Supplementary materials to "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
Department of Mathematical Sciences, Aalborg University, Denmark
and
Jesper Møller
Department of Mathematical Sciences, Aalborg University, Denmark

October 19, 2018

Appendix

Appendix A-F gather complements and additional examples to Sections 3-5. Our setting and notation are as follows. All the examples are based on a simulated point cloud $\{x_1, \ldots, x_N\} \subset \mathbb{R}^2$ as described in Section 2.1, with $x_1, \ldots, x_N$ being IID points where N is a fixed positive integer. As in Section 1.1.1, our setting corresponds to applications typically considered in TDA where the aim is to obtain topological information about a compact set $C \subset \mathbb{R}^2$ which is unobserved and where possibly noise appears: For specificity, we let $x_i = y_i + \epsilon_i$, $i = 1, \ldots, N$, where $y_1, \ldots, y_N$ are IID points with support C, the noise $\epsilon_1, \ldots, \epsilon_N$ are IID and independent of $y_1, \ldots, y_N$, and ϵ_i follows the restriction to the square $[-10\sigma, 10\sigma]^2$ of a bivariate zero-mean normal distribution with IID coordinates and standard deviation $\sigma \geq 0$ (if $\sigma = 0$ there is no noise). We denote this distribution for ϵ_i by $N_2(\sigma)$ (the restriction to $[-10\sigma, 10\sigma]^2$ is only imposed for technical reasons and is not of practical importance). We let C_t be the union of closed discs of radii t and centred at $x_1, \ldots, x_N$, and we study how the topological features of C_t changes as $t \geq 0$ grows. For this we use the Delaunay-complex mentioned in Remark 1, Section 1.1.1. Finally, we denote by $\mathcal{C}((a, b), r)$ the circle with center (a, b) and radius r.

A Transforming confidence regions for persistence diagrams used for separating topological signal from noise

As noted in Section 3, 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. We avoid presenting the technical description of these constructions and results, which depend on different choices of complexes (or more precisely so-called filtrations). For specificity, in this appendix we just consider the Delaunay-complex and discuss the transformation of such a confidence region into one for an accumulate persistence function.

We use the following notation. As in the aforementioned references, consider the persistence diagram PD_k for an unobserved compact manifold $C \subset \mathbb{R}^2$ and obtained as in Section 1.1.1 by considering the persistence as $t \geq 0$ grows of k-dimensional topological features of the set consisting of all points in $\mathbb{R}^2$ within distance t from C. Note that PD_k is considered as being non-random and unknown; of course in our simulation study presented in Example 5 below we only pretend that PD_k is unknown. Let $\widehat{\mathrm{PD}}_{k,N}$ be the random persistence diagram obtained as in Section 1.1.1 from IID points $x_1, \ldots, x_N$ with support C. Let $\mathcal{N} = \{(b,d) : b \leq d,\, l \leq 2c_N\}$ be the set of points at distance $\sqrt{2}c_N$ of the diagonal in the persistence diagram. Let $S(b,d) = \{(x,y) : |x-b| \leq c_N,\, |y-d| \leq c_N\}$ be the square with center (b,d), sides parallel to the b- and d-axes, and of side length $2c_N$. Finally, let $\alpha \in (0,1)$.

Fasy *et al.* (2014) and Chazal *et al.* (2014) suggested various ways of constructing a bound $c_N > 0$ so that an asymptotic conservative confidence region for PD_k with respect to the bottleneck distance W_∞: Briefly, for $\epsilon > 0$, we have $W_\infty(\mathrm{PD}_k^{(1)}, \mathrm{PD}_k^{(2)}) \leq \epsilon$ if for any $(b_i^{(1)}, d_i^{(1)}, c_i^{(1)}) \in \mathrm{PD}_k^{(1)}$, the multiplicity $c_i^{(1)}$ is equal to the sum of those multiplicities $c_j^{(2)}$ with $(b_j^{(2)}, d_j^{(2)}, c_j^{(2)}) \in \mathrm{PD}_k^{(2)}$ and $(b_j^{(2)}, d_j^{(2)}) \in S(b_i^{(1)}, d_i^{(1)})$. More simply, if all multiplicities of $\mathrm{PD}_k^{(1)}$ and $\mathrm{PD}_k^{(2)}$ are one, then $W_\infty(\mathrm{PD}_k^{(1)}, \mathrm{PD}_k^{(2)}) \leq \epsilon$ whenever $\mathrm{PD}_k^{(2)}$ has exactly one point $(b_j^{(2)}, d_j^{(2)})$ in each square $S(b_i^{(1)}, d_i^{(1)})$, with $(b_i^{(1)}, d_i^{(1)}, c_i^{(1)})$ a point of $\mathrm{PD}_k^{(1)}$. Then the asymptotic conservative $100(1-\alpha)\%$-confidence region for PD_k with respect to W_∞ is given by

$$\liminf_{N\to\infty} \mathrm{P}\left(W_\infty(\mathrm{PD}_k, \widehat{\mathrm{PD}}_{k,N}) \leq c_N\right) \geq 1-\alpha. \tag{11}$$

If all multiplicities of PD_k and $\widehat{\mathrm{PD}}_{k,N}$ are one, this confidence region consists of those persistence diagrams PD_k which have exactly one point (b_j, d_j) in each square $S(b_i, d_i)$, with $(b_i, d_i, 1) \in \widehat{\mathrm{PD}}_{k,N}$, and have an arbitrary number of birth-death pairs in the set $\mathcal{N}$. Fasy *et al.* (2014) considered the birth-death pairs of $\widehat{\mathrm{PD}}_{k,N}$ falling in $\mathcal{N}$ as noise and the remaining pairs as representing a significant topological feature of C.

