EKG Feature Extraction using Persistent Homology

The goal of topological data analysis (TDA) is to study shapes formed from subsets of a dataset which are invariant under continuous deformations such as twisting and stretching. One of the main strengths of TDA is that it generalizes nicely for data of an arbitrary number of dimensions, and in this case, our data is two-dimensional (voltage vs. time). Consequently, the two types of topological parameters associated with our data are the 0-dimensional and 1-dimensional homology groups (H0 and H1 features). The H0 and H1 features are equivalence classes of connected components and noncontractible loops, respectively, and we’ll focus strictly on H1 features which I’ll elaborate on over the next few pages.

First consider the example EKG signal in Figure 1 below that I constructed using a gaussian P-wave, six line segments for the QRS-complex, and a gaussian T-wave. Note that I included a baseline that connects the isoelectric segments of the signal. The reason for including this baseline will become evident soon.

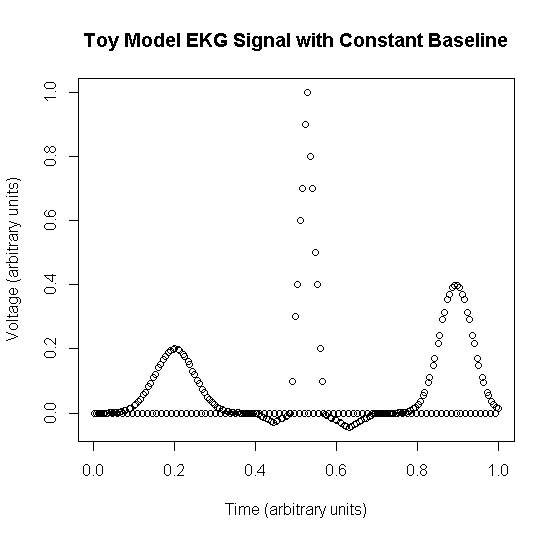


Figure 1: Example EKG signal with normalized voltage, an arbitrary time axis, and a baseline connecting isoelectric segments.

Now consider drawing a circle around each datum, each with the same radius r=0.025, as shown in Figure 2 below. The subset of the two-dimensional plane formed by the union of these circles is referred to as the radius r geometric Cech complex. In other words, the radius r geometric Cech complex is the region of space spanned by all of the circles of radius r centered at each of the data points. Notice that in Figure 2, the circles centered at points which are a distance less than r apart from one another overlap while the circles centered at points greater than r apart from one another are isolated and do not overlap. Furthermore, notice that a noncontractible loop can be drawn within the radius 0.025 Geometric Cech Complex, specifically around the P-wave and T-wave, as shown in Figure 3. We term these loops as noncontractible since the voids in the P-wave and T-wave (the white areas of the P-wave and T-wave surrounded by overlapping green circles) make it impossible to deform the loops down to a single point.

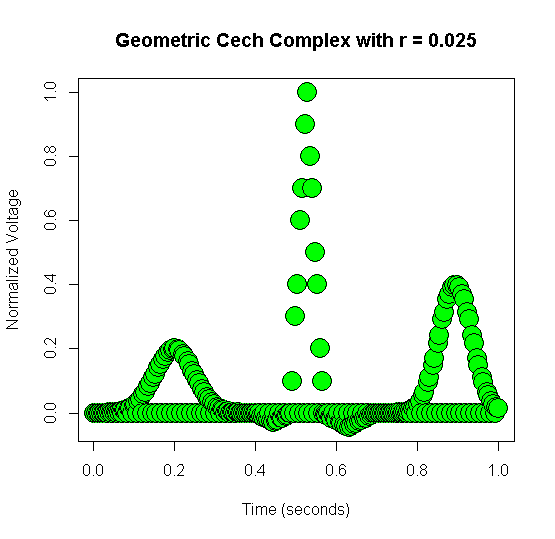


Figure 2: Circles of radius 0.025 centered around each data point are drawn. The region of the dataset shown in green is termed as the “radius 0.025 Geometric Cech Complex”.

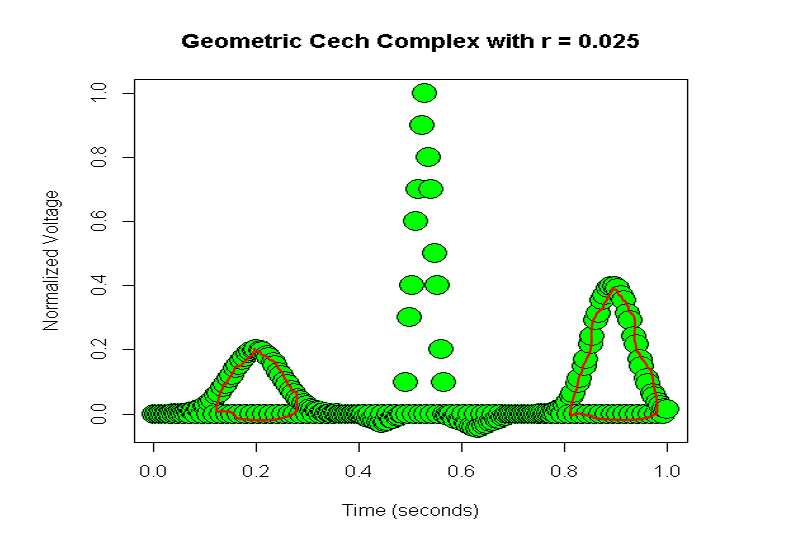


Figure 3: radius 0.025 Geometric Cech Complex with representative loops from the equivalence class of noncontractible loops around the P-wave and T-wave drawn in red.

