Reconstruction simpliciale de variété via l'estimation d'espace tangent

Eddie Aamari [1] & Clément Levrard [2]

[1] *eddie.aamari@math.u-psud.fr*
[2] *clement.levrard@inria.fr*

Résumé. On s'intéresse au problème de reconstruction de variété dans un cadre semi-asymptotique. Sous des contraintes géométriques de régularité, nous proposons un estimateur calculable $\hat{\mathcal{M}}$ du support $\mathcal{M} \subset \mathbb{R}^D$ d'une mesure inconnue dont on observe un n-échantillon *i.i.d.*. $\hat{\mathcal{M}}$ a la même topologie que $\mathcal{M}$ et on donne une vitesse de convergence pour la distance de Hausdorff. La méthode s'appuie sur la construction d'un complexe de Delaunay tangentiel. Après avoir réduit la question à l'estimation des espaces tangents de $\mathcal{M}$, le problème est traité par analyse en composantes principale locale. Si le temps le permet, nous présenterons une technique de débruitage des données par ACP locale dans le cadre d'un modèle de mélange.

Mots-clés. Approximation, complexe de Delaunay tangentiel, ACP locale.

Abstract. We look at the problem of manifold reconstruction in a semi-asymptotic framework. Under geometrical regularity constraints, we propose a computable estimator $\hat{\mathcal{M}}$ of the support $\mathcal{M}$ of an unknown measure from which we observe a *i.i.d.* n-sample. $\hat{\mathcal{M}}$ has the same topology as $\mathcal{M}$ and we give a rate of convergence for the Hausdorff distance. The method is based on a tangential Delaunay complex. After having reduced the question to estimating the tangent spaces of $\mathcal{M}$, the problem is handled with local principal components analysis. If we have enough time, we will present a denoising technique with local PCA in a mixture model.

Keywords. Approximation, tangential Delaunay complex, local PCA.

1 Introduction

La reconstruction de variété consiste à donner des approximations d'une forme inconnue $\mathcal{M} \subset \mathbb{R}^D$ à partir d'un échantillon $X_1, \ldots, X_n$ de points tirés sur celle-ci, ou dont la loi a un lien avec $\mathcal{M}$. L'approximation $\hat{\mathcal{M}}$ permet alors d'expliquer la structure géometrique ou topologique de $\mathcal{M}$; structure qui est absente du nuage de point initial composé de n points déconnectés et désorganisés. Proposer une triangulation - ou complexe simplicial - présente l'avantage de résumer la variété par un objet purement combinatoire, rendant possible le calcul d'invariants géométriques comme les groupes d'homologie.

En analyse de données déterministe, le cas des surfaces dans $\mathbb{R}^3$ est désormais bien connu (voir [1]). Des résultats récents ([2]) permettent de travailler en toute dimension et avec des algorithmes de reconstruction explicites.

La qualité d'approximation d'un estimateur $\hat{\mathcal{M}}$ de $\mathcal{M}$ est mesurée par la *distance de Hausdorff*,

$$d_H(\mathcal{M}, \hat{\mathcal{M}}) = \|d_{\mathcal{M}} - d_{\hat{\mathcal{M}}}\|_\infty,$$

où pour $K \subset \mathbb{R}^D$ $d_K(x) = \inf_{y \in K} \|x - y\|$ désigne la fonction distance à K.

1.1 Régularité

Comme dans tous les problèmes d'estimation non-paramétrique, les propriétés de régularité des objets estimés jouent un rôle crucial quand il s'agit d'obtenir des vitesses. Ici, le paramètre important est le reach. *L'axe médian* est l'ensemble des points de l'espaces qui ont au moins deux plus proches voisins sur $\mathcal{M}$: $\text{med}(\mathcal{M}) = \{x \in \mathbb{R}^D, \exists a, b \in \mathcal{M}, a \neq b, \|x - a\| = \|x - b\| = d_{\mathcal{M}}(x)\}$. Le reach de $\mathcal{M}$ est la distance minimale de $\mathcal{M}$ à son axe médian,

$$\text{reach}(\mathcal{M}) = \inf_{x \in \mathcal{M}} d_{\text{med}(\mathcal{M})}(x).$$

Le reach encode à la fois des informations locales et globales. Intuitivement, vérifier $\text{reach}(\mathcal{M}) \geq \rho > 0$ impose une courbure sectionnelle bornée par $2/\rho^2$ (Prop. 2.1, [5]) et à avoir des "composantes intrinsèquement éloignées" qui sont aussi séparées d'au moins ρ lorsqu'on les considère dans $\mathbb{R}^D$ (voir Figure 1).