Using (11) an asymptotic conservative $100(1-\alpha)\%$-confidence region for the APF_k corresponding to PD_k is immediately obtained. This region will be bounded by two functions $\widehat{A}_{k,N}^{\min}$ and $\widehat{A}_{k,N}^{\max}$ specified by $\widehat{\mathrm{PD}}_{k,N}$ and c_N. Due to the accumulating nature of APF_k,

the span between the bounds is an increasing function of the meanage. When using the Delaunay-complex, Chazal *et al.* (2014) showed that the span decreases as N increases; this is illustrated in Example 5 below.

Example 5 (simulation study). Let $C = \mathcal{C}((-1.5, 0), 1) \cup \mathcal{C}((1.5, 0), 0.8)$ and suppose each point x_i is uniformly distributed on C. Figure 8 shows C and an example of a simulated point cloud with $N = 300$ points. We use the bootstrap method implemented in the **R**-package `TDA` and presented in Chazal *et al.* (2014) to compute the 95%-confidence region for PD_1 when $N = 300$, see the top-left panel of Figure 9, where the two squares above the diagonal correspond to the two loops in C and the other squares correspond to topological noise. Thereby 95%-confidence regions for RRPD_1 (top-right panel) and APF_1 (bottom-left panel) are obtained. The confidence region for APF_1 decreases as N increases as demonstrated in the bottom panels where N is increased from 300 to 500. As noticed in Section 1.3.1, we must be careful when using results based on the bottleneck metric, because small values of the bottleneck metric does not correspond to closeness of the two corresponding APFs: Although close persistence diagram with respect to the bottleneck distance imply that the two corresponding APFs are close with respect to the L^q-norm ($1 \leq q \leq \infty$), the converse is not true. Hence, it is possible that an APF is in the confidence region plotted in Figure 9 but that the corresponding persistence diagram is very different from the truth.

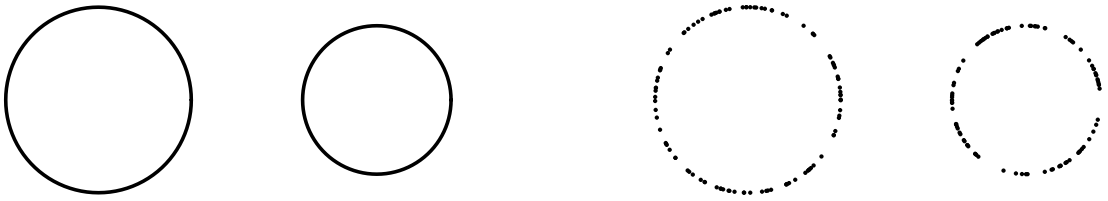

Figure 8: The set C in Example 5 (left panel) and a simulated point cloud of $N = 300$ independent and uniformly distributed points on C (right panel).

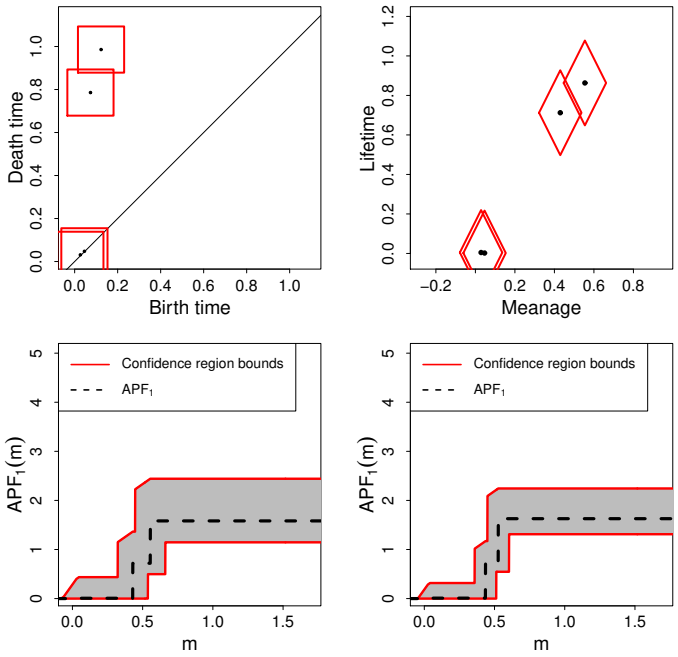

Figure 9: 95%-confidence regions obtained by the bootstrap method for PD_1 (top-left panel) and its corresponding $RRPD_1$ (top-right panel) when C and $x_1, \ldots, x_{300}$ are as in Figure 8. The bottom-left panel shows the corresponding 95%-confidence region for APF_1. The bottom-right panel shows the 95%-confidence region for APF_1 when a larger point cloud with 300 points is used.

B Additional example related to Section 4.1 "Functional boxplot"

The functional boxplot described in Section 4.1 can be used as an exploratory tool for the curves given by a sample of APF_ks. It provides a representation of the most central curve and the variation around this. It can also be used for outliers detection as illustrated in the following example.

Example 6 (simulation study). We consider a sample of 65 independent APF_ks, where the joint distribution of the first 50 APF_ks is exchangeable, whereas the last 15 play the role of outliers. We suppose each APF_k corresponds to a point process of 100 IID points,

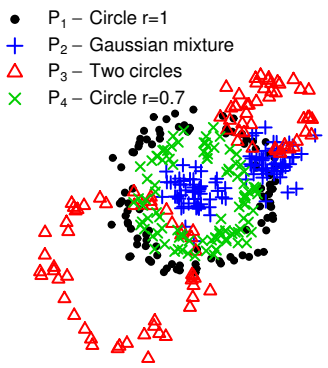

Figure 10: Simulated realizations of the four types of point processes, each consisting of 100 IID points with distribution either P_1 (black dots), P_2 (blue crosses), P_3 (red triangles), or P_4 (rotated green crosses).

where each point x_i follows one of the following distributions $P_1, \ldots, P_4$.

