Réduction de la dimension et algèbre linéaire

Alain Franc

INRA BioGeCo & INRIA Equipe Pleiade
alain.franc@inra.fr

Saint-Pierre d'Oléron
ANF "Données massives"
2017

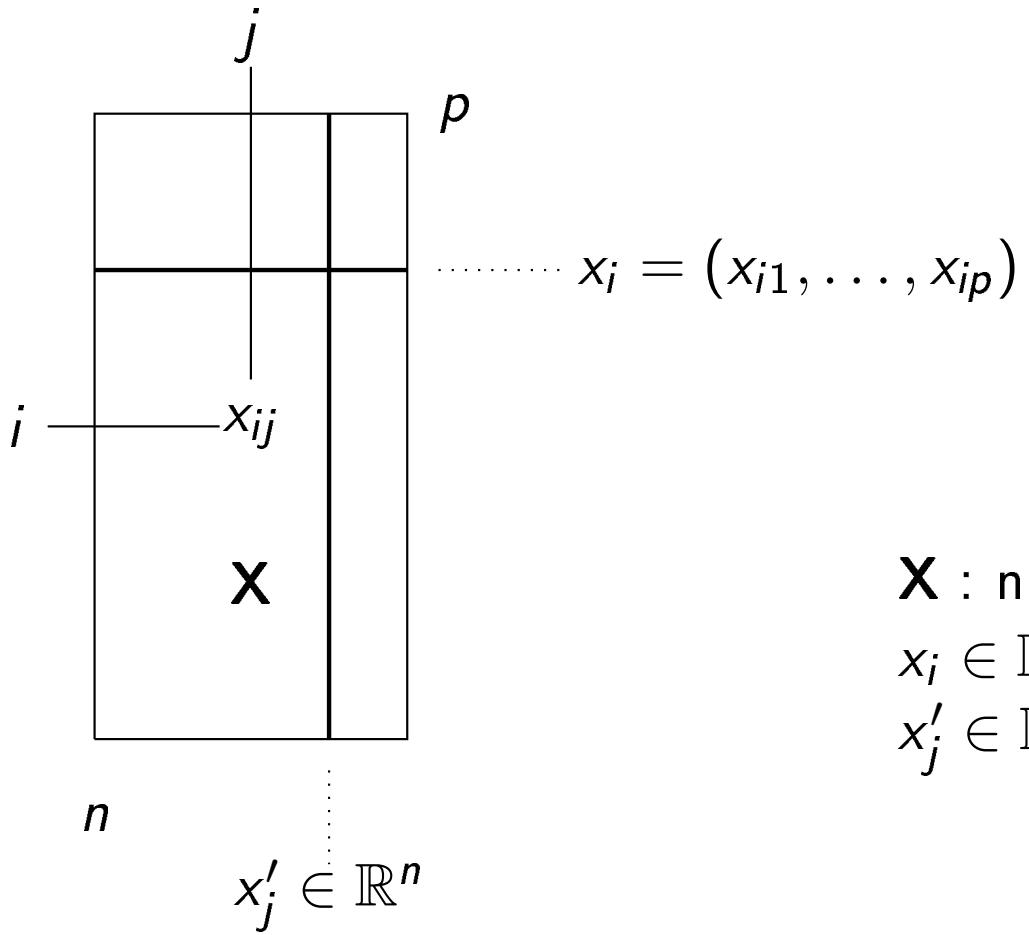
Take home message ...

- La réduction de dimension d'un nuage de points en géométrie euclidienne (analyses multivariées, "machine learning", compression, caractérisation, ...) est une sorte de diabolo
- 1. prétraitement des données
  2. diagonalisation d'une certaine matrice ou SVD
  3. visualisation et post-traitement
- il s'agit d'algèbre linéaire dense
- aujourd'hui à très grandes dimensions (possiblement plusieurs millions ...)

Plan

1. Introduction
2. ACP
3. Distance dans $\mathbb{R}^p$ et $\mathbb{R}^q$
4. ACP avec variables instrumentales
5. Analyse Canonique
6. Isométries
7. Exemples
8. MDS
9. Méthodes non linéaires

Les objets d'intérêt

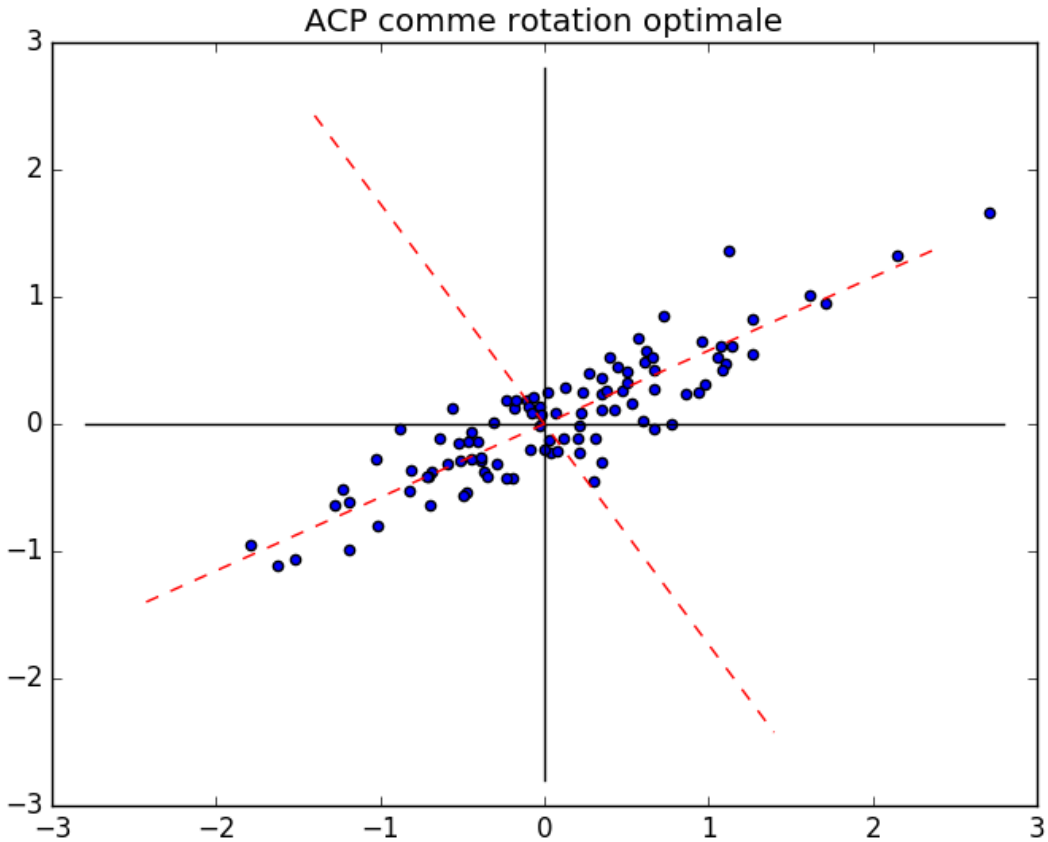


X : nuage de n points dans $\mathbb{R}^p$

$x_i \in \mathbb{R}^p$: point i

$x'_j \in \mathbb{R}^n$: variable j

Adaptation des axes à un nuage de points



Trois contextes

géométrique

approcher le mieux possible un nuage de n points dans $\mathbb{R}^p$ par un nuage dans un sous-espace affine de dimension moindre

algébrique

approcher au mieux une matrice donnée $X \in \mathcal{M}(n, p)$ par une matrice de rang $r < p$ (low rank approximation)

statistique

décrire une structure de corrélations entre variables par quelques combinaisons linéaires des variables de variance maximale

Correspondances entre distance, norme et variance.

Structure algébrique

- un espace vectoriel : $\mathbb{R}^n$
- muni d'un produit scalaire noté $\langle x, y \rangle$
- dont on déduit une norme : $\|x\| = \sqrt{\langle x, x \rangle}$
- dont on déduit une distance : $d(x, y) = \|x - y\|$

Structure algébrique

- un espace vectoriel : $\mathbb{R}^n$
- muni d'un produit scalaire noté $\langle x, y \rangle$
- dont on déduit une norme : $\|x\| = \sqrt{\langle x, x \rangle}$
- dont on déduit une distance : $d(x, y) = \|x - y\|$

Remarque

- *La structure euclidienne s'étend à des espaces de dimenson infinie : les espaces Hilbertiens.*
- *par exemple, l'ensemble des fonctions* $f : \mathbb{R} \longrightarrow \mathbb{R}$ *telles que* $\int_{\mathbb{R}} |f(x)|^2 \, dx < +\infty$ *est un espace Hilbertien noté* $L^2(\mathbb{R})$

Une petite piqûre de rappel sur les produits scalaires ...

$$\langle x, y \rangle = \sum_i x_i y_i = \|x\| \|y\| \cos\theta$$

mais aussi

$$\langle x, y \rangle_\Sigma = \sum_i \sigma_i^2 \, x_i y_i$$

et plus généralement

$$\langle x, y \rangle_P = \sum_{i,j} \pi_{ij} x_i y_j, \qquad P = [\pi_{ij}]_{i,j} \quad \text{SDP}$$

On vérifie que

$$\langle x, y \rangle_P = \langle Qx, Qy \rangle = \langle Px, y \rangle, \quad P = Q^2$$

... les distances ...

On définit la norme

$$\|x\|^2 = \langle x, x \rangle$$

et

$$d(x, y) = \|x - y\|$$

est une distance, telle que

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

On définit pour toute MSDP P

$$d_P(x, y) = \|x - y\|_P = \|Q(x - y)\|$$

... et les matrices

On se donne

$$X \in \mathcal{M}(n,p) \qquad X = [x_{ij}]$$

Alors, si $X, Y \in \mathcal{M}(n,p)$

$$\langle X, Y \rangle = \sum_{i,j} x_{ij} y_{ij} = \text{Tr } X'Y = \text{Tr } Y'X$$

On définit (norme dite de Frobenius)

$$\|X\|^2 = \langle X, X \rangle = \text{Tr } X'X$$

et

$$d^2(X,Y) = \|X - Y\|^2 = \sum_{i,j} (X_{ij} - Y_{ij})^2$$

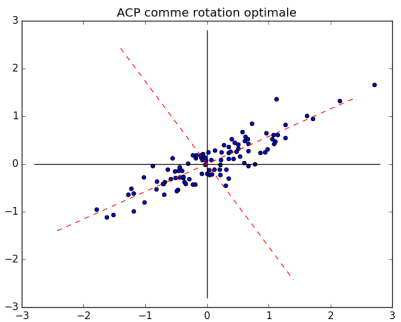
Problématique de l'ACP

Optimisation sous contrainte

Etant donné	un nuage $\mathbf{X}$ de n points dans $\mathbb{R}^p$
	une dimension $1 \leq r < p$
trouver	un sous-espace $E \subset \mathbb{R}^p$
avec	$\dim E = r$
tel que	$\|\mathbf{X} - \widetilde{\mathbf{X}}\|$ est minimum
où	$\widetilde{\mathbf{X}}$ est la projection de $\mathbf{X}$ sur E

