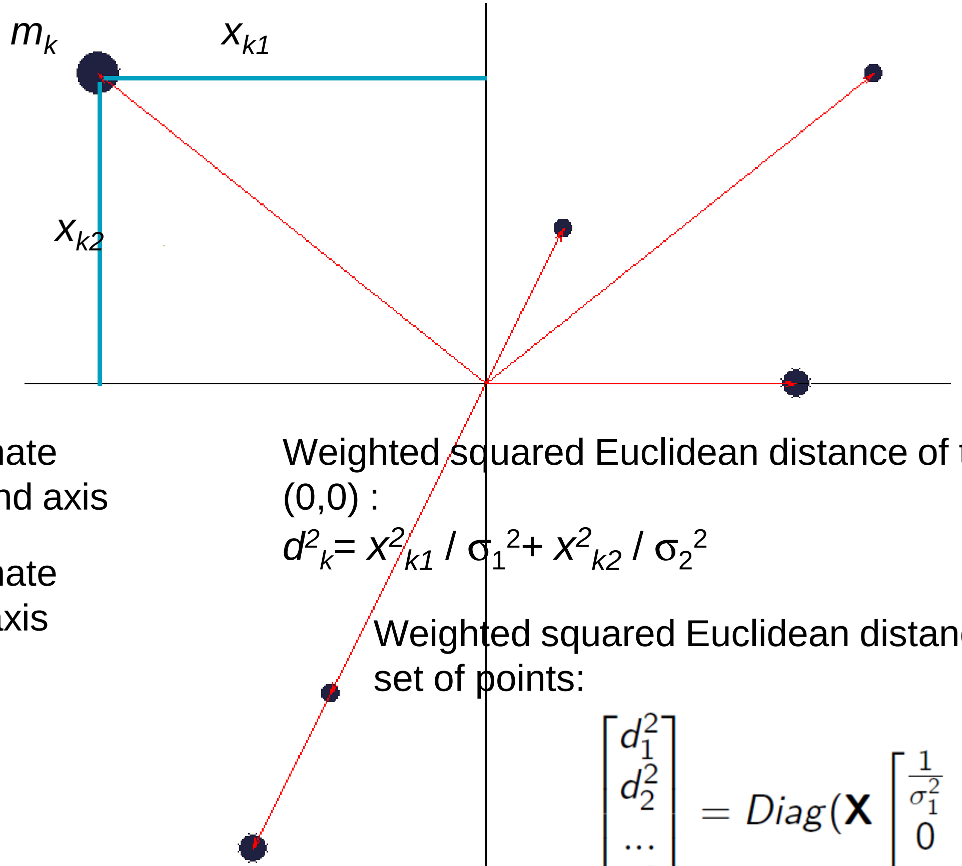
x_{k1} : Coordinate on the first axis

m_k : Mass

Euclidean distance of the point K to the origin (0,0) :

$d^2_k = x^2_{k2} + x^2_{k1}$

The weighted euclidean distance



x_{k2} : Coordinate on the second axis

x_{k1} : Coordinate on the first axis

m_k : Mass

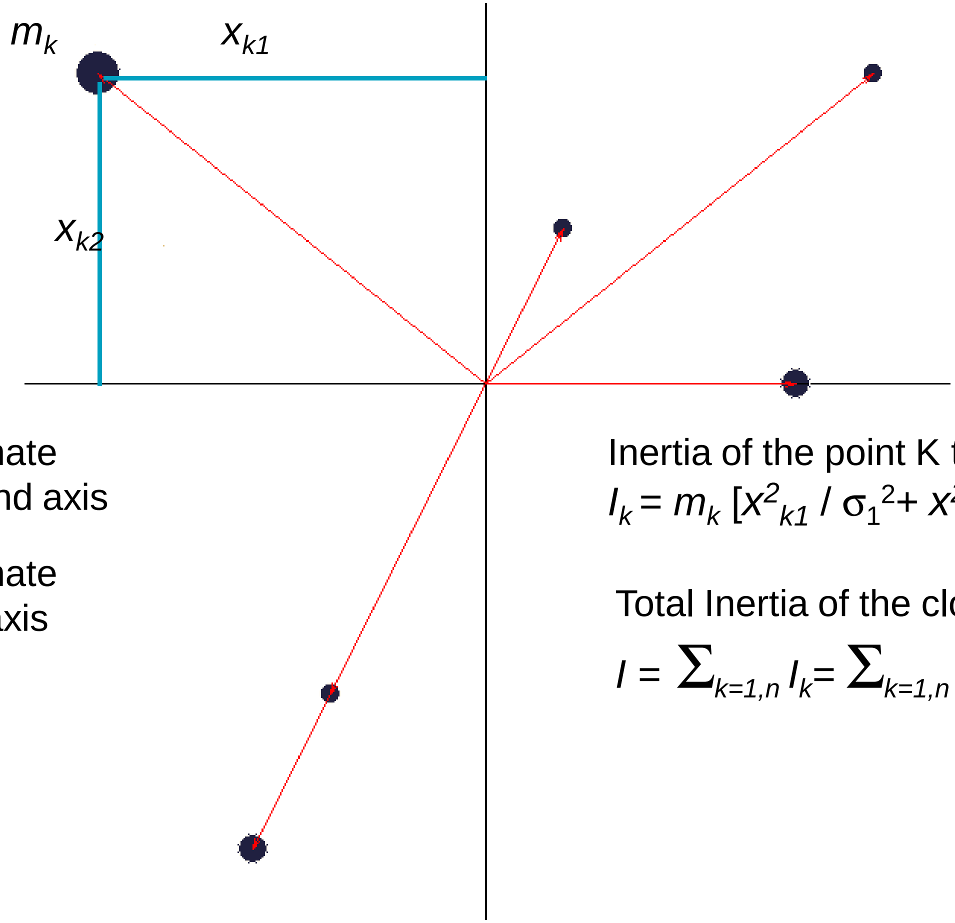
Weighted squared Euclidean distance of the point K to the origin (0,0) :

