Reconstruction de variétés

Clément Berenfeld

3 novembre 2018

Résumé

La reconstruction de variétés concerne l'estimation des caractéristiques géométriques de la loi d'un échantillon aléatoire à valeur dans $\mathbb{R}^D$, lorsque celle-ci est supportée par une sous-variété de $\mathbb{R}^D$. L'objectif est multiple : il s'agit à la fois de mieux comprendre la loi sous-jacente (dimension, régularité, courbure, homologie, etc.) mais aussi de fournir aux statisticiens des descripteurs pertinants pour l'étude de données en haute dimension.

Table des matières

1 Introduction

Lorsque la dimension ambiante D d'un jeu de données $\mathcal{X}_n \subset \mathbb{R}^D$ devient très grande devant le nombre de mesures n, le bon fonctionnement des algorithmes d'apprentissage n'est plus garanti (Friedman [1997], Giraud [2014]). Ces dernières années ont pourtant vu exploser à la fois la dimensionnalité des données récoltées et l'efficacité des méthodes algorithmiques d'inférence et d'apprentissage. Face à cette contradiction apparente, certains ont été tentés de supposer qu'une grande partie des jeux de données en grande dimension vivent en réalité sur des structures de faible dimension intrinsèque (Fefferman et al. [2016]). La reconstruction de variété (*manifold learning* en anglais) part de l'hypothèse que les données sont effectivement supportées par des sous-variétés de $\mathbb{R}^D$, et essaie de développer les outils nécessaires à l'estimation de quantités interessantes liées à ces dernières.

Ces problèmes d'estimation dépendent intimement de l'echelle à laquelle on s'autorise à se placer pour regarder les données. Ce paramètre de résolution, que l'on appellera *reach*, et initialement introduit par Federer [1959], apparaitra dans tous les modèles que l'on considerera, et controlera à la fois les vitesses d'estimation ainsi que le niveau de bruit admissible dans nos données (pour lequel une estimation précise est encore possible). Puisque c'est un coefficient important, il semble indispensable de pouvoir l'estimer (Aamari et al. [2017]).

L'homologie apparaît également comme un descripteur pertinant du support du jeu de données. Determiner l'homologie d'ordre zero par exemple, qui correspond au nombre de composantes connexes, est à la base des techniques de *clustering* (Filippone et al. [2008]). Il peut-être également intéressant de regarder l'homologie d'ordre supérieur (comme par exemple en astrophysique, voir Van de Weygaert et al. [2011]). Le caractère naturellement instable de l'homologie et le bruit naturel observé au sein de jeux de données ont amené les statisticiens à utiliser la notion d'homologie persistante (Edelsbrunner and Harer [2008]) dont les principaux outils (diagramme, paysage de persistance) sont autant de représentations dont les statisticiens peuvent se servir dans leurs algorithmes (Chazal et al. [2013], Li et al. [2014]).

Le support $\mathcal{S}$ d'une loi de probabilité P est naturellement muni d'une distance $d_{\mathcal{S}}$ dite intrinsèque (ou encore riemanienne lorsque $\mathcal{S}$ est une sous-variété de $\mathbb{R}^D$). L'estimation de cette distance est au coeur de la technique de dimension de réduction Isomap (Tenenbaum et al. [2000]). La determination de cette distance va de paire avec celle des plus courts chemins de $\mathcal{S}$ (Arias-Castro and Le Gouic [2017]) et a des applications importantes en planification de trajectoires (Karaman and Frazzoli [2011]).

Nous nous intéresserons enfin à l'estimation du support $\mathcal{S}$ lorsque celui-ci est une sous-variété de de $\mathbb{R}^D$ (Genovese et al. [2012a], Genovese et al. [2012b]) ainsi qu'aux approximations d'ordre supérieur (plan tangent, seconde forme fondamentale,etc.) de ce support (Aamari and Levrard [2017]).

2 Cadre et notation

On fixe $D \geq 1$ et on note $\|\cdot\|$ et $\langle \cdot, \cdot \rangle$ la norme et le produit scalaire euclidiens usuels sur $\mathbb{R}^D$. On notera $d(\cdot, \cdot)$ la distance associée. Les symboles $\wedge$ et $\vee$ désignent respectivement le minimum et le maximum entre deux quantités. L'introduction d'une constante est toujours suivi des paramètres dont elle dépend entre parenthèses.

2.1 Sous-variétés et courbure

Une sous-variété M de $\mathbb{R}^D$ de dimension d et de classe C^k est un sous-ensemble de $\mathbb{R}^D$ vérifiant qu'en tout point $x \in M$, il existe un voisinage V de 0 dans $\mathbb{R}^d$ et un voisinage U de x dans $\mathbb{R}^D$ ainsi qu'une immersion propre $\phi : V \to \mathbb{R}^D$ de classe C^k telle que $\phi(V) = M \cap U$. On appellera une telle application une paramétrisation de M en x. À tout point x d'une telle sous-variété, on peut associer un plan tangent $T_x M$ défini par

$$T_x M = \{\dot{\gamma}(0) \;\; | \;\; \gamma :]-\epsilon, \epsilon[\to \mathbb{R} \text{ est de classe } C^1 \text{ avec } \gamma(0) = x\}.$$

C'est un sous-espace vectoriel de $\mathbb{R}^D$. La fibre normale de M en x est l'orthogonal dans $\mathbb{R}^D$ de $T_x M$, notée $T_x M^\top$. La seconde forme fondamentale de M en x (Gallot et al. [1990], p.185) est une application bilinéaire symmétrique

$$\mathrm{I\!I}_x : T_x M \times T_x M \to T_x M^\top$$

qui encode la manière dont se courbe la sous-variété M en x. Pour tout vecteur normal unitaire $\eta \in T_x M^\top$, l'application $\langle \mathrm{I\!I}_x, \eta \rangle$ est une forme symmétrique sur $T_x M$. L'endomorphisme associé $A_{x,\eta} : T_x M \to T_x M$ est appelé endomorphisme de forme de M en x dans la direction η. Ses valeurs propres sont appelées courbures principales de M en x dans la direction η. On notera

$$\mathrm{R}_{\mathrm{loc}}(M) = \inf \left\{ \frac{1}{\|A_{x,\eta}\|_{\mathrm{op}}} \;\middle|\; x \in M \text{ et } \eta \in T_x M^\top \text{ unitaire} \right\}.$$

C'est l'inverse de la plus grande courbure principale (en valeur absolue) de M.

