Journal of Data Science and Modeling, Vol. 1, No. 1, 33-44, December 2022

Research Manuscript

Statistical Topology Using the Nonparametric Density Estimation and Bootstrap Algorithm

Soroush Pakniat*[1]

1. M.Sc. graduate, Allameh Tabataba'i University, Tehran, Iran.

Received: 12/4/2017 Accepted: 30/7/2017

Abstract:

This paper presents approximate confidence intervals for each function of parameters in a Banach space based on a bootstrap algorithm. We apply a kernel density approach to estimate the persistence landscape. In addition, we evaluate the quality distribution function estimator of random variables using the integrated mean square error (IMSE). The results of the simulation studies show a significant improvement achieved by our approach compared to the standard version of confidence intervals algorithm. Finally, the real data analysis demonstrate the accuracy of our method compared to previous works for computing the confidence interval.

Keywords: Nonparametric topological data analysis, Persistence landscape, Persistence homology, Bootstrap method, Density estimation.

Mathematics Subject Classification (2010): 57R19, 57N65, 05E45, 62Gxx.

*Corresponding author: sorush.pakniat@atu.ac.ir

1. Introduction

In recent years, the increased rate of data generation in some fields has emerged the need for some new approaches to extract knowledge from large data sets. One of the approaches for data analysis is topological data analysis (TDA), which refers to a set of methods for estimating topological structure in data (point cloud)(see the survey Carlsson (2009); Ghrist (2008); Carlsson (2014); Edelsbrunner and Harer (2009)). A persistence homology is a fundamental tool for extracting topological features in the nested sequence of subcomplexes (Edelsbrunner et al (2002)). In Chazal and Bertrand (2017), the authors introduced a TDA from the perspective of data scientists. Since the use of TDA has been limited by combining machine learning and statistic subjects, we need to create a set of real-valued random variables that satisfy the usual central limit theorem and allow us to obtain approximate confidence interval and hypothesis testing. In the present study, we propose an alternative approach to approximate the sampling distribution and compute interval without some presupposition. This approach, which is asymptotically more accurate than the computation of standard intervals, analyzes a sample data population and identify the probability distribution of data. Due to the limitation of barcode and persistence diagram with combining statistics, we use a sequence of function such that $\lambda_k(t) : \mathbb{R} \to \bar{\mathbb{R}}$ where $\bar{\mathbb{R}}$ denotes the extended real numbers and $\lambda_k(t)$ is persistence landscape (Bubenik (2015)). Next, we create a real-valued random variable by applying some functional in separable Banach space, and we obtain the list of real-valued random variables. In the present work, we aimed at proposing a nonparametric inference of data to infer an unknown quantity to keep the number of underlying assumptions as weak as possible. The remainder of this paper is organized as follows: In section 2, we review the necessary background of persistence landscape. In section 3, we provide theoretical background from nonparametric approach. Finally, in section 4, we apply our approach on a sampling of objects.

2. Background of Persistence Landscape

A simplicial complex K is defined for representing a manifold and triangulation of topological space X. K is a combinatorial object that is stored easily in computer memory and can be constructed by several methods in high dimensions with any metric space. A subcomplex L of simplicial complex K is a simplicial complex such that $L \subseteq K$. A filtration of simplicial complex K is a nested sequence of

subcomplexs such that $K^0 \subseteq K^1 \subseteq \ldots \subseteq K^m$. To create this object, you can see the (Khuyen et al (2014); Chambers et al (2010); K. Dey et al (2013) and Silva and Carlsson (2004)).