$d^2_k = x^2_{k1} / \sigma_1^2 + x^2_{k2} / \sigma_2^2$

Weighted squared Euclidean distances to the origin of the n-set of points:

$$\begin{bmatrix} d_1^2 \\ d_2^2 \\ \dots \\ d_n^2 \end{bmatrix} = Diag(\mathbf{X} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \mathbf{X^t}) = Diag(\mathbf{XQX^t})$$

The inertia



x_{k2} : Coordinate
on the second axis

x_{k1} : Coordinate
on the first axis

m_k : Mass

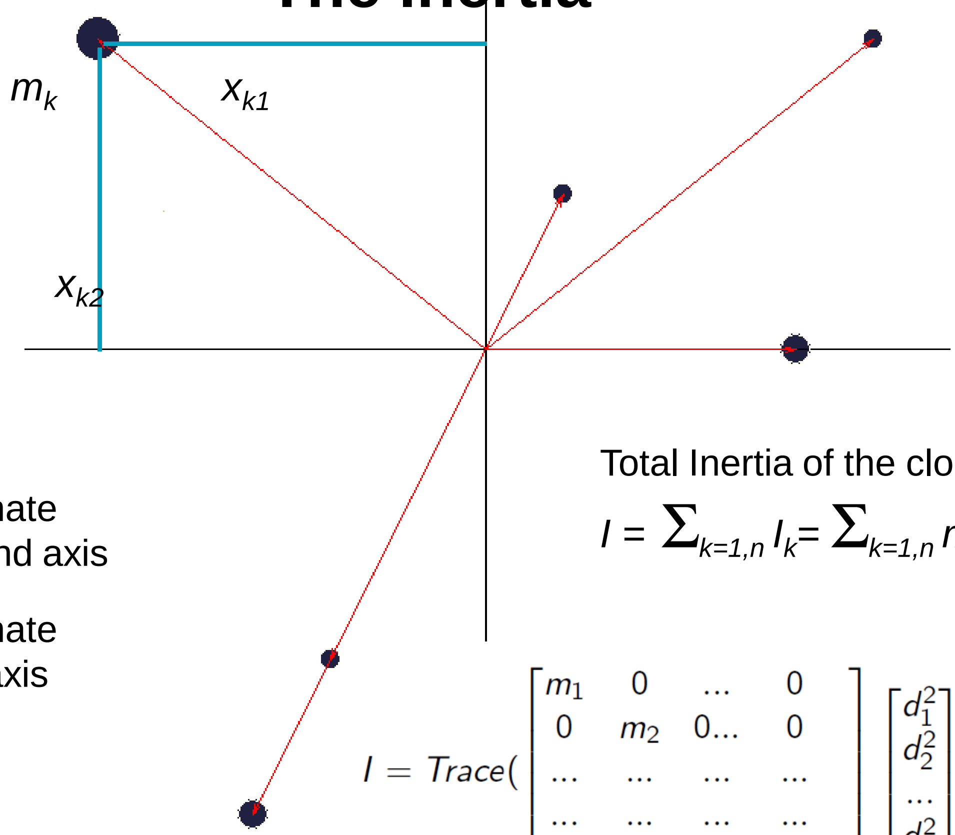
Inertia of the point K to the origin (0,0) :

$I_k = m_k\,[x^2_{k1} / \sigma_1{}^2 + x^2_{k2} / \sigma_2{}^2] = m_k\, d^2{}_k$

Total Inertia of the cloud :

