Algorithm:

Given:

- S a set of data-points with n points of d_s dimensionality
- ε the minimum distance to be considered density reachable
- minPts the minimum ε-reachable neighbors to make a core point
- t, a threshold cluster size to divide noise from correct data
- d_t a target dimensionality for S

Do:

1. Train and deploy an autoencoder for dimensionality reduction to generate set S' from S with n points of dimensionality d_t. Ignore this step is $d_s = d_t$.

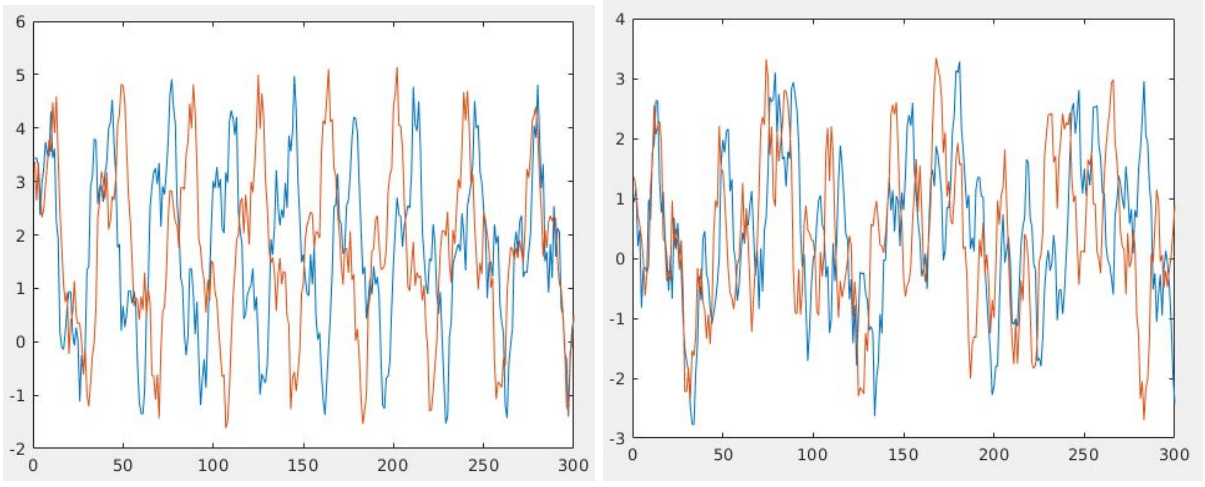
2. For each s in S' append the label l as a one-hot encoded feature which is observed for s to create S'' with n points of dimensionality d_t+L where L is the number of labels.

3. Apply DBScan(ε, minPts) to S'' to generate C, a set of clusters observed in S''.

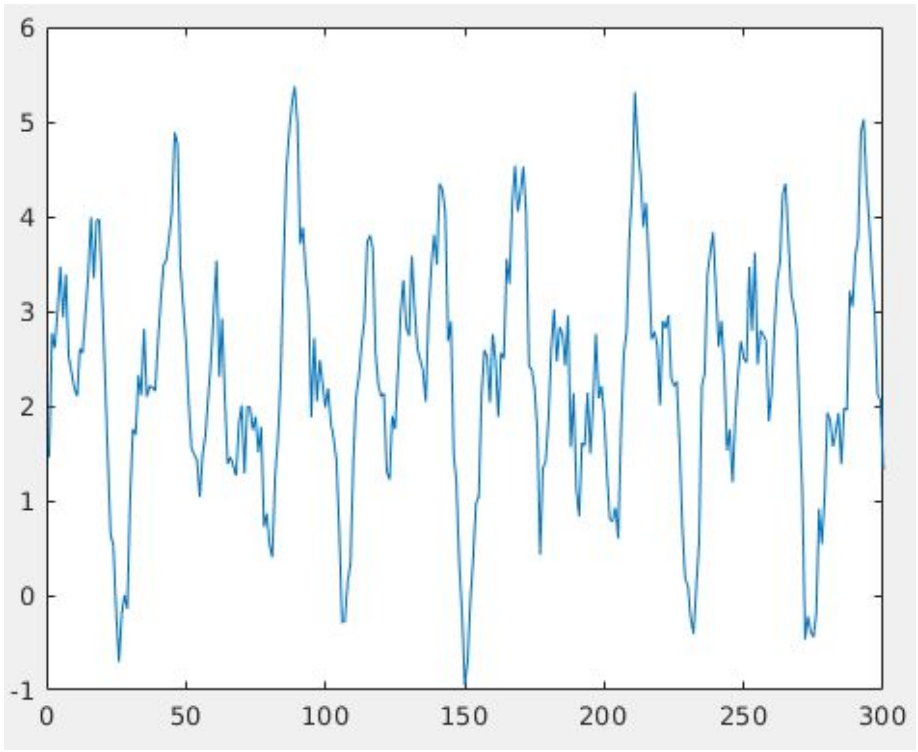
4. Add all points in C_o to M, a set of mislabeled example indexes from S.

