Statistics on Persistence Landscape

김지수 (Jisu KIM)

통계이론세미나 - 위상구조의 통계적 추정, 2023 가을학기

For a given task and data, Machine Learning / Deep Learning fits a parametrized model.

- Given data X,
- Parametrized model f_θ,
- Loss function $\mathcal{L}$ tailored to the task,
- Machine Learning minimizes $\arg\min_\theta \mathcal{L}(f_\theta, \mathcal{X})$.

For many cases, getting explicit formula for $\arg\min_\theta \mathcal{L}(f_\theta, \mathcal{X})$ is impossible or too costly (e.g., inverting a large scale matrix). So, gradient descent is used with the $\nabla_\theta \mathcal{L}(f_\theta, \mathcal{X})$:

$$\theta_{n+1} = \theta_n - \lambda \nabla_\theta \mathcal{L}(f_\theta, \mathcal{X}).$$

Application of Topological Data Analysis to Machine Learning is usually in two directions. First, a more common approach, is to use TDA as features, so that the data X is augmented with extra TDA features. Second approach is to accompany the loss function $\mathcal{L}$ with topological loss terms.

A persistence diagram is a multiset, and the space of persistence diagrams is complex. So directly applying a persistence diagram in machine learning is difficult, due to the complicated space structure, cardinality issues, computationally inefficient metrics, etc. If a persistence diagram is further summarized and embedded into a Euclidean space or a functional space, then applying in machine learning becomes much more convenient. Some examples are: persistence landscape, persistence silhouette, persistence image, etc.

The persistence landscape introduced in the study by [2] is an alternative representation of persistence diagrams. This approach aims at representing the topological information encoded in persistence diagrams as elements of a Hilbert space, for which statistical learning methods can be directly applied. The persistence landscape is a collection of continuous, piecewise linear functions $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ that summarizes a persistence diagram.

We first recall the definition of the persistence diagram:

Definition (Persistence Diagram)**.** Let $\mathcal{F}$ be a filtration and let $k \in \mathbb{N}_0$. The corresponding k-th persistence diagram $Dgm_k(\mathcal{F})$ is a finite multiset of $(\mathbb{R} \cup \{\infty\})^2$, consisting of all pairs (b, d) where either $[b, d)$, (b, d), $(b, d]$, or $[b, d]$ is the interval of filtration values for which a specific homology appears in $PH_k\mathcal{F}$. b is called a birth time and d is called a death time.

Throughout this paper, we will assume that persistence diagrams consist of finitely many points (b, d) with $-\infty < b < d < \infty$.

Let $\mathcal{D}$ be a persistence diagram. For a birth-death pair $p = (b, d) \in \mathcal{D}$, define a piecewise linear function $\Lambda_p : \mathbb{R} \to \mathbb{R}$ as

$$\begin{aligned} \Lambda_p(t) &= \max\{0, \min\{b + t, d - t\}\} \\ &= \begin{cases} t - b, & t \in \left[b, \frac{b+d}{2}\right], \\ d - t, & t \in \left(\frac{b+d}{2}, d\right], \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

In other words, a birth-death pair $p = (b, d)$ is rotated $\frac{\pi}{4}$ clockwise to become $\left(\frac{b+d}{2}, \frac{d-b}{2}\right)$, and then Λ_p is a tent function with $\left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ as its apex point.

The persistence landscape λ of $\mathcal{D}$ is a summary of the arrangement of piecewise linear curves obtained by overlaying the graphs of the functions $\{\Lambda_p\}_{p \in \mathcal{D}}$. See Figure .