2.2 Géodésiques et distance intrinsèque

Une géodésique de M est une courbe $\gamma : [a, b] \to M$ de classe C^2 paramétrée à vitesse constante telle que en tout point $\ddot{\gamma}(t)$ est normal à M en $\gamma(t)$ (Gallot et al. [1990], exemple 2.80). On peut montrer que pour tout $x \in M$ et tout $v \in T_x M$, il existe une unique géodésique maximale $\gamma_{x,v}$ définie sur un voisinage de 0 vérifiant $\gamma_{x,v}(0) = x$ et $\dot{\gamma}_{x,v}(0) = v$. Les géodésiques sont en effet solution de l'équation différentielle $\ddot{\gamma} = \mathrm{I\!I}(\dot{\gamma}, \dot{\gamma})$ et un argument de type Cauchy-Lipschitz permet de conclure. On obtient par la même occasion que $\|\ddot{\gamma}\| \leq \|\dot{\gamma}\|^2 / \mathrm{R}_{\mathrm{loc}}(M)$ pour toute géodésique de M. Pour un chemin C^1-par morceaux $\gamma : [a, b] \to M$, on définit sa longueur par

$$L(\gamma) = \int_a^b \|\dot{\gamma}(t)\| dt.$$

Pour x, y dans M, on définit la distance riemanienne de x à y comme l'infimum des longueurs de tout les chemins C^1 par morceaux à valeur dans M joignant x à y. On note cette quantité $d_M(x, y)$. On peut montrer que, lorsque M est connexe, cette quantité est atteinte (Burago et al. [2001], théorème 2.5.23) et que les chemins minimisant la longueur sont automatiquement des géodésiques (Lee [2006], théorème 6.6). En particulier, tous les résultats de régularité des géodésiques se transferent aux plus courts chemins. Réciproquement, une géodésique est toujours localement un plus court chemin. (Lee [2006], théorème 6.12).

2.3 Fonction distance et reach

Pour un compact $K \subset \mathbb{R}^D$, on définit la fonction distance à K par

$$d_K : \begin{cases} \mathbb{R}^D \to \mathbb{R}_+ \\ x \mapsto \min_{y \in K} \|x - y\| \end{cases}$$

et, pour tout $x \in \mathbb{R}^D$, on note $\Gamma_K(x)$ l'ensemble des points $y \in K$ tel que $\|x-y\| = d_K(x)$. C'est également un compact de $\mathbb{R}^D$. On note B_K l'unique boule fermée de $\mathbb{R}^D$ qui contient K et de rayon minimal. On designera par $\operatorname{rad} K$ son rayon et $\operatorname{center} K$ son centre. On définit alors le gradient généralisé de la fonction distance par

$$\forall x \in \mathbb{R}^D \setminus K, \quad \nabla d_K(x) = \frac{x - \mathrm{pr}_{\operatorname{hull} \Gamma_K(x)}(x)}{d_K(x)}$$

où hull désigne l'enveloppe convexe. Ce gradient coïncide avec le gradient usuel là où d_K est différentiable (pour les points x qui admette une unique projection sur K).

Un point critique pour la fonction d_K est un point x tel que $\nabla d_K(x) = 0$. La valeur $d_K(x)$ est alors appelée valeur critique. Le *weak feature size* de K, noté $\mathrm{R}_{\mathrm{glob}}(K)$ est alors défini comme la plus petite valeur critique de d_K

$$\mathrm{R}_{\mathrm{glob}}(K) = \inf\{d_K(x) \mid x \in \mathbb{R}^D \setminus K \text{ et } \nabla d_K(x) = 0\}.$$

Enfin, le *reach* de K, noté $\mathrm{R}(K)$, est défini de la manière suivante

$$\mathrm{R}(K) = \sup\{r \geq 0 \mid \forall x \in \mathbb{R}^D \setminus K, \quad d_K(x) < r \;\Rightarrow\; \|\nabla d_K(x)\| = 1\}.$$

On voit en particulier que $\mathrm{R}(K) \leq \mathrm{R}_{\mathrm{glob}}(K)$. Pour un réel $r \geq 0$, on note $K \oplus r$ le r-voisinage fermé de K, c'est-à-dire

$$K \oplus r = \{x + y \mid x \in K \text{ et } \|y\| \leq r\} = \{x \in \mathbb{R}^D \mid d_K(x) \leq r\}.$$

On dispose des différentes caractérisations suivantes pour le reach d'un compact.