5. For all remaining c in C, add all indexes in c to M if |c| < t.

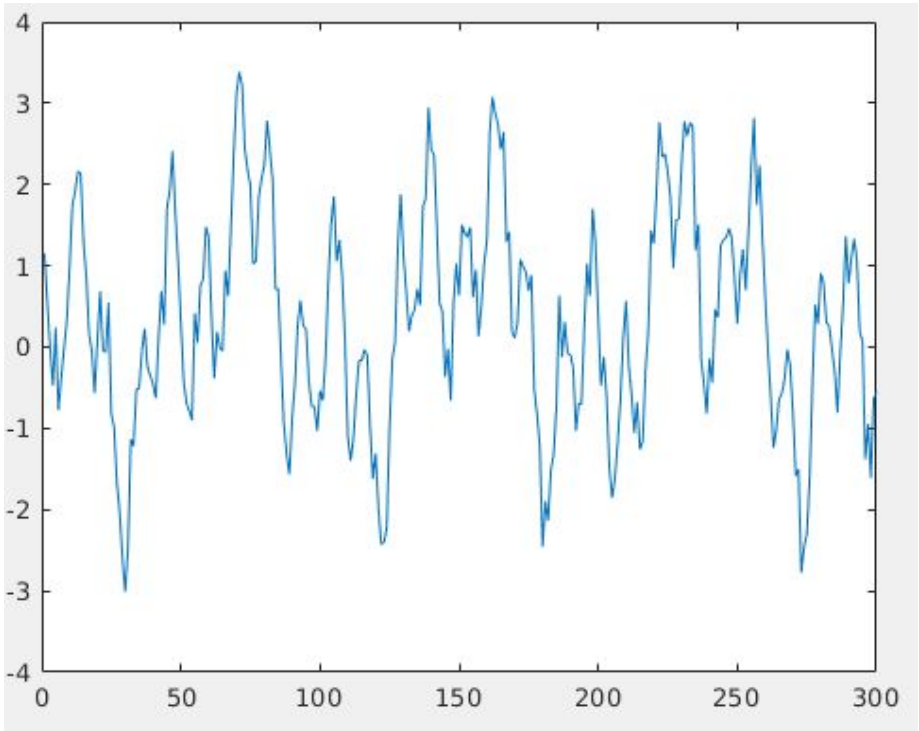
6. Return M.