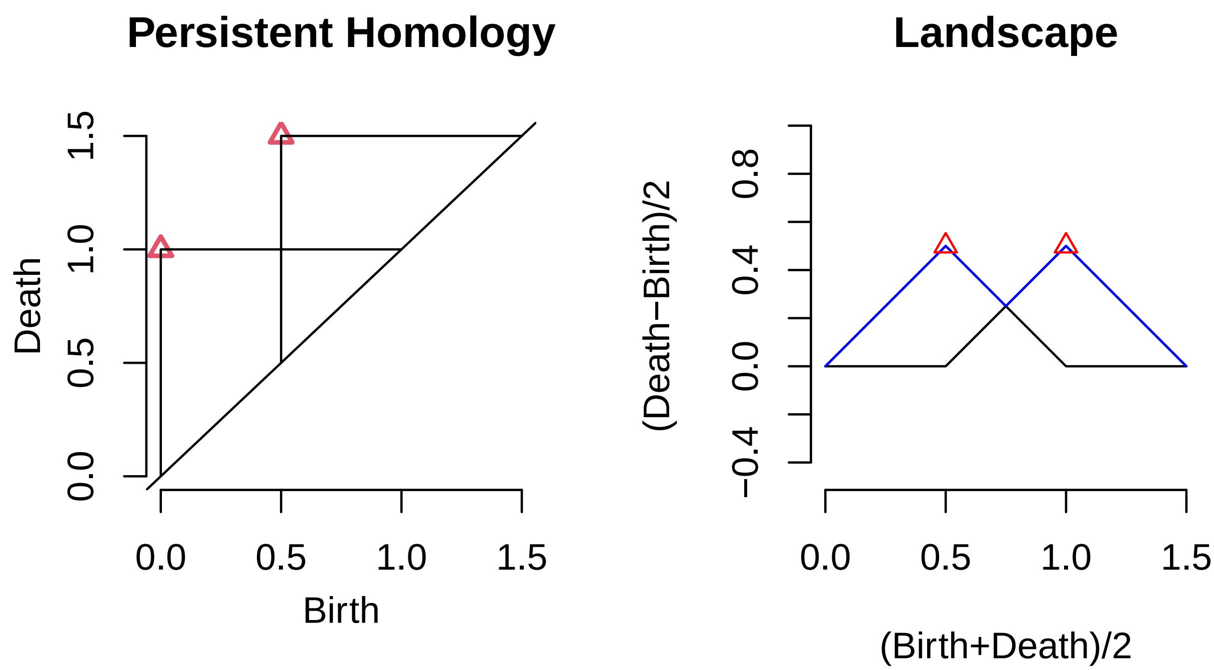

Figure 1: Persistence Landscape. The 1st landscape function is the blue curve.

Definition. For a persistence diagram $\mathcal{D}$, the corresponding persistence landscape is a function $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ defined as

$$\lambda(k,t) = \text{kmax}_{p \in \mathcal{D}} \Lambda_p(t),$$

where kmax is the kth largest value in the set; in particular, 1max is the usual maximum function. Given $k \in \mathbb{N}$, the function $\lambda_k : \mathbb{R} \to \mathbb{R}$ by $\lambda_k(t) = \lambda(k,t)$ is called the kth persistence landscape function.

Definition. The following is given informally in [2, Section 2.3]. It is proved more formally and precisely in [1].

Theorem. *The mapping from persistence diagrams to persistence landscapes is invertible.*

The advantage of the persistence landscape representation is two-fold. First, persistence diagrams are mapped as elements of a functional space, opening the door to the use of a broad variety of statistical and data analysis tools for further processing of topological features. Second, and fundamental from a theoretical perspective, the persistence landscapes share the same stability properties as those of persistence diagrams.

Stability

The persistence diagrams and persistence landscapes share the same stability properties. Recall the stability theorem for persistence modules:

Theorem ([3, Theorem 5.23])**.** *Let $\mathcal{PF}$ and $\mathcal{PG}$ be two q-tame persistence modules. Then*

$$d_B(\mathcal{PF}, \mathcal{PG}) \leq d_I(\mathcal{PF}, \mathcal{PG}).$$

We have the following stability theorem between persistence landscape's L_∞ distance and the bottleneck distance.

Theorem ([2, Theorem 13])**.** *For persistence diagrams $\mathcal{D}$and $\mathcal{D}'$, let λ_k and λ'_k be their k-th persistence landscape functions. Then,*

$$\|\lambda_k - \lambda'_k\|_\infty \leq d_B(\mathcal{D}, \mathcal{D}').$$

In particular, we have the stability theorem of persistence landscape of functions. We remark that we don't even need the tameness conditions.

Theorem ([2, Theorem 13])**.** *Let $f, g : \mathbb{X} \to \mathbb{R}$ be real-valued functions, and let $\mathcal{D}(f)$ and $\mathcal{D}(g)$ be the persistence diagrams induced from sublevel (or superlevel) filtration of f and g. Let λ_k^f and λ_k^g be their k-th persistence landscape functions. Then,*

$$\left\|\lambda_k^f - \lambda_k^g\right\|_\infty \leq \|f - g\|_\infty .$$

Asymptotic Normality and Bootstrap Confidence band

We first recall the bootstrap empirical process.

