Computational Geometry 58 (2017) 70–96

Contents lists available at ScienceDirect

Computational Geometry: Theory and Applications

www.elsevier.com/locate/comgeo

Efficient and robust persistent homology for measures

Mickaël Buchet *, Frédéric Chazal, Steve Y. Oudot, Donald R. Sheehy

ARTICLE INFO

Article history:
Received 27 May 2014
Received in revised form 11 August 2015
Accepted 3 June 2016
Available online 7 July 2016

Keywords:
Persistent homology
Topological data analysis
Distance to a measure
Power distance
Sparse rips filtration

ABSTRACT

A new paradigm for point cloud data analysis has emerged recently, where point clouds are no longer treated as mere compact sets but rather as empirical measures. A notion of distance to such measures has been defined and shown to be stable with respect to perturbations of the measure. This distance can easily be computed pointwise in the case of a point cloud, but its sublevel-sets, which carry the geometric information about the measure, remain hard to compute or approximate. This makes it challenging to adapt many powerful techniques based on the Euclidean distance to a point cloud to the more general setting of the distance to a measure on a metric space.
We propose an efficient and reliable scheme to approximate the topological structure of the family of sublevel-sets of the distance to a measure. We obtain an algorithm for approximating the persistent homology of the distance to an empirical measure that works in arbitrary metric spaces. Precise quality and complexity guarantees are given with a discussion on the behavior of our approach in practice.

1. Introduction

Given a sample of points P from a metric space $\mathbb{X}$, the distance function d_P maps each $x \in \mathbb{X}$ to the distance from x to the nearest point of P. The related fields of geometric inference and topological data analysis have provided a host of theorems about what information can be extracted from the distance function, with a particular focus on discovering and quantifying intrinsic properties of the shape underlying a data set [4,19]. The flagship tool in topological data analysis is persistent homology and the most common goal is to apply the persistence algorithm to distance functions, either in Euclidean space or in metric spaces [1,14,23]. From the very beginning, this line of research encountered two major challenges. First, distance functions are very sensitive to noise and outliers (Fig. 1 left). Second, the representations of the sublevel sets of a distance function become prohibitively large even for moderately sized data. These two challenges led to two distinct research directions. First, the distance to the data set was replaced with a distance to a measure induced by that data set [5]. The resulting theory is provably more robust to outliers, but the sublevel sets become even more complex to represent (Fig. 1 center). Towards more efficient representations, several advances in *sparse filtrations* have led to linear-size constructions [12,20,21], but all of these methods exploit the specific structure of the distance function and do not obviously generalize. In this paper, we bring these two research directions together by showing how to combine the robustness of the distance to a measure, with the efficiency of sparse filtrations.

* Corresponding author.
E-mail addresses: mickael.buchet@m4x.org (M. Buchet), frederic.chazal@inria.fr (F. Chazal), steve.oudot@inria.fr (S.Y. Oudot), don.r.sheehy@gmail.com (D.R. Sheehy).

http://dx.doi.org/10.1016/j.comgeo.2016.07.001
0925-7721/© 2016 Elsevier B.V. All rights reserved.