The fundamental group of space X ($\pi_1(X, x_0)$ at the basepoint x_0), as an important functor in algebraic topology, consist of loops and deformations of loops. The fundamental group is one of the homotopy group $\pi_n(X, x_0)$ that has a higher differentiating power from space X, however, this invariant of topological space X depends on smooth maps and is very complicated to compute in high dimensions. Thus, we must use an invariant of topological space that is computable on the simplicial complex. Homology groups show how cells of dimension n attach to subcomplex of dimension $n-1$ or describe holes in the dimension of n (connected components, loops, trapped volumes,etc.). The nth homology group is defined as $H_n = \ker \partial_n / \mathrm{im} \partial_{n+1} = Z_n / B_n$ such that ∂_n is the boundary homomorphism of subcomplexs, Z_n is the cycle group and B_n is boundary group. The nth Betti number β_n of a simplicial complex K is defined as $\beta_n = rank(Z_n) - rank(B_n)$. Through filtration step, we tend to extract invariant that remains fixed in this process, thus persistence homology satisfies this criterion for space-time analysis. Let K^l be a filtration of simplicial complex K, the pth persistence of nth homology group of K^l is $H_n^{b,d} = Z_n^b / (B_n^{b+d} \cap Z_n^b)$. The Betti number of pth persistence of nth homology group is defined as $\beta_n^{b,d}$ for the rank of free subgroup $(H_n^{b,d})$. To visualize persistence in space-time analysis, we should find the interval of (i, j) that is invariant constantly through the filtration and obtain a topological summary from the point cloud.

Now, by rewriting the Betti number of the pth persistence of nth homology group, we have:

$$\lambda(b, d) = \begin{cases} \beta^{b,d} & \text{if } b \leq d \\ 0 & otherwise \end{cases}$$

To convert $\lambda(b, d)$ function to a decreasing function, we change coordinate on it, Let $m = \dfrac{b+d}{2}$ and $h = \dfrac{d-b}{2}$, The rescaled rank function is:

$$\lambda(m, h) = \begin{cases} \beta^{m-h, m+h} & \text{if } h \geq 0 \\ 0 & otherwise \end{cases}$$

Definition 2.1. *The persistence landscape is a function $\lambda : \mathbb{N} \times \mathbb{R} \to \bar{\mathbb{R}}$ where $\bar{\mathbb{R}}$ denoted the extended real numbers (introduced by Bubenik (2015)). In the other words, persistence landscape is sequence of function $\lambda_k : \mathbb{R} \to \bar{\mathbb{R}}$ such that:*

$$\lambda_k(t) = \sup(m \geq 0 | \beta^{t-m, t+m} \geq k). \tag{2.1}$$

We assume that our persistence landscape lies in separable Banach space (L^p). Let $Y : (\Omega, \mathcal{F}, \mathcal{P}) \to \mathbb{R}$ be a real value random variable on underlying probability space, Ω is a sample space, $\mathcal{F}$ is a σ-algebra of events, and $\mathcal{P}$ is a probability measure. The expected value $E(Y) = \int Y dP$ and Λ is the corresponding persistence landscape. If f is a functional member of L^q with $\frac{1}{p} + \frac{1}{q} = 1$, let

$$Y = \int f\Lambda = \| f\Lambda \|_1 .$$

Then

$$\sqrt{n}[\bar{Y_n} - E(Y)] \xrightarrow{\mathrm{d}} N(0, Var(Y)).$$

3. Nonparametric on Persistence Landscapes

The basic idea of this approach is to use data to infer an unknown quantity without any presumption. For a more detailed exposition, we refer the reader to Wasserman (2006). The first problem is to estimate the cumulative distribution function (CDF), which is an important problem in our approach.

Definition 3.1. *Let $X_1, \ldots, X_n \sim F$ where $F(x) = P(X \leq x)$. We estimate F with the empirical distribution function $\widehat{F}_n$ which is the CDF that puts mass $\frac{1}{n}$ at each data point X_i. Formally,*

$$\widehat{F}_n = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

where

$$I(X_i \leq x) = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_1, \ldots X_n \sim F$ and let $\widehat{F}_n$ be the empirical CDF, Then, at any fixed value of x $E(\widehat{F}_n(x)) = F(x)$ and $V(\widehat{F}_n(x)) = \dfrac{F(x)(1 - F(x))}{n}$, where $V(\widehat{F}_n(x))$ denotes variance of empirical CDF.