Bootstrap empirical process can be used to find a confidence band for a function $h(t)$; that is, we find a pair of functions $a(t)$ and $b(t)$ such that the probability that $h(t) \in [a(t), b(t)]$ for all t is at least $1 - \alpha$. I refer the reader to [4], Van der Vaart and Wellner [1996], and [5] for more details.

An empirical process is a stochastic process based on a random sample. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in the measure space $(\mathbb{X}, P)$. For a measurable function $f : \mathbb{X} \to \mathbb{R}$, we denote $Pf = \int f dP$ and $P_n f = \int f dP_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$. By the law of large numbers $P_n f$ converges almost surely to Pf. Given a class $\mathcal{F}$ of measurable functions, we define the empirical process $\mathbb{G}_n$ indexed by $\mathcal{F}$ as

$$\{\mathbb{G}_n f\}_{f\in\mathcal{F}} = \{\sqrt{n}(P_n f - Pf)\}_{f\in\mathcal{F}}.$$

$\ell^\infty(\mathcal{F})$ is the collection of all bounded functions $\phi : \mathcal{F} \to \mathbb{R}$, equipped with the sup norm. We say $\{\mathbb{G}_n f\}_{f\in\mathcal{F}}$ converges in distribution (or converges weakly) to $\{\mathbb{G} f\}_{f\in\mathcal{F}}$ in the space $\ell^\infty(\mathcal{F})$ if, for any bounded continuous function $H : \ell^\infty(\mathcal{F}) \to \mathbb{R}$, $\mathbb{E}H(\{\mathbb{G}_n f\}_{f\in\mathcal{F}}) \to \mathbb{E}H(\{\mathbb{G} f\}_{f\in\mathcal{F}})$ holds.

Definition ([4, Definition 1.3][5, Section 2.1])**.** A class $\mathcal{F}$ of measurable functions $f : \mathbb{X} \to \mathbb{R}$ is called P-Donsker if the process $\{\mathbb{G}_n f\}_{f\in\mathcal{F}}$ converges in distribution to a limit process in the space $\ell^\infty(\mathcal{F})$. The limit process is a Gaussian process $\mathbb{G}$ with zero mean and covariance function $\mathbb{E}[\mathbb{G}f\mathbb{G}g] := Pfg - PfPg$; this process is known as a Brownian Bridge.

One sufficient condition for Donsker class is to assume bound on the covering number: a set $\mathcal{C} = \{f_1, \ldots, f_N\}$ is an ϵ-cover of $\mathcal{F}$ if, for every $f \in \mathcal{F}$ there exists a $f_j \in \mathcal{C}$ such that $\|f - f_j\|_{L_2(Q)} < \epsilon$, and the size of the smallest ϵ-cover is called the covering numberand is denoted by $N_p(\mathcal{F}, L_2(Q), \epsilon)$.

Theorem ([4, Lemma 2.3][5, Theorem 2.5])**.** *Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with F satisfying $f(x) \leq F(x)$ for all $f \in \mathcal{F}$ with $PF^2 < \infty$. Suppose*

$$\int_0^1 \sqrt{\log \sup_Q \mathcal{N}(\mathcal{F}, L_2(Q), \epsilon \|F\|_{Q,2})} d\epsilon < \infty,$$

then $\mathcal{F}$ is P-Donsker.

Let $P_n^* f = \frac{1}{n}\sum_{i=1}^n f(X_i^*)$ where $\{X_1^*, \ldots, X_n^*\}$ is a bootstrap sample from P_n. the measure that puts mass $1/n$ on each element of the sample $\{X_1, \ldots, X_n\}$. The bootstrap empirical process $\mathbb{G}_n^*$ indexed by $\mathcal{F}$ is defined as

$$\{\mathbb{G}_n^* f\}_{f\in\mathcal{F}} = \{\sqrt{n}(P_n^* f - P_n f)\}_{f\in\mathcal{F}}.$$

Theorem ([4, Theorem 1.4][5, Theorem 2.6, Theorem 2.7])**.** *$\mathcal{F}$ is P-Donsker if and only if $\mathbb{G}_n$ converges in distribution to $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$.*

In words, above theorem states that $\mathcal{F}$ is P-Donsker if and only if the bootstrap empirical process converges in distribution to the limit process $\mathbb{G}$. Suppose we are interested in constructing a condence band of level $1 - \alpha$ for $\{Pf\}_{f\in\mathcal{F}}$, where $\mathcal{F}$ is P-Donsker. Let $\hat{\theta} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$. We proceed as follows:

1. Draw $X_1^*, \ldots, X_n^* \sim P_n$ and compute $\hat{\theta}^* = \sup_{f\in\mathcal{F}} |\mathbb{G}_n^* f|$.
2. Repeat the previous step B times to obtain $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
3. Compute $Z_\alpha = \inf\left\{r : \frac{1}{B}\sum_{j=1}^B I(\hat{\theta}_j^* \leq r) \geq 1 - \alpha\right\}$.