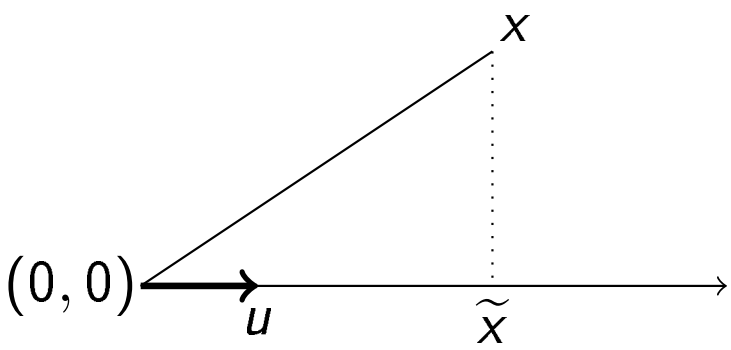
Note

$$\|\mathbf{X} - \widetilde{\mathbf{X}}\|^2 = \sum_i \|x_i - \widetilde{x}_i\|^2$$



Pythagore, tout simplement

Visualisation en 2D



Pythagore : $\|x\|^2 = \|\widetilde{x}\|^2 + \|x - \widetilde{x}\|^2$

d'où ($\|x\|$ est fixe ; u varie) :

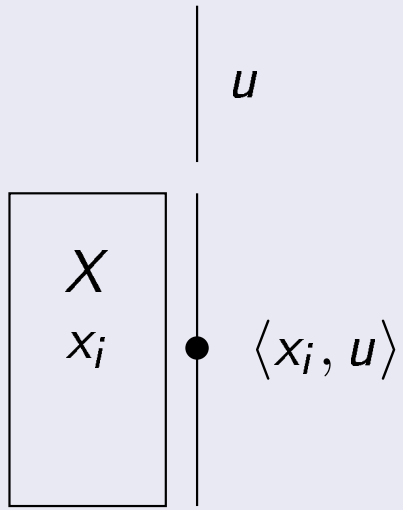
$$\|x - \widetilde{x}\| \quad \text{minimum} \quad \Rightarrow \quad \|\widetilde{x}\| \quad \text{maximum}$$

Solution pour $r = 1$

Poser le problème

- Pour tout axe défini par $u \in \mathbb{R}^p$.
- la projection de x_i sur $\mathbb{R}u$ est $\widetilde{x}_i = \langle x_i, u \rangle u$
- le vecteur $y \in \mathbb{R}^n$ tel que $y_i = \langle x_i, u \rangle$ est Xu
- On cherche u tel quel que

$$\begin{cases} \|Xu\| & \text{maximum} \\ \text{sous} & \|u\| = 1 \end{cases}$$



Solution pour $r = 1$

Solution

Multiplicateurs de Lagrange (optimum sous contraintes)

$$\nabla \|Xu\|^2 = 2X'X$$

$$\nabla \|u\|^2 = 2u$$

$$\Longrightarrow \quad X'Xu = \lambda u$$

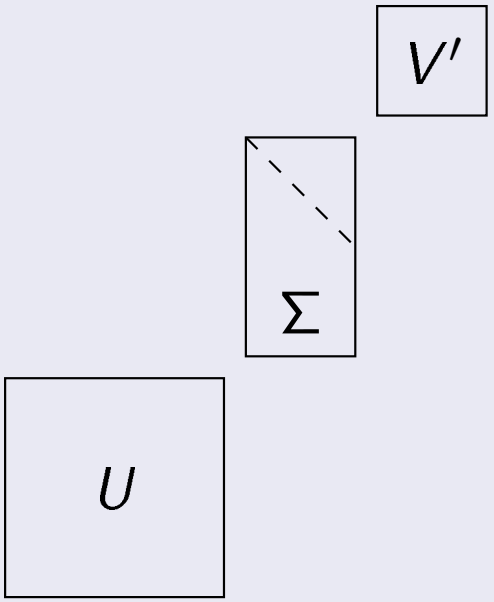
Algorithm 1 pseudocode for PCA

1: **input** $X \in \mathcal{M}(n, p), \quad r \leq p$
2: **compute** $C = X'X$
3: **compute** (Λ, U) such that $CU = U\Lambda, \quad U'U = \mathbb{I}$
4: **compute** $Y = XU$
5: **return** Λ, U, Y

Décomposition en valeurs singulières (SVD)

SVD

$$X = U\Sigma V'$$



avec

$$U \in \mathcal{M}(n,n), \quad \Sigma \in \mathcal{M}(n,p), \quad V \in \mathcal{M}(p,p)$$

et

$$U'U = \mathbb{I}_n, \qquad V'V = \mathbb{I}_p$$

On remarque que, si $X = U\Sigma V'$, alors

$$\begin{aligned} X'X &= (U\Sigma V')'(U\Sigma V') \\ &= V\Lambda V', \qquad \Lambda = \Sigma^2 \end{aligned}$$

et

$$CV = V\Lambda$$

ACP via la SVD

Remarque

La SVD est une opération de base en calcul numérique matriciel. C'est une fonction qui a été très étudiée pour l'optimisation. D'où l'intérêt numérique de son utilisation.

L'ACP se déduit naturellement de la SVD de X selon

Algorithm 2 pseudocode for PCA

1: **input** $X \in \mathcal{M}(n,p), \quad r \leq p$
2: **compute** $(U, \Sigma, V) = \texttt{SVD}(X)$
3: **compute** $Y = U\Sigma$
4: **return** Λ, V, Y

La fonction de base

fonction PCA

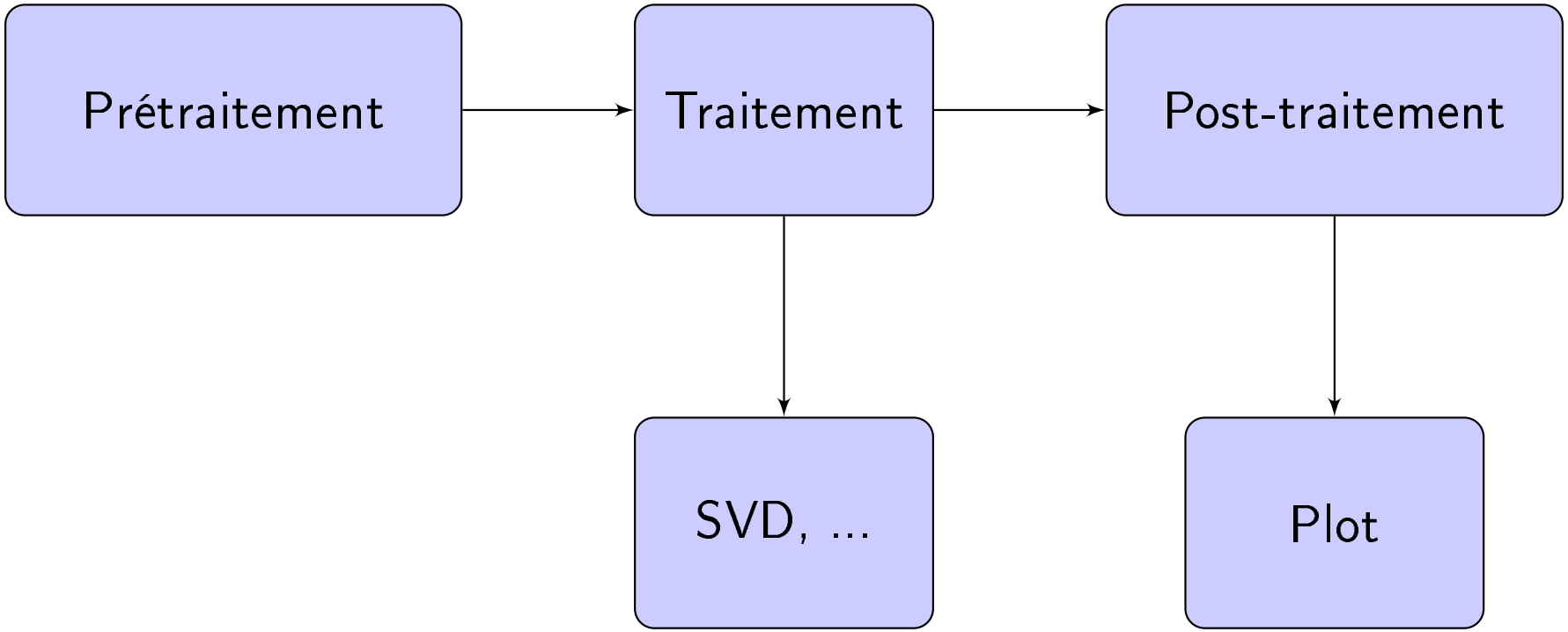
$$(X, r) \xrightarrow{\text{PCA}} (\Lambda, U, Y)$$

avec

$$\begin{cases} C &= X'X \\ CU &= U\Lambda \\ Y &= XU \\ \widetilde{X} &= YU' \end{cases}$$

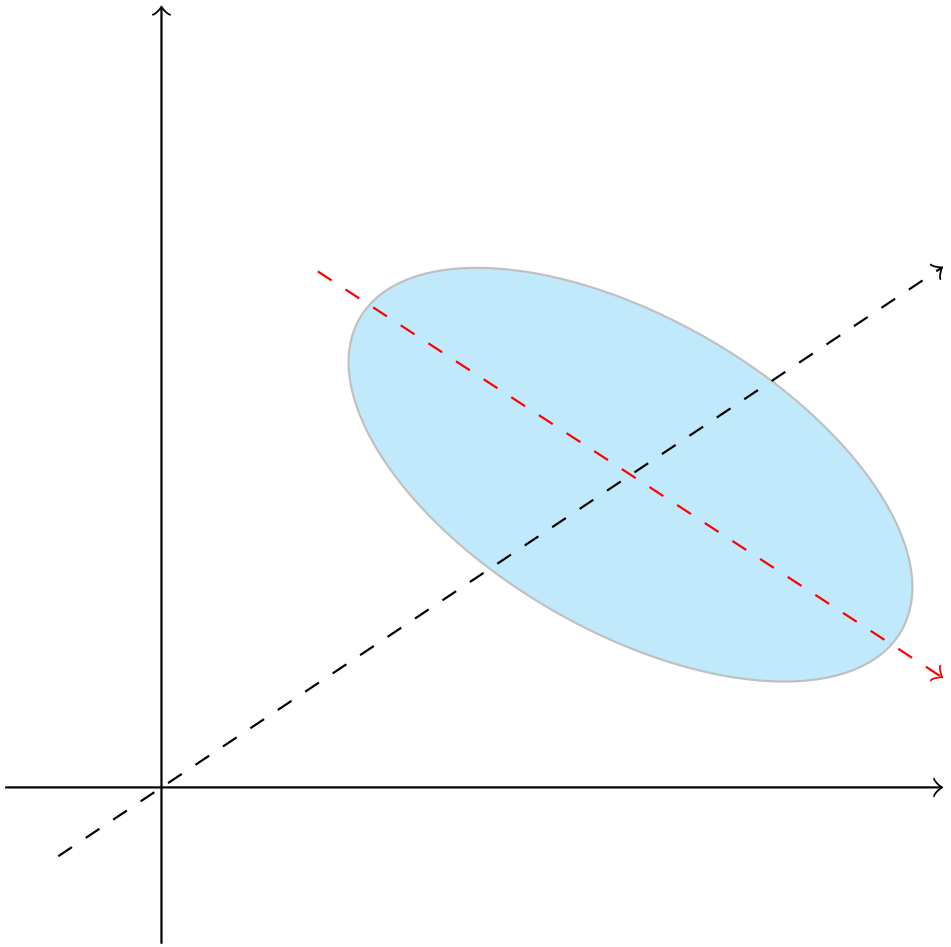
Une chaîne de traitement

Tout commence par une distance ...