- P_1 (unit circle): x_i is a uniform point on $\mathcal{C}((0,0),1)$ perturbed by $N_2\,(0.1)$-noise.
- P_2 (Gaussian mixture): Let y_i follow $N_2\,(0.2)$, then $x_i = y_i$ with probability 0.5, and $x_i = y_i + (1.5, 0.5)$ otherwise.
- P_3 (two circles): x_i is a uniform point on $\mathcal{C}((-1,-1),1) \cup \mathcal{C}((1,1),0.5)$ perturbed by $N_2\,((0,0),0.1)$-noise.
- P_4 (circle of radius 0.7): x_i is a uniform point on $\mathcal{C}((0,0),0.7)$ perturbed by $N_2\,(0.1)$-noise.

We let the first 50 point processes be obtained from P_1 (the distribution for non-outliers), the next 5 from P_2, the following 5 from P_3, and the final 5 from P_4. Figure 10 shows a simulated realization of each of the four types of point processes.

Figure 11 shows the functional boxplots when considering APF_0 (left panel) and APF_1 (right panel). The curves detected as outliers and corresponding to the distributions P_2, P_3, and P_4 are plotted in red, blue, and green, respectively. In both panels the outliers detected by the 1.5 criterion agree with the true outliers.

In the left panel, each curve has an accumulation of small jumps between $m = 0$ and $m \approx 0.1$, corresponding to the moments where the points associated to each circle are connected by the growing discs in the sequence $\{C_t\}_{t\geq 0}$. The curves corresponding to realizations of P_3 have a jump at $m \approx 0.38$ which corresponds to the moment where the points associated to the two circles used when defining P_3 are connected by the growing discs in the sequence $\{C_t\}_{t\geq 0}$. The points following the distribution P_4 are generally closer to each other than the ones following the distribution P_1 as the radius of the underlying circle is smaller. This corresponds to more but smaller jumps in APF_0 for small meanages, and hence the curves of APF_0 are lower when they correspond to realizations of P_1 than to realizations of P_4; and as expected, for large meanages, the curves of APF_0 are larger when they correspond to realizations of P_1 than to realizations of P_4. Note that if we redefine P_4 so that the $N_2\,(0.1)$-noise is replaced by $N_2\,(0.07)$-noise, then the curves would be the same up to rescaling.

In the right panel, we observe clear jumps in all APF_1s obtained from P_1, P_3, and P_4. These jumps correspond to the first time that the loops of the circles in P_1, P_3, and P_4 are covered by the union of growing discs in the sequence $\{C_t\}_{t\geq 0}$. Once again, if we have used $N_2\,(0.07)$-noise in place of $N_2\,(0.1)$-noise in the definition of P_4, the curves would be the same up to rescaling.

If we repeat everything but with the distribution P_4 redefined so that $\mathcal{C}((0,0),0.7)$ is replaced by $\mathcal{C}((0,0),0.8)$, then the support of P_4 is closer to that of P_1 and it becomes harder in the case of APF_0 to detect the outliers with distribution P_4 (we omit the corresponding plot we have produced); thus further simulations for determining a stronger criterion would be needed.

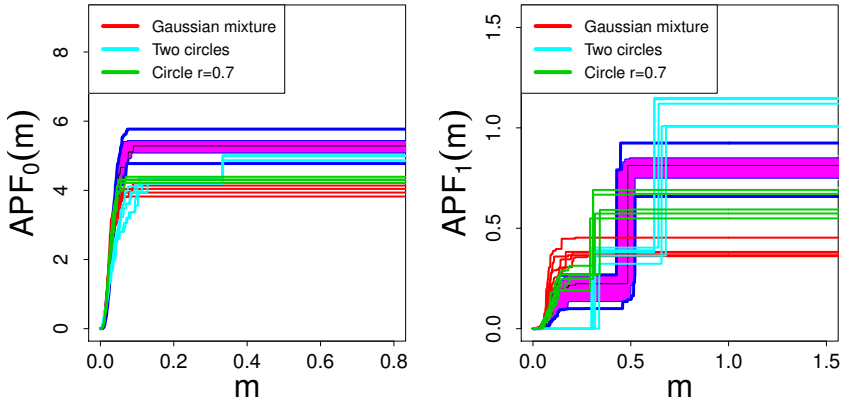

Figure 11: Functional boxplots of 65 APFs based on the topological features of dimension 0 (left panel) and 1 (right panel). In each panel, 50, 5, 5, and 5 APFs are obtained from the Delaunay-complex of 100 IID points from the distribution P_1, P_2, P_3, and P_4, respectively. The APFs detected as outliers are plotted in red, blue, and green in the case of P_2, P_3, and P_4, respectively.

C Additional example related to Section 4.2 "Confidence region for the mean function"

This appendix provides yet an example to illustrate the bootstrap method in Section 4.2 for obtaining a confidence region for the mean function of a sample of IID APF_ks.

Example 7 (simulation study). Consider 50 IID copies of a point process consisting of 100 independent and uniformly distributed points on the union of three circles with radius 0.25 and centred at $(-1,-1)$, $(0,1)$, and $(1,-1)$, respectively (these circles were also considered in the example of Section 1.1.1). A simulated realization of the point process is shown in the left panel of Figure 12, and the next two panels show simulated confidence regions for APF_0 and APF_1, respectively, when the bootstrap procedure with $B = 1000$ is used. In the middle panel, between $m = 0$ and $m = 0.2$, there is an accumulation of small jumps corresponding to the moment when each circle is covered by the union of growing discs from the sequence $\{C_t\}_{t\geq 0}$; we interpret these small jumps as topological noise. The jump at $m \approx 0.25$ corresponds to the moment when the circles

