Journal of Symbolic Computation 78 (2017) 91–114

Contents lists available at ScienceDirect

Journal of Symbolic Computation

www.elsevier.com/locate/jsc

A persistence landscapes toolbox for topological statistics ☆

Peter Bubenik [a], Paweł Dłotko [b,c]

[a] Department of Mathematics, University of Florida, Gainesville, FL, USA
[b] Geometrica, INRIA Saclay, Ile-de-France, France
[c] Institute of Computer Science, Jagiellonian University, Krakow, Poland

ARTICLE INFO

Article history:
Received 8 January 2015
Accepted 29 November 2015
Available online 26 March 2016

Keywords:
Topological data analysis
Persistent homology
Statistical topology
Topological machine learning
Intrinsic dimension

ABSTRACT

Topological data analysis provides a multiscale description of the geometry and topology of quantitative data. The persistence landscape is a topological summary that can be easily combined with tools from statistics and machine learning. We give efficient algorithms for calculating persistence landscapes, their averages, and distances between such averages. We discuss an implementation of these algorithms and some related procedures. These are intended to facilitate the combination of statistics and machine learning with topological data analysis. We present an experiment showing that the low-dimensional persistence landscapes of points sampled from spheres (and boxes) of varying dimensions differ.

1. Introduction

We provide some algorithms and computational tools for statistical topological data analysis. In particular, we give algorithms for calculating the persistence landscape, a functional summary of persistence modules. We also give algorithms for calculating the averages of such summaries, and for calculating distances between such averages. These tools also provide an alternative computational approach for calculating distances between topological summaries that may be useful when other

☆ PB is supported by AFOSR grant FA9550-13-1-0115. PD is supported by the Advanced Grant of the European Research Council GUDHI 339025 (Geometric Understanding in Higher Dimensions), DARPA grant FA9550-12-1-0416 and AFOSR grant FA9550-14-1-0012.

E-mail addresses: peter.bubenik@ufl.edu (P. Bubenik), pawel.dlotko@inria.fr (P. Dłotko).

http://dx.doi.org/10.1016/j.jsc.2016.03.009
0747-7171/

methods are computational prohibitive. In addition, we specify an implementation of these algorithms and some related tools that we have made publicly available.

We are motivated by *topological data analysis* (Ghrist, 2008; Carlsson, 2009). Its main tool, *persistent homology* provides a multiscale description of the topology of the data of interest, called either a *barcode* or a *persistence diagram*. Unfortunately this summary is difficult to work with from the point of view of statistics and machine learning. For example, it is not feasible to calculate averages. For these purposes, it is convenient to replace these summaries with a linear summary, that is, a finite- or infinite-dimensional vector. In a linear space it is easy to calculate averages. One such vector which does not lose any information is the functional summary called the *persistence landscape* (Bubenik, 2015). Since this summary may be thought of as lying in a Hilbert space, in the language of machine learning, it is a *feature map*. There is an associated *kernel* (Reininghaus et al., 2015) to which standard machine learning tools may be applied.

1.1. Background

In the simplest computational setting for topological data analysis, the data of interest is encoded in a finite filtered complex,

$$\mathcal{K}_0 \subset \mathcal{K}_1 \subset \ldots \subset \mathcal{K}_n. \tag{1}$$

This is a filtration of the complex $K = K_n$ and it is sometime convenient to add $K_{-1} = \emptyset$. *Persistent homology* (Edelsbrunner et al., 2002; Zomorodian and Carlsson, 2005) gives a multiscale representation of the topology of this complex. To be precise, one applies homology in some degree with coefficients in some field to (1) to obtain a sequence of finite-dimensional vector spaces and linear maps,

$$H(\mathcal{K}_0) \to H(\mathcal{K}_1) \to \ldots \to H(\mathcal{K}_n), \tag{2}$$

called a *persistence module*. It turns out that the persistence module can be completely described by a finite sequence of pairs $\{(b_i, d_i)\}$, with $b_i < d_i$. For each such pair (b_i, d_i) there is a choice of a nonzero homology class $\alpha_i \in H(\mathcal{K}_{b_i})$ that is not in the image of $H(\mathcal{K}_{b_i-1})$ and whose image is nonzero in $H(\mathcal{K}_{d_i-1})$ but is zero in $H(\mathcal{K}_{d_i})$. One sometimes says that α_i is born at b_i and dies at d_i. Furthermore, the homology classes $\{\alpha_i\}$ and their nonzero images under the maps in (2) give a basis for the vector spaces in (2). Considering these pairs as points in the plane, one obtains the *persistence diagram*. Considering them as intervals $[b_i, d_i)$ one obtains the *barcode*. We will often refer to them as *birth–death pairs*. In the simple setting of (1), we have $b_i, d_i \in \{0, 1, \ldots, n\}$. However we can generalize to $b_i, d_i \in \mathbb{R}$ by associated a corresponding increasing sequence of real numbers with (1). This summary is stable (Cohen-Steiner et al., 2007, 2010; Chazal et al., 2012) in that small perturbations of the data will lead to small perturbations of these pairs, under suitable choices of distance. Successful applications of topological data analysis include breast cancer data (Nicolau et al., 2011), sensor networks (de Silva and Ghrist, 2007), orthodontic data (Gamble and Heo, 2010), signal analysis (Perea and Harer, 2015), target tracking (Bendich et al., 2014a), and brain artery data (Bendich et al., 2016).

Now let us define the persistence landscape (Bubenik, 2015). First, for a birth–death pair (b, d), let us define the piecewise linear function $f_{(b,d)} : \mathbb{R} \to [0, \infty]$.

$$f_{(b,d)} = \begin{cases} 0 & \text{if } x \notin (b, d) \\ x - b & \text{if } x \in (b, \frac{b+d}{2}] \\ -x + d & \text{if } x \in (\frac{b+d}{2}, d) \end{cases} \tag{3}$$

The *persistence landscape* of the birth–death pairs $\{(b_i, d_i)\}_{i=1}^n$ is the sequence of functions $\lambda_k : \mathbb{R} \to [0, \infty]$, $k = 1, 2, 3, \ldots$ where $\lambda_k(x)$ is the k-th largest value of $\{f_{(b_i,d_i)}(x)\}_{i=1}^n$. We set $\lambda_k(x) = 0$ if the k-th largest value does not exist; so $\lambda_k = 0$ for $k > n$. Equivalently, the persistence landscape is a function $\lambda : \mathbb{N} \times \mathbb{R} \to [0, \infty]$, where $\lambda(k, t) = \lambda_k(t)$. In this definition we have assumed that b and d are finite. In the appendix we show that this definition extends to the cases where b and/or d are infinite.