The Accumulated Persistence Function, a New Useful Functional Summary Statistic for Topological Data Analysis, With a View to Brain Artery Trees and Spatial Point Process Applications

Christophe A. N. Biscio and Jesper Møller

Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark

ABSTRACT

We start with a simple introduction to topological data analysis where the most popular tool is called a persistence diagram. Briefly, a persistence diagram is a multiset of points in the plane describing the persistence of topological features of a compact set when a scale parameter varies. Since statistical methods are difficult to apply directly on persistence diagrams, various alternative functional summary statistics have been suggested, but either they do not contain the full information of the persistence diagram or they are two-dimensional functions. We suggest a new functional summary statistic that is one-dimensional and hence easier to handle, and which under mild conditions contains the full information of the persistence diagram. Its usefulness is illustrated in statistical settings concerned with point clouds and brain artery trees. The supplementary materials include additional methods and examples, technical details, and the R code used for all examples.

1. Introduction

Statistical methods that make use of algebraic topological ideas to summarize and visualize complex data are called topological data analysis (TDA). By combining algebraic topology with other mathematical tools, TDA allows a mathematically rigorous study of “shape.” In particular persistent homology and a tool called the persistence diagram are used to measure the so-called persistence of topological features, which provides a novel way of describing shape in data.

As we expect many statisticians may not be familiar with these tools and the underlying ideas and concepts, Section 1.1 discusses two examples of persistence diagrams and their interpretations without going into technical details, and Section 1.2 discusses further the use of the persistence diagram and related summary statistics. Section 1.3 introduces our new functional summary statistic called the accumulative persistence function. Both Sections 1.2 and 1.3 motivate why it is a useful alternative or supplement to the persistence diagram. Briefly, in contrast to the persistence diagram, many methods from functional data analysis apply for the accumulative persistence function. Moreover, apart from applications in TDA, the accumulative persistence function may be useful in spatial point process analysis. This is demonstrated in the remainder of the article which considers the use of the accumulative persistence function in various statistical settings concerned with point clouds and brain artery trees.

1.1. Examples of TDA

The mathematics underlying TDA uses technical definitions and results from persistent homology, see Fasy et al. (2014) and the references therein, but this theory will not be needed for the present paper. Instead we provide an understanding through the examples in Section 1.1.1 and 1.1.2 which consider the topological features of sets in the plane \( \mathbb{R}^2 \) and the space \( \mathbb{R}^3 \), respectively.

We first need to recall the following. A set \( A \subseteq \mathbb{R}^d \) is path-connected if any two points in \( A \) are connected by a curve in \( A \). Let \( C \subseteq \mathbb{R}^d \) be compact, that is, closed and bounded. A connected component of \( C \), also called a zero-dimensional topological feature of \( C \), is a maximal path-connected subset of \( C \). The meaning of a loop of \( C \), also called a one-dimensional topological feature of \( C \), is simply understood when \( d = 2 \), and we appeal to this in Remark 2 at the end of Section 1.1.1 when \( d = 3 \). As briefly noticed in Remark 1, \( k \)-dimensional topological features can be described as well for \( k = 0, 1, \ldots, d-1 \), but for this paper we need only to consider the cases \( k = 0 \) and \( k = 1 \), though our new accumulative persistence function will apply in general.

1.1.1. A Toy Example

The aim of this section is to introduce the persistence diagram for a planar set \( C \) in a very simple situation. Later, in Remark 1, we stress that in practice much more complicated cases are considered as the set \( C \) will not be fully observed.
Let \( C \subset \mathbb{R}^2 \) be the union of the three circles depicted in the top-left panel of Figure 1. We are interested in the topological features of \( C \): the three circles are the zero-dimensional topological features (the connected components) of \( C \), as any curve that goes from a circle to another will be outside \( C \). The complement \( \mathbb{R}^2 \setminus C \) has four connected components, one of which is unbounded, while the three others are the one-dimensional topological features (the loops) of \( C \) (a "loop" since the boundary of each bounded connected component is a closed curve with no crossings—here, the closed curve is just a circle). We can also think of each loop as a hole.

For \( t \) a real number, let \( C_t \) be the subset of points in \( \mathbb{R}^2 \) with distance \( \leq t \) from \( C \). Note that \( C_t \) is empty if \( t < 0 \). For \( t \geq 0 \), thinking of \( t \) as time, \( C_t \) results when each point on \( C \) grows as a disc with constant speed one. The number of connected components or loops changes exactly at the times \( t = 0, 0.5, 0.62, 0.75 \), see the first four panels of Figure 1: for each topological dimension \( k = 0, 1 \), let \( t^{(k)}_j \) denote the time of the \( j \)th change. In the first panel, \( C_0 = C \) has three connected components and three loops as specified in the previous paragraph; we say that they are born at time \( t^{(0)}_1 = t^{(1)}_1 = 0 \). In the second panel, \( C_{0.5} \) has two connected components and there is no loop because \( \mathbb{R}^2 \setminus C_{0.5} \) has only one connected component (and it is unbounded); we say that one of the connected components and the three loops from \( C_0 \) die at time \( t^{(0)}_2 = t^{(1)}_2 = 0.5 \). In the third panel, at time \( t^{(0)}_3 = t^{(1)}_3 = 0.62 \), the two connected components from \( C_{0.5} \) have grown into one connected component, namely \( C_{0.62} \), and a new loop is born (it represents the small white region bounded by \( C_{0.62} \)). The situation for connected components is unchanged after time 0.62 and it will be discarded in our analysis. In the fourth panel, the loop from \( C_{0.62} \) dies at time \( t^{(1)}_4 = 0.75 \).

Now, for each \( k = 0, 1 \) and as time \( t \) grows, we obtain from the first four panels of Figure 1 a multiset of birth–death points specifying the appearance and disappearance of each \( k \)-dimensional topological feature. The persistence diagram shown in the bottom-middle panel of Figure 1 is a scatterplot of these birth–death points for each topological dimension \( k \): for \( k = 0 \), the points corresponding to birth–death times of connected components are \((0, 0.5)\) and \((0, 0.62)\) (as \((0, \infty)\) is discarded from the diagram), both with multiplicity 1; these two points are the dots in the persistence diagram. For \( k = 1 \), the points corresponding to birth–death times of loops are \((0, 0.5)\) and \((0.62, 0.75)\), with multiplicities 3 and 1, respectively; these are the two triangles in the persistence diagram.

From the persistence diagram in Figure 1, we can extract interesting topological information about the set \( C \) as follows. One plus the number of solid dots in Figure 1 is the number of connected components of \( C \) (namely three connected components). The solid dot at \((0, 0.5)\) shows that two of the connected components are distance 0.25 apart, and the dot at \((0, 0.62)\) tells that the third connected component is at a distance larger than 0.25 from the two others and distance 0.31 apart from one of them. The triangle at \((0, 0.5)\), with multiplicity three, shows that \( C \) has three loops/holes and that in each of them it is not possible to include a circle of radius \( r > 0.5 \). The triangle at \((0.62, 0.75)\) tells that \( C_t \) for \( t \in (0.62, 0.75) \) has one loop/hole which disappears when \( t \geq 0.75 \).

The term “persistence” means a long lifetime (the difference between a corresponding death and birth time): the lifetimes for the connected components are 0.5, 0.62, and infinity, corresponding to that \( C \) has exactly three connected components. The lifetimes for the loops are first 0.5, corresponding to that \( C \) has exactly three loops, and second 0.13, corresponding to the last appearing loop; the latter is a short lifetime and the last appearing loop is considered as “noise.”