Centrage

Translation vers le centre de gravité du nuage



sans translation

après translation

Centrage

par ligne ou par colonne ou les deux ...

$$x = (x_1, \ldots, x_n) \longrightarrow \overline{x} = \frac{1}{n} \sum_i x_i$$

$$x \in \mathbb{R}^n \longrightarrow z = (x_1 - \overline{x}, \ldots, x_n - \overline{x})$$

On définit

$$H_n = \mathbb{I}_n - \frac{1}{n} \mathbf{1}_n \in \mathcal{M}(n, n)$$

Alors, $z = H_n x$ et le centrage par ligne s'écrit

$$Y = X H_p$$

et par colonne

$$Y = H_n X$$

et le bicentrage

$$Y = H_n X H_p$$

Distances

$$d(x,y) = \left(\sum_i (x_i - y_i)^2 \right)^{1/2}$$

avec des poids

$$d(x,y) = \left(\sum_i \sigma_i^2 (x_i - y_i)^2 \right)^{1/2}$$

Matriciellement

$$\|x\|_\Sigma^2 = \|\Sigma x\|^2$$

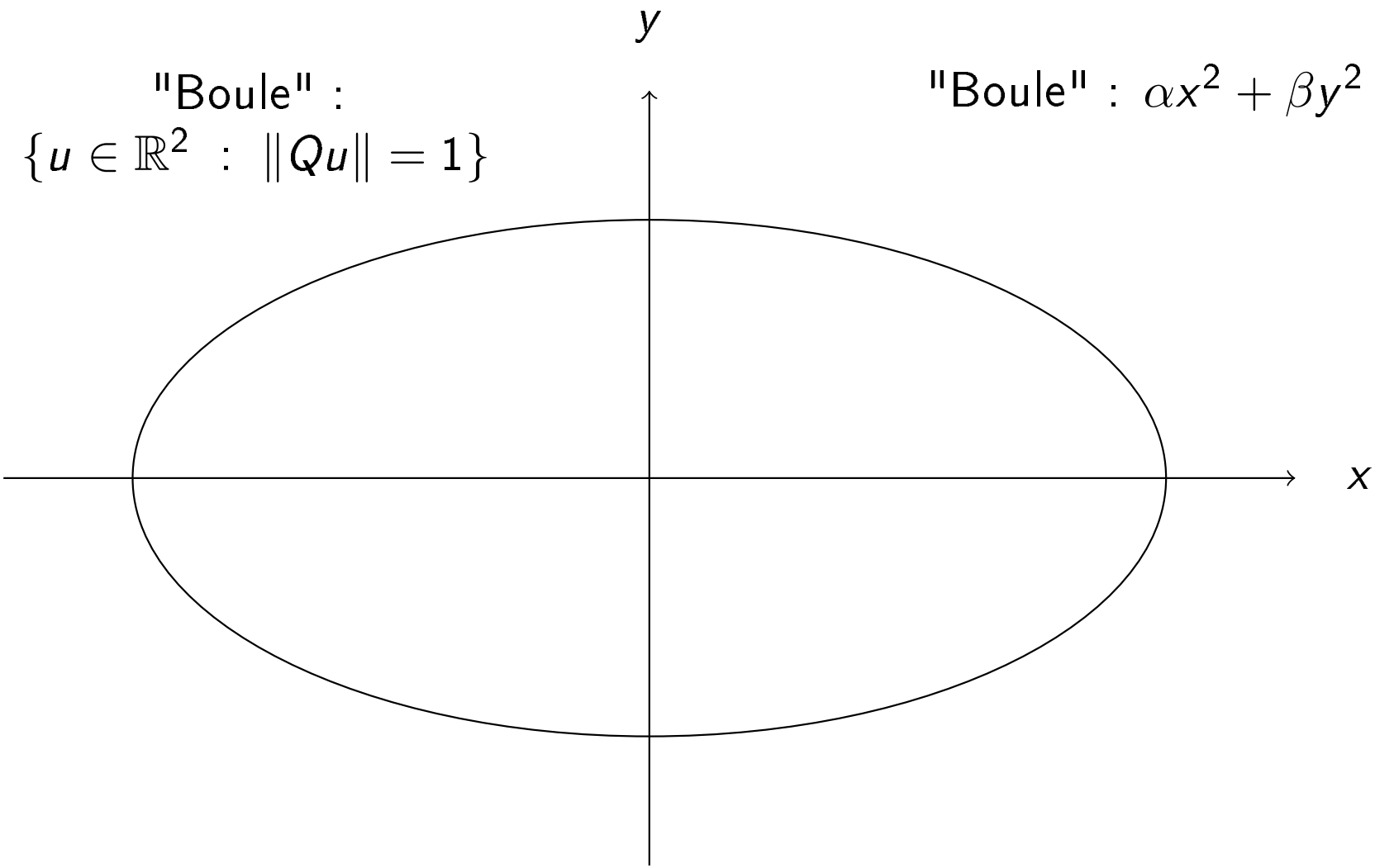
S'étend à des matrices Σ non diagonales,

$$\|x\|_Q^2 = \|Qx\|^2, \qquad Q \quad \text{s.d.p.}$$

"Boules" dans un espace euclidien

"Boule" :
$\{u \in \mathbb{R}^2 \ : \ \|Qu\| = 1\}$

"Boule" : $\alpha x^2 + \beta y^2 = 1$



Pythagore ...

Même démarche : Pythagore "marche" dans tout espace euclidien

- On se donne un vecteur $u \in \mathbb{R}^p$ tel que $\|u\|_Q = \|Qu\| = 1$
- On définit la projection de x sur u comme $\widetilde{x} \in \mathbb{R}u$ tq $\|x - \widetilde{x}\|_Q$ minimum
- La solution est $\widetilde{x} = \xi u$ avec $\xi = \langle x, u \rangle_Q = \langle Qx, Qu \rangle$
- On a

$$\|x\|_Q^2 = \|\widetilde{x}\|_Q^2 + \|x - \widetilde{x}\|_Q^2$$

Poser le problème

Optimisation sous contrainte

Etant donné	une matrice X représentant n points dans $\mathbb{R}^p$
	un produit scalaire dans $\mathbb{R}^p$ défini par une MSDP Q
	une dimension $1 \leq r < p$
trouver	un sous-espace $E \subset \mathbb{R}^p$
avec	$\dim E = r$
tel que	$\|\widetilde{X}\|_Q$ est maximum
où	$\widetilde{X}$ est la projection de X sur E

Solution pour $r = 1$

On définit

$$P = Q'Q = Q^2 \in \mathcal{M}(p, p)$$

Reposer le problème

- On définit un axe par $u \in \mathbb{R}^p$ avec $\|u\|_Q = \|Qu\| = 1$.
- la projection de x_i sur $\mathbb{R}u$ est $\widetilde{x}_i = \xi_i\, u$ avec $\xi_i = \langle x_i, u\rangle_Q$
- le vecteur $y \in \mathbb{R}^n$ tel que $y_i = \langle x_i, u\rangle_Q = \langle Px_i, u\rangle$ est $XPu = (XQ)(Qu)$
- On cherche u tel quel que

$$\begin{cases} \|(XQ)(Qu)\| & \text{maximum} \\ \text{sous} & \|Qu\| = 1 \end{cases}$$

Solution pour $r = 1$

On rappelle

$$\|X\|_Q = \|XQ\|$$

Solution

Dans

$$\begin{cases} \|(XQ)(Qu)\| & \text{maximum} \\ \text{sous} & \|Qu\| = 1 \end{cases}$$

on reconnait l'ACP de $Z = XQ$ dont la solution est v tel que

$$Z'Zv = \lambda v, \qquad v = Qu$$

D'où : $u = Q^{-1}v$ est solution de (après multiplicaton à gauche par Q^{-1})

$$X'XPu = \lambda u, \qquad \|Qu\| = 1$$

Remarque pour le calcul numérique

- $QX'XQ = (XQ)'(XQ)$ est symétrique alors que $X'XP$ en général ne l'est pas
- les méthodes de calcul des vecteurs et valeurs propres sont plus stables et efficaces si la matrice est symétrique
- d'où le conseil d'employer l'algorithme suivant

Algorithm 3 pseudocode for PCA with distances between columns

1: **input** $\{X \in \mathbb{R}^p, Q \in \mathcal{M}(p,p), r < p\}$
2: **compute** $Z = XQ$
3: **compute** $(\Lambda, V) = \text{PCA}(Z)$
4: **compute** $U = Q^{-1}V$
5: **compute** $Y = XV$
6: **return** (Λ, U, Y)

Exemple : normalisation

On note

$$x'_j = (x_{1j}, \ldots, x_{nj}) \in \mathbb{R}^n$$

La normalisation est la transformation

$$x'_j \longrightarrow x''_j = \frac{x'_j}{\|x'_j\|}$$

de telle sorte que

$$\|x''_j\| = 1$$

L'ACP normée de X est l'ACP de X'' (dont la colonne j est x''_j).