$I = \sum_{k=1,n} I_k = \sum_{k=1,n} m_k\, d^2{}_k$

The inertia



Total Inertia of the cloud :

$$I = \Sigma_{k=1,n} I_k = \Sigma_{k=1,n} m_k \, d^2_k$$

x_{k2} : Coordinate on the second axis

x_{k1} : Coordinate on the first axis

m_k : Mass

$$I = Trace\left(\begin{bmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & 0\dots & 0 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & m_n \end{bmatrix} \begin{bmatrix} d_1^2 \\ d_2^2 \\ \dots \\ d_n^2 \end{bmatrix} \right) = Trace\left(\mathbf{X} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \mathbf{X}^t \mathbf{M} \right)$$

$$I = Trace(\bar{\mathbf{X}} \mathbf{Q} \mathbf{X}^t \mathbf{M})$$

The cyclic property of the trace

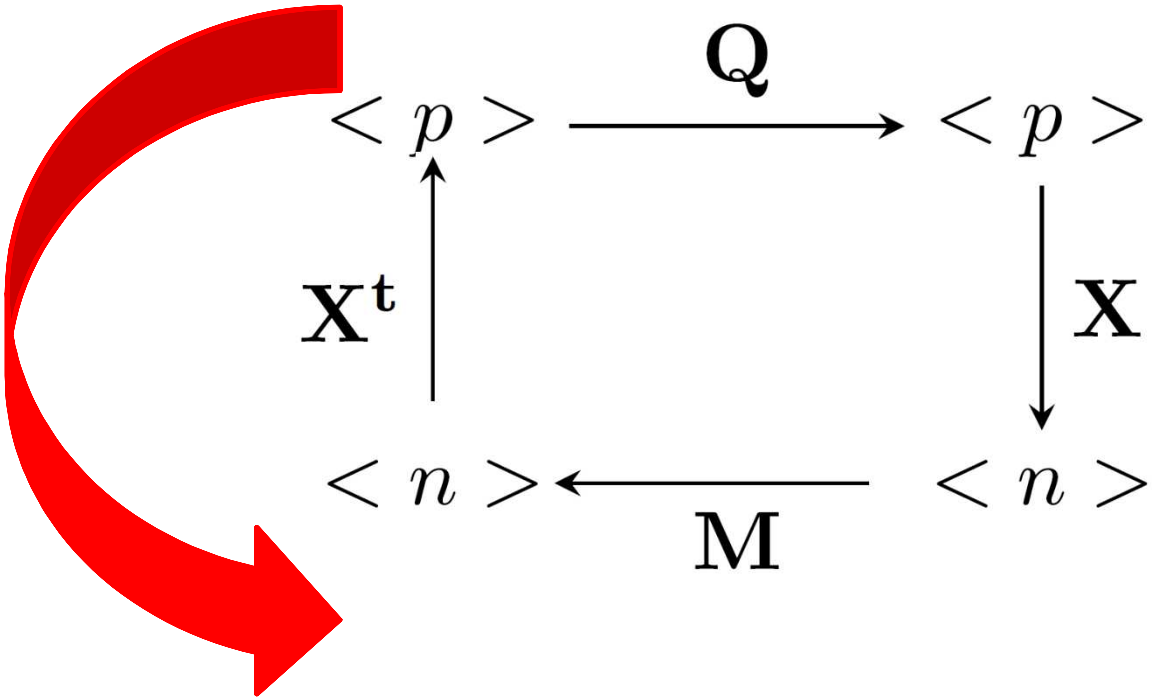
$$\mathit{Trace}\,(\mathbf{ABCD}) = \mathit{Trace}\,(\mathbf{BCDA}) = \mathit{Trace}\,(\mathbf{CDAB}) = \mathit{Trace}\,(\mathbf{DABC})$$

The inertia and the duality

$$I = \mathit{Trace}\,(\mathbf{XQX}^t\mathbf{M}) = \mathit{Trace}\,(\mathbf{X}^t\mathbf{MXQ})$$

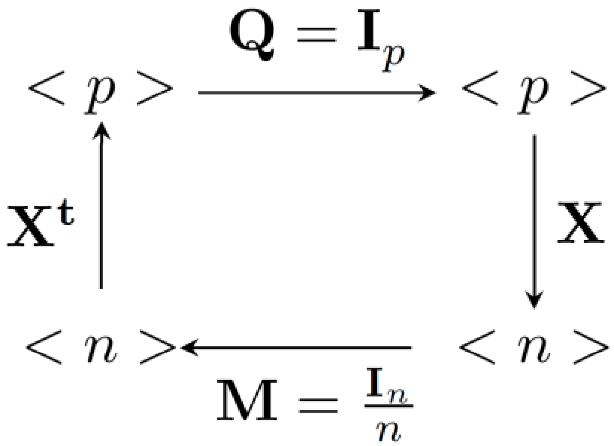
The inertia and the duality

$$I = Trace\,(\mathbf{XQX^tM}) = Trace\,(\mathbf{X^tMXQ})$$



A simple duality diagram

Principal Components Analysis on unscaled variables (Covariance)



Some references

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, *22*(4), 1-20

Holmes, S. (2008). Multivariate data analysis: the French way. In *Probability and statistics: Essays in honor of David A. Freedman* (pp. 219-233). Institute of Mathematical Statistics.

Lebart, L ; Piron, M & Morineau A (2006). Statistique exploratoire multidimensionnelle. Dunod, Paris.