Physica D 334 (2016) 99–117

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids

Vanessa Robins [a,*], Katharine Turner [b]

[a] Applied Mathematics, Research School of Physics and Engineering, The Australian National University, Canberra, Australia

[b] EPFL Chair of Mathematical Statistics and Laboratory for Topology and Neuroscience, Lausanne, Switzerland

HIGHLIGHTS

- We use the persistent homology rank function to study spatial point patterns.
- The rank function is shown to be amenable to standard statistical techniques.
- We demonstrate null hypothesis testing on simulated point patterns.
- We develop a principal component analysis method for experimental data.
- PCA of rank functions is successfully applied to colloidal and sphere-packing data.

ARTICLE INFO

Article history:
Received 6 July 2015
Received in revised form
22 December 2015
Accepted 9 March 2016
Available online 23 March 2016

Keywords:
Topological data analysis
Betti numbers
Stochastic geometry

ABSTRACT

Persistent homology, while ostensibly measuring changes in topology, captures multiscale geometrical information. It is a natural tool for the analysis of point patterns. In this paper we explore the statistical power of the persistent homology rank functions. For a point pattern X we construct a filtration of spaces by taking the union of balls of radius a centred on points in X, $X_a = \cup_{x \in X} B(x, a)$. The rank function $\beta_k(X) : \{(a, b) \in \mathbb{R}^2 : a \leq b\} \to \mathbb{R}$ is then defined by $\beta_k(X)(a, b) = \operatorname{rank}(\iota_* : H_k(X_a) \to H_k(X_b))$ where ι_* is the induced map on homology from the inclusion map on spaces. We consider the rank functions as lying in a Hilbert space and show that under reasonable conditions the rank functions from multiple simulations or experiments will lie in an affine subspace. This enables us to perform functional principal component analysis which we apply to experimental data from colloids at different effective temperatures and to sphere packings with different volume fractions. We also investigate the potential of rank functions in providing a test of complete spatial randomness of 2D point patterns using the distances to an empirically computed mean rank function of binomial point patterns in the unit square.

0. Introduction

Random point patterns arise in a wide variety of application areas from astrophysics to materials science to ecology to protein interactions. The points might represent galaxies, colloidal particles, locations of trees, or molecules, and their distribution in space is indicative of the underlying processes that created the pattern. When studying such systems a number of questions arise. For example, could a given pattern be generated from a purely random process with no underlying interactions between the objects (points)? If this null hypothesis can be rejected then we would like to say whether a proposed theoretical model generates patterns that are consistent with experimental observations. We might also need to compare a large number of point patterns and classify them into different groups, or quantitatively track changes over time to understand the dynamics of a system.

Stochastic geometry provides various tools to characterise random spatial patterns and these have mostly focussed on first and second-order techniques analogous to means and variance in single-variable statistics [1]. Statisticians have historically used functional summary statistics to study point processes with important examples including Ripley's K-function (which describes the cumulative distribution function of pairwise distances), the empty

* Corresponding author.
E-mail addresses: vanessa.robins@anu.edu.au (V. Robins), katharine.turner@epfl.ch (K. Turner).

http://dx.doi.org/10.1016/j.physd.2016.03.007

space function and the nearest neighbour function. One advantage of summaries that are functions of a distance parameter is their ability to capture information on different length scales, but it is known that there are a number of situations where more sensitive tests of structural difference are required [2]. Persistent homology is an algebraic topological tool developed for data analysis that is an intuitively appealing measure of higher-order structure and encompasses spatial correlations of all orders [3,4]. This paper demonstrates how the information encoded by persistent homology can be converted into a form that is amenable to standard statistical analysis techniques such as hypothesis testing and functional principal component analysis (PCA).

Topology is the study of spatial objects equivalent under continuous deformations—the old joke is that a topologist cannot tell the difference between a coffee mug and a bagel. The homology groups of a space X, $H_k(X)$, $k = 0, 1, 2, \ldots$ are algebraically quantified topological invariants that provide information about equivalent points, loops, and higher dimensional analogues of loops. Homology detects a k-dimensional hole as a k-dimensional loop (cycle) that does not bound a $(k+1)$-dimensional piece of the object. The ranks of the homology groups are called *Betti numbers*: β_0 counts the number of connected components, β_1 the number of independent non-bounding loops, β_2 the number of enclosed spaces in a three-dimensional object. A solid bagel and a coffee mug both have $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 0$, while their surfaces have $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$. Homology groups and their Betti numbers are inherently global properties of an object that are sensitive to some geometric perturbations (tearing and gluing) and not to others (continuous deformation).

Another topological invariant that has recently been used to summarise structure in point patterns is the Euler characteristic signature function [2,5]. The Euler characteristic is the alternating sum of the Betti numbers (for three-dimensional objects, $\chi = \beta_0 - \beta_1 + \beta_2$). It is a topological invariant but also has measure-theoretic properties that make it more amenable to statistical analysis than the Betti numbers, but it is a less sensitive topological invariant by definition [6]. Nevertheless, the above studies have demonstrated that the Euler characteristic signature function is sensitive to short-range, higher-order correlations in point patterns.