Lemme

L'ACP normée de X est l'ACP de X avec la distance définie par les poids $\sigma_j^2 = \frac{1}{\|x'_j\|}$

Distances entre lignes et entre colonnes

Distance Q dans $\mathbb{R}^p$ et M dans $\mathbb{R}^n$

Etant donnée	une matrice $X \in \mathcal{M}(n,p)$
	un produit scalaire défini par $P = Q^2$ dans $\mathbb{R}^p$
	et par $N = M^2$ dans $\mathbb{R}^n$
trouver	un axe $u \in \mathbb{R}^p$ avec $\|u\|_Q = 1$
tel que	$\|XPu\|_M$ soit maximum

En associant les résultats précédents, (λ, u) est solution de

$$X'NXPu = \lambda u, \qquad \|Qu\| = 1$$

Remarque pour le calcul numérique

- $X'NXP$ en général n'est pas symmétrique
- les méthodes de calcul des vecteurs et valeurs propres sont plus stables et efficaces si la matrice est symétrique
- d'où le conseil d'utiliser cette équation sous la forme

$$(MXQ)'(MXQ)v = \lambda v, \qquad v = Qu$$

Algorithm 4 pseudocode for PCA with distances between columns and rows

1: **input** $\{X \in \mathbb{R}^p, Q \in \mathcal{M}(p,p), M \in \mathcal{M}(n,n), r < p\}$
2: **compute** $Z = MXQ$
3: **compute** $(\Lambda, V) = \mathrm{PCA}(Z)$
4: **compute** $U = Q^{-1}v$
5: **compute** $Y = XV$
6: **return** (Λ, U, Y)

Exemple : l'AFC

- Tableau X : une table de contingence. On se donne deux variables qualitative I et J, on note $i \in I$ et $j \in J$

$$n_{ij} = |\{i(x) = i \; ; \; j(x) = j\}|$$

- On note

$$n_{+j} = \sum_i n_{ij}, \qquad n_{i+} = \sum_j n_{ij}, \qquad n_{++} = \sum_{i,j} n_{ij} = \sum_i n_{i+} = \sum_j n_{+j}$$

- On choisit les distances entre lignes et colonnes avec les poids

$$w_i = \frac{1}{n_{i+}}, \qquad w_j = \frac{1}{n_{+j}}$$

ACP sur variables instrumentales (ACP-VI)

Idée générale

On contraint les axes u à être au sein d'un sous-espace vectoriel F de $\mathbb{R}^p$

Définition

Etant donné	un nuage X de n points dans $\mathbb{R}^p$
	une dimension $1 \leq r < p$
	un sous-espace vectoriel $F \subset \mathbb{R}^p$ avec $r \leq \dim F$
Trouver	un sous-espace $E \subset \mathbb{R}^p$
	avec $E \subset F$
	et $\dim E = r$
tel que	$\|X - \widetilde{X}\|$ est minimum
	où $\widetilde{X}$ est la projection de X sur E

Exemple en écologie

On se donne n sites, et sur chaque site les présence / absence de p espèces, et q variables environnementales. On a donc deux tableaux

$$\begin{cases} X & : & n \times p & \rightarrow & \text{plantes} \\ Y & : & n \times q & \rightarrow & \text{sol-climat} \end{cases}$$

- l'ACP-VI de X avec les contraintes données par Y signifie que la projection de X doit être dans l'espace engéndré par les colonnes de Y: les points projetés sont "expliqués" par les variables instrumentales. On reconstruit la flore connaissant les conditions environnementales.

- On peut aussi réaliser l'ACP de Y avec le tableau X comme variables instrumentales : il s'agit alors de bioindication (on reconstruit la qualité evironnementale connaissant la diversité).

Solution pour $r = 1$: point de départ

- on se donne $A \in \mathcal{M}(p, q)$ où les colonnes de A sont une base de F (mieux si orthonormée ...)
- on cherche $v \in \mathbb{R}^q$ tel que, si $u = Av$, $\|Xu\|$ maximum sous $u = 1$
- soit $\|XAv\|$ maximum sous $\|Av\|$ maximum
- qui donne (multiplicateurs de Lagrange)

$$A'X'XAv = \lambda A'Av$$

- soit, par multiplication à gauche par $A(A'A)^{-1}$

$$\underbrace{A(A'A)^{-1}A'}_{\mathcal{P}_A} X'X \underbrace{Av}_{u} = \lambda \underbrace{Av}_{u}$$

Remarque pour le calcul

- Il est possible d'écrire autrement l'équation $\mathcal{P}_A X'XAv = \lambda Av$
- considérons pour cela le tableau $X_A = X\mathcal{P}_A$ (ses lignes sont les projections des x_i sur l'espace F engendré par les colonnes de A)
- on vérifie que

$$\begin{aligned}(X\mathcal{P}_A)'(X\mathcal{P}_A)Av &= \mathcal{P}_A X'X\mathcal{P}_A Av \\ &= \mathcal{P}_A X'XAv\end{aligned}$$

Lemme

L'ACP-VI de X avec contraintes définies par A est l'ACP de $X_A = \mathcal{P}_A X$

Esquisse de la preuve : comme les lignes de $\mathcal{P}_A X$ sont dans F, cela revient à faire l'ACP dans F, et u tq $X'_A X_A u = \lambda u$ vérifie $u \in F$. $\square$

Algorithm

Algorithm 5 pseudocode for PCA with instrumental variables (PCA-LV)

1: **input** $\{X \in \mathcal{M}(n,p)\,;\; A \in \mathcal{M}(p,q)\,;\; r < q\}$
2: **preprocess** A ?
3: **compute** $\mathcal{P}_A = A(A'A)^{-1}A'$
4: **compute** $X_A = \mathcal{P}_A X$
5: **compute** $(\Lambda, U) = \mathrm{PCA}(X_A)$

ACP-VI avec distances

Poser le problème

Etant donné	un tableau X
	une distance définie par N sur $\mathbb{R}^n$
	par P sur $\mathbb{R}^p$
	un ensemble de contraintes définies par A
Trouver	un axe $u \in \operatorname{span} A \subset R^p$
	avec $\|Qu\| = 1$
tel que	$\|MXPu\|$ maximum

Analyse Canonique

Idée générale

Etant donnés deux tableaux, construire des axes dans chacun des tableaux dont la corrélation soit maximale (redondance entre variables)

Poser le problème

Etant donnés	deux tableaux $A \in \mathcal{M}(n,p)$ et $B \in \mathcal{M}(n,q)$
	une dimension $r \leq \inf(p,q)$
trouver	un axe $u \in \mathbb{R}^p$ $\quad (\|u\| \neq 1)$
	un axe $v \in \mathbb{R}^q$ $\quad (\|v\| \neq 1)$
tels que	la corrélation $\mathrm{corr}\,(Au, Bv)$ soit maximale
soit	$\langle Au, Bv \rangle$ maximal sous $\|Au\| = \|Bv\| = 1$

Solution

Multiplicateurs de Lagrange

$$\left\{ \begin{array}{rcl} B'Au & = & \lambda B'Bv \\ A'Bv & = & \mu A'Au \end{array} \right.$$

soit, tous calculs faits

$$\left\{ \begin{array}{rcl} (A'A)^{-1}(A'B)(B'B)^{-1}(B'A)u & = & \lambda^2 u \\ (B'B)^{-1}(B'A)(A'A)^{-1}(A'B)v & = & \lambda^2 v \end{array} \right.$$

Remarque un : cela peut également s'écrire

$$\left\{ \begin{array}{rcl} \mathcal{P}_A \mathcal{P}_B \, Au & = & \lambda^2 Au \\ \mathcal{P}_B \mathcal{P}_A \, Bv & = & \lambda^2 Bv \end{array} \right.$$

Remarque deux : si $A'A = \mathbb{I}_p$ et $B'B = \mathbb{I}_q$, alors on retrouve l'ACP de $X = A'B$

Algorithm 6 pseudocode for Canonical Analysis

1: **input** $\{A \in \mathcal{M}(n,p)\ ;\ B \in \mathcal{M}(n,q)\ ;\ r \leq p,q\}$
2: **preprocess** A, B ?
3: **compute** $\mathcal{P}_A = A(A'A)^{-1}A'$
4: **compute** $\mathcal{P}_B = B(B'B)^{-1}B'$
5: **compute** U such that $\mathcal{P}_A\mathcal{P}_B AU = AU\Lambda^2$
6: **compute** V such that $\mathcal{P}_B\mathcal{P}_A BV = BV\Lambda^2$
7: **return** Λ, U, V, AU, BV

Note

$$(\mathcal{P}_A\mathcal{P}_B)' = \mathcal{P}_B\mathcal{P}_A$$

AC avec distances

Poser le problème

Etant donnés	deux tableaux $A \in \mathcal{M}(n,p)$ et $B \in \mathcal{M}(n,q)$
	une dimension $r \leq \inf(p,q)$
	une distance définie par Q_A sur les lignes de A
	par Q_B sur les lignes de B
	par M_A entre individus pour A et M_B pour B
trouver	un axe $u \in \mathbb{R}^p$ $\quad (\|u\|_{Q_A} \neq 1)$
	un axe $v \in \mathbb{R}^q$ $\quad (\|v\|_{Q_B} \neq 1)$
tels que	la corrélation $\mathrm{corr}\,(Au, Bv)$ soit maximale
soit	$\langle M_A A Q_A u, M_B B Q_B v \rangle$ maximal sous $\|M_A Au\| = \|M_B Bv\| =$

Pourquoi?

Par exemple pour réaliser une analyse canonique des corespondances ... ou des correspondances canoniques ...