Lemme 2.1. *Pour tout $r > 0$, on a équivalence entre*

i. $\mathrm{R}(K) > r$

ii. *d_K est de classe C^1 sur $d_K^{-1}(]0, r[)$*

iii. *Pour tout $x \in K \oplus r$, les ensembles $\Gamma_K(x)$ sont réduits à des singletons.*

En particulier, l'application projection orthogonale sur K

$$\mathrm{pr}_K : x \in K \oplus r \mapsto y \quad \text{tel que} \quad \Gamma_K(x) = \{y\}$$

est bien définie si et seulement si $r < \mathrm{R}(K)$. Lorsque K est une sous-variété compact sans bord de $\mathbb{R}^D$, Federer a montré la formule suivante

Théoreme 2.2. (Federer [1959], théorème 4.18) *Pour toute sous-variété compacte sans-bord $M \subset \mathbb{R}^D$, on a*

$$\mathrm{R}(M) = \inf_{x,y \in M} \frac{\|y - x\|^2}{2d(y - x, T_x M)}.$$

En tirant parti de cette formulation du reach, les auteurs de Aamari et al. [2017] ont pu montrer que

Théoreme 2.3. *(*Aamari et al. [2017], théorème 3.4) *Pour toute sous-variété compacte sans bord $M \subset \mathbb{R}^D$, on a*

$$\mathrm{R}(M) = \mathrm{R}_{\mathrm{loc}}(M) \wedge \mathrm{R}_{\mathrm{glob}}(M).$$

Le reach peut donc être compris comme la contribution de deux phénomènes : un reach local $\mathrm{R}_{\mathrm{loc}}(M)$ qui dépend de la courbure locale de M, et un reach global $\mathrm{R}_{\mathrm{glob}}(M)$ qui va mesurer à quel point la sous-variété M est proche de s'auto-intersecter.

2.4 Distance de Hausdorff

On note $\mathcal{K}(\mathbb{R}^D)$ l'ensemble des compacts de $\mathbb{R}^D$. Pour $K_1, K_2 \in \mathcal{K}(\mathbb{R}^D)$, on définit la distance de Hausdorff asymétrique par

$$\mathrm{H}(K_1|K_2) = \sup_{x\in K_1} d_{K_2}(x) = \inf\{r \geq 0 \mid K_1 \subset K_2 \oplus r\}$$

puis la distance de Hausdorff comme le symétrisé de $\mathrm{H}(\cdot|\cdot)$

$$\mathrm{H}(K_1, K_2) = \mathrm{H}(K_1|K_2) \vee \mathrm{H}(K_2|K_1).$$

On peut montrer que H est effectivement une distance sur $\mathcal{K}(\mathbb{R}^D)$, et que $(\mathcal{K}(\mathbb{R}^D), \mathrm{H})$ est un espace métrique complet, séparable et localement compact.

3 Modèles et risque minimax

3.1 Risque minimax

Soit $\mathcal{C}$ une classe de lois de probabilité sur $\mathbb{R}^D$. Pour toute loi $P \in \mathcal{C}$, on cherche à estimer un paramètre $\theta(P) \in \Theta$, où Θ est un espace métrique muni d'une distance d_Θ. On se donne un estimateur $\widehat{\theta}$, c'est-à-dire une collection d'applications mesurables

$$\widehat{\theta}_n : ((\mathbb{R}^D)^n, \mathcal{B}(\mathbb{R}^D)^{\otimes n}) \longrightarrow (\Theta, \mathcal{B}(\Theta)).$$

Le risque dans le pire des cas sur $\mathcal{C}$ de $\widehat{\theta}$ est défini par

$$\mathcal{R}_n(\widehat{\theta}, \mathcal{C}) = \sup_{P\in\mathcal{C}} \mathbb{E}_{P^{\otimes n}}[d_\Theta(\widehat{\theta}_n, \theta(P))].$$

Le risque minimax sur $\mathcal{C}$ correspond au meilleur risque dans le pire des cas que l'on peut atteindre avec des estimateurs mesurables :

$$\mathcal{R}_n(\mathcal{C}) = \inf_{\widehat{\theta}} \mathcal{R}_n(\widehat{\theta}, \mathcal{C}) = \inf_{\widehat{\theta}} \sup_{P\in\mathcal{C}} \mathbb{E}_{P^{\otimes n}}[d_\Theta(\widehat{\theta}_n, \theta(P))].$$

On dira qu'un estimateur $\widehat{\theta}$ est (asymptotiquement) minimax s'il vérifie que

$$\limsup_{n\to\infty} \frac{\mathcal{R}_n(\widehat{\theta}, \mathcal{C})}{\mathcal{R}_n(\mathcal{C})} < +\infty.$$

Pour une collection de classes $\mathcal{C}_\lambda$ dépendant d'un paramètre $\lambda \in \Lambda$, on dira qu'un estimateur est adaptatif par rapport à λ s'il est simultanément minimax sur tous les modèles de la collection, avec une performance majorée indépendemment de λ :

$$\sup_{\lambda\in\Lambda} \limsup_{n\to\infty} \frac{\mathcal{R}_n(\widehat{\theta}, \mathcal{C}_\lambda)}{\mathcal{R}_n(\mathcal{C}_\lambda)} < +\infty.$$

Pour majorer le risque minimax $\mathcal{R}_n(\mathcal{C})$, il suffira de trouver un bon estimateur $\widehat{\theta}$. La plupart des techniques de minorations repose sur le lemme de Le Cam.

Lemme 3.1. (Le Cam) *Pour tout couple de lois $P_1, P_2 \in \mathcal{C}$, on a*

$$\mathcal{R}_n(\mathcal{C}) \geq \frac{1}{2} d_\Theta(\theta(P_1), \theta(P_2))(1 - \mathrm{TV}(P_1, P_2))^n.$$

Il s'agira donc de trouver à chaque fois deux lois P_1 et P_2 statistiquement peu discernables (faible distance en variation totale) et dont les paramètres d'intérêt sont suffisamment éloignés les uns des autres.

