Apprentissage d'une variété fonctionnelle, Application au clustering de courbes

Benjamin Auder

CEA - UPMC

2 septembre 2009

Thèse depuis 02/2008

Directeur de thèse : Gérard Biau (UPMC)

Encadrant CEA : Bertrand Iooss (CEA)

Plan

Contexte industriel CEA

Choc thermique pressurisé

Code thermo-hydraulique coûteux en temps, déterminant les évolutions temporelles de paramètres physiques dans l'espace annulaire de la cuve.

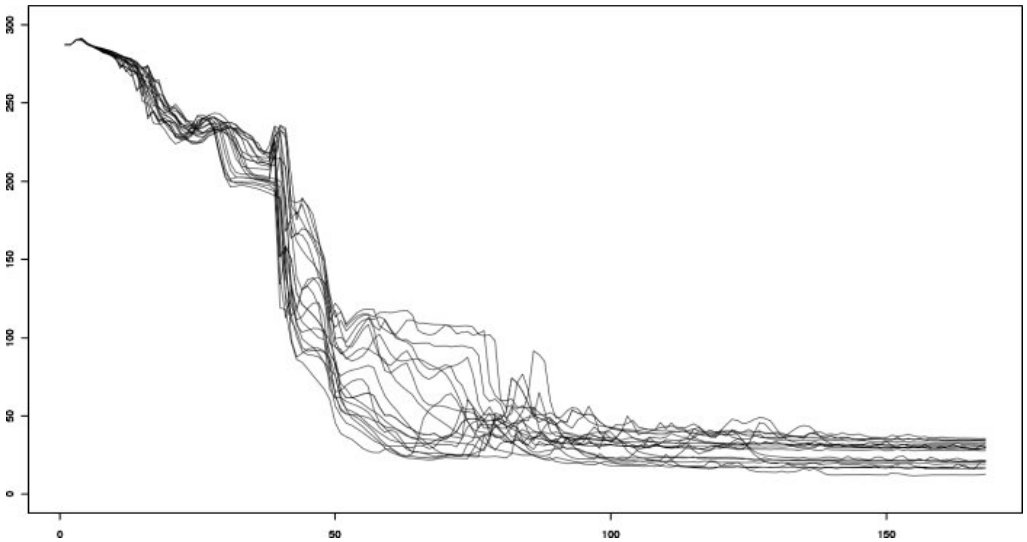


FIG.: Transitoires de température.

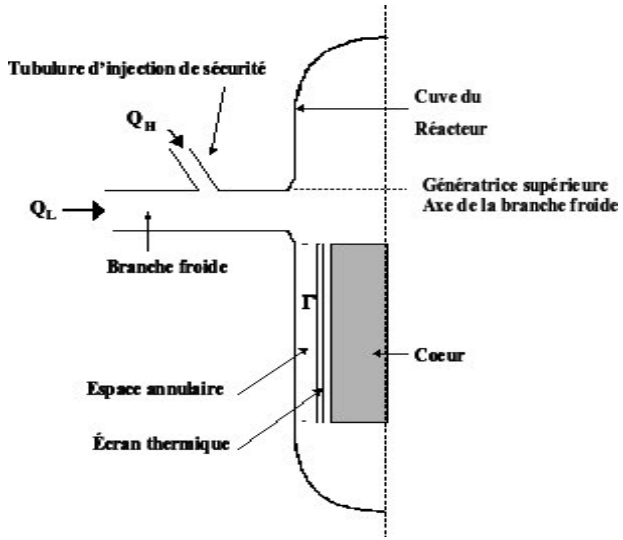


FIG.: Zone modélisée

Code de calcul Cathare :

- Entrées $z \in \mathbb{R}^p$ = état initial du système physique ;
- Sorties $y \in \mathcal{C}([a,b],\mathbb{R})$ = évolution des paramètres du système.

Résultats attendus

Code de calcul entrées vectorielles et sorties fonctionnelles.

$$\begin{pmatrix} z_{11} & \dots & z_{1p} \\ \vdots & \vdots & \vdots \\ z_{n1} & \dots & z_{np} \end{pmatrix} \longrightarrow \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix} = \begin{pmatrix} y_1(t_1) & \dots & y_1(t_D) \\ \vdots & \vdots & \vdots \\ y_n(t_1) & \dots & y_n(t_D) \end{pmatrix}$$

$i = 1..N$, $N \simeq 100\ 1000$; $z_{ij} \in \mathbb{R}$, $t \in [a, b]$.

Objectif : prédiction de données fonctionnelles via un métamodèle :

$$y^{\text{new}} \simeq \varphi(z^{\text{new}}) .$$

Résultats attendus

Code de calcul entrées vectorielles et sorties fonctionnelles.

$$\begin{pmatrix} z_{11} & \dots & z_{1p} \\ \vdots & \vdots & \vdots \\ z_{n1} & \dots & z_{np} \end{pmatrix} \longrightarrow \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix} = \begin{pmatrix} y_1(t_1) & \dots & y_1(t_D) \\ \vdots & \vdots & \vdots \\ y_n(t_1) & \dots & y_n(t_D) \end{pmatrix}$$

$i = 1..N$, $N \simeq 100\ 1000$; $z_{ij} \in \mathbb{R}$, $t \in [a, b]$.

Objectif : prédiction de données fonctionnelles via un métamodèle :

$$y^{\text{new}} \simeq \varphi(z^{\text{new}}) .$$

Sous-objectifs

→ Réduction de la dimension en sortie.
→ Clustering des entrées / sorties.

Plan

Illustration

Données "non linéaires", mais structurées en une variété (au moins) $\mathcal{C}^0$.

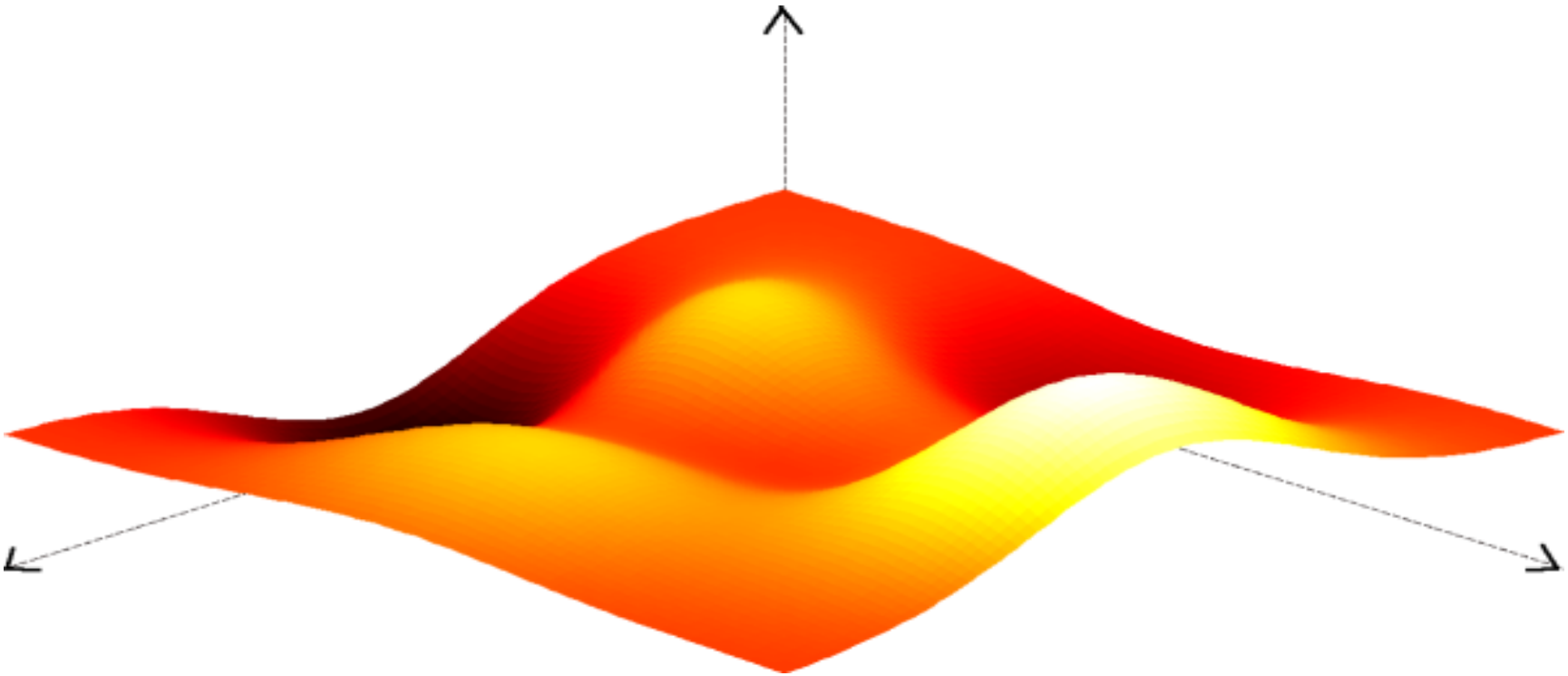


FIG.: Surface de dimension deux dans $\mathbb{R}^3$.