Definition 3.2. *A statistical functional $T(F)$ is any function of F. The plug-in estimator of $\theta = T(F)$ is defined by*

$$\widehat{\theta}_n = T(\widehat{F}_n).$$

A functional of the form $\int a(x)dF(x)$ is called a linear functional where $a(x)$ denoted a function of x. The plug-in estimator for linear functional $T(F) =$

$\int a(x)dF(x)$ *is:*

$$T(\widehat{F}_n) = \int a(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} a(X_i).$$

For an approximation of the standard error of a plug-in estimator, use the influence function as follows:

Definition 3.3. *The Gâteaux derivative of T at F in the direction G is defined by:*

$$L_F(G) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G) - T(F)}{\epsilon}$$

The empirical influence function is defined by $\widehat{L}(x) = L_{\widehat{F}_n}(x)$. Thus,

$$\widehat{L}(x) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\widehat{F}_n + \epsilon G) - T(\widehat{F}_n)}{\epsilon}.$$

Definition 3.4. *If T is Hadamard differentiable with respect to $d(F,G) = \sup_x |F(x) - G(x)|$ then*

$$\sqrt{n}(T(\widehat{F}_n) - T(F)) \rightsquigarrow N(0, \tau^2)$$

where $\tau^2 = \int L_F(x)^2 dF(x)$ and $\rightsquigarrow$ denotes convergence in distribution. Also,

$$\frac{(T(\widehat{F}_n) - T(F))}{\widehat{se}} \rightsquigarrow N(0,1)$$

Such that

$$\widehat{se} = \frac{\widehat{\tau}}{\sqrt{n}} \text{ and } \widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n} L^2(X_i).$$

3.1 Bootstrap Confidence Intervals

There are several ways to construct bootstrap confidence intervals that are difference from accuracy criterion. Let $\theta = T(F)$ and $\widehat{\theta}_n = T(\widehat{F}_n)$ be an estimator for θ. We tend to estimate a nonparametric confidence interval for functions of θ. The pivot $R_n = \widehat{\theta}_n - \theta$. Let $H(r)$ denotes the CDF of the pivot:

$$H(r) = P_F(R_n \leq r).$$

Let $C_n^* = (a, b)$ where

$$a = \widehat{\theta}_n - H^{-1}\big(1 - \frac{\alpha}{2}\big) \text{ and } b = \widehat{\theta}_n - H^{-1}\big(\frac{\alpha}{2}\big)$$

Since a and b depend on the unknown distribution H, we should form a bootstrap estimate of H as:

$$\widehat{H}(r) = \frac{1}{B}\sum_{i=1}^{B} I(R_{n,b}^* \leq r)$$

Where $R^*_{n,b} = \widehat{\theta}^*_{n,b} - \widehat{\theta}_n$. Let r^*_β denote the β sample quantile of $(R^*_{n,1}, \ldots, R^*_{n,B})$ and let θ^*_β denote the β sample quantile of $(\theta^*_{n,1}, \ldots, \theta^*_{n,B})$. Note that $r^*_\beta = \theta^*_\beta - \widehat{\theta_n}$. Follows that an approximate $1-\alpha$ confidence interval is $C_n = (\widehat{a}, \widehat{b})$ is a nonparametric confidence interval a least $(1-\alpha)$, where

$$\widehat{a} = \widehat{\theta}_n - \widehat{H}^{-1}(1 - \frac{\alpha}{2}) = \widehat{\theta}_n - r^*_{1-\alpha/2} = 2\widehat{\theta}_n - \theta^*_{1-\alpha/2}$$
$$\widehat{b} = \widehat{\theta}_n - \widehat{H}^{-1}(\frac{\alpha}{2}) = \widehat{\theta}_n - r^*_{\alpha/2} = 2\widehat{\theta}_n - \theta^*_{\alpha/2}.$$

3.2 Quality of Estimator

The goal of nonparametric density estimation is to estimate f with as few assumptions about f as possible. We denote the estimator by $\widehat{f}_n$. We will evaluate the quality of an estimator $\widehat{f}_n$ with the risk, or integrated mean squared error, $R = \mathbb{E}(L)$ where

$$L = \int (\widehat{f}_n(x) - f(x))^2 dx \tag{3.2}$$

is the integrated squared error loss function. The estimators depend on some smoothing parameter h chosen by minimizing an estimate of the risk. The loss function, which we now refer to as function of h, is:

$$\begin{aligned} L &= \int (\widehat{f}_n(x) - f(x))^2 dx \\ &= \int \widehat{f}_n^2(x) dx - 2\int \widehat{f}_n(x) f(x) dx + \int f^2(x) dx. \end{aligned}$$

The last term does not depend on h so minimizing the loss is equivalent to minimizing the expected value, therefore the cross-validation estimator of risk is:

$$\widehat{J}(h) = \int \big(\widehat{f}_n(x)\big)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i) \tag{3.3}$$

where $\widehat{f}_{(-i)}$ is the density estimator obtained after removing the i^{th} observation.