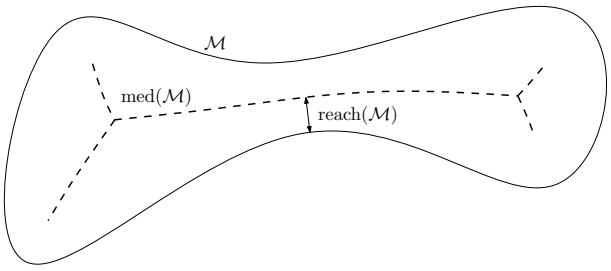


Figure 1: *Axe médian* et *reach* d'une courbe dans $\mathbb{R}^2$.

1.2 Modèle

La sous-variété inconnue $\mathcal{M} \subset \mathbb{R}^D$ est supposée lisse, compacte, sans bord, et telle que $\text{reach}(\mathcal{M}) \geq \rho > 0$. La dimension $d = \dim(\mathcal{M})$ est supposée connue.

$X_1, \ldots, X_n$ est un échantillon *i.i.d.* de loi P sur $\mathcal{M}$. On suppose que P a une densité f par rapport à la mesure de Hausdorff d-dimensionnelle. De plus, f est deux fois différentiable et vérifie pour tout $x \in \mathcal{M}$,

$$\begin{cases} p_{min} \leq f(x) \leq p_{max} \\ \|H_f(x)\| \leq H, \end{cases}$$

pour des constantes p_{min}, p_{max} et H, où $H_f(x)$ est la matrice hessienne de f en x.

2 Complexe de Delaunay tangentiel

On explique ici la construction de la triangulation $\hat{\mathcal{M}}$ lorsque les espaces tangents à $\mathcal{M}$ sont supposés connus.

Soit $\mathcal{P} = \{x_1, \ldots, x_n\} \subset \mathcal{M}$ et $\omega : \mathcal{P} \to \mathbb{R}_+$ appelée *fonction poids*. $Vor_\omega(\mathcal{P})$ est le diagramme de Voronoï à poids ω de $\mathcal{P}$. La cellule de Voronoï à poids de $p \in \mathcal{P}$ étant définie par

$$Vor^\omega(p) = \{x \in \mathbb{R}^D, \|p - x\|^2 - \omega(p)^2 \leq \|q - x\|^2 - \omega(q)^2, \forall q \in \mathcal{P}\},$$

pour tout $\tau \subset \mathcal{P}$, la face de Voronoï associée τ est $Vor^\omega(\tau) = \bigcap_{p \in \tau} Vor^\omega(p)$. Pour $p \in \mathcal{P}$ fixé, $T_p = T_p\mathcal{M}$ désigne l'espace tangent à $\mathcal{M}$ en p. La triangulation est construite par restriction à T_p du Delaunay à poids, *i.e.* pour un simplexe $\tau \subset \mathcal{P}$ avec $p \in \tau$, $\tau \in Del_p^\omega(\mathcal{P}) \Leftrightarrow Vor^\omega(\tau) \cap T_p \neq \emptyset$. Enfin, la triangulation de Delaunay tangentielle à poids ω, est $Del_{T\mathcal{M}}^\omega(\mathcal{P}) = \bigcup_{p \in \mathcal{P}} Del_p^\omega(\mathcal{P})$.

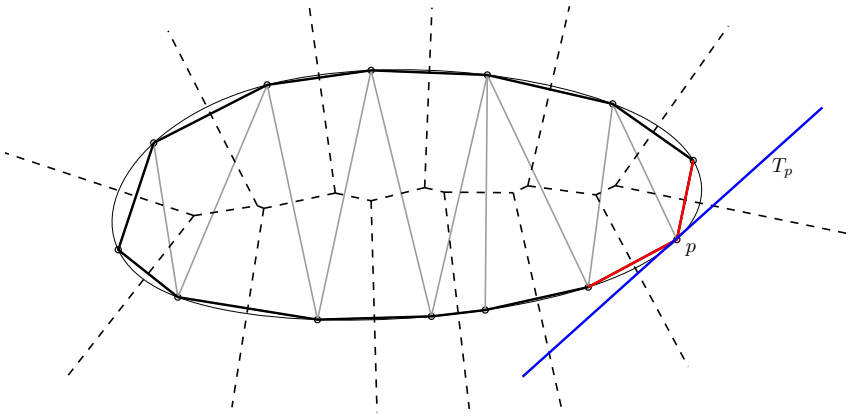


Figure 2: Construction de $Del_p^\omega(\mathcal{P})$ pour $\omega \equiv 0$.