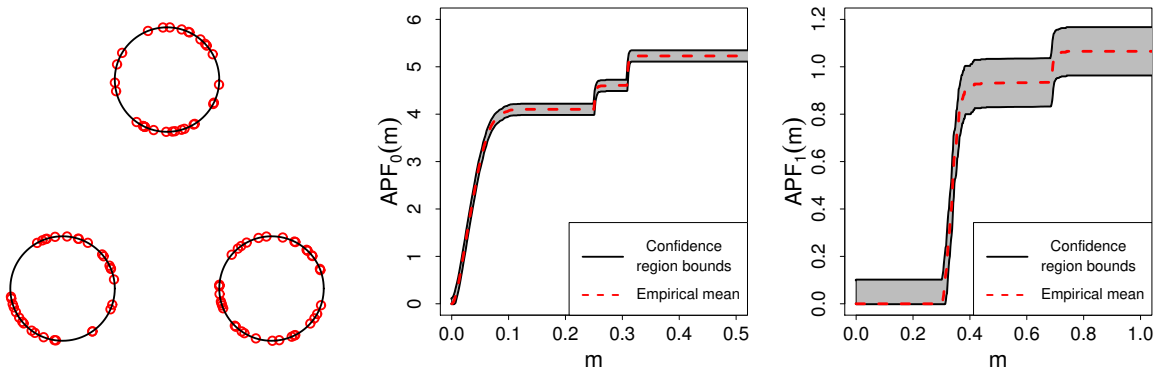

Figure 12: A simulation of 100 independent and uniformly distributed points on the union of three circles (dashed lines) with the same radius $r = 0.5$ and centred at $(-1,-1)$, $(0,1)$, and $(1,-1)$ (left panel). The 95%-confidence regions for the mean APF_0 (middle panel) and the mean APF_1 (right panel) are based on 50 IID simulations.

centred at $(-1,-1)$ and $(1,-1)$ are connected by the growing discs, and the jump at $m \approx 0.3$ to when all three circles are connected by the growing discs. In the right panel, at $m \approx 0.3$ there is an accumulation of small jumps corresponding to the moment when the three circles are connected by the growing discs and they form a loop at $m = 0.25$ in Figure 1. The disappearance of this loop at $m = 0.69$ in Figure 1 corresponds to the jump at $m \approx 0.7$ in Figure 12.

D Additional example to Section 5 "Two samples of accumulated persistence functions"

Section 5 considered two samples of independent RRPD_ks $D_1, \ldots, D_{r_1}$ and $E_1, \ldots, E_{r_2}$, where each D_i $(i = 1, \ldots, r_1)$ has distribution P_D and each E_j has distribution P_E $(j = 1, \ldots, r_2)$. Then we studied a bootstrap two-sample test to asses the null hypothesis $\mathcal{H}_0$: $\mathrm{P}_D = \mathrm{P}_E$, e.g. in connection to the brain artery trees. An additional example showing the performance of the test is presented below.

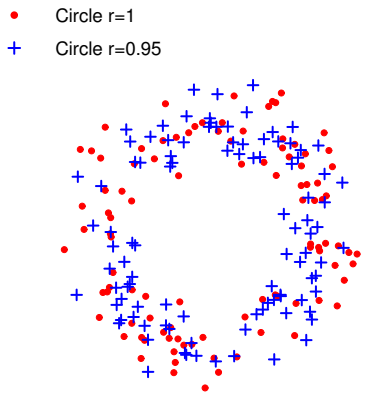

Figure 13: A simulation of 100 independent and uniformly distributed points on the circle centred at $(0,0)$ with radius 1 and perturbed by $N_2\left((0,0),0.2\right)$-noise (red dots), together with 100 independent and uniformly distributed points on the circle centred at $(0,0)$ with radius 0.95 and perturbed by $N_2\left((0,0),0.2\right)$-noise (blue crosses).

Example 8 (simulation study). Let P_D be the distribution of a RRPD_k obtained from 100 independent and uniformly distributed points on $\mathcal{C}((0,0),1)$ perturbed by $N_2\left(0.2\right)$-noise, and define P_E in a similar way but with a circle of radius 0.95. A simulated realization of each point process is shown in Figure 13; it seems difficult to recognize that the underlying circles are different. Let us consider the two-sample test statistics (6) and (10) with $I = [0,3]$, $r_1 = r_2 = 50$, and $\alpha = 0.05$. Over 500 simulations of the two samples of RRPD_k we obtain the following percentage of rejection: For M_{r_1,r_2}, 5.2% if $k = 0$, and 24.2% if $k = 1$. For KS_{r_1,r_2} much better results are observed, namely 73.8% if $k = 0$, and 93.8% if $k = 1$, where this high percentage is mainly caused by the largest lifetime of a loop.

E Further methods for two or more samples of accumulated persistence functions

E.1 Clustering

Suppose $A_1, \ldots, A_r$ are APF_ks which we want to label into $K < r$ groups by using a method of clustering. Such methods are studied many places in the literature for functional data, see the survey in Jacques and Preda (2014). In particular, Chazal *et al.* (2009), Chen *et al.* (2015), and Robins and Turner (2016) consider clustering in connection to RRPD_ks. Whereas the RRPD_ks are two-dimensional functions, it becomes easy to use clustering for the one-dimensional APF_ks as illustrated in Example 9 below.

For simplicity we just consider the standard technique known as the K-means clustering algorithm (Hartigan and Wong (1979)). For more complicated applications than considered in Example 9 the EM-algorithm may be needed for the K-means clustering algorithm. As noticed by a referee, to avoid the use of the EM-algorithm we can modify (9) or (10) and thereby construct a distance/similarity matrix for different APFs which is used to perform hierarchical clustering. However, for Example 9 the results using hierarchical clustering (omitted here) were not better than with the K-means algorithm.

Assume that $A_1, \ldots, A_r$ are pairwise different and square-integrable functions on $[0, T]$, where T is a user-specified parameter. For example, if $\mathrm{RRPD}_k \in \mathcal{D}_{k,T,n_{\max}}$ (see Section 4.2), then $\mathrm{APF}_k \in L^2([0, T])$. The K-means clustering algorithm works as follows.