But : trouver un système de coordonnées le plus réduit possible pour décrire efficacement les données.

Objectifs

Paramétrer $\mathcal{Y}$, ensemble des sorties du code :
$f(x \in \mathbb{R}^d) = y \in \mathcal{Y} \subset \mathcal{C}([a,b],\mathbb{R})$, d le plus petit possible.
En pratique : N échantillons $y_i \Rightarrow N$ vecteurs $x_i = f^{-1}(y_i)$ à déterminer.

Objectifs

Paramétrer $\mathcal{Y}$, ensemble des sorties du code :
$f(x \in \mathbb{R}^d) = y \in \mathcal{Y} \subset \mathcal{C}([a,b],\mathbb{R})$, d le plus petit possible.
En pratique : N échantillons $y_i \Rightarrow N$ vecteurs $x_i = f^{-1}(y_i)$ à déterminer.

Contraintes :

- conservation des voisinages :
  les voisins de x_i correspondent à ceux de $y_i = f(x_i)$ $(k \in \mathbb{N}^*)$;
- conservation des distances :
  $f(x_i) = y_i$ et $f(x_j) = y_j \Rightarrow \|x_i - x_j\| \simeq \|y_i - y_j\|$ (..etc)

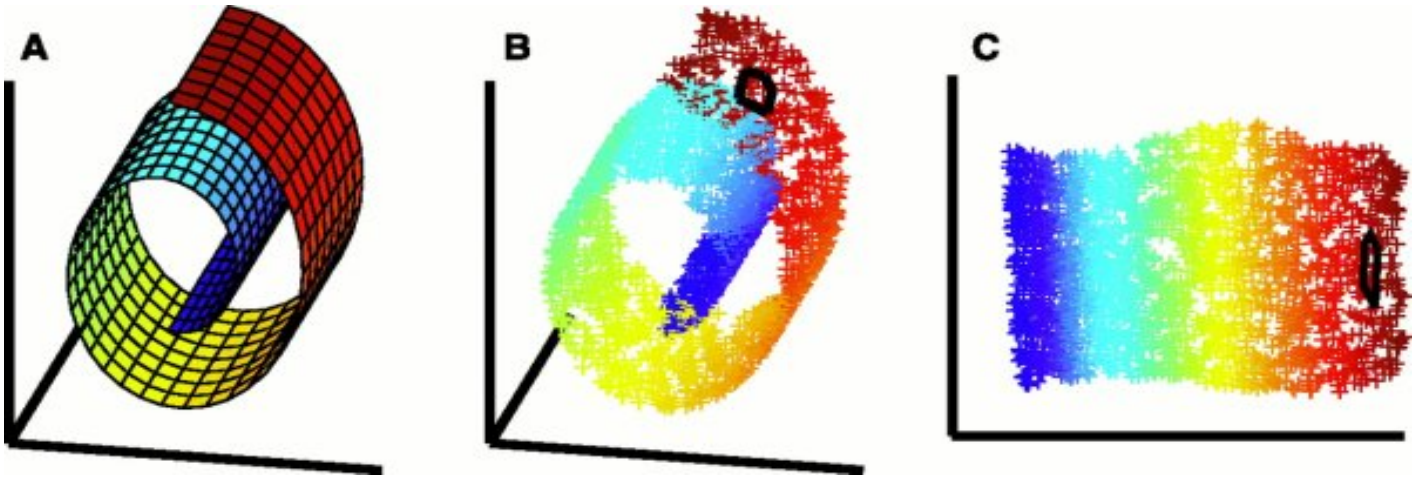


FIG.: carte 2D du jeu de données swissroll

Méthode

Recherche de la vraie dimension $\Rightarrow$ représentation non linéaire,

distances euclidiennes $\leftarrow$ distances géodésiques.

1. Estimation de la géométrie locale : graphe de voisinage.
2. Estimation de la dimension : basée sur $\mathbb{P}(Y \in B(y, r)) \propto r^d$.
3. Représentation en coordonnées globales.

Méthode

Recherche de la vraie dimension $\Rightarrow$ représentation non linéaire,

distances euclidiennes $\leftarrow$ distances géodésiques.

1. Estimation de la géométrie locale : graphe de voisinage.
2. Estimation de la dimension : basée sur $\mathbb{P}(Y \in B(y, r)) \propto r^d$.
3. Représentation en coordonnées globales.

FIG.: Exemple : un graphe des 6 plus proches voisins.

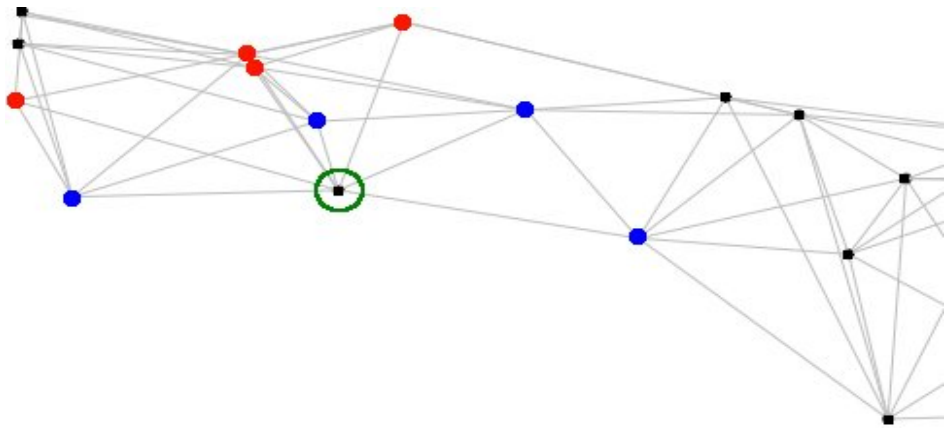
Détermination des voisinages

Méthode utilisée par T. Lin & H. Zha (articles 2006 et 2008).

Définition : visibilité depuis un noeud

v est voisin de p si aucun autre point r ne vérifie à la fois

$$\|r - p\| < \|v - p\| \text{ et}$$
$$\langle p - r, v - r \rangle < 0.$$



bleu : points testés et acceptés
rouge : points testés et refusés

FIG.: Exemple : voisinage du sommet entouré en vert.

Plan

Description (J. B. Tenenbaum et al., 2000)

Étape 1 : estimer toutes les distances géodésiques $d_{ij} = d(y_i, y_j)$.

Théorème

$D = (d_{ij})_{i,j=1..n}$ est une matrice de distances euclidiennes ssi.

$B = -\frac{1}{2}HDH$ est semi définie positive, avec $H = I - \frac{1}{n}\mathbb{1}^t\mathbb{1}$.

Dans ce cas B est la matrice de Gram associée à $x_1, \ldots, x_n$ centrés, représentant les y_i.

Description (J. B. Tenenbaum et al., 2000)

Étape 1 : estimer toutes les distances géodésiques $d_{ij} = d(y_i, y_j)$.

Théorème

$D = (d_{ij})_{i,j=1..n}$ est une matrice de distances euclidiennes ssi.