La nécessité d'ajouter la fonction poids ω se pose à partir de la dimension $D = 4$ où un phénomène d'instabilité apparaît [6], même sous des hypothèses fortes d'échantillonnage. *A contrario* ω apparaît comme un degré de liberté supplémentaire permettant d'énoncer le résultat d'approximation suivant.

Théorème 1. *(Theorem 5.3 de [2]) Si $\mathcal{P} = \{x_1, \ldots, x_n\} \subset \mathcal{M}$ est un ϵ-échantillon pour $\epsilon > 0$ assez petit, et sous de bonnes hypothèses d'échantillonnage, il existe une fonction de poids ω calculable ainsi qu'une constante explicite C telles que $\hat{\mathcal{M}} = Del^{\omega}_{T\mathcal{M}}(\mathcal{P})$ vérifie*

- *$d_H(\mathcal{M}, \hat{\mathcal{M}}) \leq C\epsilon^2$;*
- *$\hat{\mathcal{M}}$ et $\mathcal{M}$ sont isotopes. (En particulier $\hat{\mathcal{M}}$ et $\mathcal{M}$ sont homéomorphes)*

De plus C ne dépend que de $\mathrm{reach}(\mathcal{M}), \dim(\mathcal{M})$ *et des conditions d'échantillonnage de $\mathcal{P}$.*

Le calcul de ω est assuré par un algorithme dans [2] et ne dépend que des données. Cependant la construction du complexe $Del^{\omega}_{T\mathcal{M}}(\mathcal{P})$ nécessite la connaissance de chaque T_{x_i}. La méthode proposée est un plug-in qui consiste à:

(i) estimer les espaces tangents;

(ii) appliquer la construction du Delaunay tangentiel à poids.

Notons $\hat{T}_{x_i}$ un espace affine de dimension d dans $\mathbb{R}^D$ passant par x_i. L'angle entre deux espaces vectoriels de même dimension U et V étant défini par $\angle(U, V) = \max_{u \in U} \min_{v \in V} \angle(u, v)$ - et dans le cas d'espaces affines comme l'angle entre leurs espaces vectoriels parallèles - on donne le résultat de stabilité suivant.

Proposition 1. *Sous de bonnes hypothèses d'échantillonnage, la construction du théorème 1 est stable par perturbation du plan tangent, au sens où*

$$\forall i, \angle(T_{x_i}, \hat{T}_{x_i}) \leq \frac{\pi}{32} \implies Del^{\omega}_{T\mathcal{M}}(\mathcal{P}) = Del^{\omega}_{\hat{T}\mathcal{M}}(\mathcal{P}).$$

Le problème est alors réduit à l'estimation des espaces tangents.

3 Estimation d'espace tangent

On se place dans le modèle décrit Section 1.2 et on utilise une analyse en composantes principales - ACP - locale en chaque en chaque point $X_1, \ldots, X_n$. Soit K le noyau $K(x) = \mathbb{1}_{[0,1]}(x)$. Pour un point de l'échantillon X_j fixé, la matrice de covariance locale $\hat{O}_j$ est définie comme dans [7] par

$$\hat{O}_j = \frac{1}{n-1} \sum_{i \neq j} K\left(\frac{\|X_i - X_j\|}{\epsilon}\right) (X_i - X_j)(X_i - X_j)^t,$$

où $\epsilon > 0$ est un paramètre de fenêtrage. L'espace vectoriel engendré par les vecteurs propres associées au d plus grandes valeurs propres de $\hat{O}_j$ est noté $\hat{T}_{X_j}$. Nous donnons un résultat de convergence de $\hat{T}_{X_j}$ vers T_{X_j}.