**Remark 1.** Usually in practice the topological features of a compact set \( C \subset \mathbb{R}^2 \) is of interest but only a finite subset of points \( \{x_1, \ldots, x_N\} \) has been collected as a sample on \( C \), possibly with noise. Then, for \( t \geq 0 \), we redefine \( C_t \) as the union of closed discs

---

**Figure 1.** At time \( t = 0 \), the first panel shows three circles all of radius 0.5 and centered at \((-1, -1), (1, -1), \) and \((0, 1)\). The next three panels show how the circles have been growing at times \( t = 0.5, 0.62, \) and 0.75. The fifth panel shows the persistence diagram for the connected components \((k = 0)\) and the loops \((k = 1)\). The final panel shows the corresponding accumulated persistence functions defined in Section 13.1.
of radius $t$ and with centers given by the point cloud. Hence, the connected components of $C_t$ are just the points $x_1, \ldots, x_N$, and $C_0$ has no loops. For $t > 0$, it is in general difficult to directly compute the connected components and loops of $C_t$. Using methods from persistent homology, a graph in $\mathbb{R}^2$, called the Delaunay-complex (or alpha-complex), can be constructed, so that its connected components correspond to those of $C_t$. Moreover, the triangles of this graph are filled or not in a way so that the loops of the obtained triangulation correspond to those of $C_t$. The precise definition of the Delaunay-complex is unnecessary for this article and we refer the interested reader to Edelsbrunner and Harer (2010).

In fact, for any dimension $d = 1, 2, \ldots$ and points $x_1, \ldots, x_N$ sampled from a compact set $C \subset \mathbb{R}^d$, Delaunay-complexes can be defined in the case where $C_t$ is the finite union of $d$-dimensional closed balls $\cup_{i=1}^N B(x_i, r)$ of radius $t$ and with centers at the $N$ points. Thereby the notion of persistence of $k$-dimensional topological features can be defined for the sets $\{C_t\}_{t \geq 0}$ and $k = 0, \ldots, d - 1$. This is indeed a specific case of persistent homology; for a full picture, see Fasy et al. (2014) and the references therein.

Remark 2. When $d = 3$, a zero-dimensional topological feature of a compact set $C \subset \mathbb{R}^3$ is still a connected component and we may still think of a one-dimensional topological feature as a loop, that is, a closed curve with no crossings. For example, a two-dimensional sphere has no loops, while a torus in $\mathbb{R}^3$ has two.

1.1.2. Persistent Homology for Brain Artery Trees

The left panel of Figure 2 shows an example of one of the 98 brain artery trees analyzed in Bendich et al. (2016). The data for each tree specifies a graph in $\mathbb{R}^3$ consisting of a dense cloud of vertices (about $10^6$ points) together with the edges (small straight line segments) connecting the neighboring vertices; further details about the data are given in Section 2.2. As in Bendich et al. (2016), we consider only the topological dimension $k = 0$ and $k = 1$, and depending on $k$, we consider different types of data and different sets $C_t$ as specified below.

As an example, let $B \subset \mathbb{R}^3$ denote the union of the edges for the tree in Figure 2. Following Bendich et al. (2016), if $k = 0$, let $C_t = \{(x, y, z) \in B : z \leq t\}$ be the sublevel set of the height function for the tree at level $t \geq 0$ (assuming $C_t$ is empty for $t < 0$). The zero-dimensional topological features at time $t$ are the connected components of $C_t$. As illustrated in the left panel of Figure 2, instead of time we may think of $t$ as “water level”: as the water level increases, connected components of the part of $B$ surrounded by water (the darker part) may be born or die; we refer to this as sublevel persistence. As in Section 1.1.1, we represent the births and deaths of the connected component in a persistence diagram which is shown in Figure 2 (middle panel).

As in Bendich et al. (2016), if $k = 1$, we let $C$ be a cloud of 3000 points subsampled from $B$, and redefine $C_t$ to be the union of balls of radii $t \geq 0$ and centers given by the point cloud $C$. The loops of $C_t$ are then determined by the corresponding Delaunay-complex, see Remark 1. The right panel of Figure 2 shows the corresponding persistence diagram.

In Figure 1 (the toy example) we had only two birth–death points for each dimension; in Figure 2, and for the other 97 brain artery trees, it is much more complicated to see what happens. What we learn from the persistence diagrams— or more precisely from our new tool, the accumulated persistence function—will be studied later for the brain artery trees in several examples.

1.2. Further Background and Objective

As exemplified in Section 1.1, the persistence diagram is a graphical representation of the persistence of the topological features of a set $C_0$ provided by a sequence of compact sets $C_t \subset \mathbb{R}^d$, $t \geq 0$, cf. the general setting considered at the end of Remark 1. Formally, for each topological dimension $k = 0, \ldots, d - 1$, the persistence diagram is given by a finite or countable set $PD_k$ of points $(b_1, d_1, c_1), (b_2, d_2, c_2), \ldots$ (only the finite case of $PD_k$ will be of interest in this paper), where each $(b_i, d_i)$ is a pair of birth–death times for a $k$-dimensional topological feature obtained as time $t$ grows, $c_i$ is a corresponding finite multiplicity (often each $c_i = 1$), and the points $(b_1, d_1), (b_2, d_2), \ldots$ are pairwise distinct.

Let $l_i = d_i - b_i$ be the $i$th lifetime. In the majority of literature on TDA, including the analysis in Bendich et al. (2016) of brain artery trees, long lifetimes are of main interest whereas short lifetimes are considered as topological noise as exemplified in Section 1.1.1. A short lifetime $l_i$ corresponds to that $(b_i, d_i)$ is close to the diagonal $(b, d) : b = d$, see, for example, Figure 2.

However, short lifetimes may also be of interest, for example, in the study of complex structures such as branch polymers and fractals, see MacPherson and Schweinhart (2012). Moreover, for brain artery trees, Bendich et al. (2016) noticed in one case that “not-particularly high persistence have the most distinguishing power in our specific application.” In our later examples, we demonstrate that short lifetimes will also be of key interest, including when analyzing the brain artery trees dataset from Bendich et al. (2016).

![Figure 2](image-url). A brain artery tree with the “water level” indicated (left panel) and the persistence diagrams of connected components (middle panel) and loops (right panel).
Chazal et al. (2013) and Chen et al. (2015) noted that it is difficult to apply statistical methodology to persistence diagrams. Alternative functional summary statistics have been suggested: Bubenik (2015) introduced a sequence of one-dimensional functions called the persistent landscape, where his first function is denoted $\lambda_1$ and is considered to be of main interest, since it provides a measure of the dominant topological features, that is, the longest lifetimes; therefore, we call $\lambda_1$ the dominant function. Chazal et al. (2013) introduced the silhouette which is a weighted average of the functions of the persistent landscape, where the weights control whether the focus is on topological features with long or short lifetimes. Moreover, Chen et al. (2015) considered a kernel estimate of the intensity function for the persistence diagram viewed as a point cloud. The dominant function, the silhouette, and the intensity estimate are one-dimensional functions and hence easier to handle than the persistence diagram, however, they provide selected and not full information about the persistence diagram.

Hence, directly studying persistence diagrams is a difficult task and there exists no functional summary statistic which is easy to work with and which contains all the information provided by the persistence’s diagrams. Therefore, in Section 1.3, we introduce another one-dimensional functional summary statistic, our new accumulative persistence function, and discuss its advantages and how it differs from the existing functional summary statistics.

