Estimation minimax d'espaces tangents

Eddie Aamari [1] & Clément Levrard [2]

[1] *Université d'Orsay, Inria Saclay — eddie.aamari@inria.fr*
[2] *Université Paris Diderot — levrard@math.univ-paris-diderot.fr*

Résumé. Divers algorithmes utilisés en géométrie algorithmique et en inférence géométrique utilisent, explicitement ou non, la connaissance de directions tangentes. Étant donnés une sous-variété $M \subset \mathbb{R}^D$ et un point de base $x \in M$, l'espace tangent T_xM est défini comme la meilleure approximation linéaire de M en x. Le but de l'exposé est d'étudier les vitesses minimax d'estimation d'espaces tangents à partir d'un nuage de points. Après avoir motivé leur étude, on introduira une classe de sous-variétés $\mathcal{C}^k$ en analogie avec les classes de Hölder $\mathcal{C}^k(L)$. Nous proposerons un estimateur construit à partir de polynômes locaux et nous étudierons ses propriétés non asymptotiques en soulignant l'influence de la dimension et de la taille d'échantillon. Les bornes inférieures minimax seront détaillées dans les cas où le point de base est fixe et lorsqu'il est aléatoire.

Mots-clés. Inférence géométrique, plans tangents, polynômes locaux, vitesse de convergence, minimax

Abstract. Algorithms commonly used in computational geometry, for instance in the field of shape reconstruction, take advantage of some prior knowledge of tangent directions. If M is a submanifold of $\mathbb{R}^D$ and $x \in M$, the tangent space of M at x, denoted by T_xM, may be thought of as the best linear approximation of M around x. In this talk, we investigate the minimax rates of tangent space estimation from a n-sample. These minimax rates are derived over regularity classes of $\mathcal{C}^k$ submanifolds that are designed similarly to Hölder regularity classes $\mathcal{C}^k(L)$. An estimator based on local polynomials is proposed, whose non-asymptotic dependency on the sample size and ambient dimension over the regularity classes will be studied. The associated lower bounds are derived in the two cases whether the basis point x is fixed or random.

Keywords. Geometric inference, tangent spaces, local polynomials, minimax convergence rates

1 Introduction

Certains types de données, comme la répartition des galaxies dans l'univers, des points sur une surface ou encore des paramètres physiques soumis à des contraintes, peuvent être modélisés comme s'organisant autour d'une structure de dimension réduite, une sous-variété M de dimension d de l'espace ambiant $\mathbb{R}^D$. Ces structures, pour peu qu'elles

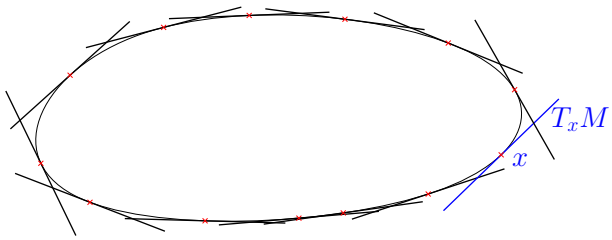

Figure 1: Droites tangentes à une courbe fermée du plan.

soient suffisamment régulières, peuvent être approchées localement par des sous-espaces vectoriels de dimension réduite. Sur cette hypothèse d'approximation linéaire locale sont basés plusieurs outils d'inférence géométrique des caractéristiques de cette variété support M, que l'objet d'intérêt soit la variété elle-même, via les algorithmes de reconstruction tels le Tangential Delaunay Complex [2], ou d'autres quantités structurelles telles le reach.

Pour beaucoup de ces algorithmes, la qualité d'estimation des espaces tangents joue un rôle déterminant. Dès lors plusieurs études ont cherché à en déterminer des estimateurs avec des garanties théoriques. Le critère communément utilisé pour témoigner de la qualité d'un estimateur $\hat{T}_xM$ du plan tangent cible T_xM est la déviation en angle

$$\angle(\hat{T}_xM, T_xM) := \|\pi_{\hat{T}_xM} - \pi_{T_xM}\|_{op} = \max_{j=1,\dots,d} |\sin(\theta_j)|,$$

où π_T est la projection orthogonale de $\mathbb{R}^D$ sur T, et les θ_j sont les angles principaux obtenus à partir de la décomposition en valeurs singulières de $\pi_T - \pi_{\hat{T}}$ (voir par exemple le chapitre 2.6 de [7]).