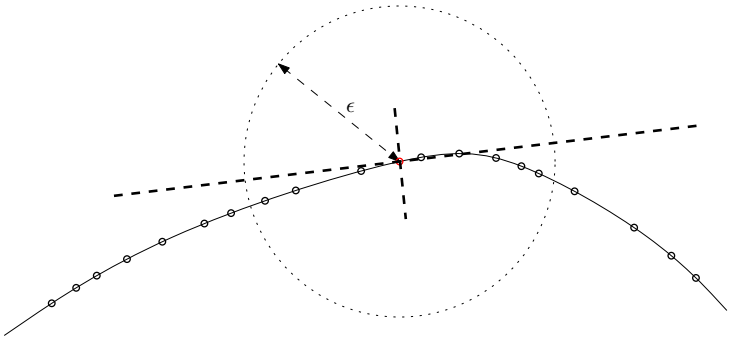
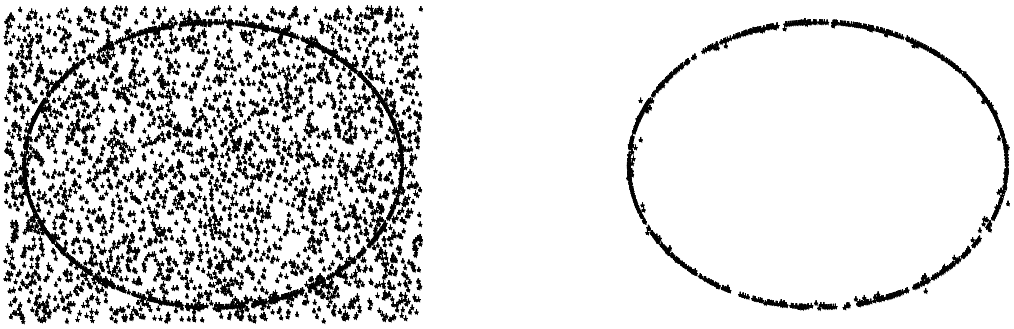


Figure 3: ACP locale au voisinage d'un point.

Proposition 2. *Supposons que $\epsilon \leq \epsilon_0$. Il existe des constantes $\kappa_1 = \kappa_1(d)$ et $C = C(d, H)$ telles que, si $n \geq 1 + \kappa_1 \frac{\rho^4}{\epsilon^{d+4}}$, avec probabilité plus grande que $1 - ne^{-\frac{2(n-1)\epsilon^{2d+4}}{\kappa_1}}$, pour tout $j \in \{1, \ldots, n\}$,*

$$\angle(T_{X_j}, \hat{T}_{X_j}) \leq \frac{C}{\rho^2}\epsilon^2.$$

Remarque. Si le temps le permet, on étudiera le cas du modèle de mélange $X_i \sim \beta P + (1-\beta)U$, où P vérifie les conditions de la Section 1.2 et U est la loi uniforme sur $B_D(0, M)$ avec $\mathcal{M} \subset B_D(0, M)$. En utilisant l'ACP locale ainsi qu'une procédure de comptage de points contenus dans des pavés centré en les données similaire à [8], il est possible de détbruiter suffisamment les données afin d'appliquer le Théorème 1. et assurer l'estimation de $\mathcal{M}$ avec des vitesses proches du cadre non bruité.



(a) Données bruitées (b) Données débruitées

Figure 4: Modèle de mélange $X_i \sim \beta P + (1-\beta)U$

Bibliographie

[1] T. K. Dey. Curve and Surface Reconstruction: Algorithms with Mathematical Analysis. Cambridge University Press, 2006.
[2] Boissonnat, J. D., & Ghosh, A. (2014). Manifold reconstruction using tangential Delaunay complexes. Discrete & Computational Geometry, 51(1), 221-267.
[3] Federer, H. (1959). Curvature measures. Transactions of the American Mathematical Society, 418-491.
[4] Niyogi, P., Smale, S., & Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. Discrete & Computational Geometry, 39(1-3), 419-441.
[5] Dey, T. K., & Li, K. (2009). Topology from Data via Geodesic complexes (p. 178). Tech Report OSU-CISRC-3/09-TR05.
[6] Boissonnat, J. D., Guibas, L. J., & Oudot, S. Y. (2009). Manifold reconstruction in arbitrary dimensions using witness complexes. Discrete & Computational Geometry, 42(1), 37-70.
[7] Singer, A., & Wu, H. T. (2012). Vector diffusion maps and the connection Laplacian. Communications on pure and applied mathematics, 65(8), 1067-1144.
[8] Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., & Wasserman, L. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. The Annals of Statistics, 40(2), 941-963.