Estimation minimax de variétés avec les complexes de Rips

Vincent DIVOL, DataShape, Inria Saclay

Il est de plus en plus courant dans les applications modernes de sciences des données de récolter des jeux de données se présentant comme un ensemble de n points dans un espace euclidien de grande dimension $\mathbb{R}^D$ (avec $n \ll D$). Une manière raisonnable d'appréhender ces données consiste à considérer qu'elles se trouvent proches d'une structure sous-jacente M de dimension d, petite par rapport à la dimension ambiante D : on suppose donc que le phénomène pertinent à expliquer est caractérisé par un "petit" nombre de variables. On supposera ici que la structure M en question est une sous-variété de $\mathbb{R}^D$, et que le nuage de points est construit comme un n-échantillon $X_1, \dots, X_n$ tiré selon une certaine loi P supportée sur la variété M. Le problème que l'on se fixe est alors celui de la reconstruction de la variété M à partir du nuage de points.
Cette question a été traitée d'un point de vue asymptotique en considérant le problème de l'estimation minimax d'une telle variété dans [1]. On y montre que, si l'on se restreint à l'ensemble $\mathcal{P}$ des lois appelées presque-uniformes, supportées sur des sous-variétés de dimension d dont le reach (un paramètre de régularité) est contrôlé, alors le risque minimax est d'ordre $\left(\frac{\ln n}{n}\right)^{-2/d}$, i.e.

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}} \mathbb{E}_{X_1,\dots,X_n \sim P}[H(\hat{M}, M)] \asymp \left(\frac{\ln n}{n}\right)^{-2/d}, \tag{1}$$

où H est la distance de Hausdorff entre sous-ensembles de $\mathbb{R}^D$, et l'infimum est pris sur l'ensemble des estimateurs de M. Ceci indique que la vitesse $\left(\frac{\ln n}{n}\right)^{-2/d}$ est la vitesse optimale d'estimation. Ce résultat reste cependant théorique, les estimateurs proposés par [1] n'étant absolument pas implémentables en pratique. D'autres estimateurs, calculables en pratique (pour d petit) ont été par la suite proposés par [2]. La performance de ces estimateurs dépendent cependant fortement de plusieurs paramètres, dont le calibrage pose de réels problèmes pratiques.
Nous proposons ici une méthode simple, basée sur les complexes de Rips pour estimer une variété. Soit $\mathbb{X}_n = \{X_1, \dots, X_n\}$ l'ensemble des données observées. On définit pour $\alpha > 0$ le complexe de Rips de paramètre α sur $\mathbb{X}_n$ par

$$\hat{M}_\alpha = \bigcup_{\sigma \subset \mathbb{X}_n, \mathrm{diam}(\sigma) \leq \alpha} \mathrm{Conv}(\sigma), \tag{2}$$

où $\mathrm{Conv}(\sigma)$ est l'enveloppe convexe de l'ensemble fini σ. Cette famille d'estimateurs peut se calculer rapidement à partir de la matrice des distances du nuage de points. On montre que pour un certain choix de α (dépendant des paramètres du modèle, en pratique inconnus), cet estimateur est minimax. Plus important, nous proposons une méthode pour calibrer ce paramètre α en pratique, inspirée par la méthode PCO de [3], et montrons que ce calibrage mène à un choix optimal, répondant ainsi à certaines limitations des estimateurs précédemment évoqués.

Références

[1] Genovese, C., Perone-Pacifico, M., Verdinelli, I., Wasserman, L., *Minimax manifold estimation*, Journal of machine learning research, 2012.

[2] Aamari, E., Levrard, C., *Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction*, Discrete & Computational Geometry, 2018

[3] Lacour, C., Massart, P., Rivoirard, V., *Estimator selection: a new method with applications to kernel density estimation*, Sankhya A, 2017

Vincent DIVOL, DataShape, Inria Saclay
vincent.divol@inria.fr