1.3. The Accumulated Persistence Function

1.3.1. Definition

As our definition of the accumulated persistence function applies for a general setting as in Remark 1 (and even for more general situations as considered in Fasy et al. (2014)), we consider a persistence diagram $PD_k$ for any compact subset of $\mathbb{R}^d$ of interest and for any topological dimension $k = 0, 1, \ldots, d - 1$. However, as in Section 1.1, the reader may just have in mind the cases $d = 2, 3$ and $k = 0, 1$.

For simplicity and specificity, we always assume that $PD_k = \{(b_1, d_1, c_1), \ldots, (b_n, d_n, c_n)\}$ is such that $n < \infty$ and $0 < b_i < d_i < \infty$ for $i = 1, \ldots, n$. This assumption will be satisfied in our examples (at least with probability one).

Often in the TDA literature, $PD_k$ is transformed to the rotated and rescaled persistence diagram ($\text{RRPD}_k$) given by $\text{RRPD}_k = \{(m_1, l_1, c_1), \ldots, (m_n, l_n, c_n)\}$, where $m_i = (b_i + d_i)/2$ is the $i$th mean age. This transformation is useful when defining our accumulative persistence function (APF) by

$$\text{APF}_k(m) = \sum_{i=1}^{n} c_i \mathbb{1}(m_i \leq m), \quad m \geq 0,$$

where $\mathbb{1}(\cdot)$ is the indicator function and we suppress in the notation that $\text{APF}_k$ is a function of $\text{RRPD}_k$.

Formally speaking, when $\text{RRPD}_k$ is considered to be random, it is viewed as a finite point process with multiplicities, see, for example, Daley and Vere-Jones (2003). In what follows it will always be clear from the context whether $PD_k$ and $\text{RRPD}_k$ are considered as being random or observed, and hence whether $\text{APF}_k$ is a deterministic or random function. In the latter case, because $\text{APF}_k(m)$ is an accumulative function, its random fluctuations typically increase as $m$ increases.

Clearly, $\text{RRPD}_k$ is in one-to-one correspondence to $PD_k$. In turn, if all $c_i = 1$ and the $m_i$ are pairwise distinct, then there is a one-to-one correspondence between $\text{RRPD}_k$ and its corresponding $\text{APF}_k$. For $k = 0$, this one-to-one correspondence would easily be lost if we had used $b_i$ in place of $m_i$ in (1).

We need to be careful with not overstating this possible one-to-one correspondence. As a referee kindly pointed out, closeness between two APFs with respect to $L^q$-norm ($1 \leq q \leq \infty$) does not always correspond to closeness between their underlying persistence diagrams which is usually understood with respect to the so-called bottleneck distance; see, for example, Fasy et al. (2014) or Appendix A.

Note that the dominant function, the silhouette, and the intensity estimate (see Section 1.2) are in general not in a one-to-one correspondence with $\text{RRPD}_k$. Like these functions, $\text{APF}_k$ is a one-dimensional function, and so it is easier to handle than the sequence of functions for the persistent landscape in Bubenik (2015) and the intensity estimate in Chen et al. (2015)—for example, confidence regions become easier to plot. Contrary to the dominant function and the silhouette, the APF provides information about the topological features without distinguishing between long and short lifetimes.

1.3.2. Interpretation

Depending on the application, the jumps and/or the shape of $\text{APF}_k$ may be of interest as demonstrated later in our examples. A large jump of $\text{APF}_k$ corresponds to a large lifetime (long persistence). In the simple example shown in Figure 1, both jumps of $\text{APF}_0$ are large and indicate the three connected components (circles), while only the first jump of $\text{APF}_1$ is large and indicates the three original loops.

For the more complicated examples considered in the following it may be hard to recognize the individual jumps and we will instead consider the shape of the APFs. For instance, as in Remark 1 at the end of Section 1.1.1, suppose $C_i$ is the union of $d$-dimensional balls of radius $t$ and with centers given by a finite point cloud $\{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ which is of interest. We may then observe the following features as illustrated later in Example 1. For small mean ages $m$, jumps of $\text{APF}_0(m)$ correspond to balls that merge together for small values of $t$. On one hand, if the point cloud is aggregated/clustered, we expect that $\text{APF}_0(m)$ has jumps and is hence large for small mean ages $m$. On the other hand, if the point cloud is regular (e.g., because of inhibition between the points), we expect the jumps of $\text{APF}_0(m)$ to happen and to be large for modest values of $m$. These two situations will later be illustrated in the middle panel of Figure 4 when considering curves for a Matérn cluster process and a determinantal point process (DPP), which are clustered and regular point processes, respectively. For large mean ages, jumps of $\text{APF}_0(m)$ are most likely to happen in the case of aggregation/clustering. Accordingly, the shape of $\text{APF}_0$ can be very different for these two cases (as illustrated later in the first panel of Figure 4). Similar considerations lead us to expect different shapes of $\text{APF}_1$ for different types of point clouds: we expect that $\text{APF}_1(m)$ is large respective small for the case of aggregation respective regularity when $m$ is small, and
the opposite happens when \( m \) is large (as illustrated later in the last panel of Figure 4).

1.4. Outline

The remainder of this paper discusses various methods based on APFs in different contexts and illustrated by simulation studies related to spatial point process applications and by reanalyzing the brain artery trees dataset previously analyzed in Bendich et al. (2016). The purpose is to demonstrate the advantages and limitations of using APFs together with graphical methods and tests from functional data analysis and spatial statistics, and to shed further light on the brain artery trees dataset.

Section 2 specifies the setting for our examples. Sections 3, 4, and 5 consider the case of a single APF, a sample of APFs, and two samples of APFs, respectively. Further examples and details appear in the supplementary materials.

2. Datasets

2.1. Simulated Data

In our simulation studies, we consider a planar point cloud \( \{x_1, \ldots, x_N\} \subset \mathbb{R}^2 \), and study as at the end of Section 1.1.1 how the topological features of \( C_t \), the union of closed discs of radii \( t \) and centred at \( x_1, \ldots, x_N \), change as \( t \) grows. Here \( \{x_1, \ldots, x_N\} \) will be a realization of a point process \( X \subset \mathbb{R}^2 \), where the count \( N \) is finite. Note that \( N \) may be random, and conditional on \( N \), the points in \( X \) are not necessarily iid.

This is a common situation in spatial statistics, for example, if the focus is on the point process \( X \) and the purpose is to assess the goodness of fit for a specified point process model of \( X \) when \( \{x_1, \ldots, x_N\} \) is observed. Note that PD\(_t\) and RRPD\(_k\) can be viewed as finite planar point processes (with multiplicities) and APF\(_k\) as a random function.

2.2. Brain Artery Trees Dataset

The dataset in Bendich et al. (2016) comes from 98 brain artery trees which can be included within a cube of side length at most 206 mm; one tree is excluded “as the java/matlab function crashed” (e-mail correspondence with Sean Skwerer). They wanted to capture how the arteries bend through space and to represent the tree is considered, cf. Section 1.1.2; then for all mean ages, \( m_i \leq 137 \); and the number of connected components is always below 3200. For \( k = 1 \), persistence of the loops for the union of growing balls with centers at a point cloud representing the tree is considered, cf. Section 1.1.2; the loops have a finite death time but some of them do not die during the allocated time \( T = 25 \) (i.e., Bendich et al. (2016) stopped the growth of balls when \( t > 25 \)). Thus, we shall only consider mean ages \( m_i \leq 25 \); then the number of loops is always below 2700.