Résumé

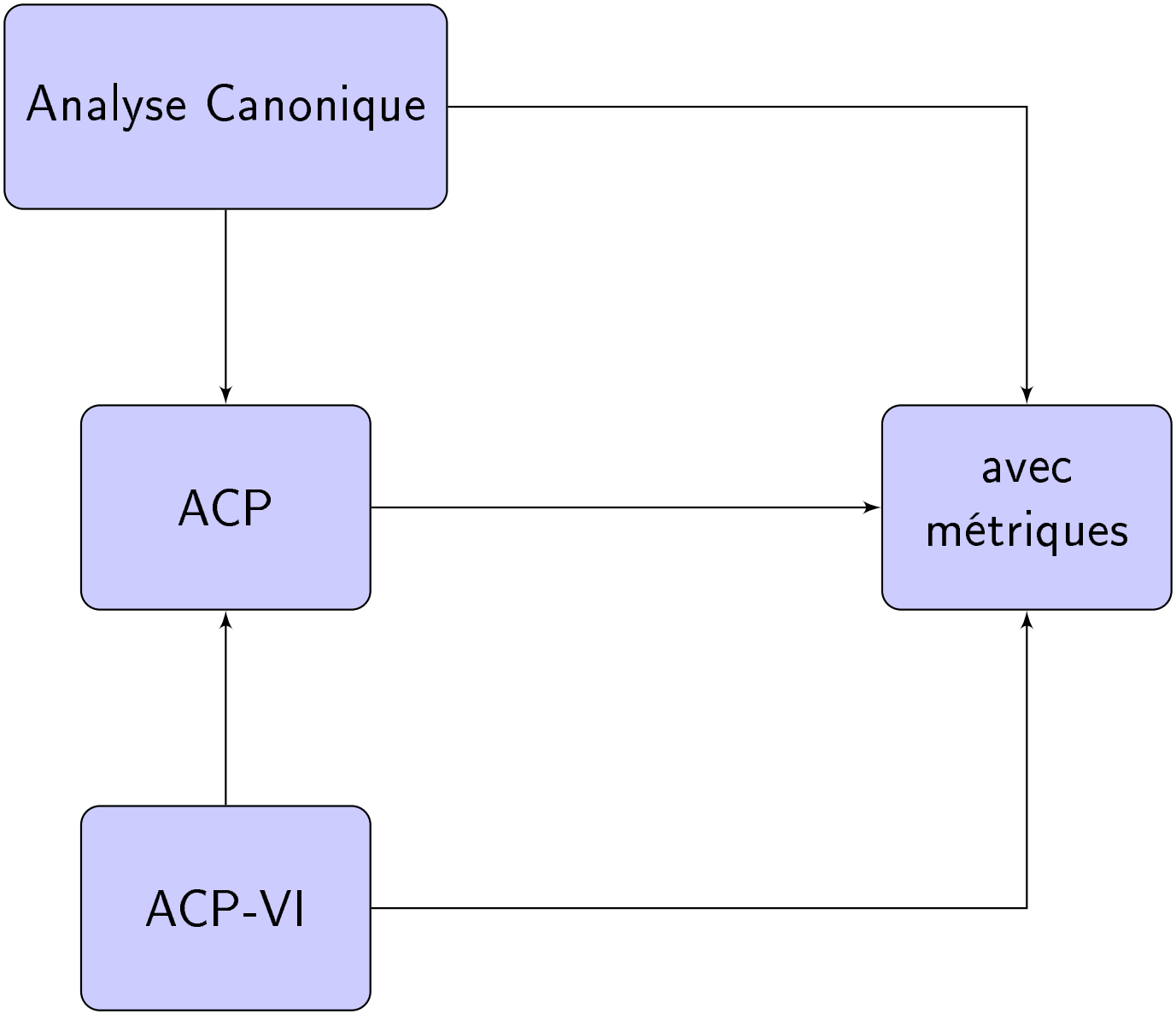
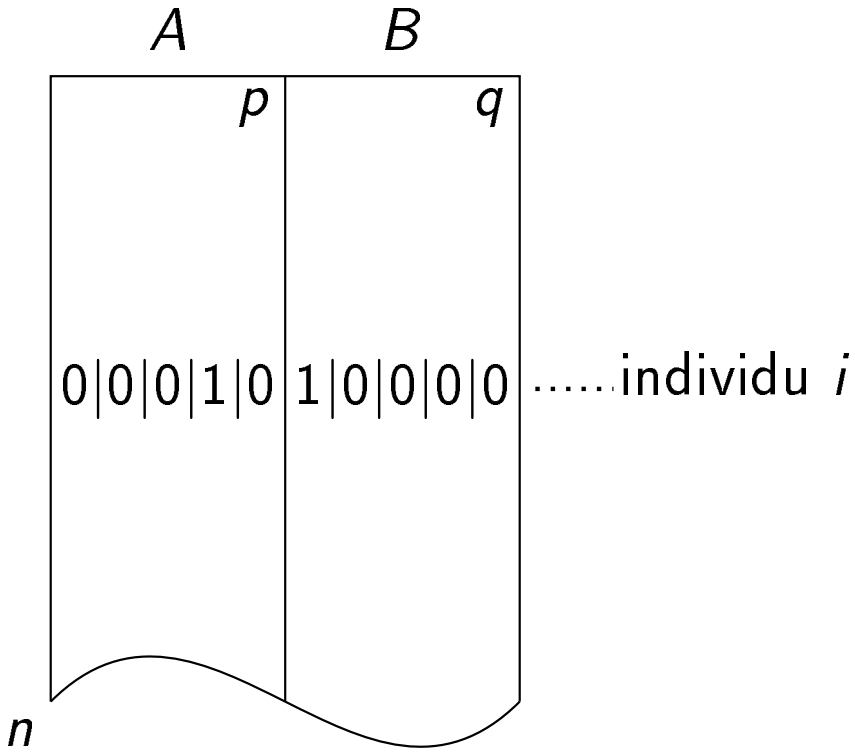


Tableau disjonctif complet

On se donne deux variables qualitatives I et J, connues sur n individus, avec $|I| = p$ et $|J| = q$. On construit le tableau disjonctif complet



$$T = [A|B]$$
$$n \times (p + q)$$

Analyse de deux TDC

Expérience

Faisons l'ACP de T avec la matrice de distance

$$P = \begin{pmatrix} (A'A)^{-1} & 0 \\ 0 & (B'B)^{-1} \end{pmatrix}$$

Alors, $T'TP[u, v] = \lambda[u, v]$ devient (tous calculs faits)

$$\begin{cases} (A'B)(B'B)^{-1}(B'A)(A'A)^{-1}u &= (\lambda-1)^2 u \\ (B'A)(A'A)^{-1}(A'B)(B'B)^{-1}v &= (\lambda-1)^2 v \end{cases}$$

où on reconnait la solution de l'analyse canonique de (A, B) qui est

$$\begin{cases} (A'A)^{-1}(A'B)(B'B)^{-1}(B'A)u' &= \lambda'^2 u' \\ (B'B)^{-1}(B'A)(A'A)^{-1}(A'B)v' &= \lambda'^2 v' \end{cases}$$

avec $\lambda' = \lambda - 1$, $u' = (A'A)^{-1}u$, $v' = (B'B)^{-1}v$

Lemme

Si A et B sont deux tableaux disjonctifs complets de taille $n \times p$ et $n \times q$ respectivement, alors, il est équivalent de faire

- *l'ACP d-u tableau $T = [A|B]$ avec la distance entre lignes définie par* $P = \begin{pmatrix} (A'A)^{-1} & 0 \\ 0 & (B'B)^{-1}) \end{pmatrix}$
- *l'analyse canonique des deux tableaux A et B*
- *l'analyse factorielle des correspondances de la table de contingence $X = A'B$*

Lebart, Morineau et Fénelon, 1982

Analyse des Correspondances Multiples

Pourquoi s'arrêter à deux ?

ACM

Etant données m variables qualitatives mesurées sur les mêmes individus

On construit les m tableaux disjonctifs complets A_k avec $k \in \{1, m\}$
où A_k a n lignes et q_k colonnes (q_k modalités)
$T = [A_1|\ldots|A_m]$ de taille $q = \sum_k q_k$
les matrices diagonales $P_k = (A_k' A_k)^{-1}$
la matrice carrée P de taille q diagonale par blocs P_k

alors,

l'ACM de $A_1, \ldots, A_m$ est l'ACP du tableau T avec la mérique définie par P entre les lignes. Elle peut également s'appeler *Analyse Canonique Multiple*.