$B = -\frac{1}{2}HDH$ est semi définie positive, avec $H = I - \frac{1}{n}\mathbb{1}{}^t\mathbb{1}$.

Dans ce cas B est la matrice de Gram associée à $x_1, \ldots, x_n$ centrés, représentant les y_i.

Étape 2 : rendre B semi définie positive en annulant ses valeurs propres négatives :

$$B \simeq U\Lambda_+{}^tU\,.$$

Description (J. B. Tenenbaum et al., 2000)

Étape 1 : estimer toutes les distances géodésiques $d_{ij} = d(y_i, y_j)$.

Théorème

$D = (d_{ij})_{i,j=1..n}$ est une matrice de distances euclidiennes ssi.
$B = -\frac{1}{2}HDH$ est semi définie positive, avec $H = I - \frac{1}{n}\mathbb{1}{}^t\mathbb{1}$.
Dans ce cas B est la matrice de Gram associée à $x_1, \ldots, x_n$ centrés, représentant les y_i.

Étape 2 : rendre B semi définie positive en annulant ses valeurs propres négatives :

$$B \simeq U\Lambda_+{}^tU\,.$$

Étape 3 : calculer les nouvelles coordonnées x_i en se limitant à d colonnes :

$$X = U\Lambda_+^{\frac{1}{2}}\,.$$

Propriétés

Sous les conditions 1 à 3, Isomap converge vers la paramétrisation optimale des n points en d dimensions :

1. la variété $\mathcal{Y}$ est isométrique à un sous-ensemble de R^D, $D \in \mathbb{N}^*$;
2. l'espace de paramétrisation de $\mathcal{Y}$ est convexe ;
3. $\mathcal{Y}$ est compacte et bien échantillonnée partout.

Bilan

Conditions 1 et 2 très restrictives, souvent non vérifiées en pratique, mais l'algorithme reste utilisable et peut donner de bons résultats sans 1 et 2.

Propriétés

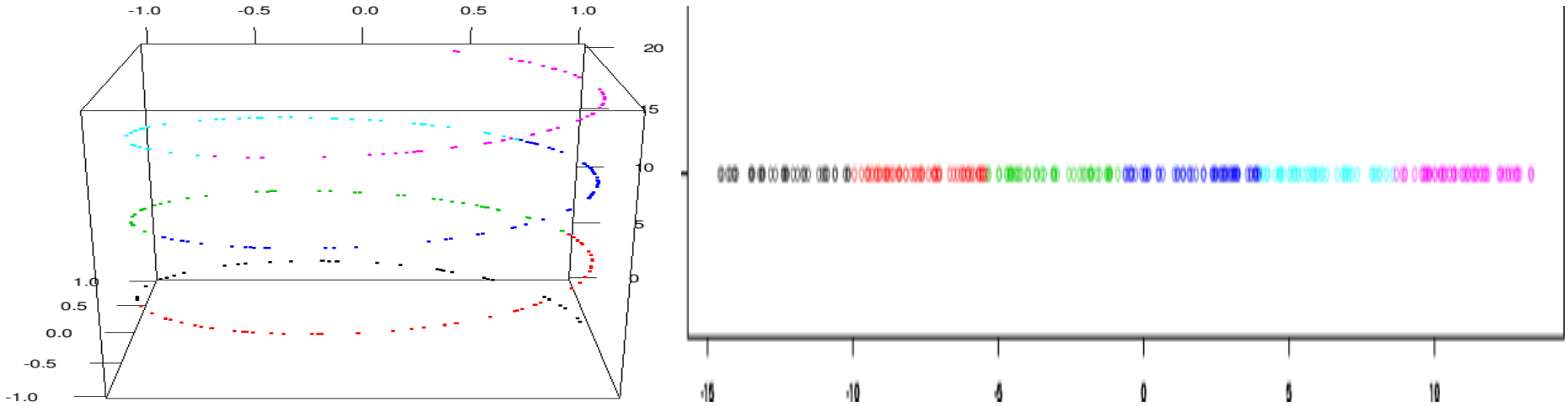
Sous les conditions 1 à 3, Isomap converge vers la paramétrisation optimale des n points en d dimensions :

1. la variété $\mathcal{Y}$ est isométrique à un sous-ensemble de R^D, $D \in \mathbb{N}^*$;
2. l'espace de paramétrisation de $\mathcal{Y}$ est convexe ;
3. $\mathcal{Y}$ est compacte et bien échantillonnée partout.

Bilan

Conditions 1 et 2 très restrictives, souvent non vérifiées en pratique, mais l'algorithme reste utilisable et peut donner de bons résultats sans 1 et 2.

Exemple : hélice 3D.



Plan

Description (M. Belkin & P. Niyogi, 2002)

W_{ij} = similarité entre y_i et y_j, p. ex. $W_{ij} = e^{-\frac{\|y_i - y_j\|^2}{\sigma^2}}$.

Définitions

D = matrice diagonale des degrés avec $D_{ii} = \sum_{j \in V(i)} W_{ij}$.

$L = D - W$, laplacien du graphe.

Description (M. Belkin & P. Niyogi, 2002)

W_{ij} = similarité entre y_i et y_j, p. ex. $W_{ij} = e^{-\frac{\|y_i - y_j\|^2}{\sigma^2}}$.

Définitions

D = matrice diagonale des degrés avec $D_{ii} = \sum_{j \in V(i)} W_{ij}$.

$L = D - W$, laplacien du graphe.

Fonction objectif naturelle à minimiser :

$$\psi(X) = \sum_{i,j=1}^{n} W_{ij} \|x_i - x_j\|^2 ,$$

sous la contrainte ${}^tXDX = 1$, avec $x_i \in \mathbb{R}^d$, en lignes dans X.

$\Rightarrow$ deux éléments similaires doivent être proches.

Solution au problème de minimisation :

$X = d$ premiers vecteurs propres de $D^{-1}L$ en colonnes.

Choix de σ.

σ déterminé localement en y_0, maximisant l'écart de similarité entre le voisin v_1 (resp. v_k) le plus proche (resp. le plus éloigné) de y_0 :

$$\sigma^2 = \arg\max_{\sigma^2} \left\{ e^{\frac{-\|y_0 - v_1\|^2}{\sigma^2}} - e^{\frac{-\|y_0 - v_k\|^2}{\sigma^2}} \right\} .$$

Après calculs :

$$\sigma^2 = \frac{\|y_0 - v_k\|^2 - \|y_0 - v_1\|^2}{\ln \|y_0 - v_k\|^2 - \ln \|y_0 - v_1\|^2} .$$

Choix de σ.

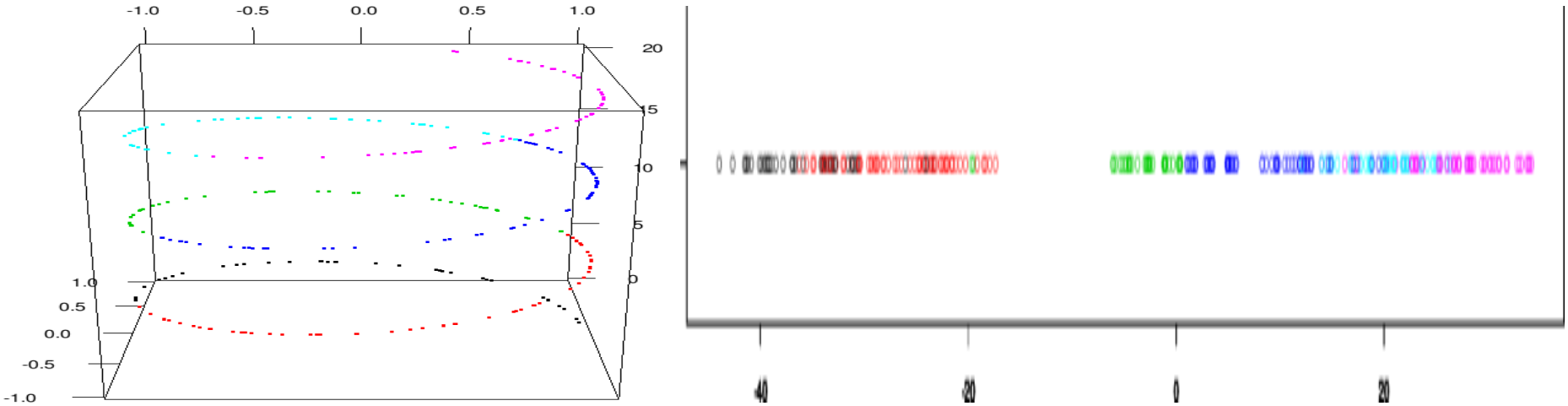
σ déterminé localement en y_0, maximisant l'écart de similarité entre le voisin v_1 (resp. v_k) le plus proche (resp. le plus éloigné) de y_0 :

$$\sigma^2 = \arg\max_{\sigma^2} \left\{ e^{\frac{-\|y_0 - v_1\|^2}{\sigma^2}} - e^{\frac{-\|y_0 - v_k\|^2}{\sigma^2}} \right\} .$$

