Topological Data Analysis

2022–2023

Lecture 12

Statistical Inference Using Landscapes

19 December 2022

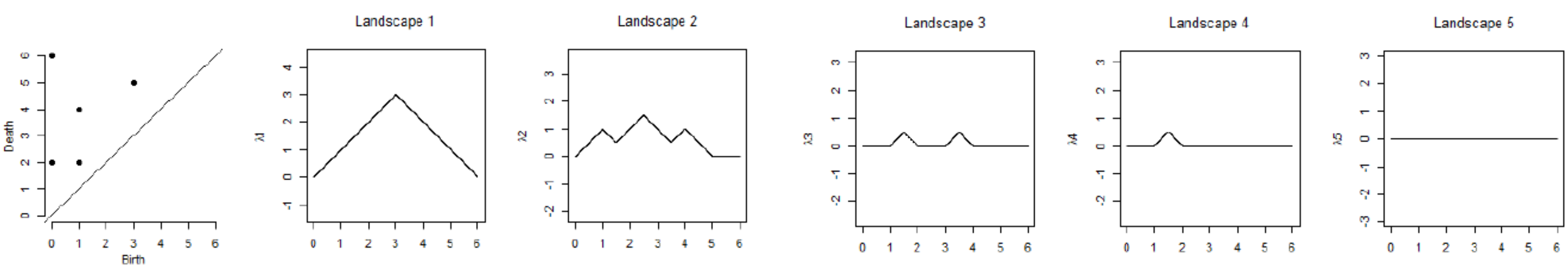
Landscape Sample Mean

For a point cloud X, let

$$\Lambda(X) = \{\lambda_k(X)\}_{k\geq 1}$$

denote the sequence of **persistence landscapes** associated to the Vietoris–Rips barcode of X.

Thus, $\lambda_k(X)\colon \mathbb{R} \to \mathbb{R}$ is a piecewise linear function with compact support for each $k \in \mathbb{N}$, and $\Lambda(X)$ may be viewed as an element of the **Banach space** $\boldsymbol{L^p(\mathbb{N} \times \mathbb{R})}$ for every $p \geq 1$.



Landscape Sample Mean

Now suppose that we treat X as a **random variable** (for example, a random point cloud on a sphere or a torus). Then $\Lambda(X)$ is a random variable with values in $L^p(\mathbb{N} \times \mathbb{R})$.

If $X_1, \ldots, X_n$ are independent, identically distributed copies of X, we may consider the **sample mean** $\overline{\Lambda(X)}_n \in L^p(\mathbb{N} \times \mathbb{R})$:

$$\overline{\Lambda(X)}_n(k, t) = \frac{1}{n} \sum_{i=1}^{n} \lambda_k(X_i)(t).$$

The **Central Limit Theorem** implies that, for $p \geq 2$, if the expected values $E\|\Lambda(X)\|_p$ and $E\|\Lambda(X)\|_p^2$ are finite, then

$$\sqrt{n}\ [\overline{\Lambda(X)}_n - E(\Lambda(X))]$$

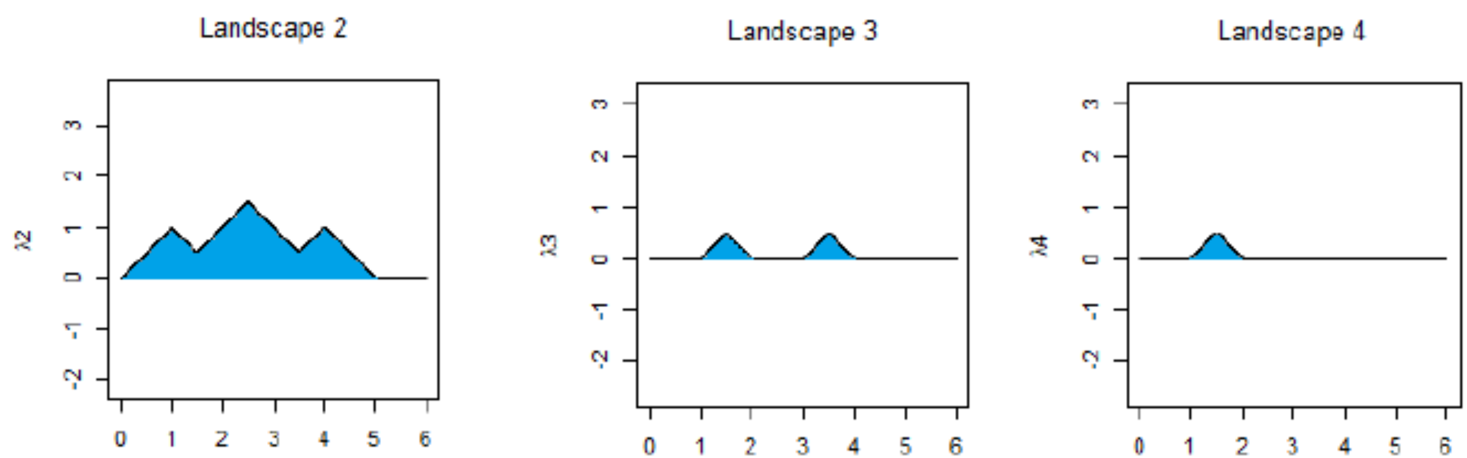
converges weakly to a Gaussian random variable.