For each tree and \( k = 0, 1 \), Bendich et al. (2016) used only the 100 largest lifetimes in their analysis. Whereas their principal component analysis clearly revealed age effects, their permutation test based on the mean lifetimes for the male and females subjects only showed a clear difference when considering PD\(_1\). Accordingly, when demonstrating the usefulness of APF\(_0\) and APF\(_1\), we will focus on the gender effect and consider the same 95 trees as in Bendich et al. (2016) (two transsexual subjects are excluded) obtained from 46 female subjects and 49 male subjects; in contrast to Bendich et al. (2016), we consider all observed mean ages and lifetimes. In accordance to the allocated time \( T = 25 \), we need to redefine APF\(_1\) by

\[
APF_1(m) = \sum_{i=1}^{n} c_i l_i(\{m_i \leq m, m_i + l_i/2 \leq T\}, m \geq 0. \tag{2}
\]

For simplicity we use the same notation APF\(_1\) in (1) and (2); although all methods and results in this article will be presented with the definition (1) in mind, they apply as well when considering (2).

3. A Single Accumulated Persistence Function

There exists several constructions and results on confidence sets for persistence diagrams when the aim is to separate topological signal from noise, see Fasy et al. (2014), Chazal et al. (2014), and the references therein. Appendix A (supplementary materials) and its accompanying Example 5 discuss the obvious idea of transforming such a confidence region into one for an accumulated persistence function, where the potential problem is that the bottleneck metric is used for persistence diagrams and this is not corresponding to closeness of APFs, see Section 1.3.1. In this section, we focus instead on spatial point process model assessment using APFs or more traditional tools.

Suppose a realization of a finite spatial point process \( X_0 \) has been observed and a specific model for \( X_0 \) has been claimed (or estimated). A common situation for model assessment is to simulate copies \( X_1, \ldots, X_r \) under the claimed model for \( X_0 \) so that the joint distribution of \( X_0, X_1, \ldots, X_r \) is exchangeable; see, for example, Baddeley, Rubak and Turner (2015) and Møller and Waagepetersen (2016). That is, for any permutation \( (\sigma_0, \ldots, \sigma_r) \) of \((0, \ldots, r)\), \((X_{\sigma_0}, \ldots, X_{\sigma_r})\) is claimed to be distributed as \((X_0, \ldots, X_r)\); for example, this is the case if \( X_0, X_1, \ldots, X_r \) are iid. Denote the APF\(_k\)s for \( X_0, \ldots, X_r \) by \( A_0, \ldots, A_r \), respectively, and the null hypothesis that the joint distribution of \( A_0, \ldots, A_r \) is exchangeable by \( \mathcal{H}_0 \). Rejection of \( \mathcal{H}_0 \) corresponds to that \( A_0 \) is extreme under the claimed model for \( X_0 \) which in turn implies that the realization of \( X_0 \) is extreme. Adapting ideas from Myllymäki et al. (2016), we will discuss how to construct a goodness-of-fit test for \( \mathcal{H}_0 \) based on a so-called global rank envelope for \( A_0 \); its usefulness will be demonstrated in Example 1.

In functional data analysis, to measure how extreme \( A_0 \) is in comparison to \( A_1, \ldots, A_r \), a so-called depth function is used for ranking \( A_0, \ldots, A_r \), see, for example, López-Pintado and Romo (2009). We suggest using a depth ordering called extreme rank in Myllymäki et al. (2016): let \( T > 0 \) be a user-specified parameter chosen such that it is the behavior of \( A_0(m) \) for \( 0 \leq m \leq T \) which is of interest. For \( l = 1, 2, \ldots \), define the \( l \)th bounding curves of \( A_0, \ldots, A_r \) by

\[
A_{low}^l(m) = \min_{i=0,\ldots,r} A_i(m) \quad \text{and}
\]
where \( \min^l \) and \( \max^l \) denote the \( l \)th smallest and largest values, respectively, and where \( l \leq r/2 \). Then, for \( i = 0, \ldots, r \), the extreme rank of \( A_i \) with respect to \( A_0, \ldots, A_r \) is

\[
R_i = \max \left\{ l : A^l_{\text{low}}(m) \leq A_i(m) \leq A^l_{\text{upp}}(m) \right\}
\]

for all \( m \in [0, T] \).

The larger \( R_i \) is, the deeper or more central \( A_i \) is among \( A_0, \ldots, A_r \).

Now, for a given \( \alpha \in (0, 1) \), the extreme rank ordering is used to define the \( 100(1-\alpha) \%- \)global rank envelope as the band delimited by the curves \( A^l_{\text{low}} \) and \( A^l_{\text{upp}} \) where

\[
I_a = \max \left\{ l : \frac{1}{r+1} \sum_{i=0}^{r} 1(R_i < l) \leq \alpha \right\}.
\]

Under \( H_0 \), with probability at least \( 1 - \alpha \),

\[
A^l_{\text{low}}(m) \leq A_0(m) \leq A^l_{\text{upp}}(m) \quad \text{for all } m \in [0, T],
\]

see Myllymäki et al. (2016). Therefore, the \( 100(1-\alpha) \%- \)global rank envelope is specifying a conservative statistical test called the extreme rank envelope test and which accepts \( H_0 \) at level \( 100\alpha\% \) if (3) is satisfied or equivalently if

\[
\frac{1}{r+1} \sum_{i=0}^{r} 1(R_i < R_0) > \alpha,
\]

see Myllymäki et al. (2016). A plot of the extreme rank envelope may allow a graphical interpretation of the extreme rank envelope test: in case of rejection, the plot may suggest an alternative model for \( X_0 \); otherwise, the more \( A_0 \) is in the “center” of the envelope, the more we may “trust” the claimed model for \( X_0 \).

To exist alternatives to the extreme rank envelope test, in particular a liberal extreme rank envelope test and a so-called global scaled maximum absolute difference envelope, see Myllymäki et al. (2016). It is also possible to combine several extreme rank envelopes, for instance by combining APF0 and APF1, see Mrkvíčka, Myllymäki, and Hahn (2016). In the following example we focus on (3)–(4) and briefly remark on results obtained by combining APF0 and APF1.

**Example 1 (Simulation study).** Recall that a homogeneous Poisson process is a model for complete spatial randomness (CSR), see, for example, Møller and Waagepetersen (2004) and the simulation in the first panel of Figure 3. Consider APFs \( A_0, A_1, \ldots, A_r \) corresponding to independent point processes \( X_0, X_1, \ldots, X_r \) defined on a unit square and where, for each \( i > 0 \), \( X_i \) is CSR with a given intensity \( \rho \) (the mean number of points). Suppose \( X_0 \) is claimed to be CSR with intensity \( \rho \), however, the model for \( X_0 \) is given by one of the following four point process models, which we refer to as the true model:

(a) CSR; hence the true model agrees with the claimed model.
(b) A Baddeley–Silverman cell process; this has the same second-order moment properties as under CSR, see Baddeley and Silverman (1984) for details. Though, from a mathematical point of view, it is a cluster process, simulated realizations will exhibit both aggregation and regularity at different scales, see the second panel of Figure 3.
(c) A Matérn cluster process; this is a model for clustering produced in two steps: first, the points of an unobserved homogeneous Poisson process are used as the centers of disks of equal radii. Second, within each disc, independent homogeneous Poisson processes of equal intensity are generated and the union of these “clusters” is then a realization of the model. See Matérn (1986) and the third panel of Figure 3.
(d) A most repulsive Bessel-type DPP; this is a model for regularity, see Biscio and Lavancier (2016) for details. See also Lavancier, Møller, and Rubak (2015) and the fourth panel of Figure 3.