Après calculs :

$$\sigma^2 = \frac{\|y_0 - v_k\|^2 - \|y_0 - v_1\|^2}{\ln \|y_0 - v_k\|^2 - \ln \|y_0 - v_1\|^2} .$$

Exemple : hélice 3D.



Plan

Description (T. Lin & H. Zha, 2006)

Tentative d'unifier des propriétés de distances globales, en respectant aussi les voisinages.

Premières étapes :

1. choisir un point origine y_0 parmi les y_i, (p.ex. le centre) ;
2. déterminer une base locale $Q_0 = (e_1, \ldots, e_d)$ de l'espace tangent en y_0 (avec les points du voisinage + SVD) ;

Description (T. Lin & H. Zha, 2006)

Tentative d'unifier des propriétés de distances globales, en respectant aussi les voisinages.

Premières étapes :

1. choisir un point origine y_0 parmi les y_i, (p.ex. le centre) ;
2. déterminer une base locale $Q_0 = (e_1, \ldots, e_d)$ de l'espace tangent en y_0 (avec les points du voisinage + SVD) ;
3. calculer les coordonnées de tous les voisins de y_0 en projection sur la base Q_0 ; un voisin y a pour coordonnées

$$x = \arg \min_{x_1,\ldots,x_d} \left\| y - \left(y_0 + \sum_{i=1}^{d} x_i e_i \right) \right\|^2 ,$$

renormalisées pour vérifier $\|y - y_0\| = \|x - x_0\|$.

Coordonnées des non voisins de y_0

Étape 4 : pour y non voisin de y_0, on cherche y_p le prédecesseur de y sur un plus court chemin issu de y_0 (Dijkstra p.ex.).

$y_{i_1}, \ldots, y_{i_d}$ sont les voisins déjà traités de y_p (parcours des y_i en largeur).

$\rightarrow$ On cherche alors x coordonnées de y, telles que les angles $\widehat{yy_py_{i_j}}$ soient $\simeq$ conservés :

$$\cos\theta = \frac{\langle y - y_p, y_{i_j} - y_p \rangle}{\|y - y_p\| \|y_{i_j} - y_p\|} \simeq \frac{\langle x - x_p, x_{i_j} - x_p \rangle}{\|x - x_p\| \|x_{i_j} - x_p\|} = \cos\theta' \,,$$

sous la contrainte $\|y - y_p\| = \|x - x_p\|$.

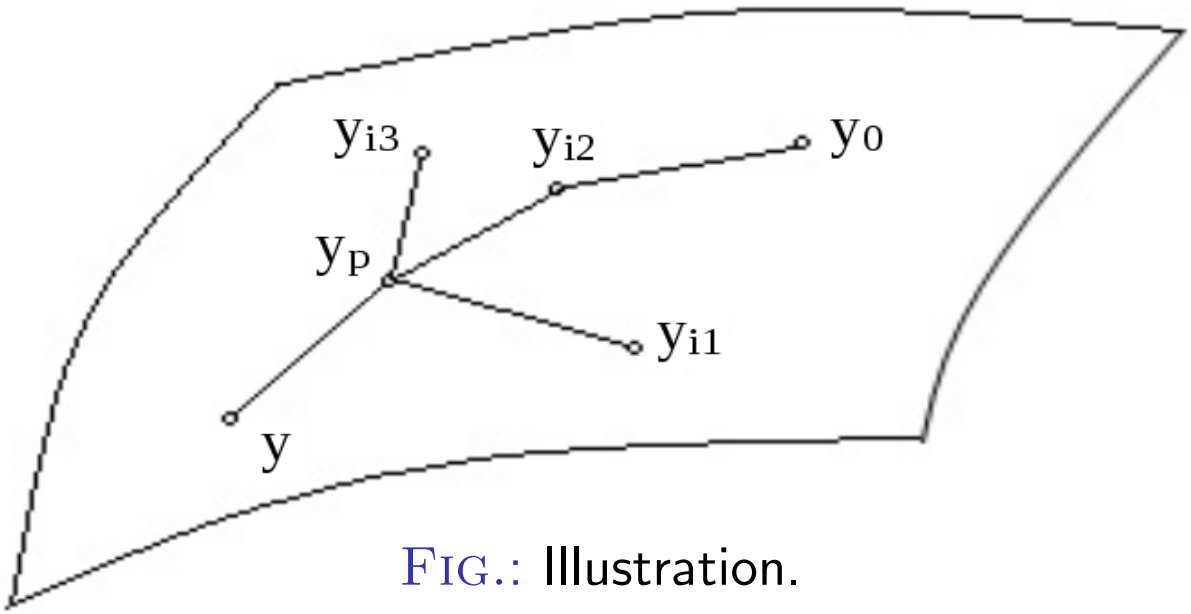


FIG.: Illustration.

Plan

Exemple analytique, 2 clusters (dimension 2)

Pour tous les exemples, $d = 2$, $N = 600$.

Pour les deux premiers, ajout d'un léger bruit gaussien.

Fonction définie sur $[0, 2\pi]$:

$$f_{\alpha,\beta,\gamma,\delta} : x \to \alpha \cos x + \beta \sin x + \gamma \cos 2x + \delta \sin 2x,$$

avec $(\alpha, \beta) \sim \mathcal{U}(\mathcal{S}(0,1)_+)$. $(\gamma, \delta) \sim \mathcal{U}(\mathcal{S}(0,1)_+)$ pour les courbes 1 à 300, puis $(\gamma, \delta) \sim \mathcal{U}(\mathcal{S}(0,2)_+)$ pour les 300 suivantes ($\mathcal{S}_+ = \mathcal{S} \cap \mathbb{R}^2_+$).