Theorem 3.5. *Suppose that f' is absolutely continuous and that $\int \big(f'(u)\big)^2 du < \infty$, Then,*

$$R(\widehat{f}_n, f) = \frac{h^2}{12}\int \big(f'(u)\big)^2 du + \frac{1}{nh} + o(h^2) + o(\frac{1}{n}). \tag{3.4}$$

Where $x_n = o(a_n)$ this means that $\lim_{n\to\infty} x_n/a_n = 0$. The value h^* that minimizes (3.5) is

$$h^* = \frac{1}{n^{1/3}}\left(\frac{6}{\int (f'(u))^2 du}\right)^{1/3}. \tag{3.5}$$

With this choice of binwidth,

$$R(\widehat{f}_n, f) \sim \frac{C}{n^{2/3}} \tag{3.6}$$

where $C = (3/4)^{2/3}\Big(\int \big(f'(u)\big)^2 du\Big)^{1/3}$.

The proof of Theorem 3.5 can be seen in appendix Wasserman (2006). We see that with an optimally chosen binwidth, the risk decreses to 0 at rate $n^{-2/3}$. Moreover, it can be seen that kernel estimators converge at the faster rate $n^{-4/5}$ and that in a certain sense no faster rate is possible.
We discuss kernel density estimators, which are smoother and can converge to the true density faster. Here, the word kernel refers to any smooth function K such that $K(x) \geq 0$ and

$$\int K(x)dx = 1, \ \int xK(x)dx = 0 \text{ and } \sigma_K^2 \equiv \int x^2 K(x)dx > 0. \tag{3.7}$$

Some commonly used kernels are the following: where

the Gaussian kernel:	$K(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2/2}$
the tricube kernel:	$K(x) = \frac{70}{81}\Big(1 - \lvert x\rvert^3\Big)^3 I(x)$

$$I(x) = \begin{cases} 1 & \text{if } |x| \ \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Definition 3.6. *Given a kernel K and a positive number h, called the bandwidth, the kernel density estimator is defined to be*

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n} K\Big(\frac{x - X_i}{h}\Big). \tag{3.8}$$

Theorem 3.7. *Assume that f is continuous at x, $h_n \to 0$, and $nh_n \to \infty$ as $n \to \infty$. Then, $\widehat{f}_n(x) \to f(x)$.*

Proof. The claim proof by weak low of large number(WLLN) states that the $\widehat{f}_n(x)$ converges with a probabiliy towards the probability density of random variables of persistence landscape. □

4. Applications

In this section, we calculated the nonparametric methods on persistence landscapes to confirm accuracy of our methods respect to another approach, using R programming language with TDA package by Terese Fasy et al (2014).

4.1 Sphere and Torus

We sample from the sphere and torus uniformly with respect to the surface. Let R be the major radius and r as the minor radius, we use an explicit equation in Cartesian coordinates for a torus, which is:

$$\left(R - \sqrt{x^2 + y^2}\right)^2 + z^2 = r^2.$$

For 1000 points, we construct a filtered simplicial complex as follows. First, we form the Vietoris-Rips complex $R(X, \epsilon)$, which consists of simplices with vertices in $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and diameter at most ϵ. The sequence of Vietoris-Rips complex obtained by gradually increasing the radius ϵ create a filtration of complexes. We denote the limit of filtration of the Vietoris-Rips complex with 5 and maximum dimension of homological feature with 1(0 for components, 1 for loops). To compute landscape function in Equation 2.1, we set $t \in [0, 5], k = 1$. We construct 100 random variables by some functional in Banach space, the logarithm function is our plug-in estimator, and the empirical influence function is different among random variables with plug-in estimator. We repeated the related algorithm for 100 times to obtain the upper and lower confidence interval. Table 1 present the nonparametric bootstrap computed using the approach for a 95% critical value with a few assumptions about persistence landsapces. As shown in Figures 1, we create 100 random variables and 500 times bootstrap sample data and replaced with orginal data. We showed that using a confidence interval such as $\bar{Y} \pm z^* \frac{\sigma}{\sqrt{n}}$, gives 0.06628939 for density estimation of the sphere and 0.02067551 for torus, which is difference between upper and lower confidence interval. On the other hand, using nonparametric method with correct kernel as the tricube kernel and h^*, we obtained 0.0004472946 for sphere and 0.0003435891 for torus points, which are significant different. Now, to evaluate the quality of an estimator $\widehat{f}_n$ with respect to f with integrated mean squared error, we apply related algorithm which is obtain Figure 2 is for 100 times with 0.002 precision of bandwidth h and Gaussian kernel for sphere points and for torus with difference between below and upper confidence interval in 100 times, is 0.00004416.

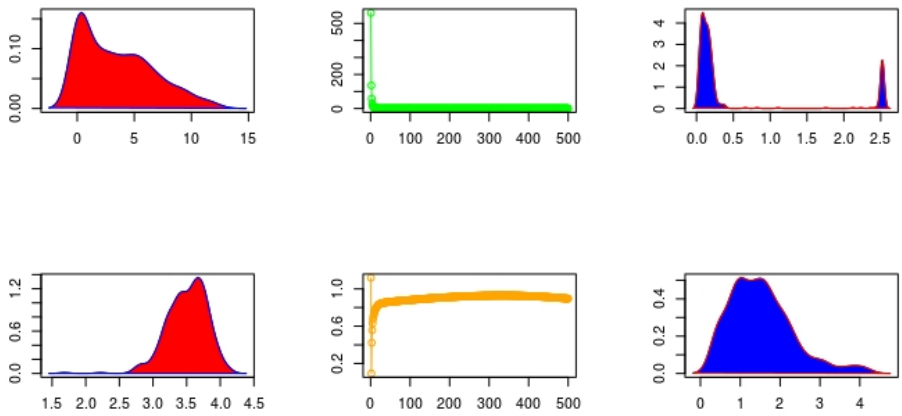

Figure 1: We sample 500 points uniformly for a sphere with radius 2, in row 1, column 1 plot density of random variables of persistence landscape, in row 1, column 2 plot $\widehat{j}(h)$ with 0.002 percision that minimum value is 0.0029, in row 1, column 3 plot kernel density estimator with bandwidth 0.056 tricube kernel. In row 2, column 1 we use bootstrap method for alternate generating random variate with persistence landscape, in row 2, column 2 plot $\widehat{j}(h)$ with 0.002 percision that minimum value is 0.0934, in row 2, column 3 plot kernel density estimator with bandwidth 0.004 and tricube kernel.

