Language in Time - 2 - Quantifying Recurrence

So far we have seen how to build recurrence plots (RPs), and briefly described the kinds of patterns that can appear on them. These patterns are the basis of recurrence quantification analysis (RQA). In essence, RQA is a set of tools to put numbers on the RPs so that you can have a quantitative, shorthand description of what is seen. These quantities are useful to describing the system's dynamics. How much regularity does the system exhibit over time? If a system is regular, is it highly repetitive, or does it engage in more strategic short bursts of repetition? These kinds of questions can be posed through the measures of RQA, as we exemplify simply here.

Like the last section, the simplest way of conveying recurrence is to compare different types of language data. In particular, transcripts of spontaneous speech will show fundamentally different dynamics than poetry or lyrics in music. These are quite distinct "linguistic modes" – each behaving through the medium of words – but showcasing quite distinct patterns in time.

Before getting into this demonstration, let's just briefly describe these recurrence measures. Our presentation is not entirely comprehensive – the use of RPs to quantify system dynamics is a veritable interdisciplinary literature unto itself – but the measures that can be implemented in `crqa` are the most common. Each is described below. In parentheses is the most common shortform nomenclature for these measures.

Recurrence Rate (RR). Recurrence rate is simply the overall density of points on the plot. It is usually calculated by taking the number of points and dividing it by the square of the length of the time series you are visualizing: $|RP|/N^2$. This reflects the overall extent of recurrence taking place, irrespective of the nature of the point distribution itself. It has been a fruitful measure in a variety of places, which we'll detail later.

Determinism (DET). Determinism is among the most commonly used measures. It starts to focus on the *distribution* of recurrence points – how are they organized on the plot? You may have multiple plots with the same RR, but fundamentally different organization to these points. DET is a measure of how much points tend to fall on diagonal lines. These diagonal lines reflect periods of time during which the system is precisely revisiting a prior sequence of states. This regularity is referred as "determinism" because a system that has many diagonal lines means that to the extent that the system is recurrent, it is recurrent on the same or some subset of paths. Systems with low DET may show punctate recurrence – transient moments of repeating the same state – but they are not organized as neatly – they are not deterministic.

Average Line Length ($\bar{l}$). This is a measure based on diagonal lines once again, and reflects the number of average points that a diagonal line extends. You may interpret this in terms of the time scale that your series is based on. If it is characters, for example, then $\bar{l}$ reflects the average number of characters ordered, if your time series is based on a 10Hz sampled systems, then $\bar{l}$ reflects the number of tenths of a second that tend to be repeated.

Maximum Line Length (MAX_l). There has been some discussion of maximum line length of diagonals as a proxy measure for the strength of the underlying attractor that governs the system. For example, you may have low DET, but a very high MAX_l, which means that the system may not revisit states in a very regular way, but that it revisits one very long portion of its trajectory in its phase space.

Entropy (ENT). There is some dispute among recurrence folks about entropy. It is a difficult-to-interpret measure, according to some, but it has also been used in some interesting spots (e.g., Dixon et al., 2009). The idea of the entropy measure is that it quantifies the orderliness of the itinerary of the system. It does so by using the distribution of diagonal lines on the plot. If the system is revisiting quite an array of different paths in its reconstructed phase space – many different diagonal line lengths – then ENT will be higher. If the system revisits sequences of states quite regularly, and so the line lengths tend to be of one or a handful of characteristic values, then ENT will be lower. In the former case, we have a more disordered system. In the latter case we have a more orderly system.

Trapping Time (TT).