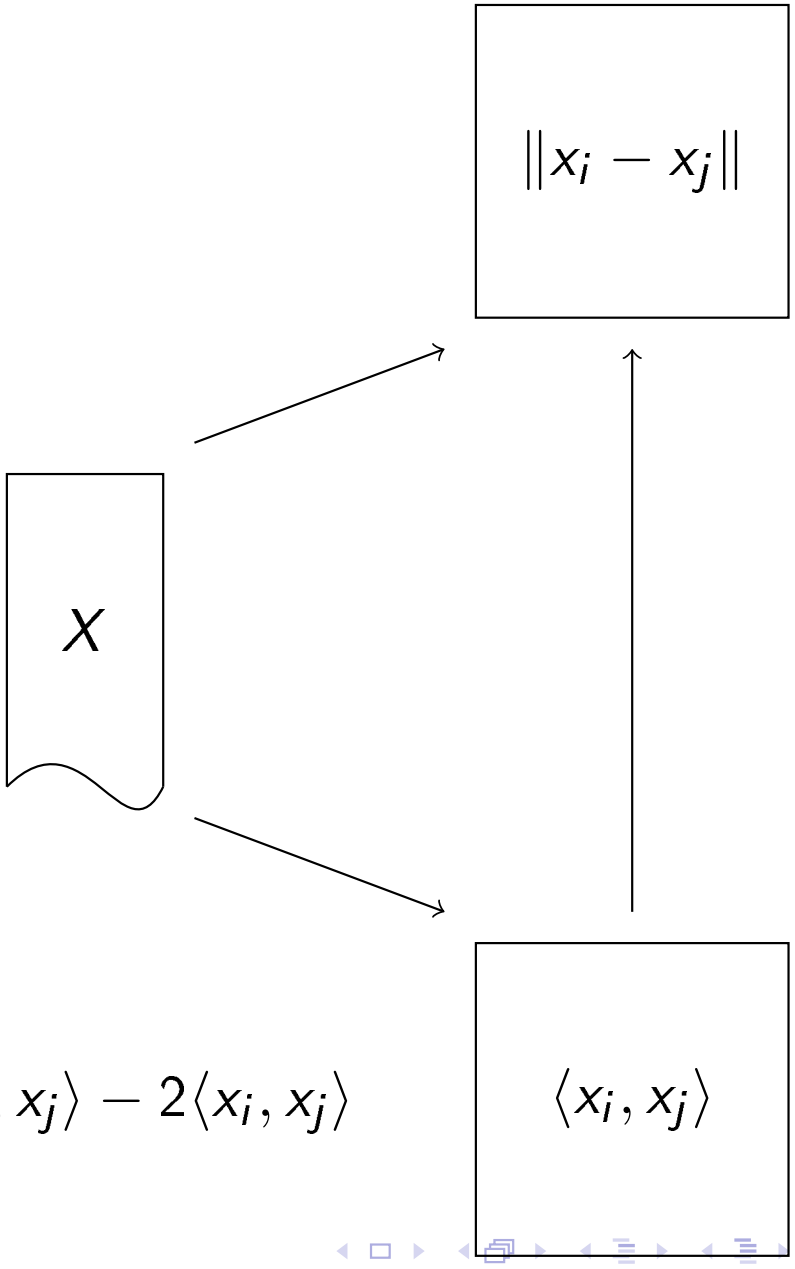
Distances et nuages de points : MDS

$\|x_i - x_j\|$

X

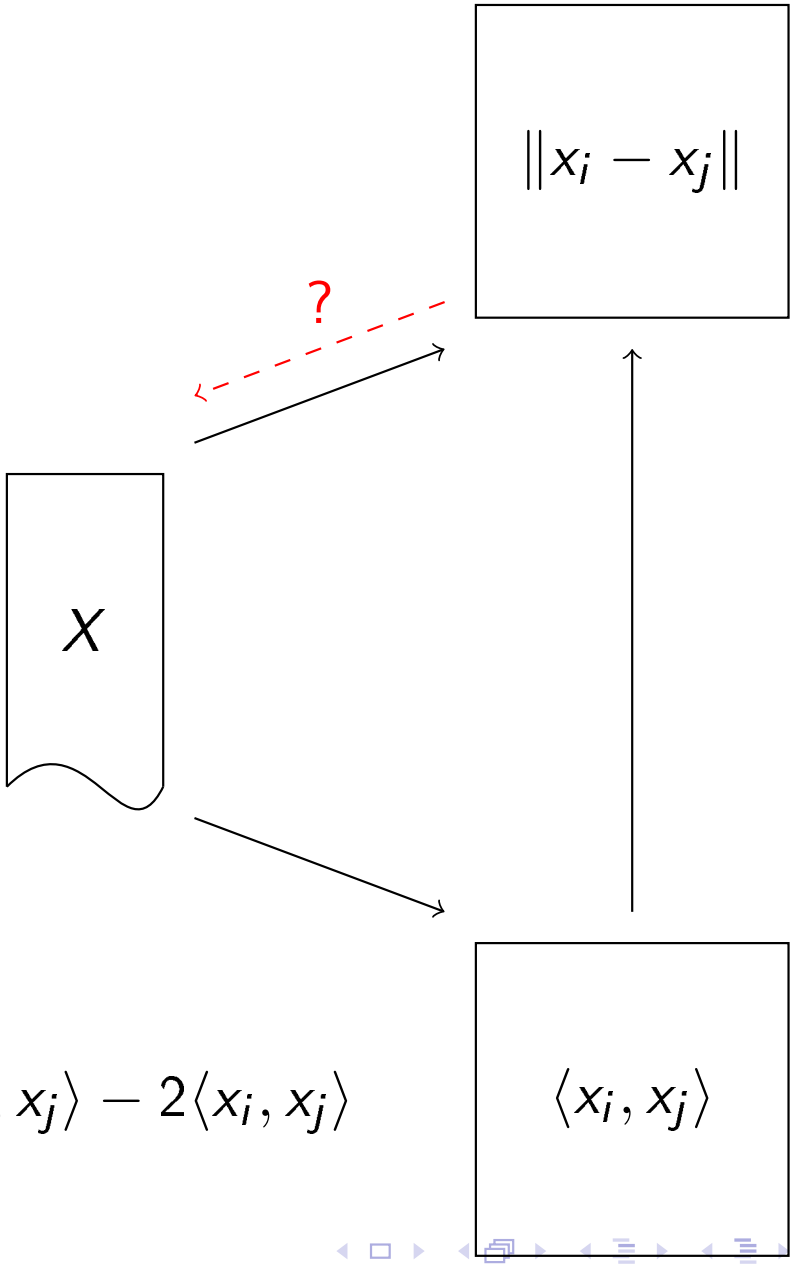
$\langle x_i, x_j \rangle$

Distances et nuages de points : MDS



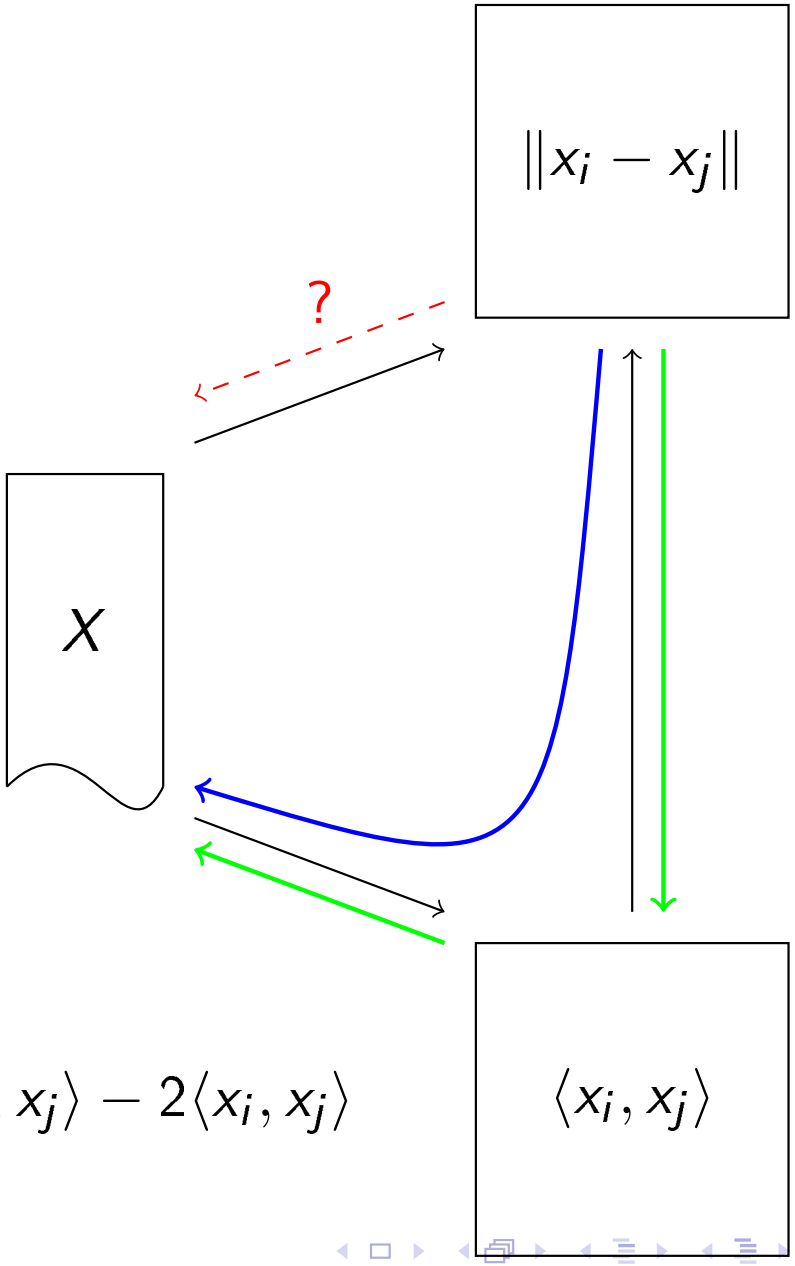
$$\|x_i - x_j\|^2 = \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2\langle x_i, x_j \rangle$$

Distances et nuages de points : MDS



$$\|x_i - x_j\|^2 = \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2\langle x_i, x_j \rangle$$

Distances et nuages de points : MDS



$$\|x_i - x_j\|^2 = \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2\langle x_i, x_j \rangle$$

MDS

poser le problème : "least square MDS"

Etant donné	un tableau de distances D entre n éléments une dimension r
Trouver	un nuage de n points X
avec	$x_i \in \mathbb{R}^r$
tel que	$\phi = \sum_{i,j}(d_{ij} - \|x_i - x_j\|)^2$ minimum

poser le problème : "classical MDS"

Etant donné	un tableau de distances D entre n éléments
	une dimension r
Trouver	un nuage de n points X
avec	$x_i \in \mathbb{R}^p$
tel que	$\forall\, i, j, \quad \|x_i - x_j\| = d_{ij}$
et	réaliser une ACP de X avec r composantes

Démarche

Matrice de Gram

$$G = [g_{ij}]_{i,j} \ : \ g_{ij} = \langle x_i, x_j \rangle = -\frac{1}{2}(d_{ij}^2 - d_{\bullet j}^2 - d_{i\bullet}^2 + d_{\bullet\bullet}^2)$$

avec

$$d_{\bullet j}^2 = \frac{1}{n} d_{ij}^2, \qquad d_{\bullet\bullet}^2 = \frac{1}{n^2} \sum_{i,j}^{2}$$

et

EVD de G

$$G = XX' \quad \text{avec} \quad X = U\Sigma \quad \text{avec} \quad Gu = \sigma^2 u$$

car

$$GU = U\Sigma^2 \quad \Longrightarrow \quad G = U\Sigma^2 U' = (U\Sigma)(U\Sigma)' = XX' \quad \text{si } X = U\Sigma$$

Classical MDS

Algorithm 7 pseudocode for classical MDS

input $\{D \in \mathcal{M}(n,n)\ ;\ r\}$
compute $d^2_{\bullet j} = \frac{1}{n} d^2_{ij}, \quad \forall j$
compute $d^2_{\bullet\bullet} = \frac{1}{n^2} \sum_{i,j}^{2}$
compte G with $g_{ij} = \langle x_i, x_j \rangle = -\frac{1}{2}(d^2_{ij} - d^2_{\bullet j} - d^2_{i\bullet} + d^2_{\bullet\bullet})$
compute (λ, u) such that $Gu = \lambda u$
keep all $\lambda > 0$
do $\sigma = \sqrt{\lambda}$
compute $X = U\Sigma$
keep the first r components
return Λ, X

Un jeu de données

Métabarcoding

- Un marqueur est un mot à un endroit du génome avec l'alphabet $\{A, C, G, T\}$ (longueur $\sim$ 300 bp)
- On prélève l'ADN d'une communauté
- On amplifie ce marqueur ($\rightarrow$ amplicons)
- on séquence des amplicons

Un jeu de données

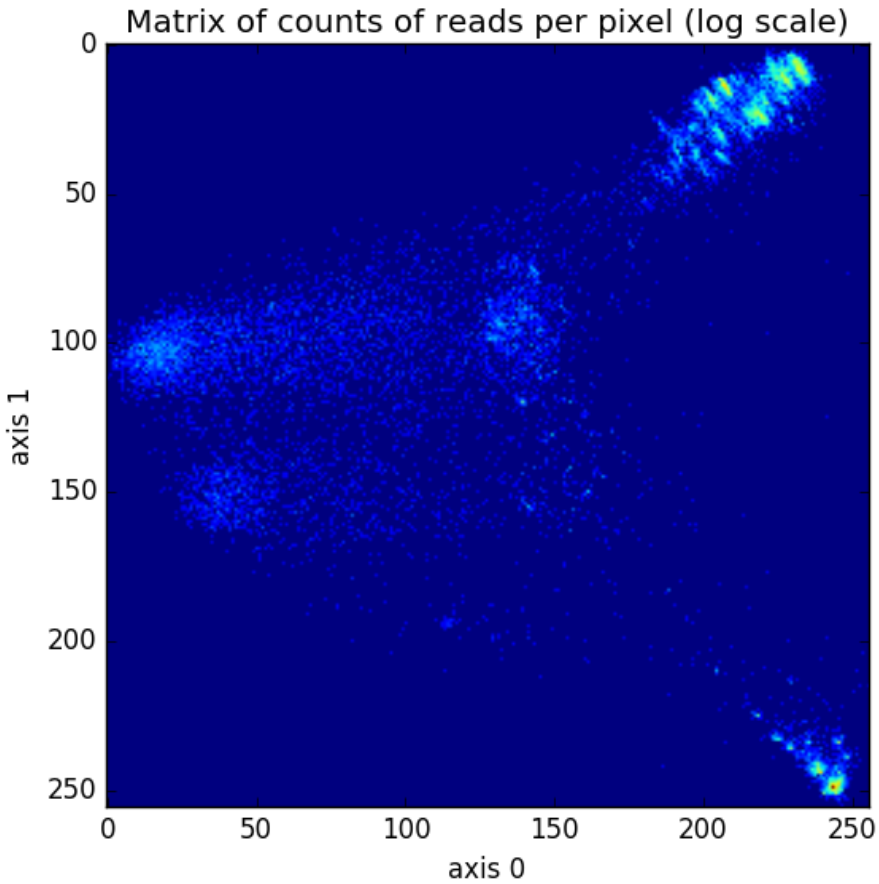
Métabarcoding

- Un marqueur est un mot à un endroit du génome avec l'alphabet $\{A, C, G, T\}$ (longueur $\sim$ 300 bp)
- On prélève l'ADN d'une communauté
- On amplifie ce marqueur ($\rightarrow$ amplicons)
- on séquence des amplicons