Continuing to increase the radius as shown in Figure 4 results in the destruction of the voids from the P-wave and T-wave that are evident in Figures 2 and 3, and consequently, the noncontractible loops which could be drawn in Figures 2 and 3 become contractible (i.e., the loops can be deformed into a single point).

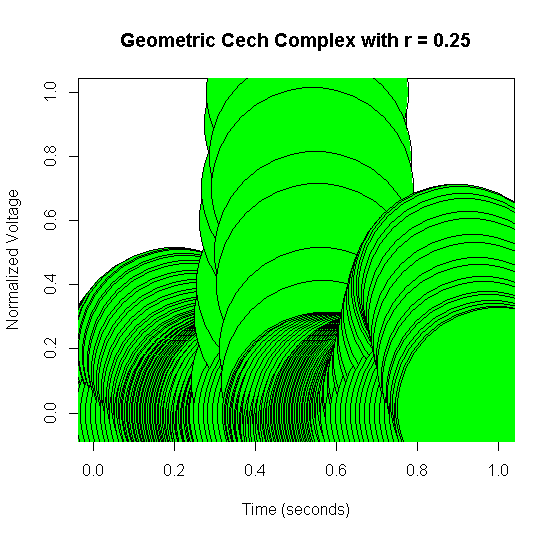


Figure 4: Increasing radius r past the death radius so that the noncontractible loops of the P-wave and T-wave H1 features become contractible.

Recall that an H1 feature is an equivalence class of noncontractible loops. To help conceptualize this, think of the red loop drawn around the P-wave in Figure 3 as being a member of the equivalence class of noncontractible loops (i.e. the H1 feature) with the P-wave as its void. Any way in which that red loop can be stretched, twisted, or deformed without tearing results in another member of the equivalence class of noncontractible loops (i.e. H1 feature). More concretely, an H1 feature is a family of noncontractible loops which can be continuously deformed into one another.

We are now equipped to define the birth radius of an H1 feature to be the smallest radius at which a noncontractible loop can be drawn and the death radius of an H1 feature to be the smallest radius at which the loops of the H1 feature become contractible. As another definition, the persistence of an H1 feature is the difference between the death radius and the birth radius of the H1 feature. Thus, the most persistent H1 features of a dataset correspond to the largest loop-like organizations of data. The persistent homological features are described in Figure 5 below by plotting their birth radius vs. their death radius. In this plot, the black dots represent H0 features (we’ll ignore these), and the red triangles represent H1 features. Note that two H1 features lie substantially further from the y=x line than the other H1 features. Since the persistence of an H1 feature is defined as the difference between its death radius and birth radius, these two H1 features are by far the most persistent H1 features of the EKG signal. Of these two H1 features, the one with the larger birth radius corresponds to the T-wave, and the one with the smaller birth radius corresponds to the P-wave.

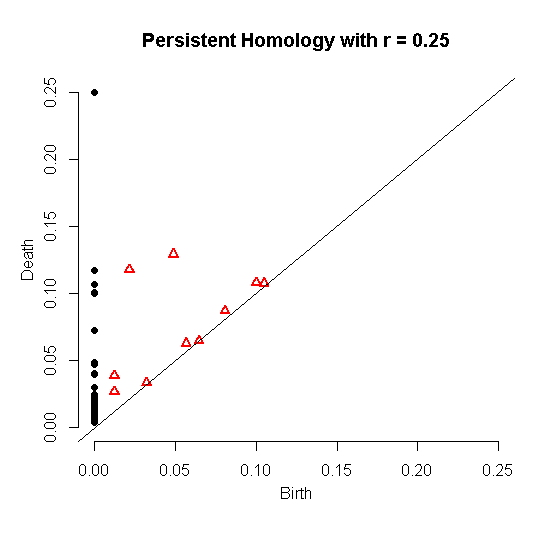


Figure 5: Persistent Homology of Toy Model EKG.

When analyzing EKG data, we will only be concerned with H1 features, but if you’re interested in hearing about H0 features, please ask and I can explain them. I don’t believe H0 features will be useful for us because they primarily describe clustering phenomenon which doesn’t seem too useful when extracting features, measuring various intervals, classifying rhythms, and describing how waveforms change over time. I should also note that in order to decrease the computations required when calculating these features, the computer actually forms the Vietoris-Rips complex, an abstract simplicial complex which is homotopy-equivalent to the geometric Cech complex. This homotopy-equivalence allows us to interpret our homology features as corresponding to the more-intuitive geometric Cech complex.

I will now show a few of the computations done with normal sinus rhythm data and briefly mention some ideas for future work. A processed recording of normal sinus rhythm along with its persistent homology is shown in Figure 6. For the left-most plot, some filtered sinus rhythm EKG data were normalized, and a solid baseline was added to allow for formation of H1 classes corresponding to P,Q,S,&T-waves. Similar to Figure 5, the black dots represent H0 features and the red triangles represent H1 features in the right-most plot. Without actually looking at the persistence of each H1 feature, it’s evident that the twelve H1 features clustered near (0.02,0.08) are by far the most persistent H1 features of this data because they are much further away from the y=x line than the other H1 features, meaning they have a larger difference between death and birth radius than the other H1 features. The code identifies which subset of data forms the representative cycle corresponding to each H1 feature, but even without that, it’s evident that these twelve H1 features correspond to the twelve T-waves of the data. This is the case because the birth radius of the geometric Cech complex does not need to be very big to create an equivalence class of noncontractible loops (H1 feature) due to the shape of the T-waves. The radius of the geometric Cech complex would also need to be relatively large to contract these loops down to a single point, thus causing them to have a relatively large death radius. With similar reasoning, we can deduce that the twelve H1 features clustered near (0.02,0.06) correspond to the twelve P-waves. Note that the code does identify which H1 feature corresponds to which subset of data, but I just wanted to point out the relation between the P-waves and T-waves of the data and their H1 features of the persistent homology output to help build intuition for how the persistent homology of the data (particularly the H1 features) relate to the shape of the data.

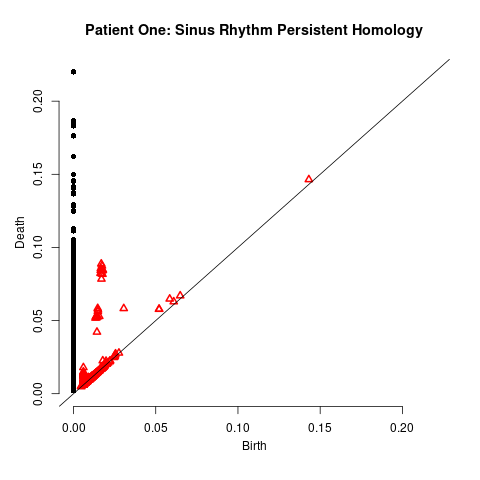
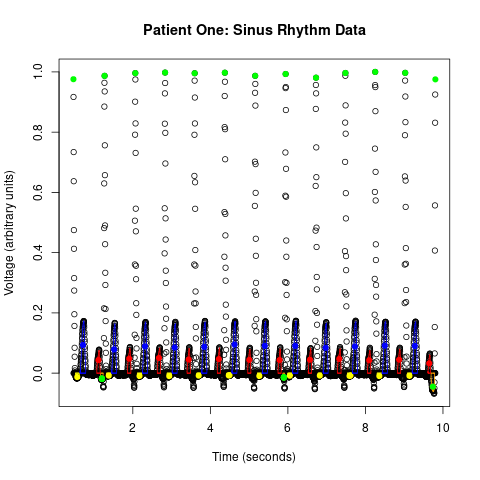


Figure 6: Persistent Homology and Center of Mass of Representative Cycles for Normal Sinus Rhythm. Representative cycles drawn in red, green, yellow, and blue are P,Q,S,&T-waves, respectively.

By conditioning on whether H1 features have a persistence within some range and utilizing the coordinates of the centers of mass of their representative cycles, the code identifies P-waves, Q-waves, S-waves, & T-waves. I haven’t actually quantified how accurately the code identifies these waves yet, but it has at least 90% accuracy for the P-waves and T-waves in the sinus rhythm, atrial fibrillation, and 1st degree AV block data that I’ve tested it on so far. A novel measure of the PR-interval is computed by taking the difference of the time coordinate of the left-most datum of the representative cycle of a detected Q-wave and the time coordinate of the left-most datum of the representative cycle of the preceding P-wave. If no Q-wave is detected before an R-wave, then the right-hand endpoint of the PR-interval is the time coordinate of the first datum below the isoelectric baseline to the left of the R-wave. Similarly, the QRS duration is computed by taking the difference of the time coordinate of the right-most datum of the representative cycle of a detected S-wave and the right-hand endpoint of the preceding PR-interval. If no S-wave is detected, then the right-hand endpoint of the QRS duration measurement is the time coordinate of the left-most data point below the isoelectric line to the right of the preceding R-wave. The QT-interval, ST-interval, P-wave duration, and T-wave duration are measured similarly. Lastly, I would like to point out that these methods could also be used to quantify the degree of irregularly of heart rhythms over time by calculating the persistent homology of small subsets of EKG recordings and then measuring how the location and persistence of H1 features corresponding to various components of a given waveform evolve over time. This topic is further explored in the word document “degree\_of\_irregularity\_description”.