Reconstruction de variétés via l'estimation d'espaces tangents

Eddie Aamari, Inria Saclay, Université d'Orsay

Résumé court:: On considère le problème de reconstruction de variété dans un cadre non-asymptotique. Sous des hypothèses de régularité géométrique, on proposera un estimateur calculable $\hat{M}$ du support M d'une distribution inconnue dont on observe un n-échantillon i.i.d. L'estimateur $\hat{M}$ possède la mme topologie que M et l'approxime, pour la perte donnée par la distance de Hausdorff, la vitesse optimale au sens minimax. La méthode développée se base sur le complexe de Delaunay tangentiel. Après avoir réduit la question l'estimation des espaces tangents de M, le problème est résolu avec des ACP locales. On examinera la robustesse de la méthode dans le cadre d'un modèle de mélange, o une technique de débruitage reposant sur l'ACP locale sera présentée.

Résumé long:

Présentation. Certains types de données, comme la répartition des galaxies dans l'univers, des points sur une surface ou encore des paramètres physiques soumis des contraintes, peuvent tre modélisés comme s'organisant autour d'une structure de dimension réduite, une sous-variété M de dimension d de l'espace ambiant $\mathbb{R}^D$.
La reconstruction de variété consiste à donner des approximations d'une forme inconnue $M \subset \mathbb{R}^D$ à partir d'un échantillon $\mathbb{X} = \{X_1, \dots, X_n\}$ de points tirés sur ou proche de celle-ci. L'approximation $\hat{M}$ permet alors d'expliquer la structure géometrique ou topologique de M; structure qui est absente du nuage de point initial composé de n points déconnectés et désorganisés. Proposer une triangulation présente l'avantage de résumer la variété par un objet purement combinatoire, rendant possible le calcul efficace de quantités géométriques ou d'invariants topologiques de M. La qualité d'approximation d'un estimateur $\hat{M}$ de M peut, par exemple, tre mesurée par la distance de Hausdorff,

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

où pour $K \subset \mathbb{R}^D$ $d_K(x) = \inf_{y \in K} \|x - y\|$ désigne la fonction distance à K. Aussi, on peut demander ce que $\hat{M}$ ait la mme topologie que M.

État de l'art. En analyse de données déterministe — lorsque le nuage $\mathbb{X}$ n'est pas considéré comme aléatoire — les résultats récents de [2] donnent un algorithme de reconstruction explicite pourvu que $\mathbb{X}$ est échantillonné de manière assez dense dans M. Celui-ci est basé sur le complexe de Delaunay tangentiel (Figure 1), une version de la triangulation de Delaunay ambiente pour laquelle on ne garde que les simplexes dont la direction est proche des espaces tangents. Schématiquement, les résultats de

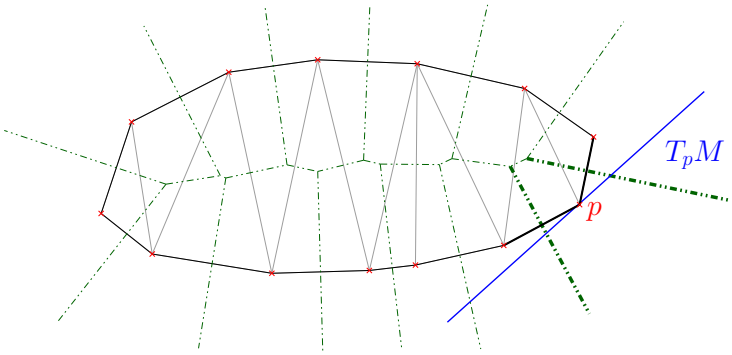

Figure 1: Complexe de Delaunay tangentiel

[2] affirment que sous une hypothèse de type $\mathcal{C}^2$ — le reach, un paramètre d'échelle introduit dans [4] encodant la fois des propriétés de régularité globale ainsi que de courbure —, si $d_H(M, \mathbb{X}) \leq \varepsilon$, alors le complexe de Delaunay tangentiel $\hat{M}_{\mathrm{TDC}}$ construit sur $\mathbb{X}$ est isotope M et vérifie $d_H(M, \hat{M}_{\mathrm{TDC}}) \leq \varepsilon^2$. Malheuresement, cette méthode n'autorise pas $\mathbb{X}$ tre bruité, et nécessite la connaissance des espaces tangents de M en chaque point de $\mathbb{X}$.

Indépendament, les auteurs de [3] ont décrit les vitesses optimales d'estimation lorsque le nuage de points $\mathbb{X}$ est tiré aléatoirement sur M, de manière indépendante et identiquement distribuée. Ici, M est supposée avoir les mmes propriétés de régularité que dans [2]. Si l'on note n le nombre de points échantillonés, la vitesse optimale d'approximation pour la distance d_H est de l'ordre de $(\log n/n)^{2/d}$. De plus, la méthode est robuste aux données aberrantes. Cependant, aucun algorithme n'est proposé pour le calcul effectif d'un estimateur $\hat{M}$ qui atteindrait la vitesse optimale.

Contribution. Dans [1], nous proposons dans un cadre aléatoire non-asymptotique un algorithme basé sur le complexe de Delaunay tangentiel $\hat{M}_{\mathrm{TDC}}$ de [2] qui atteint la vitesse optimale $(\log n/n)^{2/d}$ donnée dans [3]. Cet algorithme est basé uniquement sur le nuage de point et ne nécessite pas la connaissance des espaces tangents. Ceux-ci sont estimés par ACP locale et utilisés par plug-in dans $\hat{M}_{\mathrm{TDC}}$. Ce résultat montre que le complexe de Delaunay tangentiel peut tre considéré comme optimal d'un point de vue de la reconstruction. Inversement, il montre que la vitesse d'approximation $(\log n/n)^{2/d}$ peut tre obtenue avec des triangulations, et qu'elle est atteignable par un algorithme quadratique en n.
De plus, nous développons une méthode de débruitage lorsque des valeurs aberrantes sont présentes (Figure 2). Cette méthode se base sur un comptage de points contenus dans des pavés centrés en les points de $\mathbb{X}$, et dont les directions sont données par les espaces tangents estimés par ACP locale. Ce débruitage en une étape n'étant pas suffisant pour retrouver les vitesses optimales, une version itérative est aussi proposée.

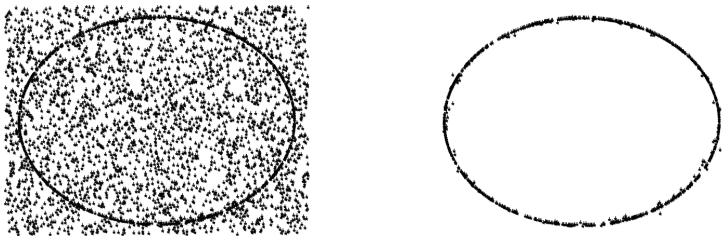

Figure 2: Nuage de point bruité (g.) et débruité (d.)

Références

[1] Eddie Aamari et Clment LevrardStability and Minimax Optimality of Tangential Delaunay Complexes for Manifold ReconstructionArXiv e-prints

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

[3] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, et Larry WassermanManifold estimation and singular deconvolution under Hausdorff loss*Ann. Statist.*, 40(2):941–963, 2012

[4] Herbert FedererCurvature measures*Trans. Amer. Math. Soc.*, 93:418–491, 1959

Eddie Aamari, Inria Saclay, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau
`eddie.aamari@inria.fr`