- Chose uniformly at random a subset of K functions from $\{A_1, \ldots, A_r\}$; call these functions centres and label them by $1, \ldots, K$.

- Assign each non-selected APF_k the label i if it is closer to the centre of label i than to any other centre with respect to the L^2-distance on $L^2([0, T])$.

- In each group, reassign the centre by the mean curve of the group (this may not be

an APF_k of the sample).

- Iterate these steps until the assignment of centres does not change.

The algorithm is known to be convergent, however, it may have several drawbacks as discussed in Hartigan and Wong (1979) and Bottou and Bengio (1995).

Example 9 (simulation study). Consider $K = 3$ groups, each consisting of 50 APF_0s and associated to point processes consisting of 100 IID points, where each point x_i follows one of the following distributions P_1, P_2, and P_3 for groups 1, 2, and 3, respectively.

- P_1 (unit circle): x_i is a uniform point on $\mathcal{C}((0,0),1)$ perturbed by $N_2\,(0.1)$-noise.
- P_2 (two circles): x_i is a uniform point on $\mathcal{C}((-1,-1),0.5) \cup \mathcal{C}((1,1),0.5)$ perturbed by $N_2\,(0.1)$-noise.
- P_3 (circle of radius 0.8): x_i is a uniform point on $\mathcal{C}((0,0),0.8)$ perturbed by $N_2\,(0.1)$-noise.

We started by simulating a realization of each of the $3 \times 50 = 150$ point processes. The left panel of Figure 14 shows one realization of each type of point process; it seems difficult to distinguish the underlying circles for groups 1 and 3, but the three APF_0s associated to these three point clouds are in fact assigned to their right groups. The right panel of Figure 14 shows the result of the K-means clustering algorithm. Here we have used the **R**-function "kmeans" for the K-means algorithm and it took only a few seconds when evaluating each $A_i(m)$ at 2500 equidistant values of m between 0 and $T = 0.5$. As expected we see more overlap between the curves of the APF_0s assigned to groups 1 and 3.

We next repeated 500 times the simulation of the 150 point processes. A clear distinction between the groups was obtained by the K-means algorithm applied for connected components: The percentage of wrongly assigned APF_0s among the $500 \times 3 \times 50 = 75000$

APF_0s had an average of 4.5% and a standard deviation of 1.6%. The assignment error was in fact mostly caused by incorrect labelling of APF_0s associated to P_1 or P_3. This is to be expected as the underlying circles used in the definitions of P_1 and P_3 are rather close, whereas the underlying set in the definition of P_2 is different with two connected components as represented by the jump at $m \approx 0.4$ in the middle panel of Figure 14.

Even better results are obtained when considering loops instead of connected components: The percentage of wrongly assigned APF_1s among the 75000 APF_1s had an average of 1.6% and a standard deviation of 1.0%. This was mainly due to the sets underlying P_1, P_2, and P_3 which have distinctive loops that results in clear distinct jumps in the APF_1s as seen in the right panel of Figure 14.

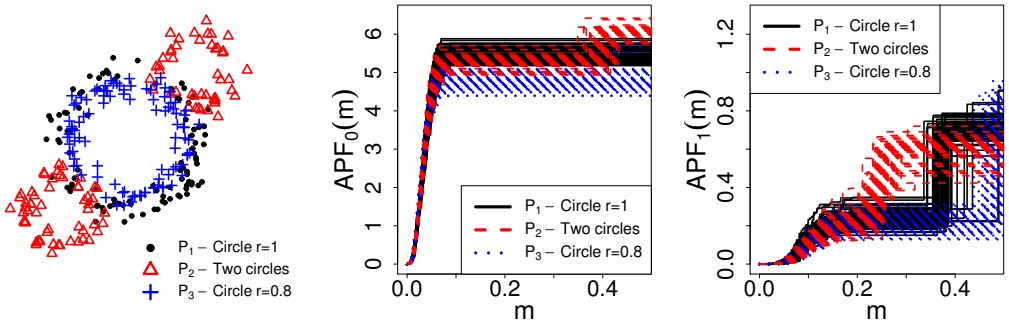

Figure 14: Left panel: Simulated example of the three point processes, each consisting of 100 IID points drawn from the distribution P_1 (black dots), P_2 (red triangles), or P_3 (blue crosses). Middle panel: The 150 APF_0s obtained from the simulation of the 150 point processes associated to P_1, P_2, or P_3, where the colouring in black, red, or blue specifies whether the K-means algorithm assigns an APF_0 to the group associated to P_1, P_2, or P_3. Right panel: As the middle panel but for the 150 APF_1s.

E.2 Supervised classification

Suppose we want to assign an APF_k to a training set of K different groups $\mathcal{G}_1, \ldots, \mathcal{G}_K$, where $\mathcal{G}_i$ is a sample of r_i independent APF_ks $A_1^i, \ldots, A_{r_i}^i$. For this purpose supervised classification methods for functional data may be adapted.

We just consider a particular method by López-Pintado *et al.* (2010): Suppose $\alpha \in [0,1]$ and we believe that at least $100(1-\alpha)\%$ of the APF_ks in each group are IID, whereas the remaining APF_ks in each group follow a different distribution and are considered as outliers (see Section 4.1). For a user-specified parameter $T > 0$ and $i = 1, \ldots, K$, define the $100\alpha\%$-trimmed mean $\overline{A}_i^{\alpha}$ with respect to $\mathcal{G}_i$ as the mean function on $[0,T]$ of the $100(1-\alpha)\%$ APF_ks in $\mathcal{G}_i$ with the largest MBD_{r_i}, see (5). Assuming $\cup_{i=1}^K \mathcal{G}_i \subset L^2([0,T])$, an APF_k $A \in L^2([0,T])$ is assigned to $\mathcal{G}_i$ if

$$i = \underset{j \in \{1,\ldots,K\}}{\operatorname{argmin}} \|\overline{A}_j^{\alpha} - A\|, \tag{12}$$

where $\|\cdot\|$ denotes the L^2-distance. Here, the trimmed mean is used for robustness and allows a control over the curves we may like to omit because of outliers, but e.g. the median could have been used instead.