Landscape Sample Mean

As a consequence of this fact, if we define

$$Y = \|\Lambda(X)\|_1 = \int_{\mathbb{N}\times\mathbb{R}} \Lambda(X) = \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} \lambda_k(X)(t)\, dt,$$

then Y has the property that $\sqrt{n}\,[\,\bar{Y}_n - E(Y)]$ converges to a normal distribution with zero mean.



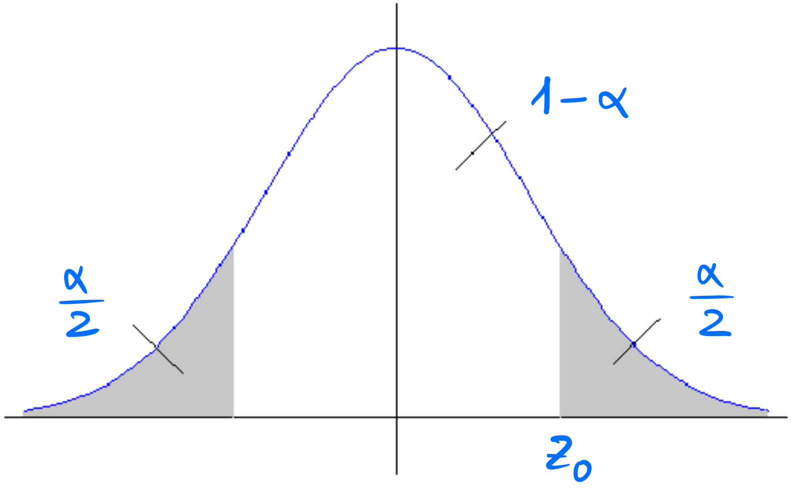
Confidence Intervals

This allows us to use $\bar{Y}_n$ as an estimator for confidence intervals.

Namely, if $(S_n^Y)^2$ is the **sample variance** of Y, then

$$\bar{Y}_n \pm z_0 \frac{S_n^Y}{\sqrt{n}}$$

is a $(1-\alpha)$ **confidence interval** for $E(Y)$, where z_0 is the upper $\alpha/2$ critical value for a $N(0,1)$ distribution.



Confidence Intervals

Assuming that $\sqrt{n}\,[\,\bar{Y}_n - E(Y)]$ has approximately a normal distribution with zero mean, our choice of z_0 ensures that

$$P\left(-z_0 \leq \frac{\bar{Y}_n - E(Y)}{S_n^Y/\sqrt{n}} \leq z_0\right) = 1 - \alpha.$$

This expression can be rewritten as

$$P\left(\bar{Y}_n - z_0\frac{S_n^Y}{\sqrt{n}} \leq E(Y) \leq \bar{Y}_n + z_0\frac{S_n^Y}{\sqrt{n}}\right) = 1 - \alpha,$$

which is the meaning of a confidence interval for $E(Y)$.

Hypothesis Testing

Let $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_m$ be samples of two random variables X and X'. Consider

$$Y = \|\Lambda(X)\|_1 \quad \text{and} \quad Y' = \|\Lambda(X')\|_1.$$

If we denote $\mu = E(Y)$ and $\mu' = E(Y')$, then the null hypothesis that $\mu = \mu'$ can be tested by means of the estimator

$$z = \frac{\bar{Y}_n - \bar{Y}'_m}{\sqrt{\dfrac{(S_n^Y)^2}{n} + \dfrac{(S_m^{Y'})^2}{m}}}$$

where S^2 stands for the sample variance. This yields hypothesis testing for point clouds by means of persistence landscapes.