Métabarcoding et MDS

- On dispose de 20 k reads d'un échantillon environnemental
- On aligne toutes les séquences deux à deux ($\frac{n(n-1)}{2} = 200k$ alignements)
- On calcule une distance selon les mismatchs d'un alignement
- On réalise la MDS de ce tableau de distances

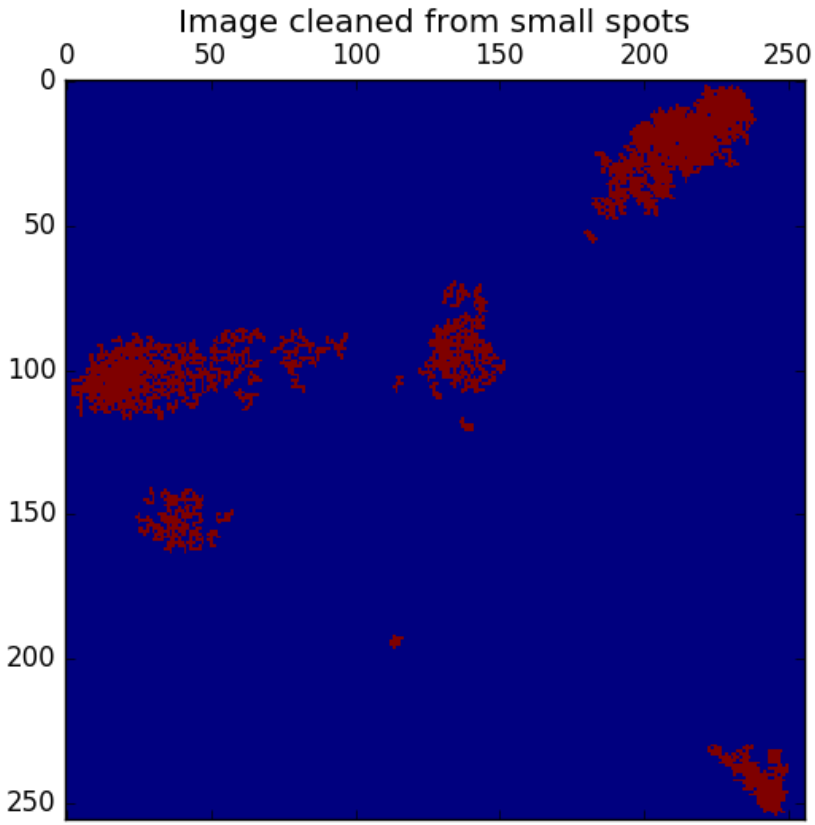
Densité des points



[1]