We let \( \rho = 100 \) or 400. This specifies completely the models in (a) and (d), whereas the remaining parameters in the cases (b)–(c) are defined to be the same as those used in Robins and Turner (2016). In all cases of Figure 3, \( \rho = 400 \). Finally, following the recommendation in Myllymäki et al. (2016), we let \( r = 2499 \).

For each value of \( \rho = 100 \) or 400, we simulated from each point process model in (a)–(d) a realization and then, for each dimension \( k = 0 \) or 1, we computed the extreme rank envelopes and performed the extreme rank envelope tests. We repeated all this 500 times, using the R package **spatstat** for the simulations and the R package **spatstat** for the extreme rank envelopes. Table 1 shows for each case (a)–(d) the percentage of rejection of the hypothesis that \( X_0 \) is a homogeneous Poisson process with known intensity \( \rho \). In case of CSR, the Type I error of the test is small except when \( k = 0 \) and \( \rho = 100 \). As expected in case of (b)–(d), the power of the test increases when \( \rho \) is increased. For both the Baddeley–Silverman process and the

---

**Figure 3.** Simulated point clouds for a homogeneous Poisson process (first panel), a Baddeley–Silverman cell process (second panel), a Matérn cluster process (third panel), and a most repulsive Bessel-type DPP (fourth panel).
DPP, when \( k = 0 \) and/or \( \rho = 400 \), the power is high and even 100% in two cases. For the Matérn cluster process, the power is 100% when both \( \rho = 100 \) and 400; this was also the case in another simulation study we performed when instead the radius of a cluster was 10 times larger and hence it was not so easy to distinguish the clusters as in the third panel of Figure 3. When we combined the extreme rank envelopes for APF0 and APF1, the results were better or close to the best results obtained when considering only one extreme rank envelope.

Figure 4 illustrates for one of the 500 repetitions and for each dimension \( k = 0 \) and \( k = 1 \) the deviation of APF \(_k\) from the extreme rank envelope obtained when the true model is not CSR. For each of the three non-CSR models, APF \(_0\) is outside the extreme rank envelope, in particular when \( k = 0 \) and both the mean age and lifetime are small, cf. the middle panel. This means that small lifetimes are not noise but of particular importance, cf. the discussion in Section 1.2. Using an obvious notation, for small \( m \), we may expect that \( \text{APF}_{0}^{\text{DPP}}(m) < \text{APF}_{0}^{\text{CSR}}(m) < \text{APF}_{0}^{\text{MC}}(m) \) which is in agreement with the middle panel. For large \( m \), we may expect that \( \text{APF}_{0}^{\text{DPP}}(m) > \text{APF}_{0}^{\text{CSR}}(m) \) and \( \text{APF}_{0}^{\text{MC}}(m) > \text{APF}_{0}^{\text{CSR}}(m) \), but only the last relation is detected by the extreme rank envelope in the left panel. Similarly, we may expect \( \text{APF}_{1}^{\text{MC}}(m) > \text{APF}_{1}^{\text{CSR}}(m) \) for small \( m \), whereas \( \text{APF}_{1}^{\text{MC}}(m) < \text{APF}_{1}^{\text{CSR}}(m) \) for large \( m \), and both cases are detected in the right panel. Note that for the Baddeley–Silverman cell process and \( k = 0,1 \), APFBS \(_k\) has a rather similar behavior as \( \text{APF}_{k}^{\text{DPP}} \), that is, as for a regular point process and probably because clustering is a rare phenomena.

A similar simulation study was discussed in Robins and Turner (2016) for the models in (a)–(c), but notice that they fixed the number of points to be 100 and they used a testing procedure based on the persistent homology rank function, which in contrast to our one-dimensional APF is a two-dimensional function and is not summarizing all the topological features represented in a persistence diagram. Robins and Turner (2016) demonstrated that a test for CSR based on the persistent homology rank function is useful as compared to various tests implemented in spatstat and which only concern first and second-order moment properties. Their method is in particular useful, when the true model is a Baddeley–Silverman cell process with the same first and second-order moment properties as under CSR. Comparing Figure 4 in Robins and Turner (2016) with the results in Table 1 when the true model is a Baddeley–Silverman cell process and \( \rho = 100 \), the extreme rank envelope test seems less powerful than the test they suggest. On the other hand, Robins and Turner (2016) observed that the latter test performs poorly when the true model is a Strauss process (a model for inhibition) or a Matérn cluster process; as noticed for the Matérn cluster process, we have a perfect power in Table 1 when using the extreme range envelope test.

### 4. A Single Sample of Accumulated Persistence Functions

#### 4.1. Functional Boxplot

This section discusses the use of a functional boxplot (Sun and Genton 2011) for a sample \( A_1, \ldots, A_r \) of APF \(_k\)s those joint distribution is exchangeable. The plot provides a representation of the variation of the curves given by \( A_1, \ldots, A_r \) around the most central curve, and it can be used for outlier detection, that is, detection of curves that are too extreme with respect to the others in the sample. This is illustrated in Example 2 for the brain artery trees dataset and in Appendix B (supplementary materials) and its accompanying Example 6 concerning a simulation study.

The functional boxplot is based on an ordering of the APF \(_k\)s obtained using a so-called depth function. For specificity, we make the standard choice called the modified band depth function (MBD), see López-Pintado and Romo (2009) and Sun and

![Figure 4](https://example.com/figure4.png) 95%-extreme rank envelope for APF, when \( i = 0 \) (left panel and the enlargement shown in the middle panel) or \( i = 1 \) (right panel) together with the curves for the three non-CSR models (Baddeley–Silverman cell process, Matérn cluster process, and Bessel-type DPP). The envelope is obtained from 2499 realizations of a CSR model on the unit square and with intensity 100.
Genton (2011): for a user-specified parameter \( T > 0 \) and \( h, i, j = 1, \ldots, r \) with \( i < j \), define

\[
B_{h,i,j} = \left\{ m \in [0, T] : \min \left\{ A_i(m), A_j(m) \right\} \leq A_h(m) \leq \max \left\{ A_i(m), A_j(m) \right\} \right\},
\]

and denote the Lebesgue measure on \([0, T]\) by \(|\cdot|\). Then the MBD of \( A_h \) with respect to \( A_1, \ldots, A_r \) is

\[
\text{MBD}_r(A_h) = \frac{2}{r(r-1)} \sum_{1 \leq i < j \leq r} |B_{h,i,j}|. \tag{5}
\]

This is the average proportion of \( A_h \) on \([0, T]\) between all possible pairs of \( A_1, \ldots, A_r \). Thus, the larger the value of the MBD of a curve is, the more central or deeper it is in the sample. We call the region delimited by the 50% most central curves the central envelope. It is often assumed that a curve outside the central envelope inflates by 1.5 times the range of the central envelope is an outlier or abnormal curve—this is just a generalization of a similar criterion for the boxplot of a sample of real numbers—and the range may be changed if it is more suitable for the application at hand, see the discussion in Sun and Genton (2011) and Example 2.

Example 2 (Brain artery trees). For the brain artery trees dataset (Section 2.2), Figure 5 shows the functional boxplots of APFs for females and males on the intervals associated with loops, the main difference is observed for females, in particular on the interval \([0, 50]\) (fourth panel) when \( k = 1 \). Moreover, we write APF\(_F\) lower bounds obtained from all curves except the outliers in black, the central envelope in purple, and the upper and lower bounds obtained from all the curves except the outliers in dark blue. Moreover, we write APF\(_F\) and APF\(_M\) to distinguish between APFs for females and males, respectively. Comparing the two left panels (concerned with connected components), the shape of the central envelope is clearly different for females and males, in particular on the interval \([40, 60]\), and the upper and lower bounds of the non-outlier are closer to the center region for females, in particular on the interval \([0, 50]\). For the right two panels (concerned with loops), the main difference is observed on the interval \([15, 25]\) where the central envelope is larger for females than for males.