4. For $f \in \mathcal{F}$ define the confidence band $C_n(f) = \left[P_n f - \frac{Z_\alpha}{\sqrt{n}}, P_n f + \frac{Z_\alpha}{\sqrt{n}}\right]$.

Let $\mathcal{D}_T$ be the space of positive, countable, T-bounded persistence diagrams; that is, for each $(b, d) \in \mathcal{D}$, we have $0 \leq b \leq d \leq T$ and there are countable number of points where $d > 0$. Let the persistence diagrams $\mathcal{D}_1, \dots, \mathcal{D}_n$ be a sample from the distribution P over the space of persistence diagrams $\mathcal{D}_T$. Let $\lambda_1^{(k)}, \dots, \lambda_n^{(k)}$ be k-th persistence landscape functions corresponding to $\mathcal{D}_1, \dots, \mathcal{D}_n$. We define the mean landscape $\mu : \mathbb{R} \to \mathbb{R}$ as $\mu(t) = \mathbb{E}_P[\lambda_i^{(k)}(t)]$, and the empirical mean landscape $\bar{\lambda}_n : \mathbb{R} \to \mathbb{R}$ as $\bar{\lambda}_n(t) = \frac{1}{n}\sum_{i=1}^n \lambda_i^{(k)}(t)$. We here show that $\left\{\sqrt{n}(\bar{\lambda}_n(t) - \mu(t))\right\}_{t \in [0,T]}$ converges to a Gaussian process, so that we can use the bootstrap confidence band above.

Let $\mathcal{F} = \{f_t : 0 \leq t \leq T\}$, where $f_t : \mathcal{D}_T \to \mathbb{R}$ is defined by $f_t(\mathcal{D}) = \lambda_{\mathcal{D}}^{(k)}(t)$, where $\lambda_{\mathcal{D}}^{(k)}$ is the k-th persistence landscape function of the persistence diagram $\mathcal{D}$. We can write $\sqrt{n}(\bar{\lambda}_n(t) - \mu(t))$ as an empirical process indexed by $t \in [0, T]$:

$$\sqrt{n}(\bar{\lambda}_n(t) - \mu(t)) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \lambda_i^{(k)}(t) - \mathbb{E}_P[\lambda_i^{(k)}(t)]\right) = \sqrt{n}(P_n f_t - P f_t) \equiv \mathbb{G}_n f_t.$$

Then by considering a constant function $F \equiv T$, we have a uniform VC type bound on $\mathcal{N}(\mathcal{F}, L_2(Q), \epsilon \|F\|_{Q,2})$, and we can conclude that $\mathcal{F}$ is P-Donsker.

Theorem ([4, Theorem 2.4]). *Let $\mathbb{G}$ be a Brownian bridge with covariance function $\kappa(t, u) = \int_{\mathcal{D}_T} f_t(\mathcal{D}) f_u(\mathcal{D}) dP(\mathcal{D}) - \int_{\mathcal{D}_T} f_t(\mathcal{D}) dP(\mathcal{D}) \int_{\mathcal{D}_T} f_u(\mathcal{D}) dP(\mathcal{D})$. Then, $\mathbb{G}_n$ converges in distribution to $\mathbb{G}$.*

This theorem gives the asymptotic normality of persistence landscapes. Moreover, we can follow the bootstrap confidence band procedure. See Figure :

1. Draw $\mathcal{D}_1^*, \dots, \mathcal{D}_n^* \sim P_n$, construct corresponding k-th persistence landscape functions $\lambda_1^*, \dots, \lambda_n^*$.
2. Let $\bar{\lambda}_n^* = \frac{1}{n}\sum_{i=1}^n \lambda_i^*$ and compute $\hat{\theta}^* = \sup_{0 \leq t \leq T} \left|\sqrt{n}\left(\bar{\lambda}_n^*(t) - \bar{\lambda}_n(t)\right)\right|$.
3. Repeat step 1 and 2 B times to obtain $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$.
4. Compute $Z_\alpha = \inf\left\{r : \frac{1}{B}\sum_{j=1}^B I(\hat{\theta}_j^* \leq r) \geq 1 - \alpha\right\}$.
5. Define the confidence band $C_n(t) = \left[\bar{\lambda}_n(t) - \frac{Z_\alpha}{\sqrt{n}}, \bar{\lambda}_n(t) + \frac{Z_\alpha}{\sqrt{n}}\right]$.

