TDA for lookout

Rob J Hyndman

Topological data analysis (TDA) uses tools from topology to study data. Using TDA, we can infer high-dimensional structure from low-dimensional representations of data such as individual points. For example, one concept from topology is "persistent homology": a method for computing topological features of a space at different spatial resolutions. Features that persist for a wider range of spatial resolutions represent important, intrinsic features of the data, while features that sporadically change are more likely due to random noise.

Simplicial complexes

Suppose we have a set of bivariate observations. These observations can be used to construct a graph where the individual points are considered vertices and the edges are determined by the distance between the points. Given a proximity parameter ε, two vertices are connected by an edge if the distance between these two points is less than or equal to ε. Starting from this graph, a simplicial complex — a space built from simple pieces — is constructed. A simplicial complex is a finite set of k-simplices, where k denotes the dimension; for example, a point is a 0-simplex, an edge a 1-simplex, a triangle a 2-simplex, and a tetrahedron a 3-simplex. Suppose S denotes a simplicial complex that includes a k-simplex. Then all non-empty subsets of the k-simplex are also included in S. For example, if S contains a triangle pqr, then the edges pq, qr and rs, and the vertices p, q and r, are also in S.

The *Vietoris-Rips* complex is one type of k-simplicial complex. Given a set of points and a proximity parameter $\varepsilon > 0$, $k+1$ points within a distance of ε to each other form a k-simplex. For example, consider the five points p, q, r, s and t shown on the left of Figure 1, and suppose we choose $\varepsilon = 0.5$. Then the distance between any two points other than t is less than ε, and the distance between t and any other point is greater than ε. Then we can construct the edges pq, pr, ps, qr, qs and rs. From the edges pq, qr and rp we can construct the triangle pqr; from pq, qs and sp the triangle pqs; and so on, because the distance between any two points p, q, r and s is bounded by ε. By constructing the four triangles pqr, qrs, rsp and spq we can construct the tetrahedron $pqrs$. The vertex t is not connected to this 3-simplex because the distance between t and the other vertices is greater than ε. The simplicial complex resulting from these five points consists of the tetrahedron $pqrs$ and all the subset k-simplices and the vertex t.

A second example is shown on the right of Figure 1, where there are eight points, and $\varepsilon = 1.3$. Here, f is a vertex, disconnected from all other points because it is further than ε from any point. The pair g and b are connected to each other, but not to any other points. The points ade and adc form connected triangles (but not a tetrahedron), while h is connected to them via c.