[1] Données : A. Chenuil & A. de Jode, IMBE, Marseille

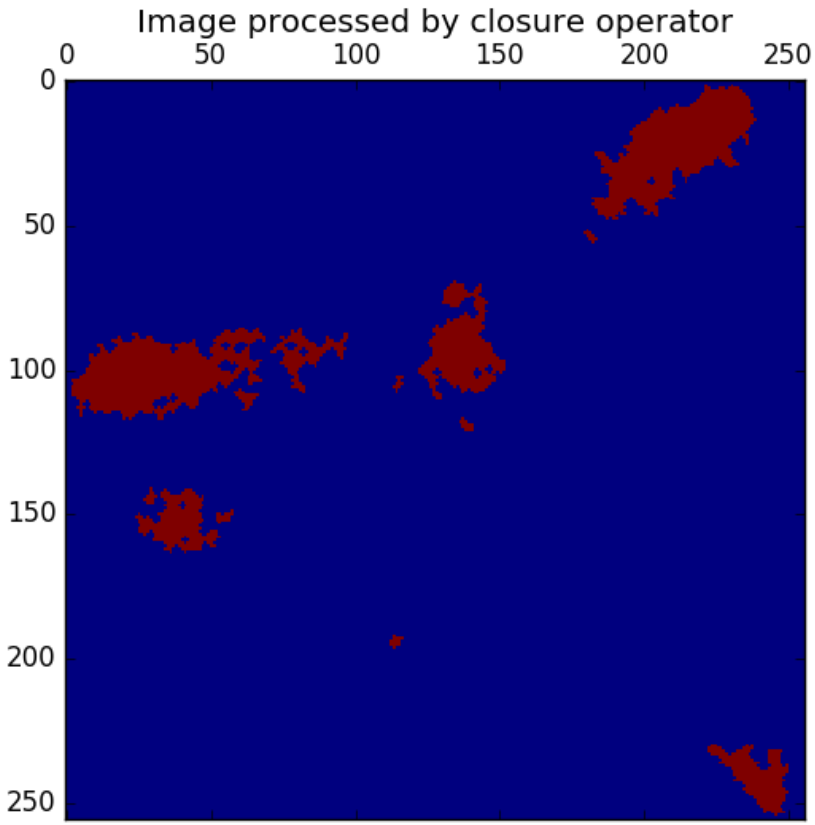
Densité des points



1

[1] Données : A. Chenuil & A. de Jode, IMBE, Marseille

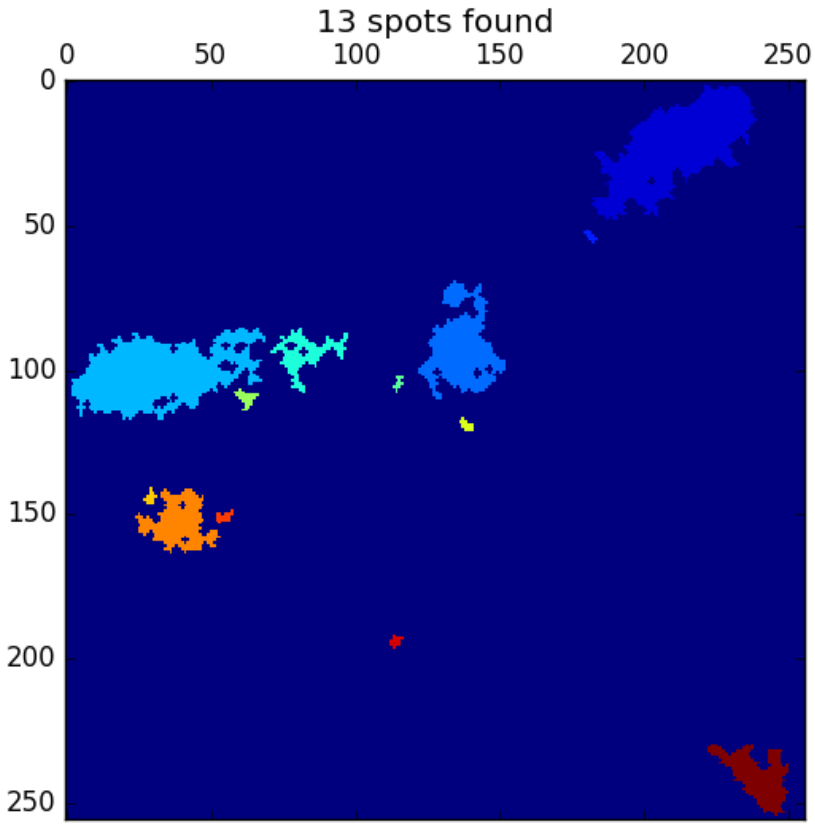
Densité des points



1

[1] Données : A. Chenuil & A. de Jode, IMBE, Marseille

Densité des points

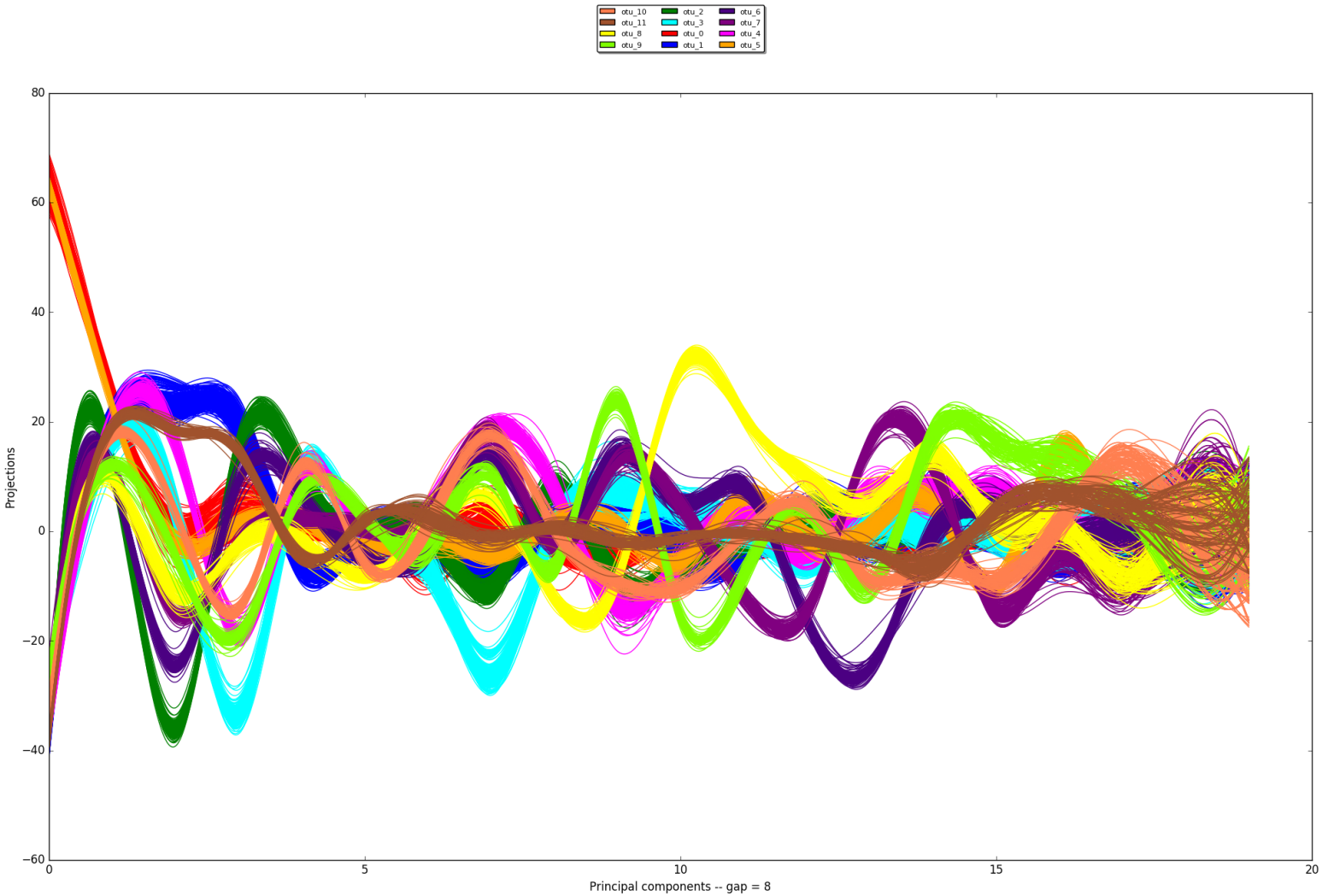


1

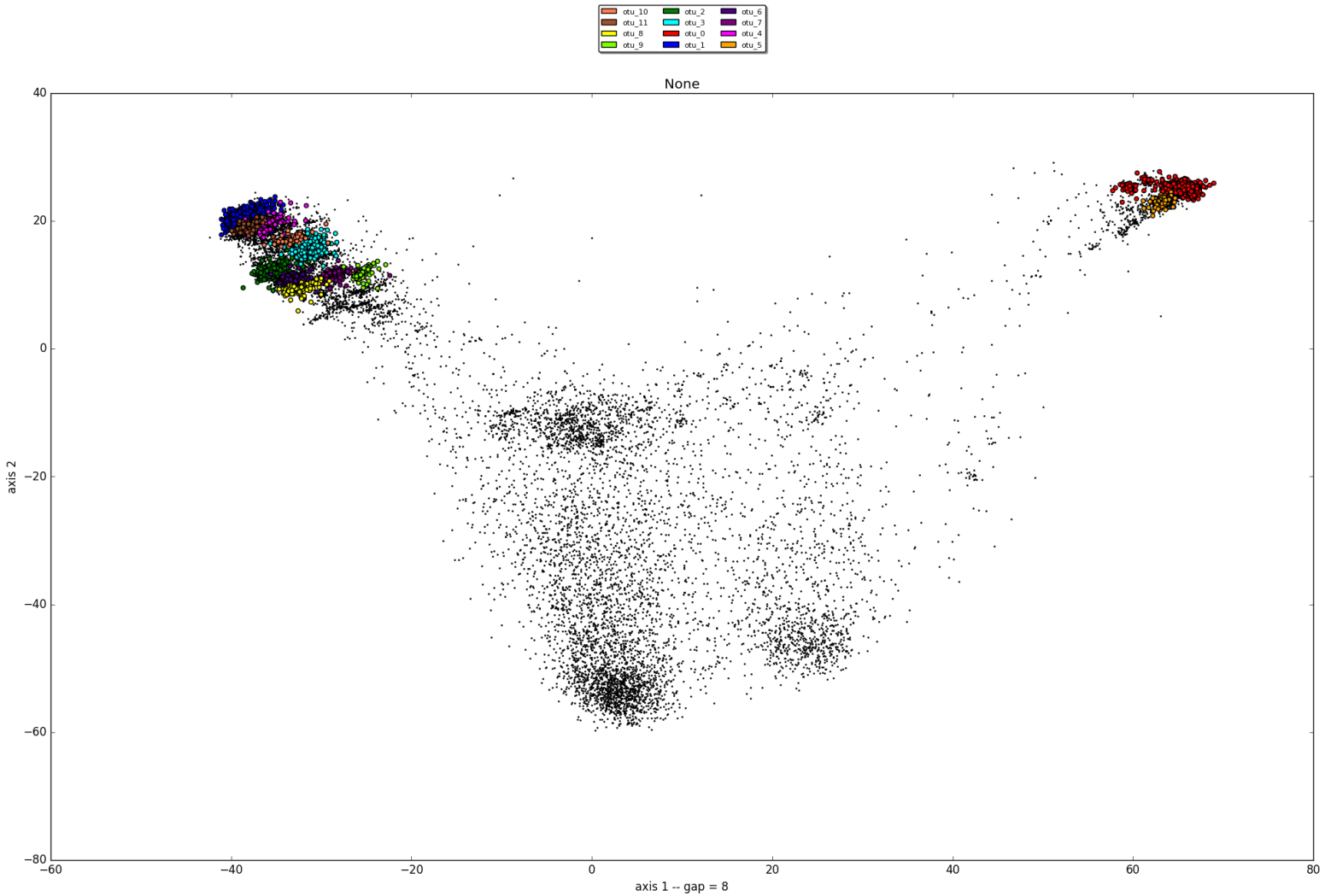
[1] Données : A. Chenuil & A. de Jode, IMBE, Marseille

Coordonnées parallèles

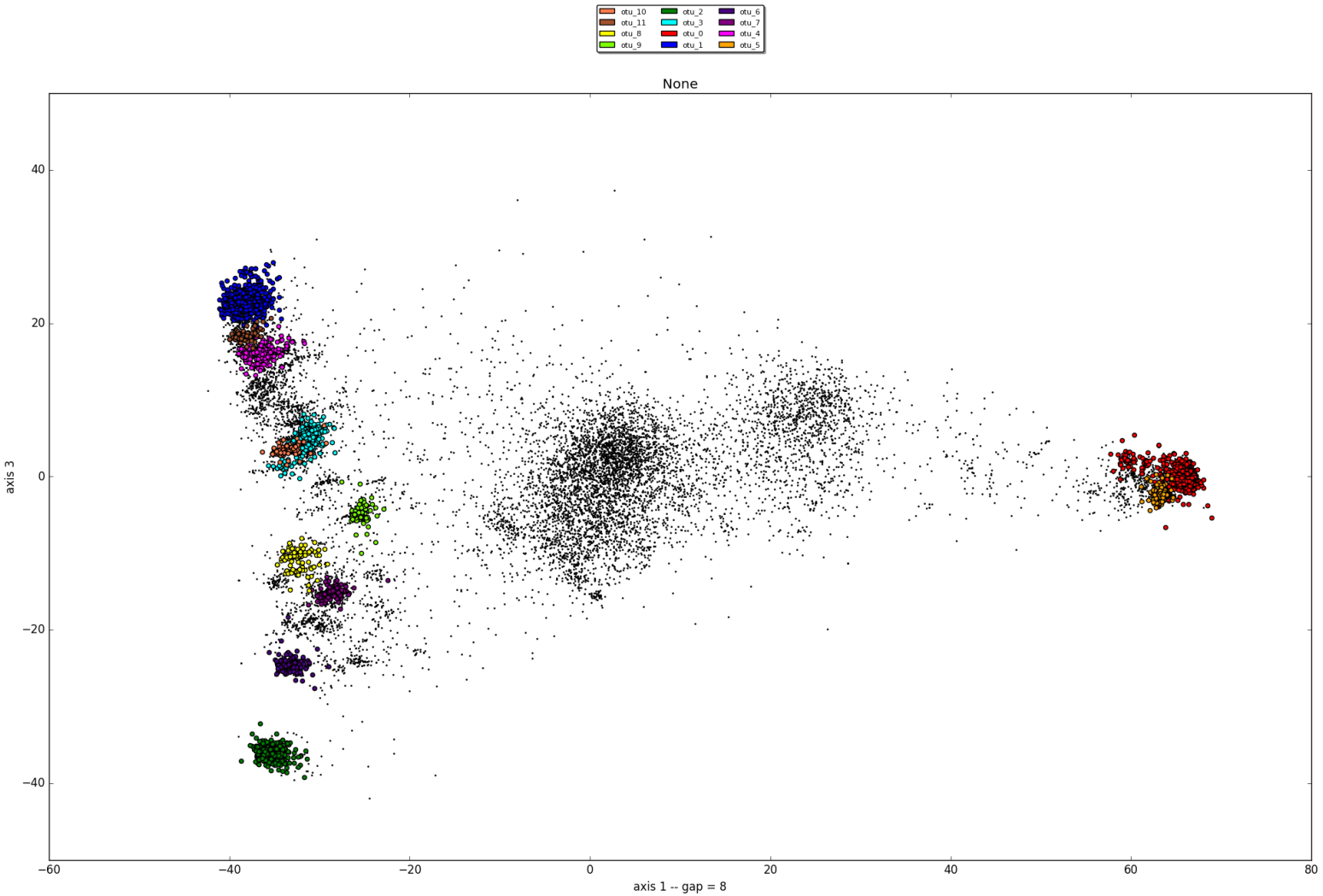
sur les 20 premiers axes de la MDS



Densité des points



Densité des points



Méthodes non linéaires

- Méthode spectrale (linéaire) $\longrightarrow$ ACP
- Méthodes non linéaires : foisonnement depuis les années 1990 ...
  1. ACP avec noyau (Kernel PCA)
  2. Manifold learning
  3. Isomap
  4. Laplacian eigenmaps
  5. ...

En gros ...

1. travailler de façon linéaire dans un espace de plus grande dimension ...
2. approcher une variété par une collection d'espaces tangents ...
3. naviguer d'un point à un point voisin par une promenade aléatoire ..

Méthode *Isomap*

Tennebaum J. B. & *al.*, 2000, *Science*, **290**:2319–2323

On se donne un ensemble de distances $d(i,j)$ entre certaines paires de points (ou tous!).

Algorithm 8 pseudocode for Isomap method

1: Have $G = (V, E)$, with weights $w(i,j) = d(i,j)$ on E
2: Select k nearest neighbors of each vertex $i \in V$
3: Build the graph $G' = (V, E')$ induced by edges of k nearest neighbors
4: Compute the shortest path in G' for any pair of vertices (Dijkstra, Floyd)
5: Build D: the pairwise distance matrix with these shortest path lengths
6: Run MDS with D

Bilan:

1. des faiblesses connues: instabilité topologique, sensible à la non convexité de la variété, ...
2. mais plusieurs réussites dans plusieurs domaines ...

Bibliographie sommaire - les classiques ...

- T. W. Anderson. An introduction to Multivariate Statistical Analysis. John Wiley & Sons, 1958.
- T.F. Cox et M. A. A. Cox. Multidimensional Scaling - Second edition, volume 88 of Monographs on Statistics and Applied Probability. Chapman & al., 2001.
- F. Cailliez et J.-P. Pagès. Introduction à l'Analyse des Données. S.M.A.S.H. Editions, 1979.
- R. Gittins. Canonical Analysis: A Review with Applications in Ecology, volume 12 of Biomathematics. Springer, 1985.
- M. Greenacre. Theory and Applications of Correspondence Analysis. Academic Press, 1984.
- A. J. Izenman. Modern Multivariate Statistical Techniques. Springer, NY, 2008.

Bibliographie sommaire (suite)

- L. Lebart, A. Morineau, et J.-P. Fénelon. Traitement des données statistiques. Dunod, Paris, 1982.
- J. A. Lee and M. Verleysen. Nonlinear Dimensionality Reduction. Springer, NY, 2007.
- K. V. Mardia, J.T. Kent, et J. M. Bibby. Multivariate Analysis. Probability and Mathematical Statistics. Academic Press,, 1979.
- C. R. Rao. The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhya, 26(4):329–368, 1964.
- C. R. Rao. Linear sstatistical Inference and its Applications. Wiley Series in Probability and Mathematical Statistics. Wiley, second edition, 1973.
- J. Wang. Geometric structure of high-dimensional data and dimensionality reduction. Springer & Higher Education Press, 2012.