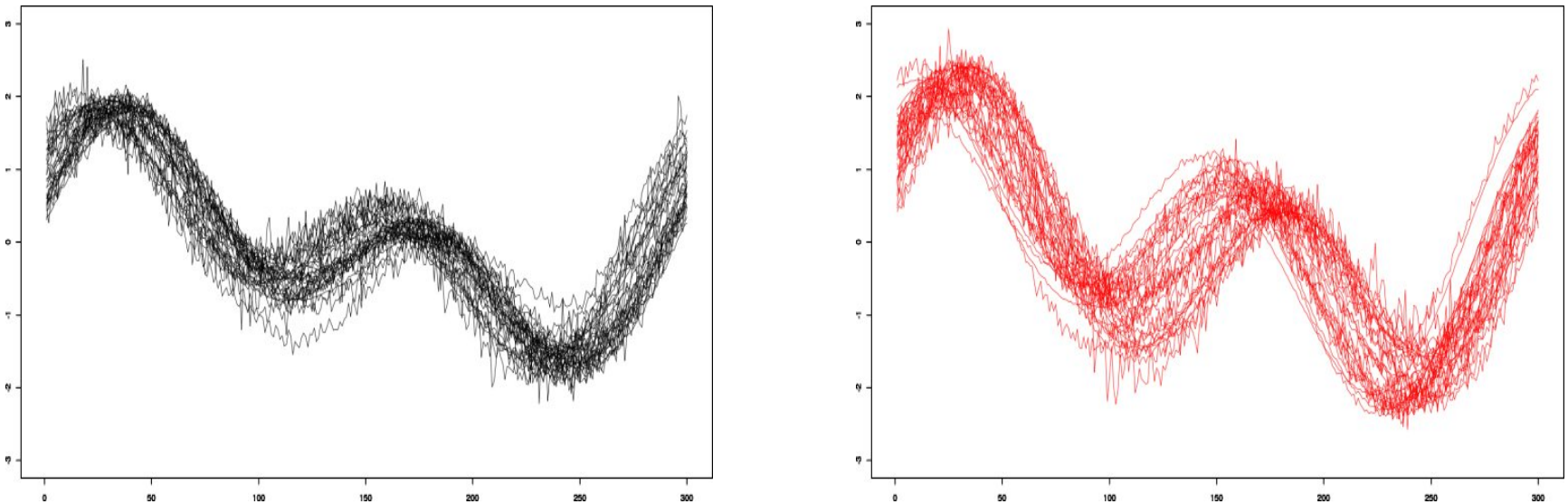


FIG.: 32 courbes du cluster $i = 1..300$ à gauche, $i = 301..600$ à droite.

Résultats

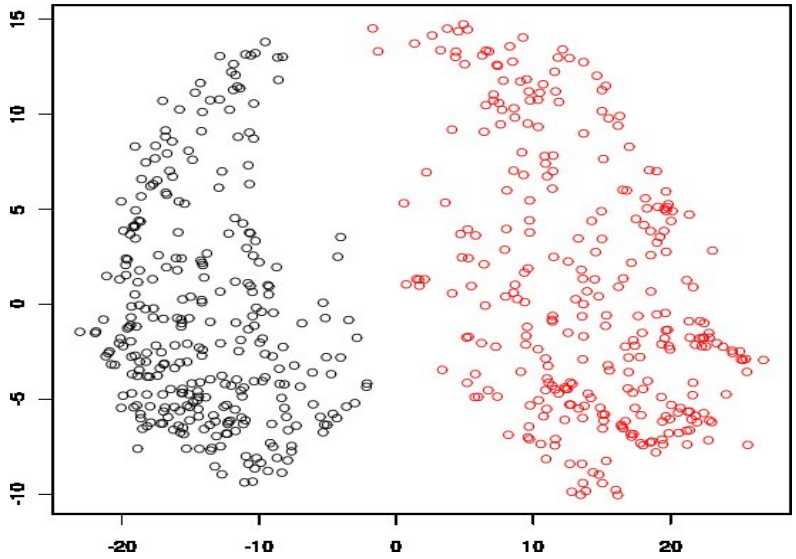


FIG.: Isomap.

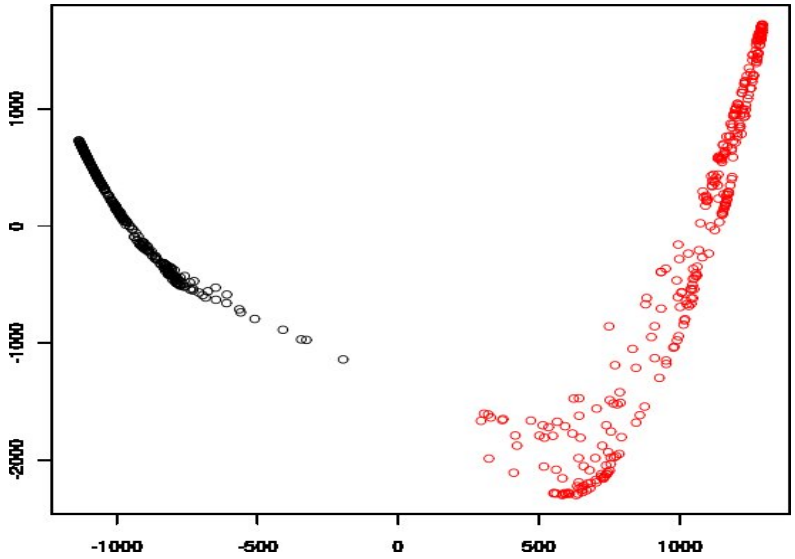


FIG.: Laplacian eigenmaps.

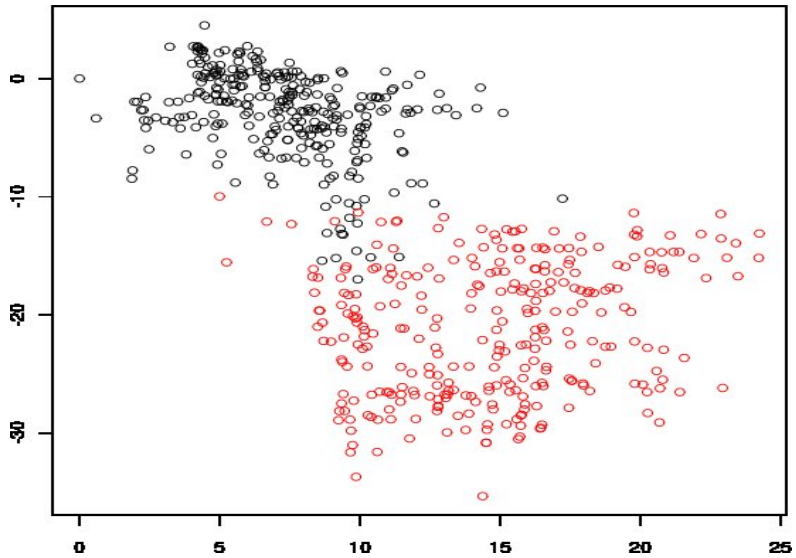


FIG.: RML.

Méthode	Réussite
K-Means	51%
Hiérarchique Ward	51%
Clustering spectral	100%
Isomap + k-means	99%
Lap. eig. + k-means	100%
RML + k-means	97%

FIG.: Homogénéité des clusters.

Plan

Exemple analytique, 3 clusters (dimension 1)

Fonction définie sur $[1, 5]$:

$$f_{\alpha,\beta,\gamma} : x \to (\frac{\sin \alpha x}{x} + e^{-\beta x}) \cos \gamma x \, ;$$

$\beta[1 : 200,] \sim \mathcal{U}(1, 2)$, $\beta[201 : 400,] \sim \mathcal{U}(0, 1)$, $\beta[401 : 600,] \sim \mathcal{U}(0.4, 1)$.

$\alpha = 3\beta$, et $\gamma[1 : 200,] = (4 - \beta[1 : 200,]^2)^{\frac{1}{2}}$,

$\gamma[201 : 400,] = (1 - \beta[201 : 400,]^2)^{\frac{1}{2}}$,

$\gamma[401 : 600,] = 3(1 - \beta[401 : 600,]^2)^{\frac{1}{2}} + 3$.

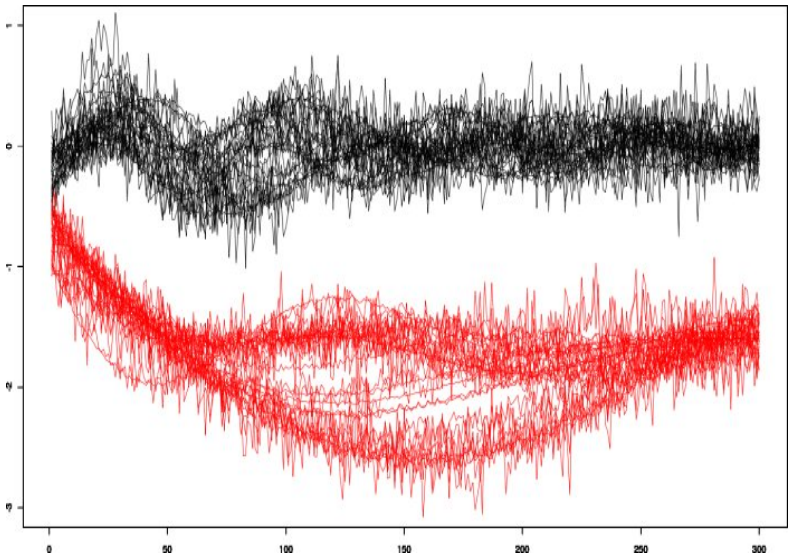


FIG.: 32 courbes des 2 1^{ers} clusters.