The dashed lines in Figure 5 show the APFs detected as outliers by the 1.5 criterion, that is 6 APF\(_F\)'s (first panel), 3 APF\(_F\)'s (third panel), 6 APF\(_M\)'s (second panel), and 4 APF\(_M\)'s (fourth panel). For the females, only for one point cloud both APF\(_F\) and APF\(_M\) are outliers, where APF\(_F\) is the steep dashed line in the third panel. For the males, only for two point clouds both APF\(_F\) and APF\(_M\) are outliers. For one of these two point clouds APF\(_M\) is the steep dashed line in the fourth panel, and the left panel in Figure 6 shows its corresponding brain artery tree. Comparing this brain artery tree with another brain artery tree which is not detected as an outlier (right panel in Figure 6) reveals that a large part on the right is missing!

Examples 3 and 4 discuss to what extent our analysis of the brain artery trees will be sensitive to whether we include or exclude the detected outliers.

4.4. Confidence Region for the Mean Function

This section considers an asymptotic confidence region for the mean function of a sample \( A_1, \ldots, A_r \) of iid APF\(_k\)'s. We assume that \( D_1, \ldots, D_r \) are the underlying iid RRPD\(_k\)'s for the sample so that with probability one, there exists an upper bound \( T < \infty \) on the death times and there exists an upper bound \( n_{\max} < \infty \) on the number of \( k \)-dimensional topological features. Note that the state space for such RRPD\(_k\)'s is

\[
D_{h,T,n_{\max}} = \{(m_1, l_1, c_1), \ldots, (m_n, l_n, c_n) : \sum_{i=1}^n c_i \leq n_{\max}, m_i + l_i/2 \leq T, i = 1, \ldots, n\}
\]

and only the existence and not the actual values of \( n_{\max} \) and \( T \) play a role when applying our method below. Moreover, in the setting of Section 2.1, it suffices to assume that \( X \) is included in a bounded region of \( \mathbb{R}^2 \) and that the number of points \( N \) is bounded by a constant (this follows from the two versions of the
nerve theorem presented in Fasy et al. (2014) and Edelsbrunner and Harer (2010), respectively, but we omit the details.

We adapt an empirical bootstrap procedure (see, e.g., van der Vaart and Wellner 1996) which in Chazal et al. (2013) was used for a confidence region for the mean of the dominant function of the persistent landscape and which in our case works as for a confidence region for the mean of the APF, that is

\[ \hat{\mu}(m) = E \{ \hat{A}_1(m) \} \]

and estimated by the empirical mean function \( \overline{\mathcal{A}}_r(m) = \frac{1}{n} \sum_{i=1}^{n} A_i(m) \). Let \( A^*_1, \ldots, A^*_r \) be independent uniform draws with replacement from the set \( \{ A_1, \ldots, A_r \} \) and set \( \overline{\mathcal{A}}_r^* = \frac{1}{n} \sum_{i=1}^{n} A^*_i(m) \) and \( \theta^* = \sup_{m \in [0,T]} \sqrt{T} | \overline{\mathcal{A}}_r(m) - \overline{\mathcal{A}}_r^*(m) | \). For a given integer \( B \), independently repeat this procedure \( B \) times to obtain \( \theta^*_1, \ldots, \theta^*_B \). Then, for \( 0 < \alpha < 1 \), the \( 100(1-\alpha)\% \)-quantile in the distribution of \( \theta^* \) is estimated by

\[ \hat{\theta}^*_B = \inf \{ q \geq 0 : \frac{1}{B} \sum_{i=1}^{B} 1(\theta^*_i > q) \leq \alpha \}. \]

The following theorem is verified in Appendix F (supplementary materials).

**Theorem 4.1.** Let the situation be as described above. For large values of \( r \) and \( B \), the functions \( \overline{\mathcal{A}}_r \pm \hat{\theta}^*_B / \sqrt{r} \) provide the bounds for an asymptotic confidence region for the mean of the APF, that is

\[ \lim_{r \to \infty} \lim_{B \to \infty} P \left( \mu(m) \in [ \overline{\mathcal{A}}(m) - \hat{\theta}^*_B / \sqrt{r}, \overline{\mathcal{A}}(m) + \hat{\theta}^*_B / \sqrt{r} ] \right) = 1 - \alpha. \]

**Example 3 (Brain artery trees).** The brain artery trees are all contained in a bounded region and presented by a bounded number of points, so it is obvious that \( T \) and \( n_{\text{max}} \) exist for \( k = 0,1 \). To establish confidence regions for the mean of the APF\(_k\)'s, we apply the bootstrap procedure with \( B = 1000 \). The result is shown in Figure 7 when all 95 trees are considered: in the left panel, \( k = 0 \) and approximately half of each confidence region overlap with the other confidence region; it is not clear if there is a difference between genders. In the right panel, \( k = 1 \) and the difference is more pronounced, in particular on the interval [15, 25]. Similar results and conclusions are obtained if we exclude the APFs detected as outliers in Example 2. Of course we should supply with a statistical test to assess the gender effect and such a test is established in Section 5 and applied in Example 4.

Appendix C (supplementary materials) provides the additional Example 7 for a simulated dataset along with a discussion on the geometrical interpretation of the confidence region obtained.

## 5. Two Samples of Accumulated Persistence Functions

This section concerns a two-sample test for comparison of two samples of APFs. Appendix E (supplementary materials) presents both a clustering method (Appendix E.1, including Example 9) and a supervised classification method (Appendix E.2, including Example 10) for two or more samples.

Consider two samples of independent RRPD\(_k\)'s \( D_1, \ldots, D_{r_1} \) and \( E_1, \ldots, E_{r_2} \), where each \( D_i \) (\( i = 1, \ldots, r_1 \)) has distribution \( P_D \) and each \( E_j \) (\( j = 1, \ldots, r_2 \)), and suppose we want to test the null hypothesis \( H_0 \): \( P_D = P_E = P \). Here, the common distribution \( P \) is unknown and as in Section 4.2 we assume it is concentrated on \( D_k \). The symmetrical test statistic is studied in Ponestgaard (1995) and van der Vaart and Wellner (1996).

Let \( r = r_1 + r_2 \). Let \( A_1, \ldots, A_r \) be the APF\(_k\) corresponding to \( (D_1, \ldots, D_{r_1}, E_1, \ldots, E_{r_2}) \), and denote by \( \overline{\mathcal{A}}_r \) and \( \overline{\mathcal{A}}^*_r \) the empirical means of \( A_1, \ldots, A_r \) and \( A^*_1, \ldots, A^*_r \), respectively. Let \( I = \{ T_1, T_2 \} \) be a user-specified interval with \( 0 \leq T_1 < T_2 \leq T \) and used for defining a two-sample test statistic by

\[ KS_{r_1,r_2} = \frac{r_1 r_2}{r} \sup_{m \in I} \left| \overline{\mathcal{A}}_{r_1}(m) - \overline{\mathcal{A}}_{r_2}(m) \right|, \]

where large values are critical for \( H_0 \). This may be rewritten as

\[ KS_{r_1,r_2} = \sup_{m \in I} \left| \frac{r_1}{r} G^1_{r_1}(m) - \frac{r_2}{r} G^2_{r_2}(m) \right|, \]