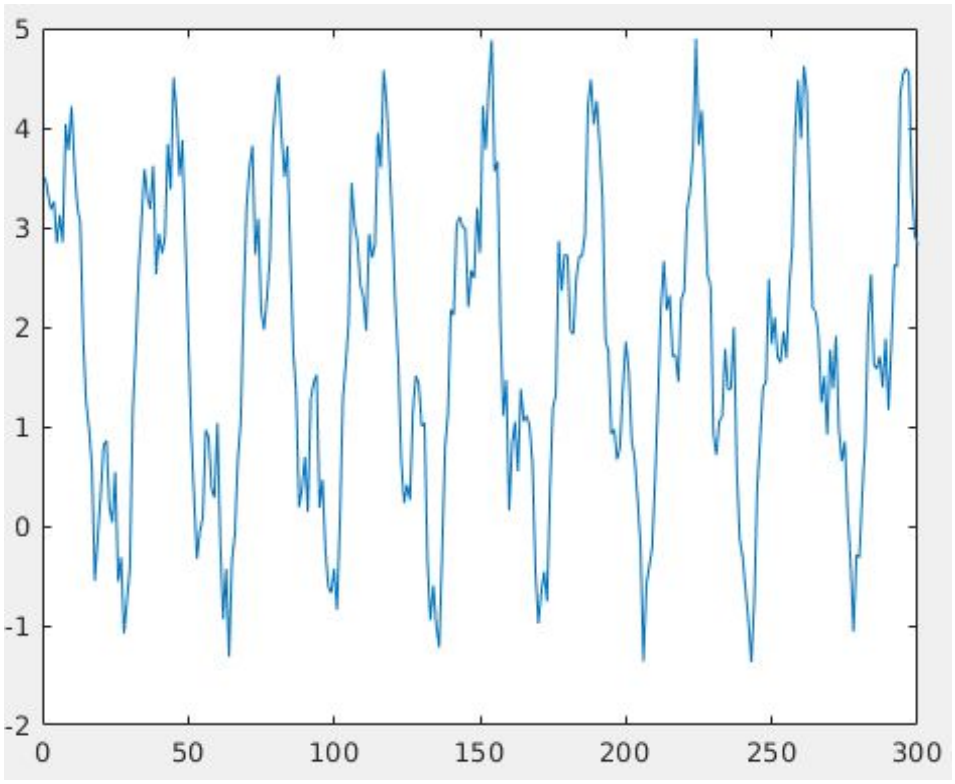
4 synthetic samples with same-labeled examples overlaid on each other.