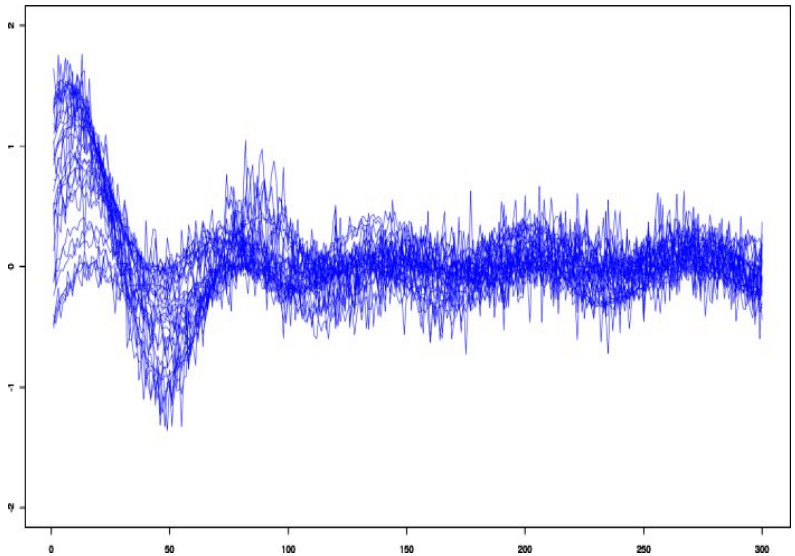


FIG.: 32 courbes du 3^{eme} cluster.

Résultats

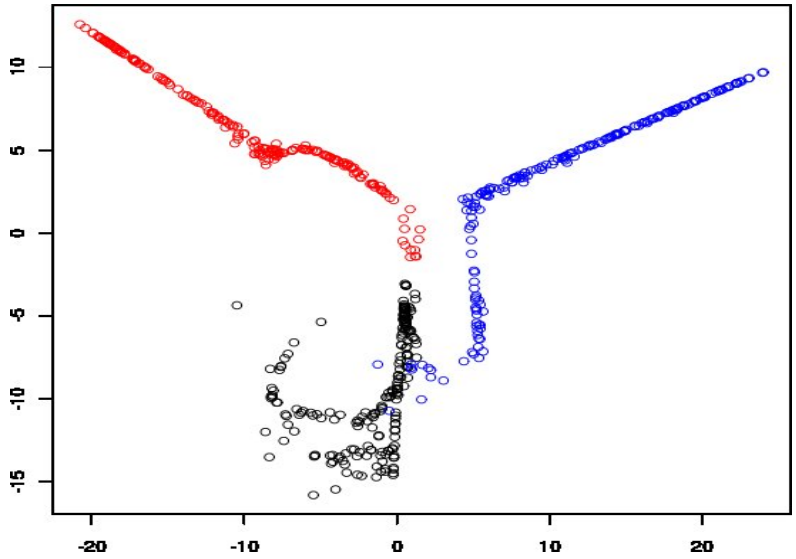


FIG.: Isomap.

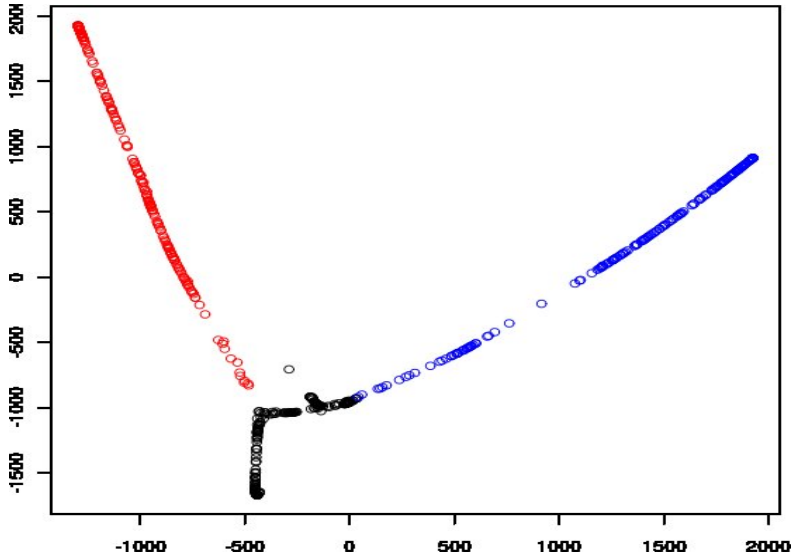


FIG.: Laplacian eigenmaps.

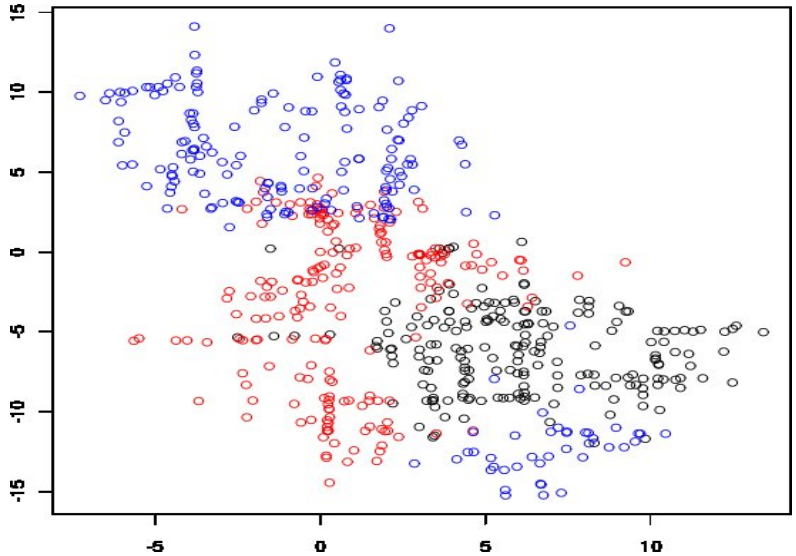


FIG.: RML.

Méthode	Réussite
K-Means	67%
Hiérarchique Ward	74%
Clustering spectral	98%
Isomap + k-means	90%
Lap. eig. + k-means	91%
RML + k-means	67%

FIG.: Homogénéité des clusters.

Plan

Exemple analytique, 6 clusters (dimension ∞)

Séries temporelles affichant les évolutions de variables physiques.

Fonction définie sur $[0, 60]$, D points de discrétisation. Génération :

- *cpt. normal* : $y(t) = m + rs$; $m = 30$, $s = 2$, $r \sim \mathcal{U}(-3, 3)$ (noir) ;
- *cpt. cyclique* : $y(t) = m + rs + a \sin \frac{2\pi t}{T}$ où $a, T \sim \mathcal{U}(10, 15)$ (rouge) ;
- *(dé)croissant* : $y(t) = m + rs \pm gt$; $g \sim \mathcal{U}(0.2, 0.5)$ (vert, bleu) ;
- *saut haut/bas* : $y(t) = m + rs \pm kx$; $x \sim \mathcal{U}(7.5, 20)$, $k = \mathbb{1}_{[t_0, D]}$. $t_0 \sim \mathcal{U}\left(\frac{D}{3}, \frac{2D}{3}\right)$ (bleu ciel, violet) ;

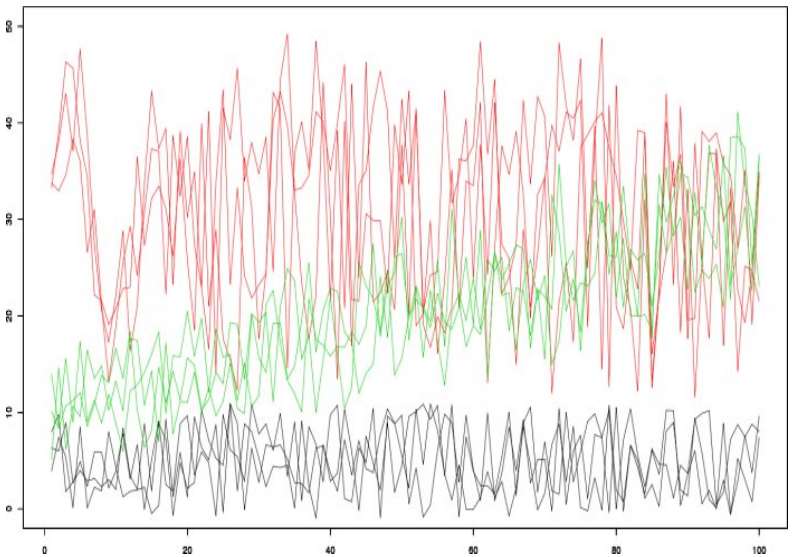


FIG.: 3 courbes des 3 1^{ers} clusters.