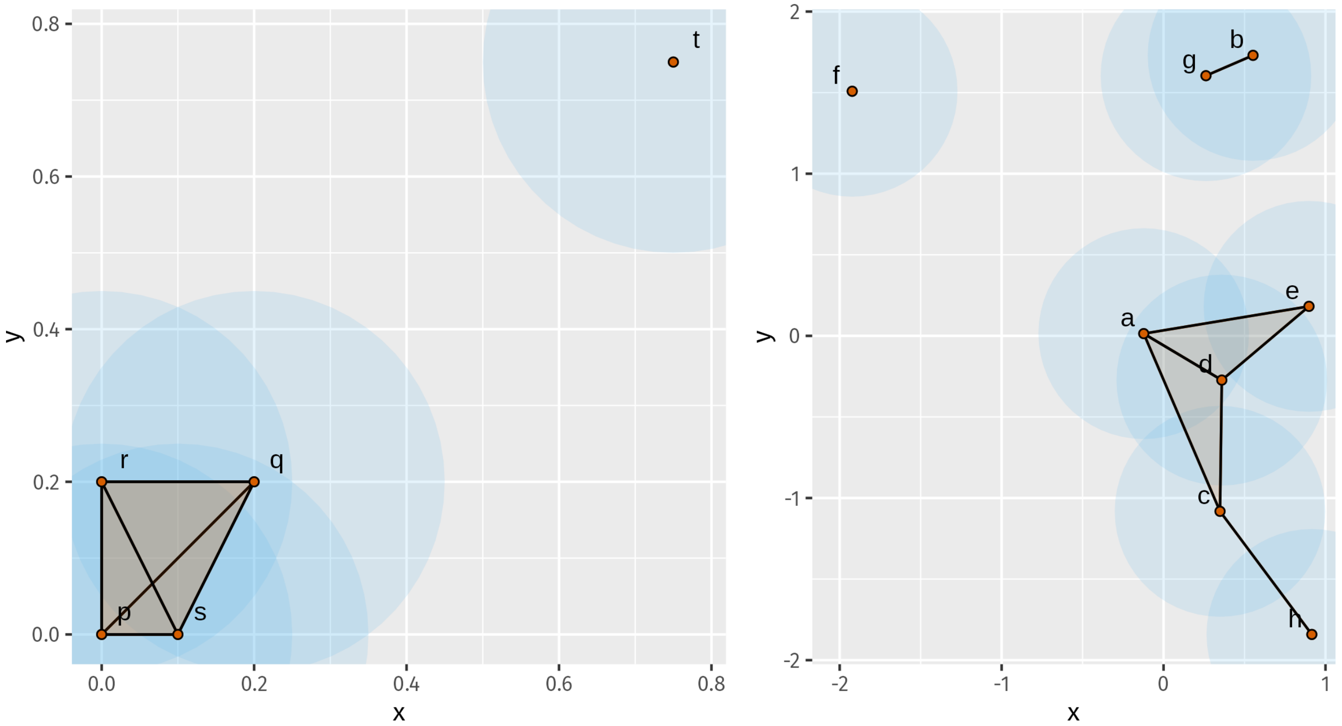

Figure 1: Two examples of Vietoris-Rips complexes. Left: points p, q, r, s and t, with a proximity parameter $\varepsilon = 0.5$. The resulting complex consists of the tetrahedron $pqrs$, triangles pqr, qrs, prs, pqs, edges pq, qr, rs, sp, qs, pr, and vertices p, q, r, s and t. Right: eight points with $\varepsilon = 1.5$. The resulting complex consists of the triangles ade, acd, edges ad, ae, de, ac, cd, ch, bg, and vertices $a, \dots, h$.

Persistent homologies

Given a point cloud of data, the resulting Vietoris-Rips complex depends on the value of the proximity parameter ε. As we increase ε, topological features such as connected components and holes appear and disappear.

Taking the small example on the right of Figure 1, we explore what happens as ε increases from 0.5 to 3.5. On the left-hand side, with a small value of ε, all points are disconnected, and the Vietoris-Rips complex consists of 8 vertices. As ε increases, more points are connected to each other, and eventually, the complex will consist of a single connected component containing all possible connections up to the 8-simplex.

To take a larger, more interesting, example, in Figure 3, we start with a point cloud of 50 points sampled uniformly from an annulus. As ε increases from 0.005 to 1.4, the number of connected components decreases from 50 to 1. At $\varepsilon = 0.005$, each point is disconnected from all others, and the Vietoris-Rips complex consists of 50 vertices. As ε increases, the points start to connect to each other, and the number of connected components decreases. By $\varepsilon = 0.7$, the connected components have merged, and the complex consists of a single connected component in the shape of the annulus. As ε increases further, the hole disappears, and the complex is now in the shape of a ball.

The appearances and disappearances of these topological features are referred to as births and deaths, and can be illustrated using a *barcode* or a *persistence diagram*.