L'estimateur le plus fréquemment proposé dans un cadre probabiliste pour T_xM ([1], [9], [8]) est basé sur une ACP locale du nuage de points au voisinage du point x. De ce point de vue les directions tangentes sont assimilées aux directions capturant au mieux l'étalement du nuage de point constitué par les voisins de x. Les différents résultats obtenus pour cet estimateur $\hat{T}_{PCA,x}$ mettent en relief l'influence de la régularité de la variété M. Informellement, pour une variété de classe $\mathcal{C}^2$, [1] montre que $\angle(\hat{T}_{PCA,x}M, T_xM) \lesssim n^{-1/d}$, tandis que pour une variété de classe $\mathcal{C}^3$, [8] obtient $\angle(\hat{T}_{PCA,x}M, T_xM) \lesssim n^{-3/(d+2)}$.

Cependant, des estimateurs prenant en compte, pour des régularités plus élevées, le caractère localement polynomial des paramétrisations de ces variétés semblent mener vers des vitesses de convergence plus élevées ([4], [3]). Par exemple, [4] prouve qu' un estimateur $\hat{T}_xM$ basé sur la l'interpolation locale des points y autour de x par une formule de type $y-x = \pi(y-x)+(y-x)^tT_2(y-x)+O(\|y-x\|^3)$ mène à une vitesse de convergence de type $\angle(\hat{T}_xM, T_xM) \lesssim (1/n)^{-2/d}$, où l'estimateur $\hat{T}$ est le sous-espace vectoriel associé au projecteur interpolant π dans la formule ci-dessus. En nous basant sur cette idée, nous nous sommes interrogés sur l'optimalité de ce type de procédure lorsqu'un ordre de régularité k est fixé. Par ailleurs, les résultats proposés dans [4] et [3] exhibant une

dépendance quadratique de la déviation en angle en la dimension ambiante, nous avons cherché à déterminer le rôle précis de la dimension ambiante dans les vitesses d'estimation.

Etant donné un ordre de régularité k, et un n-échantillon $X_1, \dots, X_n$ tiré sur M suivant la loi $P \sim f\lambda_M$, avec $0 < c < \inf_{x\in M} f(x) \leq \sup_{x\in M} f(x) \leq C$, où λ_M désigne la mesure uniforme sur M, nous proposons comme estimateur de l'espace tangent au point X_i l'espace associé au projecteur $\hat{\pi}_i$ défini par

$$\hat{\pi}_i \in \underset{\pi, \sup_{p\geq 2}\|T_p\|_{op}\leq h^{-1}}{\arg\min} P_n^{(i)} \left[\|x - \pi(x) - T_2(\pi(x), \pi(x)) - \ldots - T_{k-1}(\pi(x)^{\otimes k-1})\|^2 \mathbb{1}_{B(h)} \right], \tag{1}$$

où $h > 0$ est une fenêtre à calibrer, $P_n^{(i)}$ est l'intégration par rapport à la mesure $1/(n-1)\sum_{j\neq i} \delta_{X_j - X_i}$, $x^{\otimes k}$ représente le vecteur $(x, \dots, x)$ de dimension $k \times D$, $B_i(h)$ est la boule Euclidienne de rayon h centrée en X_i, et T_p est un tenseur d'ordre p. Pour que cet estimateur ait une chance d'approcher l'espace tangent, il semble nécessaire que localement autour de X_i les points de la variété s'ajustent à un tel type de paramétrisation. Ce point est l'objet des classes de régularité définies dans les sections suivantes.

