4 Topological data analysis: Building SCs from data

The 'natural' topic for this section would be to set up a few useful definitions, but I'm going to put this bit off until we've got something to apply them to. The next idea I'm interested in is applying simplicial complexes for data, of the type you might be interested in finding structure in for research purposes. This is the first bit of *topological data analysis* (TDA), what I was interested in. Let's start with: what is data? We might just say: anything we want to study because it might be significant, and that we can describe in Cartesian coordinates.

Work-In-Progress Definition Data refers to any subset of $\mathbb{R}^n$ we are interested in.

There's a few things wrong with this. When we're collecting data, it's uncommon to actually measure an infinite set of points, since infinity is such a big number. Further, we don't need them to be in $\mathbb{R}^n$. It might feel a bit restrictive to only *build* simplicial complexes out of points in $\mathbb{R}^n$ when we went to all this trouble to embed them in $\mathbb{R}^n$. (Besides, there's easier things to do if you're already in $\mathbb{R}^n$, like principal component analysis.) We built that machinery to deal with any kinds of points, let's use it!

Work-In-Progress Definition Data is a finite set.

Now we've got the opposite problem; this definition is too permissive. We don't necessarily need the specific structure of $\mathbb{R}^n$ to analyze data, but we do need *some* structure that determines how these things relate to one another, otherwise I could say the set $A = \{3, \text{orange}, \text{bicycle}, \square\}$ are data. Maybe they *could* be, but I don't know how any of these things relates to the others, so there's not a lot I can do.

A more general way to put a set of conditions that are similar enough to what we're used to in $\mathbb{R}^n$, but specific enough that we have some sense of structure, is the *metric space*. Call the set containing our data A, and make some choice for a larger space it lives in, M such that $A \subset M$. (It doesn't really matter exactly what M is, you can make a 'natural' choice based on what the data are.) Let's say we define a function taking in any two points in our data or in the ambient space, $d : M \times M \to \mathbb{R}$, giving us the distance $d(x, y)$ between points $x, y \in M$. A metric space is M together with d (we're talking about these things going 'together' in the same way that we defined a simplicial complex 'on' a vertex set; one needs the other to do anything, the other only exists because the one's making it make sense.)

d can't be just any function, though; there's four conditions we want it to follow. Here, I'm going to violate my rule of 'make everything intuitive' a bit, because covering why metric spaces are defined the way they are would be a long digression and would detract from our main point here. Even if I tried, I think it'd take actually doing some analysis exercises (the kind where you prove certain types of functions are continuous) to make you fully believe me, so we'll take the following definition more or less on faith. (We have to pick *some* point at which we assume an existing result, or we'd go the way of Russell and Whitehead.) I'll simply say that these are nice things to know about a space, and if there's an issue with any of them, the thing we're considering is probably conceptually distant (pardon the pun) from how we intuit distance in Euclidean space, so it wouldn't be all that useful to consider anyway.

Definition A metric space (M, d) is a set M together with a function ("the metric") $d : M \times M \to \mathbb{R}$, satisfying

1. **identity**: $d(x,x) = 0$ for each x in M,
2. **positivity**: $d(x,y) > 0$ for each $x, y \in M$ that are not the same,
3. **symmetry**: $d(x,y) = d(y,x)$ for each x, y in M; and
4. **triangle inequality**: $d(x,y) + d(y,z) \geq d(x,z)$ for all x, y, z in M.

This next bit is honestly skippable; read it if you're unconvinced by why a metric space should be defined in this way.

The first two conditions say distances can't go negative, and if a point is distance zero from another, those two points have to be the same; these are just nice things to know, and can be useful when we're making arguments about getting closer and closer to a point. A common logical chain in analysis goes something like: "only one point p has this property, because suppose not, then there's a point q that has it, but another argument shows $d(p,q) < \epsilon$, so $p = q$ and so p is unique." Without those first two points, we wouldn't be certain about this! Symmetry just agrees with our notion of what makes something feel 'distance-y', but this is sometimes not fulfilled, e.g. for the Kullback-Leibler divergence/relative entropy in statistics. The triangle inequality is similarly something it's nice to have by analogy with Euclidean space, but not something I know how to *a priori* justify the same way I'm trying to do with everything else. A lot of arguments in analysis turn out to rely on the triangle inequality. There's a few templates for this: one of them goes something like "a property holds at two reference points, we can relate this to a third new point, and we know how to upper-bound the distance to the third point because we can relate it to distances we already know with the triangle inequality".

I don't really like talking about math using the phrase '"it turns out", because it's just begging me to dig into *why* it turns out that way. But ultimately, these conditions aren't handed down from on high so much as being what turned out to be convenient after a lot of trying stuff out, and if you have a function that doesn't meet one or more of them, that makes it not a metric, but it doesn't necessarily make it not useful.

Anyway, this is a very long-winded way of saying: we can write down a working definition of what it means to be a space that isn't necessarily $\mathbb{R}^n$, but has all the modern luxuries we've grown accustomed to from there.

Definition Data is a finite set in a metric space.

Alright, so we could imagine taking data on any metric space, looking at its connections, and building a simplicial complex to represent these connections in $\mathbb{R}^n$. To me, the immediate weak point in this action plan is "looking at its connections", because the distance function on this particular set contains a wealth of information that we should be looking to mine to make the connections in the SC. It's not really obvious how to do this!

We have the set of pairwise distances $d(x_i, x_j)$ between every pair of points in our data, and we'd like to relate this to connections in a simplicial complex. Just based on the numbers, it's not possible to tell whether things 'should' be connnected. We also need a cutoff distance: if two points are closer than some ϵ, we can say they appear together in the SC, and they don't if they're further apart than that. This doesn't cover links between three or more points at once, but it's a start. Importantly, this suggests we should adopt the idea of looking at different *length scales*. We can fix an ϵ and develop (in a way we're going to iron out next) a simplicial complex from data that considers distances less than ϵ to be important. Then we can vary ϵ and see what larger-scale structure sticks around and what doesn't.

A way you might do this in $\mathbb{R}^n$ is via a *point cloud*. If we've got a set of points in $\mathbb{R}^n$, we can imagine drawing spheres centered at each point, and looking at their intersections. Each of these spheres would have the same radius, which we'll say is $\epsilon > 0$ (or if we wanted to be consistent with the last paragraph, $\epsilon/2$, but we're not going to be that strict here), and we can build networks of links in the simplicial complex based on which clouds overlap. As we increase ϵ, more and more connections are built as the point clouds grow larger and overlap with one another more and more.

This is a nice idea, but it's kind of hard to work with computationally. Imagine trying to find a path between two points within a point cloud; you might have to trace out a bunch of different potential directions and have almost all of them not go anywhere. This gets worse and worse in higher dimensions and isn't that convenient to begin with, so let's try and define something nicer. Earlier, we touched on the idea of cutoff distances for saying two points are connected in an SC. If we extend that to collections of n points by saying it has to hold for each pair inside the collection, we can define the *Vietoris-Rips filtration*:

Definition Let (M, d) be a finite metric space. The Vietoris-Rips filtration of M is an increasing family of simplicial complexes $\mathbf{VR}_\epsilon(M)$ indexed by the real numbers $\epsilon \geq 0$, defined as follows: a subset $\{x_0, x_1, \ldots, x_k\} \subset M$ forms a k-dimensional simplex in $\mathbf{VR}_\epsilon(M)$ if and only if the pairwise distances satisfy $d(x_i, x_j) \leq \epsilon$ for all i, j.

This is a *filtration*, which is a thing I didn't define before but is easy enough to qualitatively describe: it's a sequence of simplicial complexes on the same vertices in a specific order, so that as we count up we only ever add simplexes. That is, each one is a subset of the next, until we've built up the whole simplicial complex we're interested in. (We also say a family of simplicial complexes is 'increasing' if it has this property.)

Note that the Vietoris-Rips filtration isn't a finite collection; there's as many $\mathbf{VR}_\epsilon(M)$s as there are real numbers $\epsilon > 0$. However, since M is finite, there's some ϵ at which nothing is connected (anything less than the smallest pairwise distance), some ϵ at which everything is connected (anything greater than the largest pairwise distance), and only a finite number of steps in between. So if we wanted, we could describe a Vietoris-Rips filtration in a finite number of steps as a sort of piecewise function. As we would bopefully expect, a larger ϵ means more connections, so $\epsilon_1 < \epsilon_2$ means $\mathbf{VR}_{\epsilon_1}(M) \subseteq \mathbf{VR}_{\epsilon_2}(M)$.

I'll finish for today by describing the other type of filtration; the Čech filtration. This is roughly what you get if you try to express the same idea of everything being close enough together, but with reference to a common central point rather than each pairwise distance. This is conceptually similar to the point cloud idea from before, but that means it's also harder to do computations with.

For those familiar with analysis, it's the equivalent of being convergent where Vietoris-Rips is being Cauchy; if you're not, the common thread is "can we put these points together in a sphere of radius ϵ?", where in Vietoris-Rips one of the points has to be the center, and in Čech you're free to choose the center from anywhere that might contain the data in its sphere. This means we need some information about the ambient space, whereas with V-R we just needed to know about the data. The tradeoff is: we use the extra information about the space to make a simpler-looking definition.

Definition Let M be a finite subset of a metric space (Z, d). The Čech filtration of M with respect to Z is the increasing family of simplicial complexes C_ϵ indexed by $\epsilon \geq 0$ defined as follows: a subset $\{x_0, x_1, \ldots, x_k\} \subset M$ forms a k-dimensional simplex in $C_\epsilon(M)$ if and only if $d(z, x_i) \geq \epsilon$ for all i and for some $z \in Z$.

We have to check k points now instead of k^2, but that's assuming we've first been able to sweep through all the points it's possible for z to be. I don't yet know how this works computationally, but I might find out in a future note!