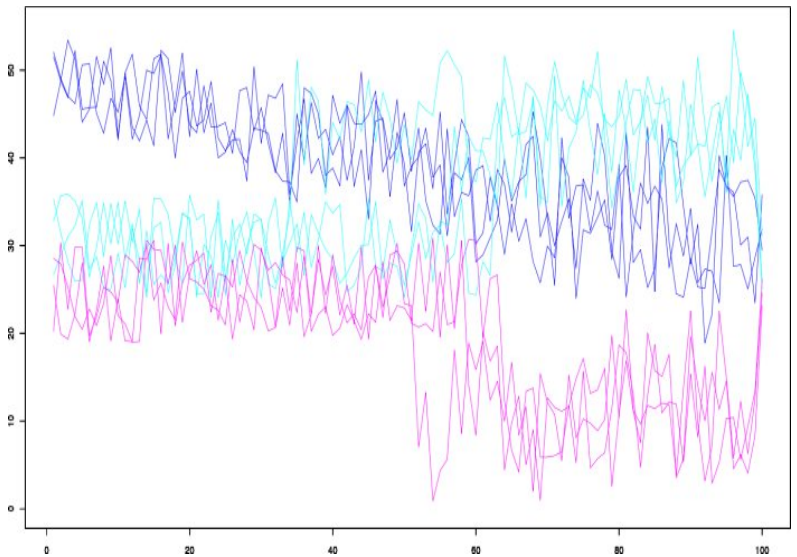


FIG.: 3 courbes des 3 derniers clusters.

Résultats

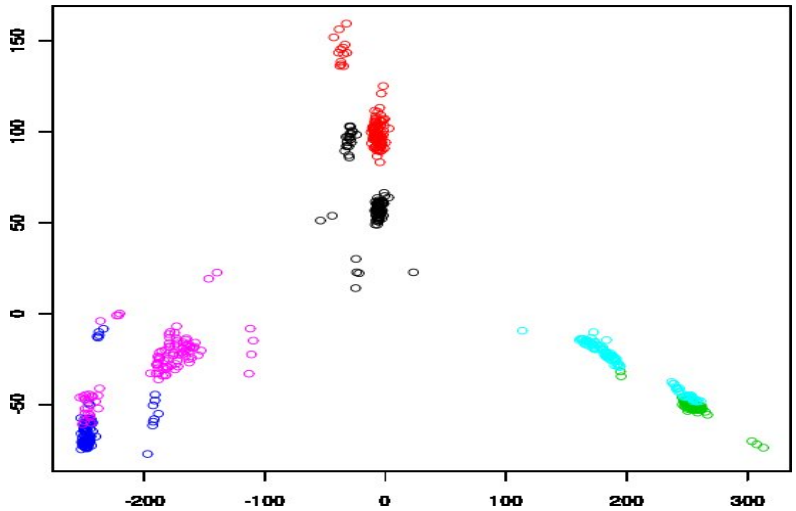


FIG.: Isomap.

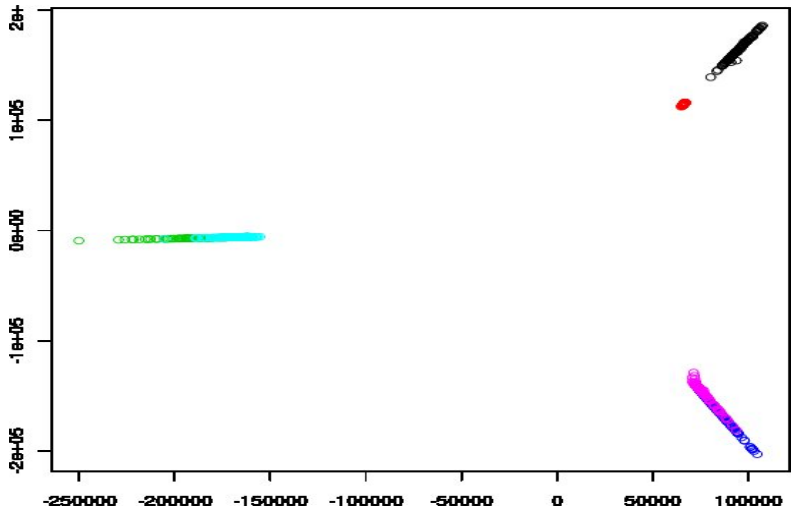


FIG.: Laplacian eigenmaps.

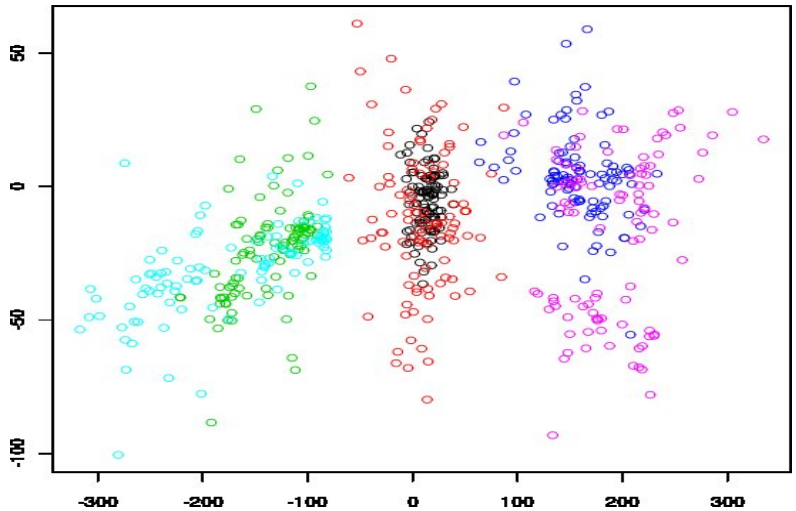


FIG.: RML.

Méthode	Réussite
K-Means	91%
Hiérarchique Ward	96%
Clustering spectral	94%
Isomap + k-means	86%
Lap. eig. + k-means	80%
RML + k-means	58%

FIG.: Homogénéité des clusters.

Note : isomap et $d = 4 \Rightarrow 92\%$, RML et $d = 8 \Rightarrow 75\%$

Exemple sur données réelles

Code Cathare (CEA) : évolution du coefficient d'échange fluide-paroi.

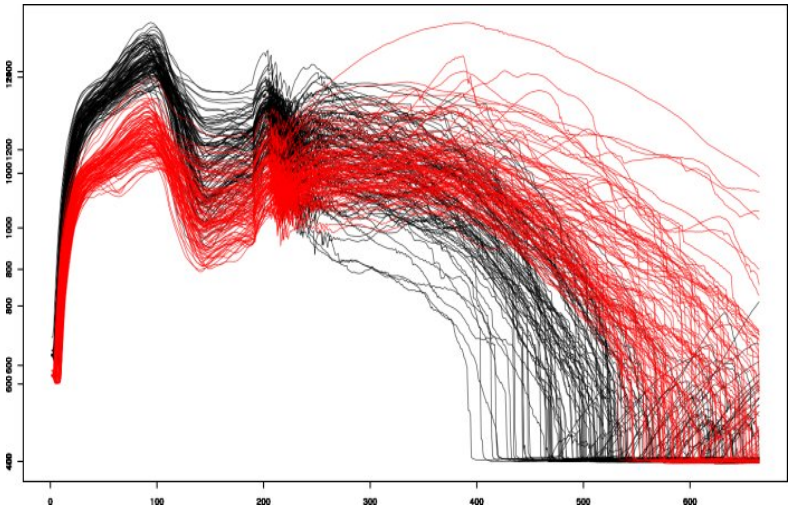


FIG.: Les 200 évolutions.