To determine the homology groups, the topological space must be represented by simple building blocks of points, line segments, surface patches, and so on, that are joined together in a specific way, i.e., as a cell complex. The homology groups are then defined via a boundary operator, ∂_k, that maps each cell of dimension k onto the cells of dimension $k-1$ in its boundary:

$$\partial_k : C_k \to C_{k-1}. \tag{0.1}$$

The kernel of ∂_k is called the cycle group Z_k and the image of ∂_{k+1} is called the boundary group B_k. All boundaries are cycles, so we can form the homology group as the quotient $H_k = Z_k/B_k$. See the text by Hatcher for a comprehensive treatment of homology theory [7], or [8] for a concise overview aimed at physicists.

When examining point data, a parameter must be introduced to define which points are connected to one another and so build the cell complex. A crucial lesson learnt in the early days of topological data analysis is that instead of trying to find a single best value for this parameter, much more is learned by looking at how the homology evolves over a sequence of parameter values. Rather than working with a single cell complex we work with a *filtration*, a family of spaces K_a such that $K_a \subset K_b$ whenever $a \le b$. Often this parameter a is a length scale, so that although we are measuring topological quantities, the way these change tells us about the geometrical features of the data set. For example, if we used a filtration of $\mathbb{R}^3$ defined by lower level sets of the distance function to the surface of an ideally-smooth bagel, we would be able to read the radius of the hole of the bagel from the function $\beta_1(K_a)$ as it is at that radius that the space of loops changes from one to zero dimensional. If the bagel is a real one with irregular cross-section, bumps and splits, the Betti number function $\beta_1(K_a)$ will not be a clean step function as it is in the ideal case, but may jump around and obscure the exact point of the large-scale change from bagel to blob.

The issue of topological noise, i.e., the lack of stability of the Betti numbers, motivated the development of persistent homology in the 1990s [9–11]. The inclusion of $K_a \subset K_b$ for $a < b$ induces a homomorphism between the homology groups $H_k(K_a)$ and $H_k(K_b)$ that tells us which topological features persist from K_a to K_b and which disappear (i.e., get filled in). The persistent homology group is the image of $H_k(K_a)$ in $H_k(K_b)$, it encodes the k-cycles in K_a that are independent with respect to boundaries in K_b:

$$H_k(a, b) := Z_k(K_a)/(B_k(K_b) \cap Z_k(K_a)). \tag{0.2}$$

Algorithms for computing persistent homology from a given filtration are quite simple in their most basic form [11,12], and are now implemented efficiently in a number of freely-downloadable packages [13–17].

The two most common ways of representing persistent homology information are the barcode [18] and the persistence diagram [19]. The barcode is a collection of intervals $[b, d)$ each representing the birth, b, and death, d, values of a persistent homology class. Equivalently, the persistence diagram is a set of points (b, d) in the plane. The main problem with these objects is that they are very difficult to work with statistically. Distances between persistence diagrams and definitions of their means and variance require advanced analytical techniques [20–22].

In this paper we return to the above definition of persistent homology groups and quantify them via their rank,

$$\beta_k(a, b) := \operatorname{rank} H_k(a, b), \quad \text{for } a < b. \tag{0.3}$$

This *persistent homology rank function* is an integer-valued function of two real variables and can be thought of as a cumulative distribution function of the persistence diagram. Since the persistent homology rank function is just a function we can apply standard statistical techniques to analyse distributions of them. This rank function is related to the size function [9] and has also been defined for multidimensional persistence, in the case of filtrations that are built using two or more parameters [23,24]. Other functional summaries of persistence diagrams have also been proposed recently and termed *persistence landscapes and silhouettes*, see [25,26]. The persistence landscapes $\lambda_k(i, t)$ are a family of functions so that for each $i = 1, 2, 3, \ldots, \lambda_k(i, t)$ is a function of a single variable (the subscript k is the homology dimension as above). The variable t is related to the persistence diagram coordinates by $t = (b + d)/2$. We argue that the rank functions used here, $\beta_k(a, b)$, retain a more direct connection to the geometry and topology of the original data, although they are functions of two variables. In particular they are more suitable for analysing distributions of local point configurations.

This functional approach to persistent homology (either using landscapes or rank functions) greatly simplifies the business of "doing statistics" with persistent homology. It provides a framework where it is simple to compute averages and variances of these topological signatures from many data sets generated by a particular system, and to make statistically rigorous statements about whether an experimentally-observed pattern is compatible with a theoretical model. In particular the pointwise average of $\beta_k(a, b)$ is a function on $\mathbb{R}^{2+}$ which tells us the expected ranks of the corresponding persistent homology groups. We can also perform functional principal component analysis where the principal component functions are also functions on $\mathbb{R}^{2+}$ containing topological information about regions in the persistence parameter space with the greatest variation.