Theorem ([4, Theorem 2.5]). *The bootstrap confidence band $C_n(t) = \left[\bar{\lambda}_n(t) - \frac{Z_\alpha}{\sqrt{n}}, \bar{\lambda}_n(t) + \frac{Z_\alpha}{\sqrt{n}}\right]$ is a confidence band for $\mu(t)$:*

$$P\left(\mu(t) \in C_n(t) \text{ for all } t\right) \geq 1 - \alpha.$$

References

[1] Leo Betthauser, Peter Bubenik, and Parker B. Edwards. Graded persistence diagrams and persistence landscapes. *Discrete Computational Geometry*, 67(1):203–230, July 2021.

[2] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16:77–102, 2015.

[3] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules.* SpringerBriefs in Mathematics. Springer, [Cham], 2016.

[4] Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Aarti Singh, and Larry A. Wasserman. On the bootstrap for persistence diagrams and landscapes. *CoRR*, abs/1311.0376, 2013.

[5] Michael R. Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer Series in Statistics. Springer, New York, 2008.

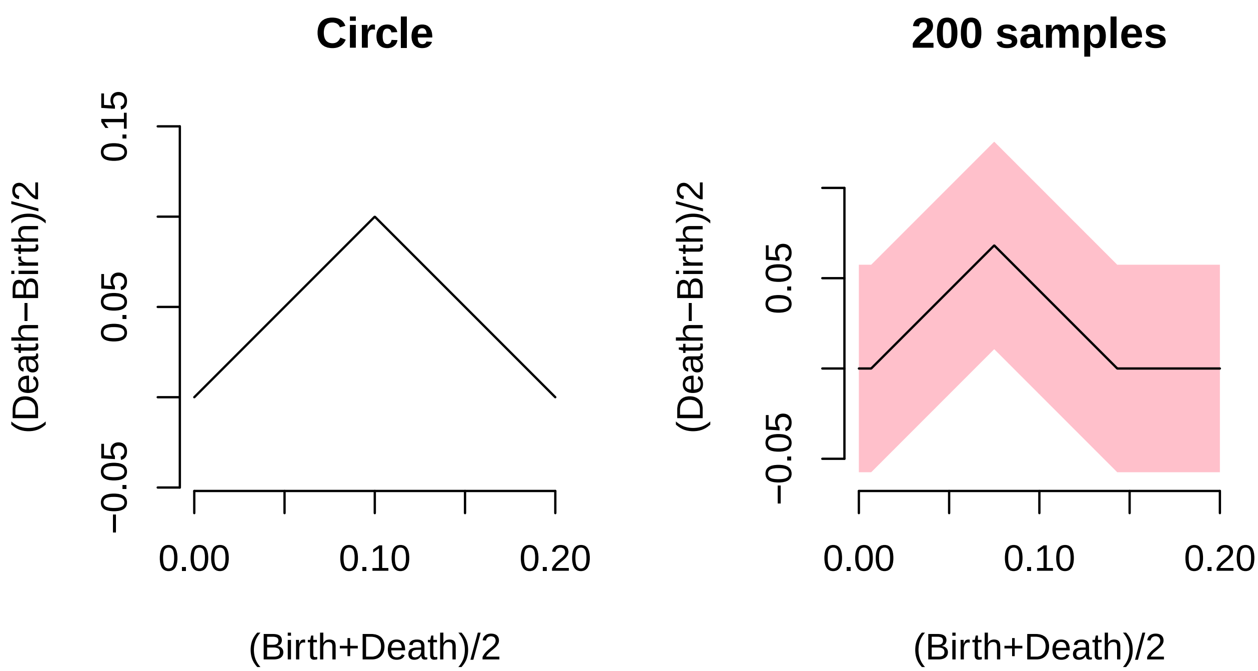

Figure 2: Confidence Band for Persistence Landscape.