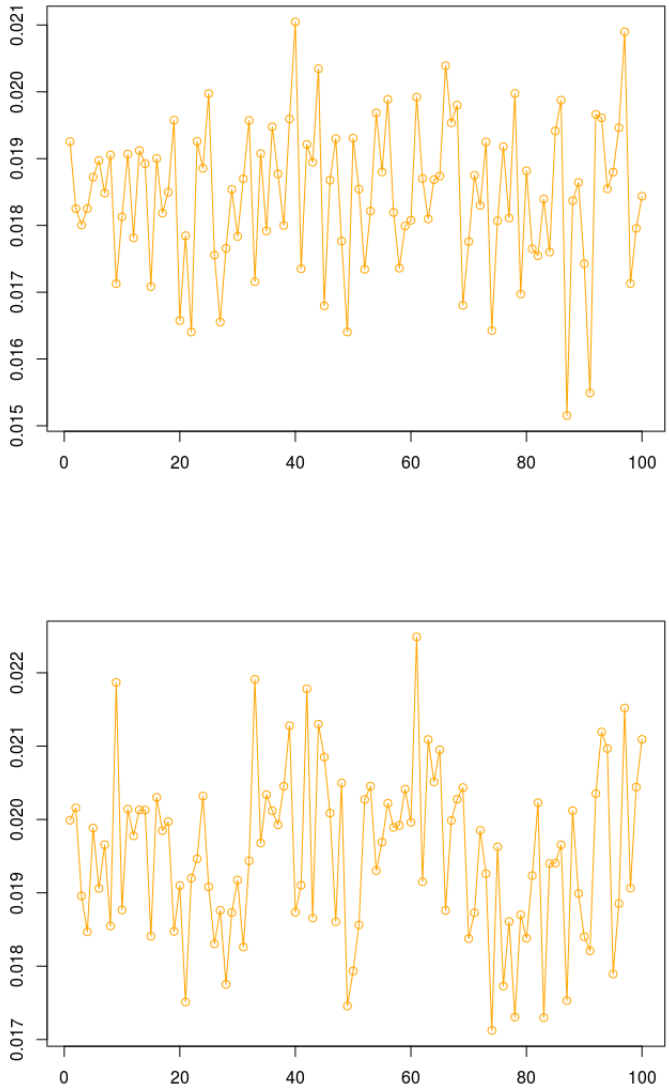

Figure 2: We run 100 times to evaluate minimize estimated risk $R(\widehat{f}_n, f)$ for points on sphere in row 1 and torus in row 2 with Guassian kernel.

Method	95% Interval
pivotal	(4.248799, 4.319345)
normal	(3.929793, 4.406627)
studentize	(4.297297, 4.418873)
percentile	(4.017076, 4.087621)

Method	95% Interval
pivotal	(1.498743, 1.540386)
normal	(1.241105, 1.566729)
studentize	(2.415224, 2.95656)
percentile	(1.267447, 1.309091)

Table 1: In the left column, we calculated bootstrap confidence interval with four commonly used accurate approaches. We sampled 1000 points for the sphere. The right column is the same for torus points.

Acknowledgment

The authors gratefully acknowledge the support of the center of statistical learning and its application at Allameh Tabatabai University (under grant No. P/H/040). We would like to thank Naiereh Elyasi for her helpful discussions.

References

Ghrist, R. 92008), Barcodes: The persistent topology of data, *BULLETIN (New Series) OF THE AMERICAN MATHEMATICAL SOCIETY* , **45**.

Carlsson, G. (2014), Topological pattern recognition for point cloud data, *Acta Numerica,* **23**, 289-368.

Edelsbrunner, H. and Harer, J. (2009), *computational topology an introduction*, American Mathematical Society.

Edelsbrunner, H., Letscher D. and Zomorodian, A. (2002), Topological persistence and simplification, *Discrete and Computational Geometry,* **28**, 511–533.

Zomorodian, A. (2005), Topology for Computing, *Cambridge University Press.*

Carlsson, G. (2009), Topology and Data, *Bulletin of the American Mathmatical society,* **2**, 255-308.

Chazal, F. and Bertrand, M. (2017), An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists, *arxiv:1710.04019v1.*

Bubenik, P. (2015), Statistical Topological Data Analysis using Persistence Landscapes, *Journal of Machine Learning Research,* **16**, 77-102.

Le N. K., Martins, P., Decreusefond, L. and Vergne, A. (2014), Construction of the generalized Cech complex, *arXiv:1409.8225,2014.*

Chambers, E.W., de Silva, V., Erickson, J. and Ghrist, R. (2010), Vietoris Rips Complexes of Planar Point Sets, *Discrete and Computational Geometry*, **44**(1), 75-90.

Dey, T.K., Fan, F. and Wang, Y., (2013), Graph Induced Complex on Point Data, *In Proceedings of the Twenty-ninth Annual Symposium on Computational Geometry*, 107–116.

De Silva, V. and Carlsson, G. (2004), Topological estimation using witness complexes, *Proc. Sympos. Point-Based Graphics.*

Wasserman , L. (2006), *All of Nonparametric Statistics*, Springer.

Fasy, B.T., Kim, J., Lecci, F. and Maria, C. (2014), Introduction to the R package TDA, *arXiv:1411.1830* 2014.