3.2 Modèles de variétés

On fixe un reach minimal $\mathrm{R}_{\min} > 0$ (qui correspond à la résolution avec laquelle on regarde les données). Pour $L = (L_\perp, L_3, \dots, L_k)$ un $(k-2)$-uplet de réels strictement positifs, on définit le sous-ensemble $\mathcal{M}_d^k(\mathrm{R}_{\min}, L) \subset \mathcal{K}(\mathbb{R}^D)$ (introduit par Aamari and Levrard [2017]) de toutes les sous-variétés $M \subset \mathbb{R}^D$ compactes et sans bord vérifiant

i. M est de dimension d, de classe C^k et $\mathrm{R}(M) \geq \mathrm{R}_{\min}$,

ii. pour tout $x \in M$ il existe une paramétrisation Φ_x de classe C^k de la forme

$$\begin{aligned} \Phi_x : T_xM \cap B(x,r) &\to M \\ v &\mapsto x + v + \mathrm{N}_x(v) \end{aligned}$$

où $r \geq 1/4L_\perp$ et où N_x vérifie

$$\begin{cases} \mathrm{N}_x(0) = 0 \\ d\,\mathrm{N}_x(0) = 0 \\ \forall v \in T_xM \cap B(x, 1/4L_\perp), \quad \|d^2\,\mathrm{N}_x(v)\|_{\mathrm{op}} \leq L_\perp \\ \quad \text{et} \qquad\qquad \forall 3 \leq i \leq k \quad \|d^i\,\mathrm{N}_x(v)\|_{\mathrm{op}} \leq L_i. \end{cases}$$

Comme noté par les auteurs Aamari and Levrard [2017], si M est une sous-variété compacte de $\mathbb{R}^D$ de classe C^k, une telle collection de paramétrisations existe toujours pourvu que L soit assez grand. On note alors

$$\mathcal{M}_d^k(\mathrm{R}_{\min}, \infty) = \bigcup_L \mathcal{M}_d^k(\mathrm{R}_{\min}, L)$$

l'ensemble de toutes les sous-variétés compactes de $\mathbb{R}^D$ de dimension d et de classe C^k dont le reach est plus grand que $\mathrm{R}_{\min}$.

3.3 Modèles statistiques

Modèle sans bruit Pour $M \in \mathcal{M}_d^k(\mathrm{R}_{\min}, \infty)$, on note μ_M la mesure volume sur M, définie pour tout A borélien de $\mathbb{R}^D$ par :

$$\mu_M(A) = \mathcal{H}^d(A \cap M)$$

où $\mathcal{H}^d$ est la mesure de Hausdorff d-dimensionelle. On définit alors $\mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$ l'ensemble des lois de probabilités P sur $\mathbb{R}^D$ vérifiant

i. il existe $M \in \mathcal{M}_d^k(\mathrm{R}_{\min}, L)$ telle que $P \ll \mu_M$,

ii. la densité $dP/d\mu_M$ est μ_M-p.p. encadrée par $f_{\min}$ et $f_{\max}$.

Pour une loi $P \in \mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$, on note M_P la sous-variété qui en est le support. Ces deux objets respectent automatiquement un certain nombre de contraintes.

Lemme 3.2. (Aamari [2017], lemmes III.23, III.24, proposition III.25)
Soit $P \in \mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$. Il existe des constantes $C_i(d)$ tels que

i. $\forall x \in M_P, \ \forall r \leq \mathrm{R}_{\min}/4$,

$$C_1 f_{\min} r^d \leq P(B(x,r)) \leq C_2 f_{\max} r^d,$$

ii. $\operatorname{diam} M_P \leq \frac{C_3}{\mathrm{R}^{d-1}(M_P) f_{\min}}$,

iii. $\mathrm{R}^d(M_P) \leq \frac{C_4}{f_{\min}}$.

En choisissant la probabilité uniforme sur M, on voit qu'en particulier

$$\mathcal{H}^d(M_P) \geq C_5\,\mathrm{R}^{d-1}(M_P) \operatorname{diam} M_P \geq 2C_5\,\mathrm{R}^d(M_P).$$

Enfin, le premier point du lemme précedent permet de contrôler la densité d'un échantillon de points $\mathcal{X}_n = \{X_1, \dots, X_n\}$ de loi $P \in \mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$

$$P^{\otimes n}(\mathrm{H}(M_P|\mathcal{X}_n) \geq r) \leq \frac{C_6}{f_{\min} r^d} \exp\{-C_1 f_{\min} n r^d\}.$$

La quantité $\mathrm{H}(M_P|\mathcal{X}_n)$ mesure à quel point $\mathcal{X}_n$ recouvre bien toute la sous-variété M_P.

Modèles bruités On se propose de lister les quelques modèles d'intérêt de la littérature faisant intervenir du bruit.

Bruit désordonné : Soit Φ une loi de probabilité sur $\mathbb{R}^D$. Pour $\alpha \in [0,1]$, on note

$$\mathcal{D}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})[\Phi, \alpha] = \{\alpha P + (1-\alpha)\Phi \mid P \in \mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})\}.$$

Ceci correspond à la situation où une fraction $1-\alpha$ des données est tirée selon la loi Φ. Un cas fréquent est celui où Φ est la loi uniforme sur un compact de $\mathbb{R}^D$.