2 Le cas particulier de l'ordre 2

Quand l'ordre de régularité est fixé à 2, la paramétrisation naturelle par les coordonnées géodésiques est adaptée pour l'interpolation que nous nous proposons d'effectuer. Définissons $\mathcal{C}^2_{\tau_{min}}$ comme la classe des variétés $\mathcal{C}^2$ dont le reach τ_M est minoré par $\tau_{min} > 0$ (la condition de reach minoré est légèrement plus forte qu'une condition de courbure uniformémement bornée, voir [5] par exemple). Alors, si $M \in \mathcal{C}^2_{\tau_{min}}$, on peut écrire, pour y suffisamment proche de x de manière à ce que $y = \exp_x(v)$ pour un vecteur v dans T_xM,

$$y = x + v + \mathbf{N}_x(v), \tag{2}$$

où $\mathbf{N}_x$ est une fonction $C^{1,1}$ vérifiant

$$\mathbf{N}_x(0) = 0, \quad d_0\mathbf{N}_x = 0, \quad \|d_v\mathbf{N}_x\|_{op} \leq L_\perp \|v\|,$$

avec $L_\perp = 1/(2\tau_{min})$. Une telle décomposition étant acquise, on peut alors obtenir le résultat de convergence suivant pour les plans tangents obtenus par interpolation à l'ordre 1 (ce qui revient à une ACP locale) avec une fenêtre de type $h = (C\log(n)/n)^{1/d}$,

$$\mathbb{E} \max_{i=1,\dots,n} \angle(\hat{T}_i, T_{X_i}M) \leq C \left(\frac{\log(n)}{n}\right)^{\frac{1}{d}},$$

où la constante C ne dépend pas de la dimension ambiente. Cette borne supérieure est similaire à celle obtenue dans [1]. De plus, on montre que cette vitesse de convergence est optimale à un facteur log près sur la classe $C^2_{\tau_{min}}$, c'est à dire

$$\inf_{\hat{T}} \sup_{M \in C^2_{\tau_{min}}} \mathbb{E} \angle \hat{T}, T_{X_1} M \geq c \left(\frac{1}{n}\right)^{\frac{1}{d}},$$

où la constante c dépend aussi des paramètres des densités considérées (minoration et majoration). À l'ordre de régularité 2, les estimateurs d'espaces tangents par ACP locale sont donc optimaux.

3 Ordres de régularité supérieurs à 3

Si l'on se donne une variété de classe $\mathcal{C}^k$, où $k \geq 3$, la paramétrisation exponentielle donnée n'est plus suffisamment régulière pour permettre de capturer les variations d'ordre $k-2$ autour d'un point x. Par conséquent, pour $k \geq 3$ et des paramètres $\tau_{min}, L_\perp, L_3, \ldots, L_k$, nous définissons par analogie avec (2) la classe de régularité $\mathcal{C}^k_{\tau_{min}, L_\perp, L_3, \ldots, L_k}$ comme l'ensemble des variétés $M \in \mathcal{C}^k$ de reach minoré par τ_{min} telles qu'il existe pour tout x de M une paramétrisation $\Psi_x : T_x M \to M$ vérifiant, pour tout y assez proche tel que $y = \Psi_x(v)$,

$$y = x + v + \mathbf{N}_x(v), \tag{3}$$

où cette fois-ci $\mathbf{N}_x$ est de régularité $\mathcal{C}^{k-1,1}$ et vérifie

$$\mathbf{N}_x(0) = 0, \quad d_0 \mathbf{N}_x = 0, \quad \left\| d_v^2 \mathbf{N}_x \right\|_{op} \leq L_\perp, \quad \left\| d_v^i \mathbf{N}_x \right\|_{op} \leq L_i \text{ pour } 3 \leq i \leq k.$$

Une famille de paramétrisations de cette forme existe toujours pour une variété compacte M de $\mathbb{R}^D$ de classe $\mathcal{C}^k$, pourvu que l'on autorise $\tau_{min}^{-1}, L_\perp, \ldots, L_k$ à être assez grands. Pour $M \in \mathcal{C}^k_{\tau_{min}, L_\perp, L_3, \ldots, L_k}$, les espaces tangents donnés par interpolation locale à l'ordre $k-1$ avec le même type de fenêtre $h = (C \log(n)/n)^{1/d}$ donnent le résultat suivant,

$$\mathbb{E} \max_{i=1,\ldots,n} \angle(\hat{T}_i, T_{X_i} M) \leq C \left(\frac{\log(n)}{n}\right)^{\frac{k-1}{d}},$$