Figure 4 shows the barcode and the persistence diagram of the point cloud shown in Figure 3. The barcode comprises a set of horizontal line segments, each denoting a feature that starts at its birth diameter and ends at its death diameter. These line segments are grouped by their dimension. The orange lines in Figure 4 denote the connected components, and the blue lines denote holes. The same information is

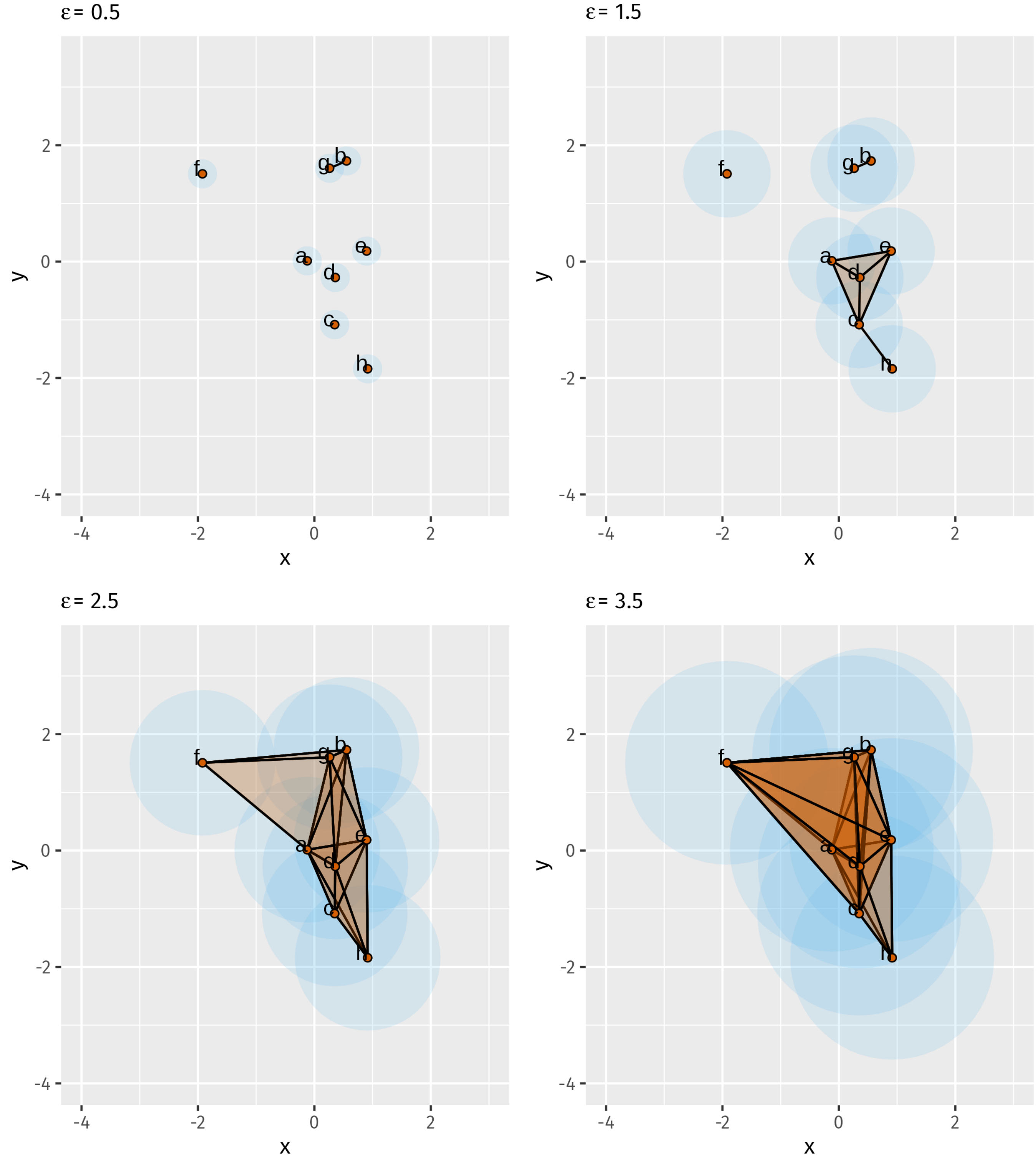

Figure 2: Vietoris-Rips complexes resulting from different ε values.