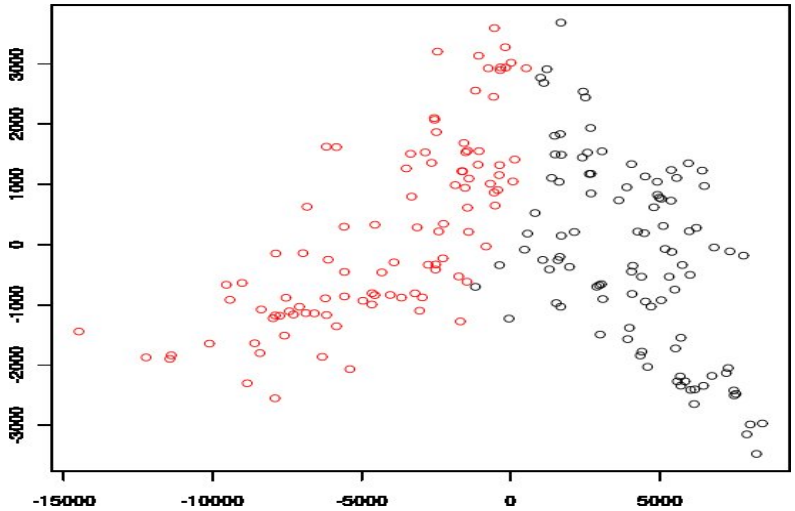


FIG.: Isomap.

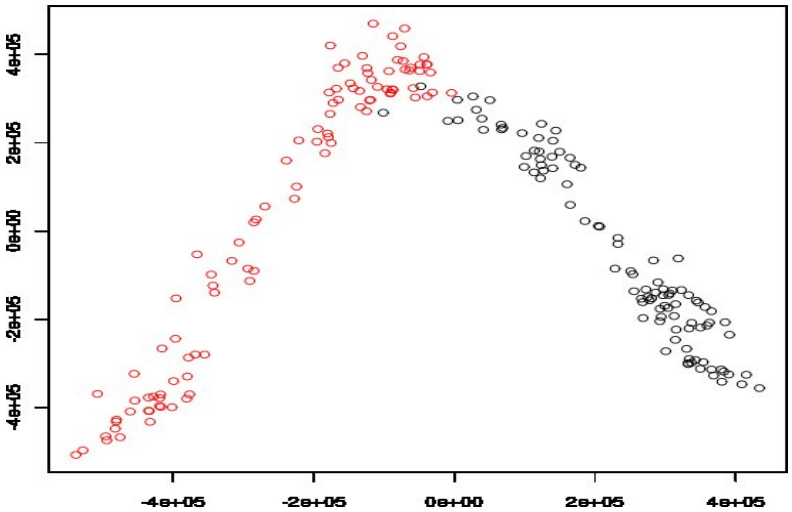


FIG.: Laplacian eigenmaps.

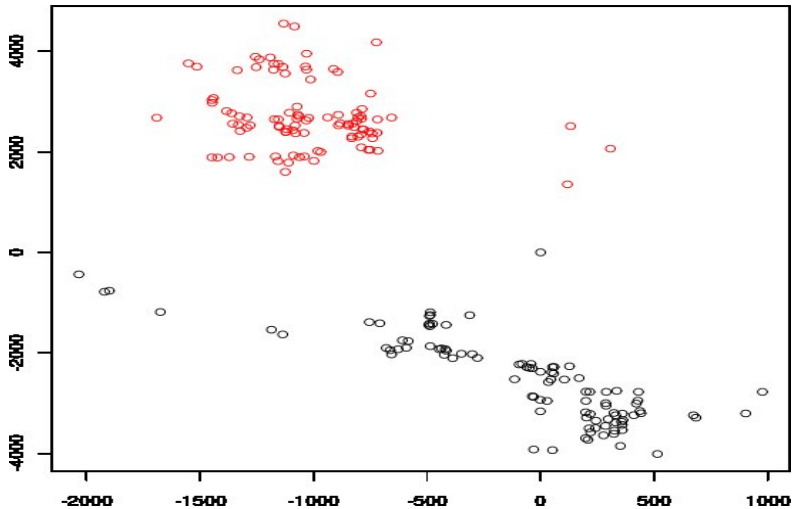


FIG.: RML.

Conclusion

- Isomap : très bons résultats (mais pas conçu pour cette tâche..).
- Laplacian eigenmaps : beaucoup plus orienté vers le clustering, comparable à Isomap en performance.

Algorithme RML à améliorer / mieux adapter au cadre fonctionnel.
Version actuelle : trop de chevauchements inter-classes.

Conclusion

- Isomap : très bons résultats (mais pas conçu pour cette tâche..).
- Laplacian eigenmaps : beaucoup plus orienté vers le clustering, comparable à Isomap en performance.

Algorithme RML à améliorer / mieux adapter au cadre fonctionnel.
Version actuelle : trop de chevauchements inter-classes.

Méthode "intermédiaire" à explorer : courbes principales (T. Hastie, 1984).

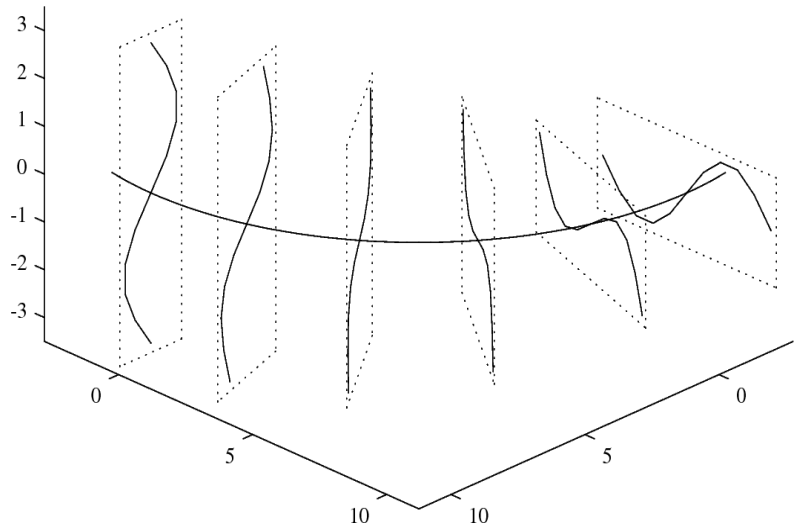


FIG.: Exemple de surface principale en $2D$.

..dans un espace de fonctions.

Bibliographie

Isomap : *A global geometric framework for nonlinear dimensionality reduction* ; J. B. Tenenbaum, V. de Silva & J. C. Langford (2000). in Science, vol. 290, pp. 2319-2323.

Lap. eig. : *Laplacian eigenmaps and spectral techniques for embedding and clustering* ; M. Belkin & P. Niyogi (2002).

RML : *Riemannian Manifold Learning for Nonlinear Dimensionality Reduction* ; T. Lin, H. Zha & S. U. Lee (2006), et *Riemannian Manifold Learning* ; T. Lin & H. Zha (2008) in IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30, pp. 796-809.

CCT time series : *Time-Series Similarity Queries Employing a Feature-Based Approach* ; R. J. Alcock & Y. Manolopoulos (1999).

Courbes principales : *Principal curves* ; T. Hastie & W. Stuetzle (1989). in Journal of the American Statistical Association, vol. 84, pp. 502-516. *Another look at principal curves and surfaces* ; P. Delicado (2001). in Journal of Multivariate Analysis, vol. 77, pp. 84-116 (..etc).