généralisant ainsi la borne obtenue à l'ordre 2 et garantissant une meilleure vitesse de convergence que les estimateurs obtenus par ACP locale. Là encore, la constante C ne dépend pas de la dimension ambiante. Enfin, cette vitesse est quasi optimale sur la classe de régularité considérée, pour laquelle on a

$$\inf_{\hat{T}} \sup_{M \in \mathcal{C}^k_{\tau_{min}, L_\perp, L_3, \ldots, L_k}} \mathbb{E} \angle(\hat{T}, T_{X_1} M) \geq c \left(\frac{1}{n}\right)^{\frac{k-1}{d}},$$

où c dépend des constantes du modèle géométrique $\mathcal{C}^k_{\tau_{min},L_\perp,L_3,\dots,L_k}$ ainsi que des bornes sur la densité par rapport à la mesure uniforme sur M. On peut noter que les bornes minimax proposées prennent une espérance en le point de base X_1 sur lequel l'estimation est effectuée. Les même bornes sont valides si on fixe le point d'intérêt $x \in M$.

4 Conclusion et perspectives

On peut résumer les résultats obtenus de la manière suivante: pour une variété M dont la paramétrisation est régulière à l'ordre k, la vitesse d'estimation optimale des espaces tangents est de l'ordre de $(1/n)^{(k-1)/d}$ et ne dépend pas de la dimension ambiante. Cette vitesse est atteinte pour les ordres de régularité supérieurs à 3 par interpolation polynomiale locale, là où les estimateurs par ACP locale offrent des garanties moindres.

Le sujet de l'estimation des quantités d'ordre 2 et leurs liens avec la courbure de la variété semble aussi découler naturellement de cette procédure d'interpolation par polynomes locaux. Des résultats préliminaires tendent à prouver que regarder l'ordre 2 dans notre interpolation polynomiale fournit un estimateur de la seconde forme fondamentale (vue comme une forme quadratique) à l'ordre $(1/n)^{(k-2)/d}$, pour une variété dans la classe $\mathcal{C}^k_{\tau_{min},L_\perp,L_3,\dots,L_k}$. Par ailleurs, d'autres résultats obtenus par E. Aamari, J. Kim, F. Chazal, B. Michel, A. Rinaldo et L. Wassermann suggèrent que la vitesse d'estimation du reach est de cet ordre pour $k = 3$. Il semble alors pertinent de chercher à déterminer si la vitesse d'estimation du reach s'améliore de la même manière quand la régularité croît, ou si ces deux problèmes sont de natures fondamentalement différentes.

Une dernière perspective ouverte par cette procédure d'interpolation locale est la reconstruction effective de la variété source par recollement de morceaux de polynômes locaux. Là encore, quelques résultats préliminaires suggèrent qu'un tel estimateur de variété atteindrait une vitesse de convergence en terme de distance de Hausdorff de l'ordre de $(1/n)^{k/d}$, généralisant ainsi aux ordres de régularité plus élevés les résultats d'estimation de variété à l'ordre 2 présentés dans [6].

Bibliographie

[1] E. Aamari and C. Levrard. Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction. ArXiv e-prints, December 2015.

[2] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. Discrete Comput. Geom., 51(1):221–267, 2014.

[3] Frédéric Cazals and Marc Pouget. Estimating differential quantities using polynomial fitting of osculating jets. Computer Aided Geometric Design, 22:121–146, 2005.

[4] Siu-Wing Cheng and Man-Kwun Chiu. Tangent estimation from point samples. Discrete & Computational Geometry, 56(3):505–557, 2016.

[5] Herbert Federer. Curvature measures. Trans. Amer. Math. Soc., 93:418–491, 1959.

[6] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. J. Mach. Learn. Res., 13:1263–1291, 2012.

[7] Gene H. Golub and Charles F. Van Loan. Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[8] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. Comm. Pure Appl. Math., 65(8):1067–1144, 2012.

[9] Hemant Tyagi, Elif Vural, and Pascal Frossard. Tangent Space Estimation Bounds for Smooth Manifolds. In Proceedings of SAMPTA, 2013.