Example 10 (simulation study). Consider the following distributions $\text{P}_1, \ldots, \text{P}_4$ for a point x_i.

- P_1 (unit circle): x_i is a uniform point on $\mathcal{C}((0,0),1)$ which is perturbed by $N_2\,(0.1)$-noise.
- P_1' (two circles, radii 1 and 0.5): x_i is a uniform point on $\mathcal{C}((0,0),1) \cup \mathcal{C}((1.5,1.5),0.5)$ and perturbed by $N_2\,(0.1)$-noise.
- P_2 (circle of radius 0.8): x_i is a uniform point on $\mathcal{C}((0,0),0.8)$ which is perturbed by $N_2\,(0.1)$-noise.
- P_2' (two circles, radii 0.8 and 0.5): x_i is a uniform point on $\mathcal{C}((0,0),0.8) \cup \mathcal{C}((1.5,1.5),0.5)$ and perturbed by $N_2\,(0.1)$-noise.

For $k = 0, 1$ we consider the following simulation study: $K = 2$ and $r_1 = r_2 = 50$; $\mathcal{G}_1$ consists of 45 APF_ks associated to simulations of point processes consisting of 100 IID points with distribution P_1 (the non-outliers) and 5 APF_ks obtained in the same way but from P_1' (the outliers); $\mathcal{G}_2$ is specified in the same way as $\mathcal{G}_1$ but replacing P_1 and P_1' with

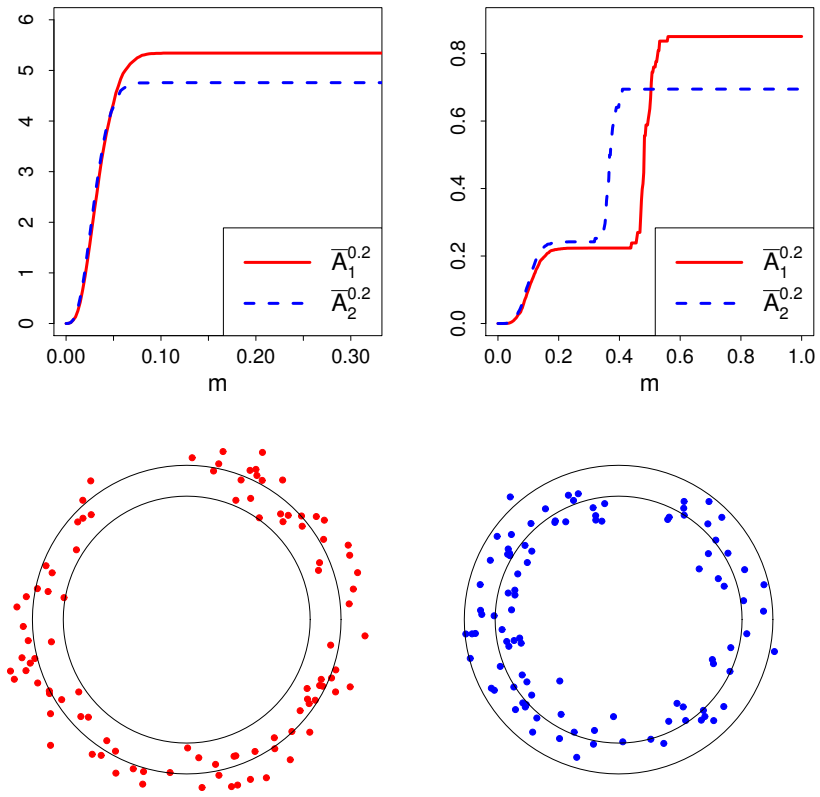

Figure 15: Top panels: The 20%-trimmed mean functions with respect to $\mathcal{G}_1$ and $\mathcal{G}_2$ when considering APF_0s (left) and APF_1s (right) obtained from the Delaunay-complex and based on 100 IID points following the distribution P_1 (solid curve) or P_2 (dotted curve). Bottom panels: Examples of point clouds with associated APF_0s assigned to the wrong group, together with the circles of radius 0.8 and 1.

P_2 and P_2', respectively; and we have correctly specified that $\alpha = 0.2$. Then we simulated 100 APF_ks associated to P_1 and 100 APF_ks associated to P_2, i.e. they are all non-outliers. Finally, we used (12) to assign each of these 200 APF_ks to either $\mathcal{G}_1$ or $\mathcal{G}_2$.

The top panels in Figure 15 show the 20%-trimmed means $\overline{A}_1^{0.2}$ and $\overline{A}_2^{0.2}$ when $k = 0$ (left) and $k = 1$ (right). The difference between the 20%-trimmed means is clearest when $k = 1$ and so we expect that the assignment error is lower in that case. In fact wrong assignments happened mainly when the support of P_1 or P_2 was not well covered by the point cloud as illustrated in the bottom panels.

Repeating this simulation study 500 times, the percentage of APF_0s wrongly assigned among the 500 repetitions had a mean of 6.7% and a standard deviation of 1.7%, whereas for the APF_1s the mean was 0.24% and the standard deviation was 0.43%. To investigate how the results depend on the radius of the smallest circle, we repeated everything but with radius 0.9 in place of 0.8 when defining the distributions P_2 and P_2'. Then for the APF_0s, the proportion of wrong assignments had a mean of 23.2% and a standard deviation of 2.9%, and for the APF_1s, a mean of 5.7% and a standard deviation of 1.9%. Similar to Example 9, the error was lowest when $k = 1$ and this is due to the largest lifetime of a loop.