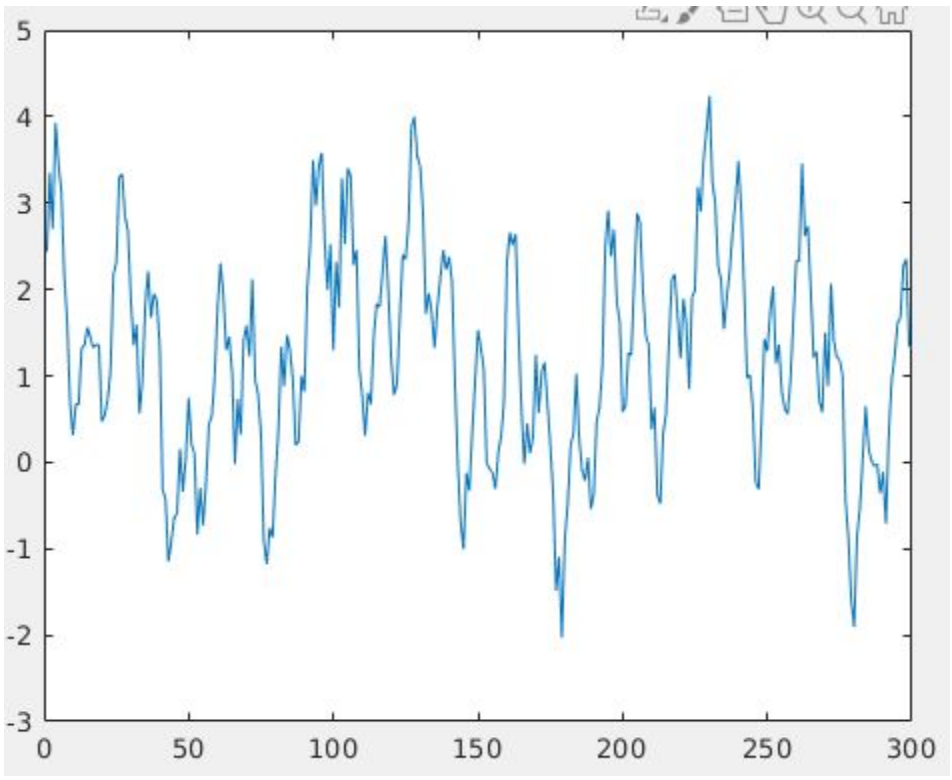
Label 0



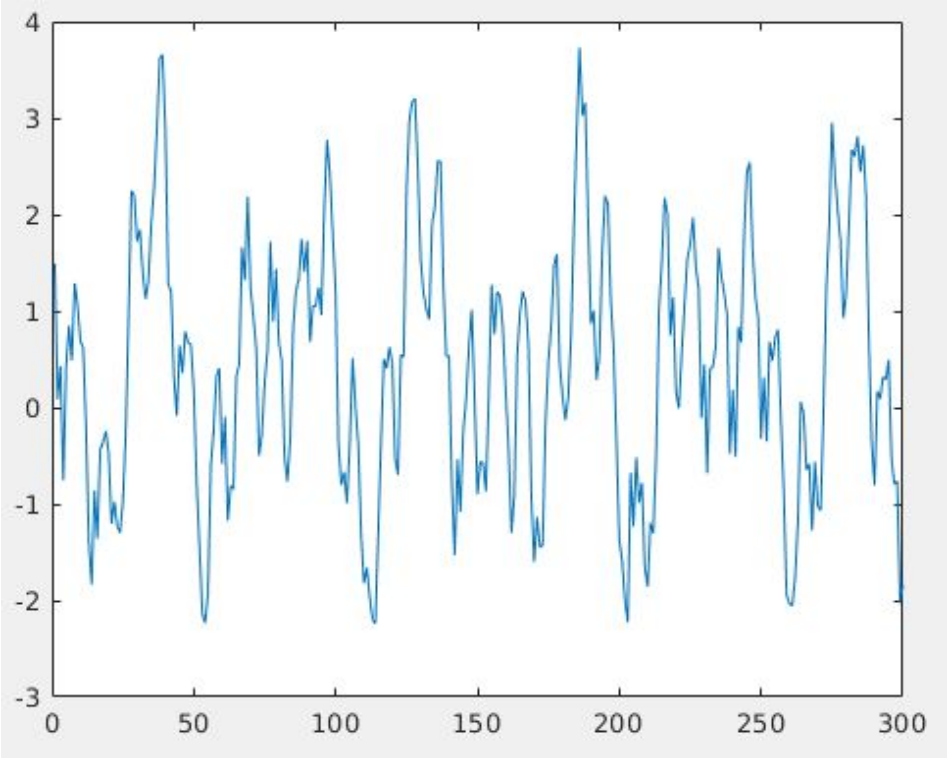
Label 1



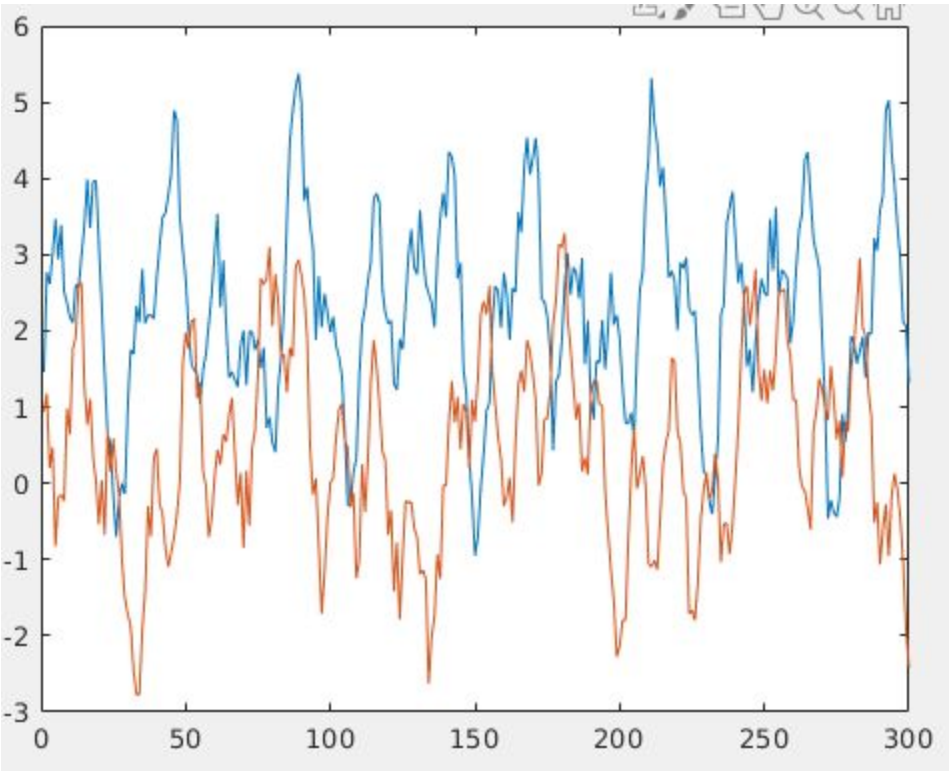
Label 2



Label 3



Label 4



A 0-labeled sample overlayed with a 1-labeled sample.