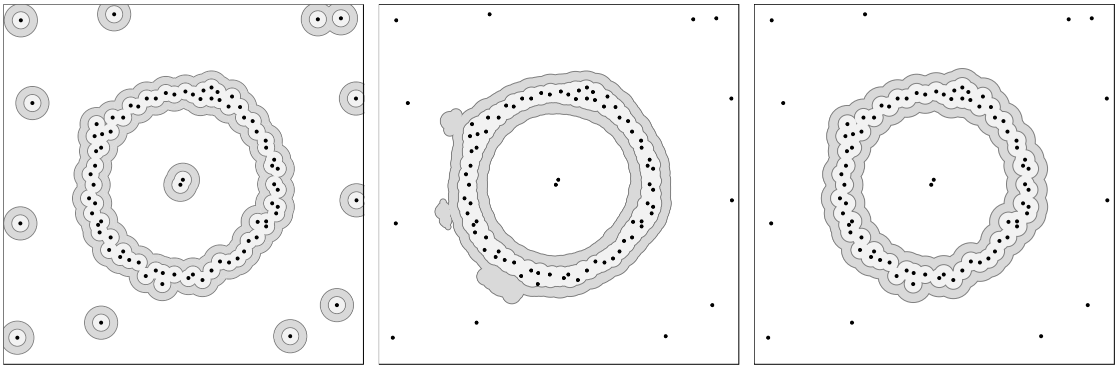

Fig. 1. From left to right, two sublevel sets for d_P, $d_{\mu_P,m}$, and $d^P_{\mu_P,m}$ with $m = \frac{3}{|P|}$. The first is too sensitive to noise and outliers. The second is smoother, but substantially more difficult to compute. The third is our approximation, which is robust to noise, efficient to compute, and compact to represent.

Contributions:

1. **A Generalization of the Wasserstein stability and persistence stability of the distance to a measure for triangulable metric spaces.**
2. **A general method for approximating the sublevel sets of the distance to a measure by a union of balls.** Our method uses $O(n)$ balls for inputs of n samples. Known methods for representing the exact sublevel sets can require $n^{\Theta(d)}$ balls. Existing approximations using a linear number of balls are only applicable in Euclidean space [15].
3. **A linear size approximation to the weighted Rips filtration.** For intrinsically low-dimensional metric spaces, we construct a filtration of size $O(n)$ that achieves a guaranteed quality approximation. Specifically, if the doubling dimension of the metric is d then the size complexity is $2^{O(dk)}n$ if one considers simplices up to dimension k (see Definition 2.1 for the formal definition of doubling dimension). This is a significant improvement over the full weighted Rips filtration, which has size 2^n in general or size n^{k+1} if one considers only simplices up to dimension k. It also has the advantage that the sparsification is independent of the weights. Thus, the (geo)metric preprocessing phase can be reused for any weighting of the points. If one attempted to use previous methods directly, this preprocessing phase would have to be repeated for each set of weights. This is especially useful if one is interested in several different weight functions such as when approximating the distance to a measure for several different values of the mass parameter.
4. **An effective implementation with experimental results.**

Overview of the paper Originally, the distance to a measure was introduced to capture information about both scale and density in a Euclidean point cloud. We extend the distance to a measure to any metric space $\mathbb{X}$. We write $\bar{B}(x,r)$ to denote the closed ball with center x and radius r. The distance to a measure is then defined as follows.

Definition 1.1. Let μ be a probability measure on a metric space $\mathbb{X}$ and let $m \in]0,1]$ be a mass parameter. We define the distance $d_{\mu,m}$ to the measure μ as

$$d_{\mu,m} : x \in \mathbb{X} \mapsto \sqrt{\frac{1}{m}\int_0^m \delta_{\mu,l}(x)^2 dl},$$

where $\delta_{\mu,m}$ is defined as

$$\delta_{\mu,m} : x \in \mathbb{X} \mapsto \inf\{r > 0 \mid \mu(\bar{B}(x,r)) > m\}.$$

The distance to a measure has interesting inference and stability results in the Euclidean setting [5]. That is, the sublevel sets of the function can be used to infer the topology of the support of the underlying distribution (inference), and also, the output for similar inputs will be similar (stability). In Section 3, we extend these stability results to any metric space. The results about the stability of persistence diagrams apply to any triangulable metric space, i.e. metric spaces homeomorphic to a locally finite simplicial complex (the persistence diagram may not exist for non-triangulable metric spaces).

We then give a new way to approximate the distance to a measure. Using a sampling of the support of a measure, we are able to compute accurately the sublevel sets of the distance to a measure in any metric space, using power distances. We show in Section 4.1 that these functions have adequate stability and approximation properties. Then, in Section 4.2, we give the practical implications for computing persistence diagram for finite samples.

The *witnessed k-distance* is another approach to approximating the distance to a measure proposed in [15]. This approach works only in Euclidean spaces as it relies on the existence of barycenters of points. The analysis links the quality of the