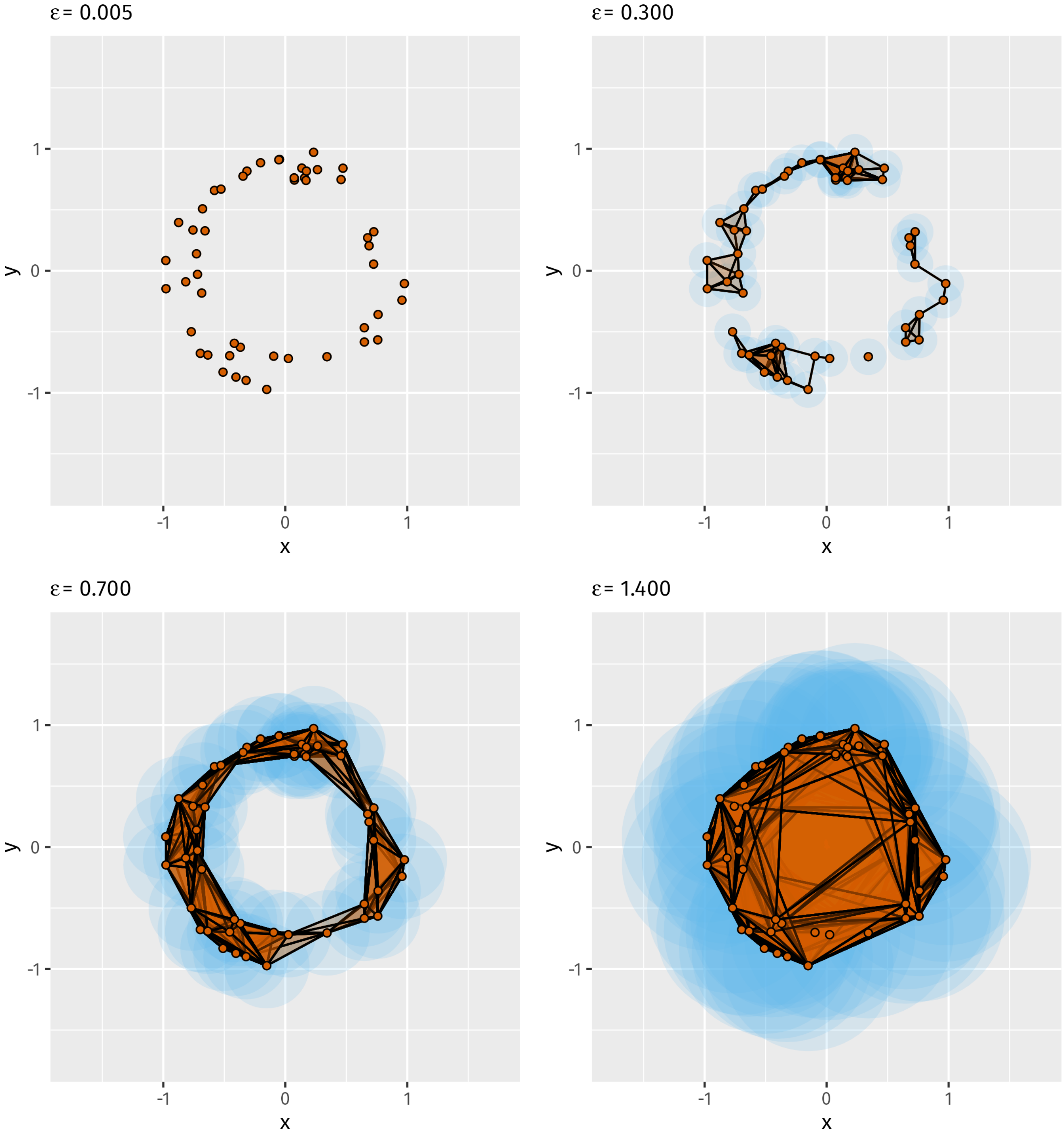

Figure 3: Vietoris-Rips complexes resulting from different ε values.