where \( G^1_{r_1} = \sqrt{r_1} (\overline{\mathcal{A}}_{r_1} - E(A_{D_{r_1}})) \) and \( G^2_{r_2} = \sqrt{r_2} (\overline{\mathcal{A}}_{r_2} - E(A_{E_{r_2}})) \). By Lemma F.2 in Appendix F (supplementary materials) and by the independence of the samples, \( G^1_{r_1} \) and \( G^2_{r_2} \) converge in distribution to two independent zero-mean Gaussian processes on \( I \), denoted \( G_D \) and \( G_E \), respectively. Assume that \( r_1/r \to \lambda \in (0,1) \) as \( r \to \infty \). Under \( H_0 \), in the sense of convergence in distribution,

\[ \lim_{r \to \infty} KS_{r_1,r_2} = \sup_{m \in I} \left| \sqrt{T - T_1} G_D(m) - \sqrt{T - T_2} G_E(m) \right|, \]

where \( \sqrt{T - T_1} G_D - \sqrt{T - T_2} G_E \) follows the same distribution as \( G_D \). If \( H_0 \) is not true and \( \sup_{m \in I} E[|A_1 - A_{r_1}|(m)] > 0 \), then \( KS_{r_1,r_2} \to \infty \) as \( r \to \infty \), see van der Vaart and Wellner (1996). Therefore, for \( 0 < \alpha < 1 \) and letting \( q_\alpha = \inf \{ q : P(\sup_{m \in I} |G(m)| > q) \leq \alpha \} \), the asymptotic test that rejects \( H_0 \) if \( KS_{r_1,r_2} \leq q_\alpha \) is of level 100% and of power 100%.

As \( q_\alpha \) depends on the unknown distribution \( P \), we estimate \( q_\alpha \) by a bootstrap method: let \( A^*_1, \ldots, A^*_r \) be independent uniform draws with replacement from \( \{ A_1, \ldots, A_r \} \).
For a given integer \( 0 \leq m \leq T \), define the empirical mean functions \( \overline{A}_r(m) = \frac{1}{m} \sum_{i=1}^{m} A_i(m) \) and \( \overline{A}_r^{*}(m) = \frac{1}{m} \sum_{i=1}^{m+r} A_i^{*}(m) \), and compute
\[
\theta^* = \frac{\sqrt{r} \overline{r}_2}{r} \sup_{m \in I} \left| \overline{A}_r(m) - \overline{A}_r^{*}(m) \right|.
\]

For a given integer \( B > 0 \), independently repeat this procedure \( B \) times to obtain \( \theta_1^*, \ldots, \theta_B^* \). Then we estimate \( \hat{q}_0 \) by the 100\((1 - \alpha)\)%-quantile of the empirical distribution of \( \theta_1^*, \ldots, \theta_B^* \), that is
\[
\hat{q}_0^B = \inf \{ q \geq 0 : \frac{1}{B} \sum_{i=1}^{B} 1(\theta_i^* > q) \leq \alpha \}.
\]

The next theorem is a direct application of Theorem 3.7.7 in van der Vaart and Wellner (1996) noticing that the APF\(_k\)'s are uniformly bounded by \( T_{\max} \) and they form a so-called Donsker class, see Lemma F.2 and its proof in Appendix F (supplementary materials).

**Theorem 5.1.** Let the situation be as described above. If \( r \to \infty \) such that \( r_1/r \to \lambda \) with \( \lambda \in (0, 1) \), then under \( \mathcal{H}_0 \)
\[
\lim_{r \to \infty} \lim_{B \to \infty} P \left( KS_{r_1,r_2} > \hat{q}_0^B \right) = \alpha,
\]
while if \( \mathcal{H}_0 \) is not true and \( \sup_{m \in I} E \left| A_1 - A_{r_1+1} \right| (m) \) > 0, then
\[
\lim_{r \to \infty} \lim_{B \to \infty} P \left( KS_{r_1,r_2} > \hat{q}_0^B \right) = 1.
\]

Therefore, the test that rejects \( \mathcal{H}_0 \) if \( KS_{r_1,r_2} > \hat{q}_0^B \) is of asymptotic level 100\% and power 100\%. As remarked in van der Vaart and Wellner (1996), by their Theorem 3.7.2 it is possible to present a permutation two-sample test so that the critical value \( \hat{q}_0^B \) for the bootstrap two-sample test has the same asymptotic properties as the critical value for the permutation test.

Other two-sample test statistics than (6) can be constructed by considering other measurable functions of \( A_r - A_{r_2} \), for example, we may consider the two-sample test statistic
\[
M_{r_1,r_2} = \int_{I} \left| A_{r_1}(m) - A_{r_2}(m) \right| dm.
\]

Then by similar arguments as above but redefining \( \theta^* \) in (9) by
\[
\theta^* = \frac{\sqrt{r} \overline{r}_2}{r} \int_{m \in I} \left| A_{r_1}(m) - A_{r_2}(m) \right| dm,
\]
the test that rejects \( \mathcal{H}_0 \) if \( M_{r_1,r_2} > \hat{q}_0^B \) is of asymptotic level 100\% and power 100\%.

**Example 4 (Brain artery trees).** To distinguish between male and female subjects of the brain artery trees dataset, we use the two-sample test statistic \( KS_{r_1,r_2} \) under three different settings:

(A) For \( k = 0, 1 \), we let \( PD'_k \) be the subset of \( PD_k \) corresponding to the 100 largest lifetimes. Then \( D_1, \ldots, D_{16} \) and \( E_1, \ldots, E_{10} \) are the RRPD\(_k\)'s obtained from the PD\(_k\)'s associated to female and male subjects, respectively. This is the setting used in Bendich et al. (2016).

(B) For \( k = 0, 1 \), we consider all lifetimes and let \( D_1, \ldots, D_{16} \) and \( E_1, \ldots, E_{10} \) be the RRPD\(_k\)'s associated to female and male subjects, respectively.

(C) The samples are as in setting (B) except that we exclude the RRPD\(_k\)'s where the corresponding APF\(_k\) was detected as an outlier in Example 2. Hence, \( r_1 = 40 \) and \( r_2 = 43 \) if \( k = 0 \), and \( r_1 = 43 \) and \( r_2 = 45 \) if \( k = 1 \).

Bendich et al. (2016) performed a permutation test based on the mean lifetimes for the male and female subjects and concluded that gender effect is recognized when considering \( PD_1 \) (p-value = 3\%) but not \( PD_0 \) (p-value = 10\%). For comparison, under each setting (A)–(C), we perform the two-sample test for \( k = 0, 1 \), different intervals \( I \) and \( B = 10,000 \). In each case, we estimate the p-value, that is, the smallest \( \alpha \) such that the two-sample test with significance level 100\% does not reject \( \mathcal{H}_0 \), by \( \hat{p} = \frac{1}{B} \sum_{i=1}^{B} 1(\theta_i^* > KS_{r_1,r_2}) \). Table 2 shows the results. Under each setting (A)–(C), using APF\(_0\) we have a smaller p-value than in Bendich et al. (2016) if \( I = [0, 137] \) and an even smaller p-value if \( I = [0, 60] \); and for \( k = 1 \) under setting (B), our p-value is about seven times larger than the p-value in Bendich et al. (2016) if \( I = [0, 25] \), and else it is similar or smaller. For \( k = 1 \) and \( I = [0, 25] \), the large difference between our p-values under settings (B) and (C) indicates that the presence of outliers violates the result of Theorem 5.1 and care should hence be taken. In our opinion we can better trust the results without outliers, where in contrast to Bendich et al. (2016) we see a clear gender effect when considering the connected components. Notice also that in agreement with the discussion of Figure 5 in Example 2, for each setting A, B, and C and each dimension \( k = 0, 1 \), the p-values in Table 2 are smallest when considering the smaller interval \( I = [0, 60] \) or \( I = [15, 25] \).