Confidence Bands

Let $\xi_1, \ldots, \xi_n$ be Gaussian random variables with mean 0 and variance 1. For each k, consider the **multiplier bootstrap**

$$\mathbb{G}_n(k,t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \Big(\lambda_k(X_i)(t) - \overline{\lambda_k(X)}_n(t) \Big)$$

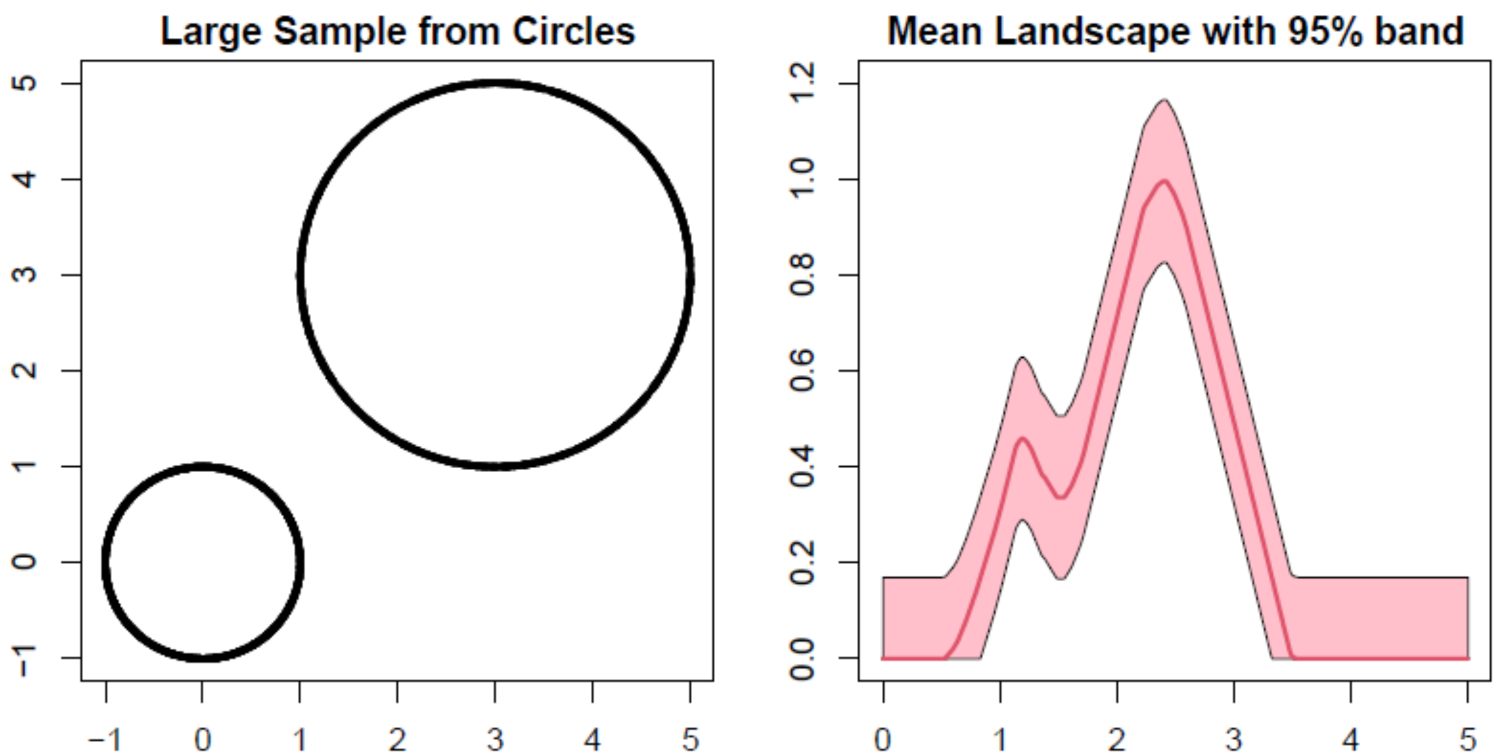
and determine Z_α (by Monte Carlo simulation) so that, for a fixed k,

$$P\left(\sup_t |\mathbb{G}_n(k,t)| > Z_\alpha\right) = \alpha.$$

Then a $(1-\alpha)$ **confidence band** for $E(\lambda_k(X))$ is given by

$$\boldsymbol{\overline{\lambda_k(X)}_n \pm \frac{Z_\alpha}{\sqrt{n}}}.$$

Confidence Bands



From a sample of 4000 points from two circles, 10 subsamples of size 80 have been extracted and the first landscape λ_1 for homological dimension 1 has been averaged and depicted with a 95% confidence band.