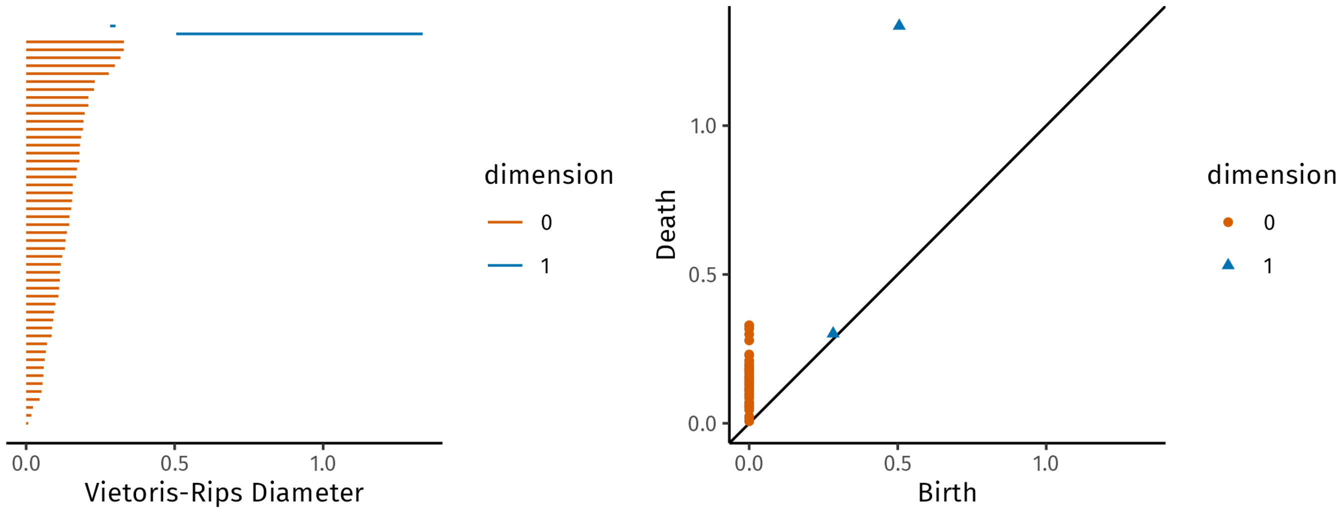

Figure 4: Left: the barcode of the point cloud in Figure 3. Each line denotes a feature, and spans from its birth diameter to its death diameter. The length of the line indicates the persistence of the feature. The 0-dimensional features (in orange) are connected components. The long blue line denotes the 1-dimensional hole that is born at $\varepsilon = 0.51$ and disappears at $\varepsilon = 1.33$. Right: the corresponding persistence diagram. Each point denotes a feature, and its coordinates indicate its birth and death diameters. Points far from the diagaonal are the most persistent.

shown in the persistence diagram, where each point corresponds to one line, with the coordinates of the point equal to the birth and death diameters of the feature.

If there are n observations, then there are n connected components (or 0-dimensional features), each born at diameter 0, and which die when the corresponding observation is connected to one or more other observations. The first death occurs when the closest two points merge, so the corresponding two features have the same birth and the same death diameters. Rather than have a repeated bar, only $n - 1$ bars are shown. As ε increases, the connected components disappear one by one, as they merge with neighbouring features, and eventually the number of connected components decreases to 1 when all observations are connected.

The 1-dimensional features are born when a hole appears in the point cloud, and die when the hole disappears. The long blue line in Figure 4 is born at 0.51 and dies at 1.33, and corresponds to the hole at the centre of the point cloud in Figure 3.

Features that continue for a large range of ε represent structural properties of the data that are of interest to us. These points lie well above the diagonal in the persistence diagram, while points closer to the diagonal are probably perturbations related to noise. In this plot, the triangle near the top represents the same feature as the long blue line in the left plot.