F Proof of Theorem 4.1

The proof of Theorem 4.1 follows along similar lines as in Chazal *et al.* (2013) as soon as we have verified Lemma F.2 below. Note that the proof of Lemma F.2 is not covered by the approach in Chazal *et al.* (2013).

We first need to recall the following definition, where $\mathcal{B}_T$ denotes the topological space of bounded real valued Borel functions defined on $[0, T]$ and its topology is induced by the uniform norm.

Definition F.1. *A sequence $\{X_r\}_{r=1,2,\ldots}$ of random elements in $\mathcal{B}_T$ converges in distribution to a random element X in $\mathcal{B}_T$ if for any bounded continuous function $f : \mathcal{B}_T \mapsto \mathbb{R}$, $\text{E}f(X_r)$ converges to $\text{E}f(X)$ as $r \to \infty$.*

Lemma F.2. *Let the situation be as in Section 4.2. As $r \to \infty$, $\sqrt{r}\left(\overline{A}_r - \mu\right)$ converges in distribution towards a zero-mean Gaussian process on $[0, T]$ with covariance function $c(m, m') = \text{Cov}\left(A_1(m), A_1(m')\right)$, $m, m' \in [0, T]$.*

Proof. We need some notation and to recall some concepts of empirical process theory. For $D \in \mathcal{D}_T^{k,n_{\max}}$, denote A_D the APF_k of D. Let $\mathcal{F} = \{f_m : 0 \leq m \leq T\}$ be the class of

functions $f_m : \mathcal{D}_T^{k,n_{\max}} \mapsto [0,\infty)$ given by $f_m(D) = A_D(m)$. To see the connection with empirical process theory, we consider

$$\mathbb{G}_r(f_m) = \sqrt{r}\left(\frac{1}{r}\sum_{i=1}^{r} f_m(D_i) - \mu(m)\right)$$

as an empirical process. Denote $\|\cdot\|$ the L^2 norm on $\mathcal{F}$ with respect to the distribution of D_1, i.e. $\|f_m(\cdot)\|^2 = \mathrm{E}\left\{A_{D_1}(m)^2\right\}$. For $u, v \in \mathcal{F}$, the bracket $[u,v]$ is the set of all functions $f \in \mathcal{F}$ with $u \leq f \leq v$. For any $\epsilon > 0$, $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the smallest integer $J \geq 1$ such that $\mathcal{F} \subset \cup_{j=1}^{J}[u_j, v_j]$ for some functions $u_1, \ldots, u_J$ and $v_1, \ldots, v_J$ in $\mathcal{F}$ with $\|v_j - u_j\| \leq \epsilon$ for $j = 1, \ldots, J$. We show below that $\int_0^1 \sqrt{\log\left(N_{[]}\left(\epsilon, \mathcal{F}, \|\cdot\|\right)\right)}\, \mathrm{d}\epsilon$ is finite. Then, by Theorem 19.5 in van der Vaart (2000), $\mathcal{F}$ is a so-called Donsker class which implies the convergence in distribution of $\mathbb{G}_r(f_m)$ to a Gaussian process as in the statement of Lemma F.2.

For any sequence $-\infty = t_1 < \ldots < t_J = \infty$ with $J \geq 2$, for $j = 1, \ldots, J-1$, and for $D = \{(m_1, l_1, c_1), \ldots, (m_n, l_n, c_n)\} \in \mathcal{D}_T^{k,n_{\max}}$, let $u_j(D) = \sum_{i=1}^{n} c_i l_i 1(m_i \leq t_j)$ and $v_j(D) = \sum_{i=1}^{n} c_i l_i 1(m_i < t_{j+1})$ (if $n = 0$, then D is empty and we set $u_j(D) = v_j(D) = 0$). Then, for any $m \in [0,T]$, there exists a $j = j(m)$ such that $u_j(D) \leq f_m(D) \leq v_j(D)$, i.e. $f_m(D) \in [u_j, v_j]$. Consequently, $\mathcal{F} \subset \cup_{j=1}^{J-1}[u_j, v_j]$.

We prove now that for any $\epsilon \in (0,1)$, the sequence $\{t_j\}_{1 \leq j \leq J}$ can be chosen such that for $j = 1, \ldots, J-1$, we have $\|v_j - u_j\| \leq \epsilon$. Write $D_1 = \{(M_1, L_1, C_1), \ldots, (M_N, L_N, C_N)\}$, where N is random and should not to be confused with N in Sections 2.1 and A (if $N = 0$, then D_1 is empty). Let $n \in \{1, \ldots, n_{\max}\}$ and conditioned on $N = n$, let I be uniformly selected from $\{1, \ldots, n\}$. Then

$$\begin{aligned}
\mathrm{E}\left\{\left(v_j(D_1) - u_j(D_1)\right)^2 1(N=n)\right\} &= n^2 \mathrm{E}\left\{1(N=n)\frac{1}{n}\sum_{i=1}^{n} C_i L_i 1\left(M_i \in \left(t_j, t_{j+1}\right)\right)\right\}^2 \\
&\leq T^2 n_{\max}^4 \mathrm{E}\left\{1(N=n)\, 1\left(M_I \in \left(t_j, t_{j+1}\right)\right)\right\}^2 \\
&\leq T^2 n_{\max}^4 \mathrm{P}\left(M_I \in \left(t_j, t_{j+1}\right) | N = n\right),
\end{aligned}$$

as $n \leq n_{\max}$, $C_i \leq n_{\max}$, and $L_i \leq T$. Further,

$$\mathrm{E}\left\{\left(v_j(D_1) - u_j(D_1)\right)^2 1(N=0)\right\} = 0.$$

Hence

$$\mathrm{E}\left\{v_j(D_1) - u_j(D_1)\right\}^2 = \sum_{n=0}^{n_{\max}} \mathrm{E}\left\{\left(v_j(D_1) - u_j(D_1)\right)^2 1(N=n)\right\}$$
$$\leq T^2 n_{\max}^5 \max_{n=1,\ldots,n_{\max}} \mathrm{P}\left(M_I \in \left(t_j, t_{j+1}\right) | N = n\right). \tag{13}$$