Confidence Bands

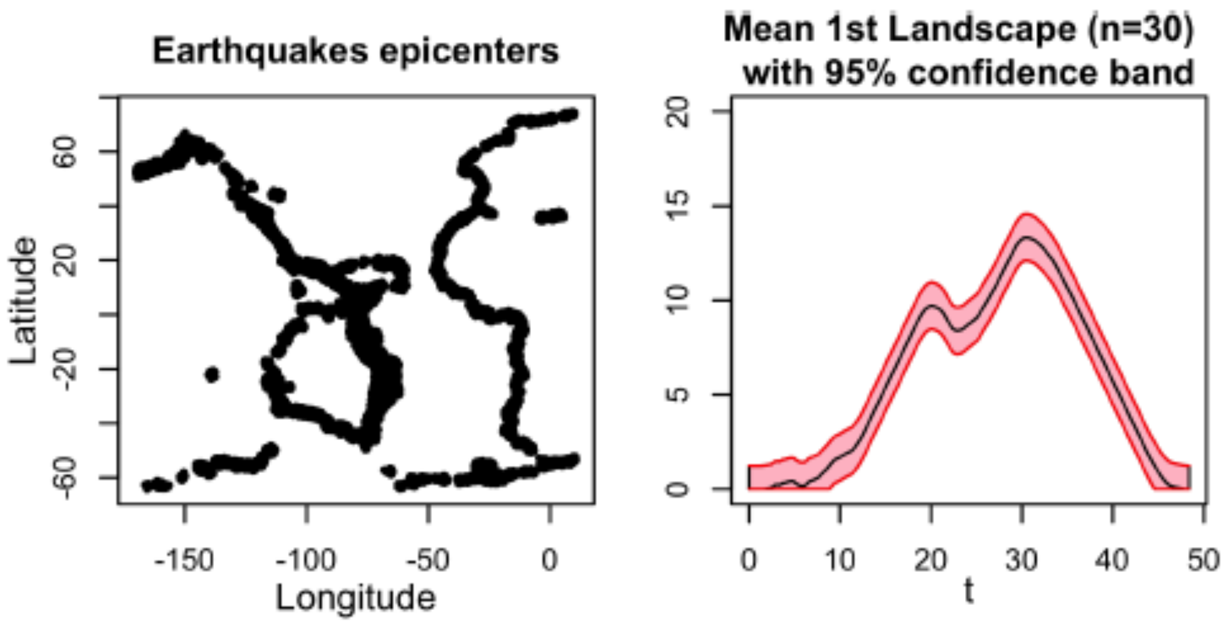


Figure 7: The plot on the left shows 8000 epicenters of earthquakes in the latitude/longitude rectangle $[-75, 75] \times [-170, 10]$ of magnitude greater than 5.0 recorded between 1970 and 2009 (USGS data). We randomly sampled $m = 400$ epicenters and computed the approximated persistence diagram of the distance function (Betti 1). We repeated this procedure $n = 30$ times and computed the empirical average landscape $\overline{\lambda}_n$. Using the multiplier bootstrap described in Chazal et al. (2013b), we obtained a uniform 95% confidence band for the average landscape $\mu(t)$ (right).

Source: F. Lezzi, PhD thesis proposal, Carnegie Mellon University (2014)

References

- Landscapes were introduced in **[P. Bubenik, Statistical topological data analysis using persistence, *J. Machine Learning Res.* 16 (2015), 77–102],** arXiv:1207.6437 (2015).
- Confidence bands for average landscapes were described in **[F. Chazal, B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, Stochastic convergence of persistence landscapes and silhouettes, *J. Comput. Geom.* 6 (2015), 140–161],** arXiv:1312.0308 (2013).
- Other estimators can be found in **[F. Chazal et al., Robust topological inference: distance to a measure and kernel distance, *J. Machine Learning Res.* 18 (2018), 1–40],** arXiv:1412.7197 (2014).