Kernel bandwidth selection using TDA

These topological concepts can be used to determine a bandwidth for a kernel density estimate designed for anomaly detection. First we construct the barcode of the data cloud for dimension zero using Vietoris-Rips complexes with increasing diameter ε. From the barcode we obtain the sequence of death diameters $\{d_i\}_{i=1}^{n}$ for the connected components.

Consider the example shown in Figure 5, comprising $n = 1000$ observations where most points lie on an annulus, with a few points near the centre. The left panel shows a scatterplot of the data, while the barcodes for the connected components are shown in the centre. The right panel displays only the first 20 barcodes, with the dashed line drawn at the second largest death diameter.

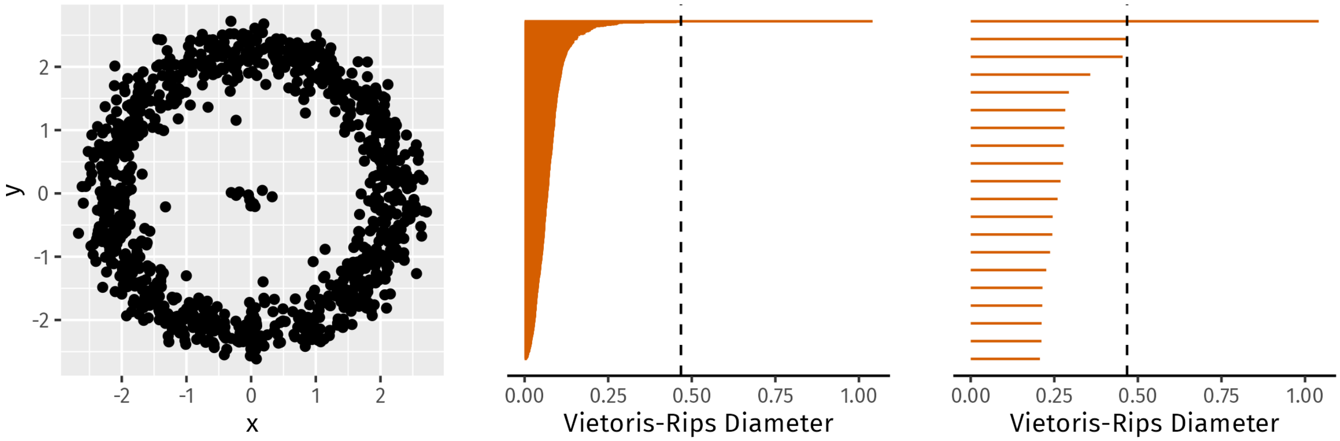

Figure 5: Left: A scatterplot of 1000 observations with most points falling on an annulus and some points near the centre. The other panels show the barcodes for the connected components, with the dashed line drawn at the second largest death diameter. The right panel is a zoomed-in version of the top few barcodes from the centre panel.

The plot on the right in Figure 5 shows the largest 20 death diameters (out of the 999 diameters shown in the centre panel). A vertical dashed line is drawn at diameter $h^* = 0.468$, the second largest death diameter. The largest death diameter is 1.041. Any diameter between these two values gives the same number of connected components. For this data set, $(0.468, 1.041)$ is the largest diameter range for which the number of components stays the same. Thus, it signifies a global structural property of the point cloud, and we want a bandwidth that will help us detect this structure. In this example, an appropriate choice would be $\boldsymbol{H} = h_*^2\boldsymbol{I}$, which ensures that points within a distance of h_* contribute to the kernel density estimate.

From $\boldsymbol{Y}^*$, we can compute the Vietoris-Rips death diameters, $d_1, \dots, d_n$. These are then ordered to give $d_{(1)}, \dots, d_{(n)}$, and we find the largest interval by computing successive differences $\Delta_i = d_{(i+1)} - d_{(i)}$, for $i = 1, \dots, n-1$. Following @lookout2021, we choose $h_* = d_{(i)}$ to be the diameter corresponding to the largest Δ_i.