Appendix D (supplementary materials) provides an additional Example 8 illustrating the use of two-sample test in a simulation study.

**Supplementary Materials**

The supplemental files concern an appendix and our code used for the various examples. The code uses R from R Core Team (2018) with the R packages: plotrix, spatstat, TDA, devtools, doParallel, fda, R.matlab, depthTools, and scatterplot3d. described in Lemon (2006), Baddeley, Rubak and Turner (2015), Wickham, Hester, and Chang (2018), Corporation and Weston (2018), Ramsay et al. (2018), Bengtsson (2018), López-Pintado and Romo (2009), and Ligges and Ligges and Mächler (2003).
Appendix: Contains the following.
A Transforming confidence regions for persistence diagrams used for separating topological signal from noise.
B Additional example related to Section 4.1.
C Additional example related to Section 4.2.
D Additional example related to Section 5.
E Further methods for two or more samples of accumulated persistence functions.
E.1 Clustering.
E.2 Supervised classification.
F Proof of Theorem 4.1.
(Appendix.pdf)

Code to reproduce the examples: Contains files for the R-code of all examples, see the file Readme. (Rcode.zip)

Acknowledgments
Helpful discussions with Lisbeth Fajstrup on persistent homology are acknowledged. We thank Henrike Häbel and Achmad Choiruddin for their helpful comments. In connection to the brain artery trees dataset we thank James Stephen Marron and Sean Skwerer for helpful discussions and the CASiLaB at The University of North Carolina at Chapel Hill for providing the data distributed by the MIDAS Data Server at Kitware, Inc. We are grateful to the editors and the referees for useful comments.

Funding
Supported by The Danish Council for Independent Research—Natural Sciences, grant 7014-00074B, “Statistics for point processes in space and beyond,” and by the “Centre for Stochastic Geometry and Advanced Bioimaging,” funded by grant 8721 from the Villum Foundation.

References
Baddeley, A., Rubak, E., and Turner, R. (2015), Spatial Point Patterns: Methodology and Applications With R, Boca Raton, FL: Chapman and Hall/CRC Press. doi: 10.1201/b19708 [675,680]
Baddeley, A. J., and Silverman, B. W. (1984), “A Cautionary Example on the Use of Second-Order Methods for Analyzing Point Patterns,” Biometrics, 40, 1089–1093. doi: 10.2307/2531159 [676]
Bendich, P., Marron, J., Miller, E., Pieloch, A., and Skwerer, S. (2016), “Persistent Homology Analysis of Brain Artery Trees,” The Annals of Applied Statistics, 10, 198–218. doi: 10.1214/15-AOAS886 [673,675,680]
Bengtsson, H. (2018), “R.matlab: Read and Write MATLAB Files and Call MATLAB From Within R,” R Package Version 3.6.2. [680]
Biscio, C. A. N., and Lavancier, F. (2016), “Quantifying Repulsiveness of Determinantal Point Processes,” Bernoulli, 22, 2001–2028. doi: 10.3150/15-BEJ718 [676]
Bubenik, P. (2015), “Statistical Topological Data Analysis Using Persistence Landscapes,” Journal of Machine Learning Research, 16, 77–102. doi: 10.1016/j.jmlr.2013.06.009 [674]
Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. (2014), “Robust Topological Inference: Distance to a Measure and Kernel Distance,” arXiv no. 1412.7197. [675]
Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. (2013), “On the Bootstrap for Persistence Diagrams and Landscapes,” Modeling and Analysis of Information Systems, 20, 111–120. doi: 10.18255/1818-1015-2013-6-111-120 [674,679]
Chen, Y.-C., Wang, D., Rinaldo, A., and Wasserman, L. (2015), “Statistical Analysis of Persistence Intensity Functions,” arXiv no. 1510.02502. [674]
Corporation, M., and Weston, S. (2018), “doParallel: Foreach Parallel Adapter for the ‘Parallel’ Package,” R Package Version 1.0.14. [680]
Daley, D. J., and Vere-Jones, D. (2003), An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods (2nd ed.), New York: Springer-Verlag. doi: 10.1007/97877777 [674]
Edelsbrunner, H., and Harer, J. L. (2010), Computational Topology, Providence, RI: American Mathematical Society. doi: 10.1090/mbk/069 [673,679]
Fasy, B., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014), “Confidence Sets for Persistence Diagrams,” The Annals of Statistics, 42, 2301–2339. doi: 10.1214/14-aos1252 [671,673,674,675,679]
Lavancier, F., Møller, J., and Rubak, E. (2015), “Determinantal Point Process Models and Statistical Inference,” Journal of the Royal Statistical Society, Series B, 77, 853–877. doi: 10.1111/rssb.12096 [676]
Lemon, J. (2006), “Plotrix: A Package in the Red Light District of R,” R-News, 6, 8–12. [680]
Ligges, U., and Mächler, M. (2003), “Scatterplot3d—An R Package for Visualizing Multivariate Data,” Journal of Statistical Software, 8, 1–20. doi: 10.18637/jss.v008.i11 [680]
López-Pintado, S., and Romo, J. (2009), “On the Concept of Depth for Functional Data,” Journal of the American Statistical Association, 104, 718–734. doi: 10.1198/jasa.2009.010 [675,677,680]
MacPherson, R., and Wineberg, D. (2012), “Measuring Shape With Topology,” Journal of Mathematical Physics, 53, 073516. doi: 10.1063/1.4737391 [673]
Matern, B. (1986), Spatial Variation, Lecture Notes in Statistics (Vol. 36), Berlin: Springer-Verlag. doi: 10.1007/978-1-4615-7892-5 [676]
Møller, J., and Waagepetersen, R. P. (2004), Statistical Inference and Simulation for Spatial Point Processes, Boca Raton, FL: Chapman and Hall/CRC. doi: 10.1201/9780203496930 [676]
Spearman, D. (2016), “Global Envelope Tests for Spatial Processes,” Journal of the Royal Statistical Society, Series B, 79, 381–404. doi: 10.1111/rssb.12172 [675,676]
Praestgaard, J. T. (1995), “Permutation and Bootstrap Kolmogorov–Smirnov Tests for the Equality of Two Distributions,” Scandinavian Journal of Statistics, 22, 305–322. doi: 10.2307/4616362 [679]
R Core Team (2018), R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. [680]
Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2018), “ida: Functional Data Analysis,” R Package Version 2.4.8. [680]
Robins, V., and Turner, K. (2016), “Principal Component Analysis of Persistent Homology Rank Functions With Case Studies of Spatial Point Patterns, Sphere Packing and Colloids,” Physica D: Nonlinear Phenomena, 334, 99–117. doi: 10.1016/j.physd.2016.03.007 [676,677]
Sun, Y., and Genton, M. G. (2011), “Functional Boxplots,” Journal of Computational and Graphical Statistics, 20, 316–334. doi: 10.1198/jcgs.2011.09224 [677,678]
van der Vaart, A. W., and Wellner, J. A. (1996), Weak Convergence and Empirical Processes, Springer Series in Statistics, New York: Springer-Verlag. doi: 10.1007/978-1-4757-2545-2 [679,680]
Wickham, H., Hester, J., and Chang, W. (2018), “devtools: Tools to Make Developing R Packages Easier,” R Package Version 2.0.1. [680]