Seismogram Clustering Suggestions

Lay Kuan Loh

June 5, 2014

1 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 $\{t_1^i, \cdots, t_{60}^i\}, \{t_{61}^i, \cdots, t_{120}^i\}, \ldots, \{t_{60\lfloor \frac{m}{60} \rfloor + 1}^i, \cdots, t_m^i\}$ 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 $\{t_1^i, \cdots, t_{60}^i\}, \{t_{61}^i, \cdots, t_{120}^i\}, \ldots, \{t_{60\lfloor \frac{m}{60} \rfloor + 1}^i, \cdots, t_m^i\}$ in s_i and $\{t_1^j, \cdots, t_{60}^j\}, \{t_{61}^j, \cdots, t_{120}^j\}, \ldots, \{t_{60\lfloor \frac{m}{60} \rfloor + 1}^j, \cdots, t_m^j\}$ 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_i^c and s_j^c in s_i and s_j giving the highest affinity, and let $A(i,j) = \max(B)$. Remember to save where s_i^c and s_j^c start and end for each s_i and s_j pair.

5. Perform clustering using A. Try k-means first, then Spectral Clustering.

2 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`

- Alternatively, instead of doing all that, for each pair of seismograms s_i and s_j, transform the time signal data to frequency signal data, and do clustering on that instead. Filter and get write of peaks below a certain minimum amplitude. Find the covariance between the transformed seismograms.