Moreover, by Lemma F.3 below, there exists a finite sequence $\{t_{n,j}\}_{1\leq j\leq J_n}$ such that $\mathrm{P}(M_I \in (t_{n,j}, t_{n,j+1})|N=n) \leq \epsilon^2/\left(T^2 n_{\max}^5\right)$ and $J_n \leq 2 + T^2 n_{\max}^5/\epsilon^2$. Thus, by choosing

$$\{t_j\}_{1\leq j\leq J} = \bigcup_{n=1,\ldots,n_{\max}} \{t_{n,j}\}_{1\leq j\leq J_n},$$

we have $J \leq 2n_{\max} + T^2 n_{\max}^6/\epsilon^2$ and

$$\max_{n=1,\ldots,n_{\max}} \mathrm{P}\left(M_I \in \left(t_j, t_{j+1}\right) | N = n\right) \leq \frac{\epsilon^2}{T^2 n_{\max}^5}.$$

Hence by (13), $\|v_j - u_j\| \leq \epsilon$, and so by definition, $N_{[]}\left(\epsilon, \mathcal{F}, \|\cdot\|\right) \leq 2n_{\max} + T^2 n_{\max}^6/\epsilon^2$. Therefore

$$\int_0^1 \sqrt{\log\left(N_{[]}\left(\epsilon, \mathcal{F}, \|\cdot\|\right)\right)}\, \mathrm{d}\epsilon \leq \int_0^1 \sqrt{\log\left(2n_{\max} + T^2 n_{\max}^6/\epsilon^2\right)}\, \mathrm{d}\epsilon < \infty.$$

This completes the proof. □

Proof of Theorem 4.1. By the Donsker property established in the proof of Lemma F.2 and Theorem 2.4 in Gine and Zinn (1990), $\sqrt{r}(\overline{A_r} - \overline{A_r^*})$ and $\sqrt{r}\left(\overline{A}_r - \mu\right)$ converge in distribution to the same process as $r \to \infty$, so the quantile of $\sup_{m\in[0,T]} \sqrt{r}\left|\overline{A_r}(m) - \overline{A_r^*}(m)\right|$ converges to the quantile of $\sup_{m\in[0,T]} \sqrt{r}\left|\overline{A_r}(m) - \mu(m)\right|$. Therefore, $\hat{q}_\alpha^B$ provides the bounds for the asymptotic $100(1-\alpha)$%-confidence region stated in Theorem 4.1.

Lemma F.3. *Let X be a positive random variable. For any $\epsilon \in (0,1)$, there exists a finite sequence $-\infty = t_1 < \ldots < t_J = \infty$ such that $J \leq 2 + 1/\epsilon$ and for $j = 1, \ldots, J-1$,*

$$\mathrm{P}\left(X \in (t_j, t_{j+1})\right) \leq \epsilon.$$

Proof. Denote by F the cumulative distribution function of X, by $F(t-)$ the left-sided limit of F at $t \in \mathbb{R}$, and by F^{-1} the generalised inverse of F, i.e. $F^{-1}(y) = \inf\{x \in$

$\mathbb{R} : F(x) \geq y\}$ for $y \in \mathbb{R}$. We verify the lemma with $J = 2 + \lfloor 1/\epsilon \rfloor$, $t_J = \infty$, and $t_j = F^{-1}((j-1)\epsilon)$ for $j = 1, \ldots, J-1$. Then, for $j = 1, \ldots, J-2$,

$$\mathrm{P}\left(X \in \left(t_j, t_{j+1}\right)\right) = F\left(F^{-1}(j\epsilon)-\right) - F\left(F^{-1}\left(\left(j-1\right)\epsilon\right)\right) \leq j\epsilon - \left(j-1\right)\epsilon = \epsilon.$$

Finally,

$$\mathrm{P}\left(X \in \left(t_{J-1}, t_J\right)\right) = \mathrm{P}(X > F^{-1}((J-2)\epsilon)) = 1 - F\left(F^{-1}((J-2)\epsilon)\right) \leq 1 - \lfloor 1/\epsilon \rfloor \epsilon < \epsilon.$$

□

References

Bottou, L. & Bengio, Y. (1995). Convergence properties of the k-means algorithms. In: *Advances in Neural Information Processing Systems*, volume 7, MIT Press, 585–592.

Chazal, F., Cohen-Steiner, D., Guibas, L. J., Mémoli, F. & Oudot, S. Y. (2009). Gromov-Hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum* **28**, 1393–1403.

Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., Singh, A. & Wasserman, L. (2013). On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems* **20**, 111–120.

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A. & Wasserman, L. (2014). Robust topological inference: Distance to a measure and kernel distance. Available on arXiv:1412.7197.

Chen, Y.-C., Wang, D., Rinaldo, A. & Wasserman, L. (2015). Statistical analysis of persistence intensity functions. Available on arXiv: 1510.02502.

Fasy, B., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S. & Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics* **42**, 2301–2339.

Gine, E. & Zinn, J. (1990). Bootstrapping general empirical measures. *The Annals of Probability* **18**, 851–869.

Hartigan, J. A. & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **28**, 100–108.

Jacques, J. & Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification* **8**, 231–255.

López-Pintado, S., Romo, J. & Torrente, A. (2010). Robust depth-based tools for the analysis of gene expression data. *Biostatistics* **11**, 254–264.

Robins, V. & 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.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge.