Seismogram Clustering Suggestions

Lay Kuan Loh
June 5, 2014

First Try

1. Given n seismograms, called $s_1, s_2, \cdots, s_n$, each of which has m datapoints in it, we want to correlate each interval along the seismograms with the rest of them, with intervals of, for instance, 60 seconds each. So, numbering the datapoints in s_i by $\{t_1, t_2, \ldots, t_m\}$, we can choose groups of $\{t^i_1, \cdots, t^i_{60}\}, \{t^i_{61}, \cdots, t^i_{120}\}, \ldots, \{t^i_{60\lfloor \frac{m}{60} \rfloor + 1}, \cdots, t^i_m\}$ in s_i.
2. To do clustering on the seismograms $s_1, s_2, \cdots, s_n$ for an event, it is best to only have one affinity matrix A for the seismograms. You tweak the metric used to compute affinity between s_i and s_j, not compute more affinity matrices.
3. Find the affinity between s_i and s_j by first separating them into groups $\{t^i_1, \cdots, t^i_{60}\}, \{t^i_{61}, \cdots, t^i_{120}\}, \ldots, \{t^i_{60\lfloor \frac{m}{60} \rfloor + 1}, \cdots, t^i_m\}$ in s_i and $\{t^j_1, \cdots, t^j_{60}\}, \{t^j_{61}, \cdots, t^j_{120}\}, \ldots, \{t^j_{60}$ respectively. To find the affinity, try finding the covariance between signals for the groups chosen.
4. Now find the local affinity matrix B between these groups for s_i and s_j. Choose the two portions s^c_i and s^c_j in s_i and s_j giving the highest affinity, and let $A(i, j) = \max(B)$. Remember to save where s^c_i and s^c_j start and end for each s_i and s_j pair.
5. Perform clustering using A. Try k-means first, then Spectral Clustering.

Possible tweaks

- Make the intervals a bit smaller, < 60 seconds. Need to experiment with this a bit.
- Replace (4) above with doing clustering on matrix B. Find the cluster in s_i and s_j that has the highest affinity, and set that to be A_{ij}. So, double clustering.

- In (3), instead of covariance, other possible metrics are correlation. Other metrics are here: `http://docs.scipy.org/doc/scipy/reference/spatial.distance.html`

  Try and transform the time signal data to frequency signal data, and do clustering on that instead. The closer the peaks and their intensity are, the higher the affinity.