Bruit normal : Le bruit associé à une observation est astreint à rester dans la fibre normale du support au dessus du point. Pour $\sigma > 0$, on introduit $\mathcal{T}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})[\sigma]$ l'ensemble des lois des variables aléatoire $Y + Z$ où la loi P de Y est dans $\mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$ et où Z vérifie que

$$Z \in T_Y M_P^\perp \quad \text{p.s.}, \quad \|Z\| \leq \sigma \text{ p.s.} \quad \text{et} \quad \mathbb{E}[Z|Y] = 0.$$

Bruit additif : Soit Φ une loi de probabilité sur $\mathbb{R}^D$. On note

$$\mathcal{A}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})[\Phi] = \{P * \Phi \mid P \in \mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})\}.$$

Le cas le plus fréquent correspond à choisir un bruit additif gaussien.

On pourra parfois demander à ce que les supports des lois de $\mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$ restent dans un certain compact $K \subset \mathbb{R}^D$. On notera dans ce cas $\mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max}, K)$ et on définit de la même manière les modèles bruités associés. Lorsqu'il n'y a pas d'ambiguïté sur le choix des paramètres $(\mathrm{R}_{\min}, L, f_{\min}, f_{\max}, K)$, on pourra tout simplement noter les modèles par $\mathcal{P}_d^k$, $\mathcal{D}_d^k[\Phi, \alpha]$, $\mathcal{T}_d^k[\sigma]$, $\mathcal{A}_d^k[\Phi]$.

Pour toute loi Q dans un des modèles précedemment introduits, on notera de même M_Q la sous-variété associée à cette loi.

4 Quelques résultats

4.1 Homologie

Pour une sous-variété $M \subset \mathbb{R}^D$, on note $\mathcal{H}(M) \in \mathbb{N}^{\mathbb{N}}$ la suite de ses nombres de Betty (les dimensions de ses groupes d'homologie). En combinant les travaux de Niyogi et al. [2008] et Balakrishnan et al. [2012], on obtient les bornes suivantes pour le risque minimax.

Proposition 4.1. *Pour le jeu de paramètres $(\mathrm{R}_{\min}, \infty, f_{\min}, f_{\max}, [0,1]^D)$, il existe des constantes C_i dépendant de $(f_{\min}, f_{\max}, d)$, telles que, si $\mathrm{R}_{\min} \leq C_0$, on a*

$$\frac{1}{2}\exp\{-C_1 n \mathrm{R}_{\min}^d\} \leq \inf_{\widehat{\mathcal{H}}} \sup_{P \in \mathcal{P}_d^2} P^{\otimes n}[\widehat{\mathcal{H}} \neq \mathcal{H}(M_P)] \leq \frac{C_2}{\mathrm{R}_{\min}^d}\exp\{-C_3 n \mathrm{R}_{\min}^d\}.$$

L'estimateur minimax proposé par Niyogi et al. [2008] se calcule en regardant l'homologie d'une union de boules centrées en les points de $\mathcal{X}_n$ pour un rayon bien choisi. Balakrishnan et al. [2012] obtiennent des bornes minimax similaires pour les autres modèles bruités, avec à chaque fois des conditions sur le niveau de bruit relatives à la résolution $\mathrm{R}_{\min}$.

4.2 Distance riemanienne et plus courts chemins

La distance riemanienne d'une sous-variété $M \subset \mathbb{R}^D$ est une application $d_M : M \times M \to \mathbb{R}$. Un estimateur de cette distance sera donc une application $\widehat{d} : K \times K \to \mathbb{R}$ avec $M \subset K$ et mesurable en les données pour la topologie de la norme infini. On mesurera la performance par $\|\widehat{d} - d_M\|_\infty$ où la supremum est pris sur $M \times M$. En regroupant les résultats de Bernstein et al. [2000] et Arias-Castro and Le Gouic [2017] et en les adaptant un peu, on obtient facilement

Théoreme 4.2. *Pour un jeu de paramètres $(\mathrm{R}_{\min}, \infty, f_{\min}, f_{\max}, K)$, il existes des constantes C_0, C_1 dépendant des paramètres et de d telles que pour tout $n \geq C_0$*

$$\inf_{\widehat{d}} \sup_{P \in \mathcal{P}_d^2} \mathbb{E}_{P^{\otimes n}}[\|\widehat{d} - d_M\|_\infty] \leq C_1 \left(\frac{\log n}{n}\right)^{1/d}.$$

L'estimateur réalisant cette performance s'obtient en regardant la longueur des plus courts chemins dans un graphe géométrique construit sur les données. En s'inspirant des sous-variétés construites pas Genovese et al. [2012a] (démonstration du théorème 2.), on peut montrer une inégalité réciproque : il existe existe de même C_2 et C_3 dépendant de d et des paramètres tels que, pour $n \geq C_2$

$$C_3 \left(\frac{1}{n}\right)^{1/d} \leq \inf_{\widehat{d}} \sup_{P \in \mathcal{P}_d^2} \mathbb{E}_{P^{\otimes n}}[\|\widehat{d} - d_{M_P}\|_\infty].$$

L'estimateur proposé est donc minimax, au terme en log près. En regardant cette fois des estimateurs $\widehat{\gamma} : K \times K \to ([0,1] \to \mathbb{R}^D)$ qui à $x, y \in M$ associe une estimé $\widehat{\gamma}$ d'un plus court chemin entre x et y, on peut montrer le résultat suivant

Proposition 4.3. *On se place, ici seulement, dans le cas où $D = 2$. On note $\mathcal{B} = \mathcal{B}_2^2(\mathrm{R}_{\min}, \infty, f_{\min}, f_{\max}, K)$ l'ensemble des mesures à densité contre la mesure volume d'une sous-variété $M \subset K$ à bord, compact et connexe de $\mathbb{R}^2$ telle que $\mathrm{R}(M) \geq \mathrm{R}_{\min}$. Il existe des constantes C_0, C_1, C_2 dépendant des paramètres telles que pour tout $n \geq C_0$, on a*

$$C_1 \left(\frac{1}{n}\right)^{1/2} \leq \inf_{\widehat{\gamma}} \sup_{P \in \mathcal{B}} \mathbb{E}_{P^{\otimes n}} \left[\sup_{x,y \in M_P} \inf_{\gamma^*} \mathrm{H}\left(\widehat{\gamma}(x,y), \gamma^*\right) \right] \leq C_2 \left(\frac{\log n}{n}\right)^{1/4}.$$

où l'infimum est pris sur toute les plus courts chemins γ^* de M liant x à y. La distance de Hausdorff est à comprendre comme la distance de Hausdorff entre les images des courbes.

La borne supérieure paraît extrèmement peu optimale en l'état. Le défaut d'optimalité semble moins venir de l'estimateur (les chemins polygonaux issus des graphes mentionnés ci-dessus) que des techniques employées pour contrôler le risque.

4.3 Support

L'idée astucieuse introduite par Aamari and Levrard [2017] consiste à faire des regression polynomiales locale autour des points du nuage $\mathcal{X}_n$. Soit $t, h > 0$ des réels. Pour tout $j \in \{1, \dots, n\}$, on regarde

$$(\widehat{\pi}_j, \widehat{T}_{j,2}, \dots, \widehat{T}_{j,k-1}) \in \operatorname*{arg\,min}_{\pi, T_i} \widehat{P}_n^{(j)} \left[\left\| x - \pi(x) - \sum_{i=2}^{k-1} T_i(\pi(x)^{\otimes i}) \right\| \mathbb{1}_{B(0,h)}(x) \right]$$

où $\widehat{P}_n^{(j)} = \sum \delta_{X_i - X_j}$. L'application π est prise parmi toute les projection orthogonale de $\mathbb{R}^D$ de rang d. Les applications T_i sont prises parmi les applications multilinéaires symmétriques de $(\mathbb{R}^D)^i$ vers $\mathbb{R}^D$ vérifiant $\|T_i\|_{\mathrm{op}} \leq t$. On note alors, pour $v \in \mathbb{R}^D$

$$\widehat{\Psi}_j(v) = X_j + \widehat{\pi}_j(x) + \sum_{i=2}^{k-1} \widehat{T}_{j,i}(\widehat{\pi}_j(x)^{\otimes i}).$$

En considérant l'estimateur $\widehat{M} = \bigcup_{j=1}^n \widehat{\Psi}_j(B(0, 7h/8))$ pour t et h bien choisi, Aamari and Levrard [2017] (théorème 6. et 7.) ont pu obtenir la performance suivante

Théoreme 4.4. *Pour un jeu de paramètres $(\mathrm{R}_{\min}, L, f_{\min}, f_{\max}, \mathbb{R}^D)$, il existe des constantes C_0, C_1 et $C_2(d, k, \mathrm{R}_{\min}, L, f_{\min}, f_{\max})$, $C_3(d,k)$ et $C_4(d, k, \mathrm{R}_{\min})$ telles que, si*

$$\sigma \leq C_0 \left(\frac{\log n}{n-1}\right)^{1/d}$$
$$\text{et} \quad \min\{\mathrm{R}_{\min} L_\perp, \dots, \mathrm{R}_{\min}^{k-1} L_k, (\mathrm{R}_{\min}^d f_{\min})^{-1}, \mathrm{R}_{\min}^d f_{\max}\} \geq C_3$$

alors, pour tout $n \geq C_1$

$$C_4 \left\{ \left(\frac{1}{n}\right)^{\frac{k}{d}} \vee \left(\frac{\sigma}{n}\right)^{\frac{k}{k+d}} \right\} \leq \sup_{Q \in \mathcal{T}_d^k[\sigma]} \mathbb{E}_{Q^{\otimes n}}[\mathrm{H}(\widehat{M}, M_Q)] \leq C_2 \left\{ \left(\frac{\log n}{n-1}\right)^{\frac{k}{d}} \vee \sigma \right\}.$$

Les deux bornes sont du même ordres de grandeurs quand $\sigma \ll n^{-k/d}$. En particulier, l'estimateur $\widehat{M}$ introduit ci-dessus est minimax sur $\mathcal{P}_d^k(\mathrm{R}_{\min}, L, f_{\min}, f_{\max})$, avec une vitesse en $(\log n/n)^{k/d}$. Pour les deux autres modèles de bruit désordonné $\mathcal{D}_d^k$ et additif $\mathcal{A}_d^k$, on dispose des résultats de Genovese et al. [2012a], qui n'ont malheureusement pas fait l'étude en considérant la régularité k du support. On fixe $K = B(0,\rho)$ avec $\rho > \mathrm{R}_{\min}$.

Théoreme 4.5. *Pour un jeu de paramètres $(\mathrm{R}_{\min}, \infty, f_{\min}, f_{\max}, K)$, il existe des constantes $C_0(\mathrm{R}_{\min}, f_{\min}, f_{\max}, K)$ et $C_1, C_2(\mathrm{R}_{\min}, f_{\min}, f_{\max}, K, \alpha)$ tel que, pour tout $n \geq C_2$*

$$C_0\left(\frac{1}{\alpha n}\right)^{2/d} \leq \inf_{\widehat{M}} \sup_{Q \in \mathcal{D}_d^2[\Phi,\alpha]} \mathbb{E}_{Q^{\otimes n}}[\mathrm{H}(\widehat{M}, M_Q)] \leq C_1\left(\frac{\log n}{n}\right)^{2/d}$$

où Φ est la loi uniforme sur K.

La borne supérieur est obtenue en exhibant un estimateur hautement non constructif. On obtient de même des bornes pour le risque additif gaussien, en utilisant cette fois des outils de déconvolution pour estimer M_Q.

Théoreme 4.6. *Pour un jeu de paramètres $(\mathrm{R}_{\min}, \infty, f_{\min}, f_{\max}, \mathbb{R}^D)$, et pour tout $0 < \delta < 1/2$ il existe des constantes C_0, C_1, C_2 dépendant de tous les paramètres et de K (et de δ pour C_1 et C_2) telles que, pour $n \geq C_2$*

$$\frac{C_0}{\log n} \leq \inf_{\widehat{M}} \sup_{Q \in \mathcal{A}_d^\infty[\Phi]} \mathbb{E}_{Q^{\otimes n}}[\mathrm{H}(\widehat{M} \cap K, M_Q \cap K)] \leq C_1\left(\frac{1}{\log n}\right)^{\frac{1-\delta}{2}}$$

où Φ est la loi gaussienne standard $\mathcal{N}(0, I_D)$.

4.4 Quantités d'ordre supérieure

Les estimateurs définis en 4.3 capturent par essence les approximations polynomiales locales du support jusqu'à l'ordre $k-1$. En particulier, $\widehat{T}_j = \operatorname{Im}\widehat{\pi}_j$ semble être un bon candidat pour estimer le plan tangent de M_P en $\mathrm{pr}_{M_P} X_j$. On mesure l'écart entre deux sous-espaces vectoriels F, G de $\mathbb{R}^D$ par leur angle défini par

$$\angle(F,G) = \|\mathrm{pr}_F - \mathrm{pr}_G\|_{\mathrm{op}}$$

où pr_F et pr_G sont les projections orthogonales sur F et G.

Théoreme 4.7. *En reprenant les notations du théorème 4.4, et les constantes $C_0, \ldots, C_4$ avec les mêmes dépendances, on a, sous les même conditions*

$$C_4\left\{\left(\frac{1}{n-1}\right)^{\frac{k-1}{d}} \vee \left(\frac{\sigma}{n-1}\right)^{\frac{k-1}{k+d}}\right\} \leq \sup_{Q \in \mathcal{T}_d^k[\sigma]} \mathbb{E}_{Q^{\otimes n}}[\angle(\widehat{T}_1, T_{X_1} M_Q)]$$
$$\leq C_2\left(\frac{\log n}{n-1}\right)^{-\frac{1}{d}}\left\{\left(\frac{\log n}{n-1}\right)^{\frac{k}{d}} \vee \sigma\right\},$$

où $T_{X_1} M_Q$ est le plan tangent à M_Q en $\mathrm{pr}_{M_Q} X_1$.

De la même manière, $\widehat{T}_{j,2}$ est un bon candidat pour estimer la seconde forme fondamentale de M en $\mathrm{pr}_{M_P} X_j$, et Aamari and Levrard [2017] montrent en effet que l'on obtient un estimateur avec la bonne vitesse minimax en $(\log n/n)^{(k-2)/d}$.

En utilisant le théorème (2.2) de Federer, Aamari et al. [2017] ont pu exhiber un estimateur du reach. Celui n'était pas minimax, parce qu'il ne prenait pas en compte la régularité du support. Avec l'estimateur de la seconde forme fondamentale qu'il exhibe et grâce à la décomposition (théorème 2.3), E. Aamari est capable de construire un estimateur du reach minimax sur chaque modèle $\mathcal{P}_d^k$. On voit cependant apparaître deux régimes de convergence, selon que $\mathrm{R}(M) = \mathrm{R}_{\mathrm{loc}}(M)$ ou $\mathrm{R}(M) = \mathrm{R}_{\mathrm{glob}}(M)$.

5 Perspectives

5.1 Questions ouvertes

Les récents modèles $\mathcal{M}_d^k(\mathrm{R}_{\min}, L)$ introduits par Aamari and Levrard [2017] permettent d'étudier finement l'effet de la régularité sur l'estimation de grandeurs géométriques. Puisque d et k sont les seuls paramètres qui gouvernent la vitesse de convergence, on peut se demander s'il est possible de proposer des estimateurs qui s'adaptent à la régularité du support, ou encore qui s'adaptent à sa dimension (travaux en cours de V. Divol).

Presque tous les résultats présentés ci-dessus sont valables pour des lois à support sur une variétés sans bord. Il serait bon de pouvoir les généraliser à des variétés à bord - surtout lorsque les données presentent naturellement des bords - mais aussi de pouvoir detecter et estimer ces derniers lorsqu'ils existent. Les resultats les plus développés en la matière semblent être ceux de Cuevas and Rodríguez-Casal [2004], et se restreignent au bord topologique d'un compact de $\mathbb{R}^D$. De même, certaine quantité semble échapper à l'estimation. C'est le cas par exemple du volume $\mathcal{H}^d(M)$ d'une sous-variété. Les résultats existants (Arias-Castro et al. [2016]) concernent seulement l'estimation du volume d'un compact d'interieur non vide de $\mathbb{R}^D$.

Il resterait enfin toute une théorie asymptotique à développer autour de chacun des estimateurs existant dans la littérature, dans la veine de ce qu'ont pu faire Chen et al. [2015].

