Consistency of Persistent Homology

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

We first recall the consistency:

Suppose we obtain a sample $X_1, \ldots, X_n \sim P$. Let $\theta(P)$ be a parameter, which is some function of P. Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ denote an estimator for $\hat{\theta}$, which is a function of a sample. Consistency is about, whether the estimator $\hat{\theta}$ converge in probability to θ, i.e. $\hat{\theta} \xrightarrow{P} \theta$. More precisely, can we find some function $f(n)$ of the sample size n such that $d(\hat{\theta}, \theta) = O_P(f(n))$? This is analogous to the Law of Large Number.

Let $\mathbb{X} \subset \mathbb{R}^d$ be the target geometric structure, and P be a distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) = \mathbb{X}$. Let $X_1, \ldots, X_n$ be i.i.d. samples from P and $\mathcal{X} = \{X_1, \ldots, X_n\}$. For the consistency of persistent homology, the distance is the bottleneck distance d_B, and $\theta(P)$ and $\hat{\theta}(\mathcal{X})$ should be appropriate persistent homologies of P and $\mathcal{X}$, respectively. We consider two cases:

1. Persistent homologies from Čech complexes and Vietoris-Rips complexes. Let $\mathcal{PC}(\mathbb{X})$ and $\mathcal{PC}(\mathcal{X})$ be the persistent homologies induced from Čech complexes $\left\{H_k \check{\mathrm{C}}\mathrm{ech}_{\mathbb{R}^d}(\mathbb{X}, r)\right\}_{r \in \mathbb{R}}$ and $\left\{H_k \check{\mathrm{C}}\mathrm{ech}_{\mathbb{R}^d}(\mathcal{X}, r)\right\}_{r \in \mathbb{R}}$, respectively. Similarly, let $\mathcal{PR}(\mathbb{X})$ and $\mathcal{PR}(\mathcal{X})$ be the persistent homologies induced from Vietoris-Rips complexes $\{H_k \mathrm{Rips}(\mathbb{X}, r)\}_{r \in \mathbb{R}}$ and $\{H_k \mathrm{Rips}(\mathcal{X}, r)\}_{r \in \mathbb{R}}$, respectively. We would like to know $d_B(\mathcal{PC}(\mathbb{X}), \mathcal{PC}(\mathcal{X})) = O_P(f(n))$ and $d_B(\mathcal{PR}(\mathbb{X}), \mathcal{PR}(\mathcal{X})) = O_P(f(n))$.

Consistency of Čech complexes and Vietoris-Rips complexes

Assume $\mathbb{X}$ is compact. Recall the stability theorem for Čech complexes and Vietoris-Rips complexes:

Corollary. *For a compact set $\mathbb{X} \subset \mathbb{R}^d$ and $\mathcal{X} \subset \mathbb{X}$,*

$$d_B(\mathcal{PC}_{\mathbb{R}^d}(\mathbb{X}), \mathcal{PC}_{\mathbb{R}^d}(\mathcal{X})) \leq d_H(\mathbb{X}, \mathcal{X}).$$
$$d_B(\mathcal{PR}(\mathbb{X}), \mathcal{PR}(\mathcal{X})) \leq d_H(\mathbb{X}, \mathcal{X}).$$

For a distribution P, we assume (a, b) assumption:

Definition. P satisfies (a, b) assumption if there exists $r_0 > 0$ such that for all $x \in \mathrm{supp}(P)$ and for all $r < r_0$,

$$P(\mathcal{B}(x, r)) \geq a r^b.$$

Recall that under (a, b) assumption, we have probabilistic bound on the Hausdorff distance between $\mathbb{X}$ and $\mathcal{X}$:

Proposition ([2, Proposition 7.2][1, Theorem 2])**.** *Let P be a distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) = \mathbb{X}$, and assume P satisfies (a, b) assumption with $a, b > 0$. Let $X_1, \ldots, X_n$ be i.i.d. samples from P, and let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Then there exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$,*

$$P(d_H(\mathbb{X}, \mathcal{X}) < \epsilon) \geq 1 - a^{-1} \epsilon^{-b} \exp(-n a \epsilon^b). \tag{1}$$

This directly implies that with probability $1 - \delta$, with large enough n,

$$d_H(\mathbb{X}, \mathcal{X}) < C \left(\frac{\log n}{n}\right)^{1/b},$$

and hence

$$d_H(\mathbb{X}, \mathcal{X}) = O_P\left(\left(\frac{\log n}{n}\right)^{1/b}\right).$$

Then this implies both that

$$d_B(\mathcal{PC}_{\mathbb{R}^d}(\mathbb{X}), \mathcal{PC}_{\mathbb{R}^d}(\mathcal{X})) = O_P\left(\left(\frac{\log n}{n}\right)^{1/b}\right),$$
$$d_B(\mathcal{PR}(\mathbb{X}), \mathcal{PR}(\mathcal{X})) = O_P\left(\left(\frac{\log n}{n}\right)^{1/b}\right).$$

(1) not only gives the probabilistic bound as above, but this also gives the bound on the expectation as well. Roughly speaking, this is deduced from

$$\mathbb{E}\left[d_H(\mathbb{X}, \mathcal{X})\right] = \int_0^\infty P\left(d_H(\mathbb{X}, \mathcal{X}) > \epsilon\right) d\epsilon.$$

Theorem ([1, Theorem 4]). *Let P be a distribution on $\mathbb{R}^d$ with $\mathrm{supp}(P) = \mathbb{X}$, and assume P satisfies (a, b) assumption with $a, b > 0$. Let $X_1, \ldots, X_n$ be i.i.d. samples from P, and let $\mathcal{X} = \{X_1, \ldots, X_n\}$. Then,*

$$\mathbb{E}\left[d_H(\mathbb{X}, \mathcal{X})\right] \leq C\left(\frac{\log n}{n}\right)^{1/b},$$

where C only depends on a and b. And correspondingly,

$$\mathbb{E}\left[d_B(\mathcal{PC}_{\mathbb{R}^d}(\mathbb{X}), \mathcal{PC}_{\mathbb{R}^d}(\mathcal{X}))\right] \leq C\left(\frac{\log n}{n}\right)^{1/b},$$
$$\mathbb{E}\left[d_B(\mathcal{PR}(\mathbb{X}), \mathcal{PR}(\mathcal{X}))\right] \leq C\left(\frac{\log n}{n}\right)^{1/b}.$$

The convergence rate $\left(\frac{\log n}{n}\right)^{1/b}$ of Čech complexes and Vietoris-Rips complexes is in fact minimax up to a logarithmic term.

Theorem ([1, Theorem 4]). *Let $\mathcal{P}$ be a set of distributions P with $\mathrm{supp}(P)$ being compact and satisfying (a, b) assumption with fixed $a, b > 0$. Then for any estimator $\hat{\mathrm{dgm}}_n$ (that is, a function of data $X_1, \ldots, X_n$),*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[d_B(\mathcal{PC}_{\mathbb{R}^d}(\mathbb{X}), \mathcal{PC}_{\mathbb{R}^d}(\mathcal{X}))\right] \geq Cn^{-1/b},$$
$$\sup_{P \in \mathcal{P}} \mathbb{E}_P\left[d_B(\mathcal{PR}(\mathbb{X}), \mathcal{PR}(\mathcal{X}))\right] \geq Cn^{-1/b}.$$

References

[1] Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.*, 16:3603–3635, 2015.

[2] 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.