5.2 Mesure du defaut de convexité

Soit $M \subset \mathbb{R}^D$ une sous-variété. Le défaut de convexité de M, introduit par Attali et al. [2013], est la fonction h_M définie de $\mathbb{R}^+$ dans $\mathbb{R}^+$ par

$$h_M(t) = \mathrm{H}(M | \operatorname{hull}(M,t)) \quad \text{où} \quad \operatorname{hull}(M,t) = \bigcup_{\substack{\sigma \subset M \\ \operatorname{rad}\sigma \leq t}} \operatorname{hull}\sigma.$$

Les Prs. Marc Hoffmann, John Harvey et Krishnan Shankar ont remarqué que, lorsque M était suffisemment régulière, h_M était deux fois dérivable en 0 et que $h_M''(0) = 1/\mathrm{R}_{\mathrm{loc}}(M)$, et qu'elle présentait un point de discontinuité en $\mathrm{R}_{\mathrm{glob}}(M)$ lorsque $\mathrm{R}_{\mathrm{glob}}(M) < \mathrm{R}_{\mathrm{loc}}(M)$. Les techniques d'estimation qui peuvent découler de cette observation, bien différentes de celles proposées par Aamari et al. [2017], offrent de nouvelles perspectives interessantes pour l'étude du reach. La fonction h_M semble encoder beaucoup d'informations sur M.

5.3 Système dynamique et inférence géométrique

Etant donné un système dynamique $(\Phi^t)_t$ sur $\mathbb{R}^D$, on peut s'interesser à l'évolution de $\Phi^t(A)$ où A est une certaine partie de $\mathbb{R}^D$. On observe un nuage de points qui évolue selon la dynamique Φ^t, c'est à dire que l'on dispose de $\mathcal{D}_n = \{(t_i, \Phi^{t_i}(\mathcal{X}_n)) \mid i = 0, \dots, T\}$ où $t_0 < \cdots < t_T$ sont des réels et où $\mathcal{X}_n \subset A$ est un certain échantillon de A. On peut alors se poser un certain nombre de questions sur l'évolution de la géométrie de $t \mapsto \Phi^t(A)$ (évolution de l'homologie, de la courbure, detection de rupture...), sur ce que cela nous apporte comme connaissance de Φ^t et enfin sur la manière de tirer parti de la dépendance de nos données. Il semblerait aussi interessant d'étudier la réunion des images des flots

$$M = \bigcup_{t \in I} \{t\} \times \Phi^t(A) \subset \mathbb{R} \times \mathbb{R}^D$$

d'un point de vue géométrique, lorsque cela a du sens, en particulier lorsque M est une sous-variété de $\mathbb{R} \times \mathbb{R}^D$.

Références

Eddie Aamari. *Convergence Rates for Geometric Inference*. PhD thesis, Université Paris-Saclay, 2017.

Eddie Aamari and Clément Levrard. Non-asymptotic rates for manifold, tangent space, and curvature estimation. *arXiv preprint arXiv :1705.00989*, 2017.

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *arXiv preprint arXiv :1705.04565*, 2017.

Ery Arias-Castro and Thibaut Le Gouic. Unconstrained and curvature-constrained shortest-path distances and their approximation. *arXiv preprint arXiv :1706.09441*, 2017.

Ery Arias-Castro, Beatriz Pateiro-López, and Alberto Rodríguez-Casal. Minimax estimation of the volume of a set with smooth boundary. *arXiv preprint arXiv :1605.01333*, 2016.

Ery Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local pca. *The Journal of Machine Learning Research*, 18(1) :253–309, 2017.

Dominique Attali, André Lieutier, and David Salinas. Vietoris–rips complexes also provide topologically correct reconstructions of sampled shapes. *Computational Geometry*, 46(4) :448–465, 2013.

Sivaraman Balakrishnan, Alesandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *Artificial Intelligence and Statistics*, pages 64–72, 2012.

Mira Bernstein, Vin De Silva, John C Langford, and Joshua B Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Technical report, Department of Psychology, Stanford University, 2000.

Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

Frederic Chazal, Andre Lieutier, and Jarek Rossignac. Normal-map between normal-compatible manifolds. *International Journal of Computational Geometry & Applications*, 17(05) :403–421, 2007.

Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for measures based on distance functions. *Foundations of Computational Mathematics*, 11(6) :733–751, 2011.

Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6) :41, 2013.

Yen-Chi Chen, Christopher R Genovese, Larry Wasserman, et al. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5) :1896–1928, 2015.

Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36(2) :340–354, 2004.

Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary mathematics*, 453 : 257–282, 2008.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3) :418–491, 1959.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4) :983–1049, 2016.

Maurizio Filippone, Francesco Camastra, Francesco Masulli, and Stefano Rovetta. A survey of kernel and spectral methods for clustering. *Pattern recognition*, 41(1) :176–190, 2008.

Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1) :55–77, 1997.

Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*, volume 3. Springer, 1990.

Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of machine learning research*, 13(May) :1263–1291, 2012a.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, Larry Wasserman, et al. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, 40(2) :941–963, 2012b.

Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

Sertac Karaman and Emilio Frazzoli. Sampling-based algorithms for optimal motion planning. *The international journal of robotics research*, 30(7) :846–894, 2011.

John M Lee. *Riemannian manifolds : an introduction to curvature*, volume 176. Springer Science &

Business Media, 2006.

Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.

Chunyuan Li, Maks Ovsjanikov, and Frederic Chazal. Persistence-based structural recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1995–2002, 2014.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3) :419–441, 2008.

Csaba Szepesvári, Jean-Yves Audibert, et al. Manifold-adaptive dimension estimation. In *Proceedings of the 24th international conference on Machine learning*, pages 265–272. ACM, 2007.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500) :2319–2323, 2000.

Rien Van de Weygaert, Gert Vegter, Herbert Edelsbrunner, Bernard JT Jones, Pratyush Pranav, Changbom Park, Wojciech A Hellwing, Bob Eldering, Nico Kruithof, EGP Bos, et al. Alpha, betti and the megaparsec universe : on the topology of the cosmic web. In *Transactions on Computational Science XIV*, pages 60–101. Springer-Verlag, 2011.