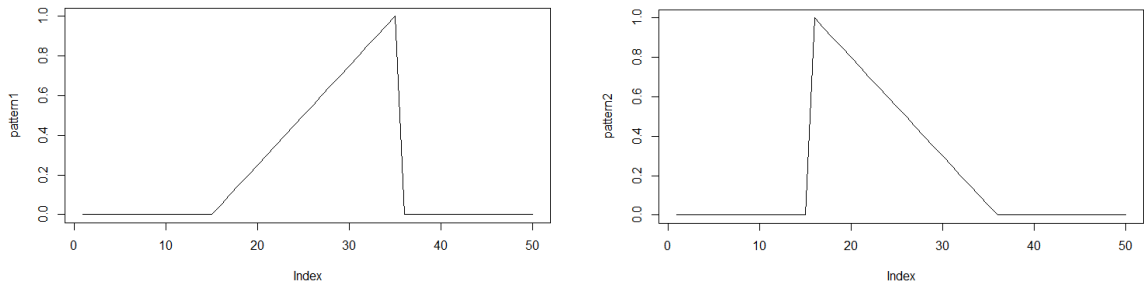
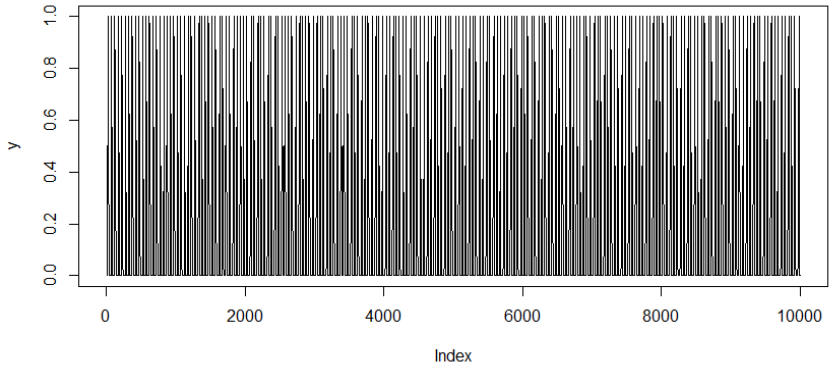
Setup

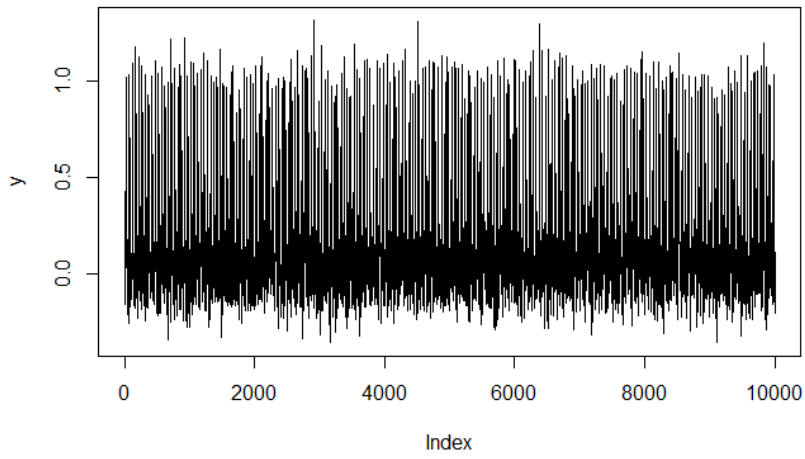
Start with 2 patterns:



Create a series of 200 of these patterns, same number of occurrence for each

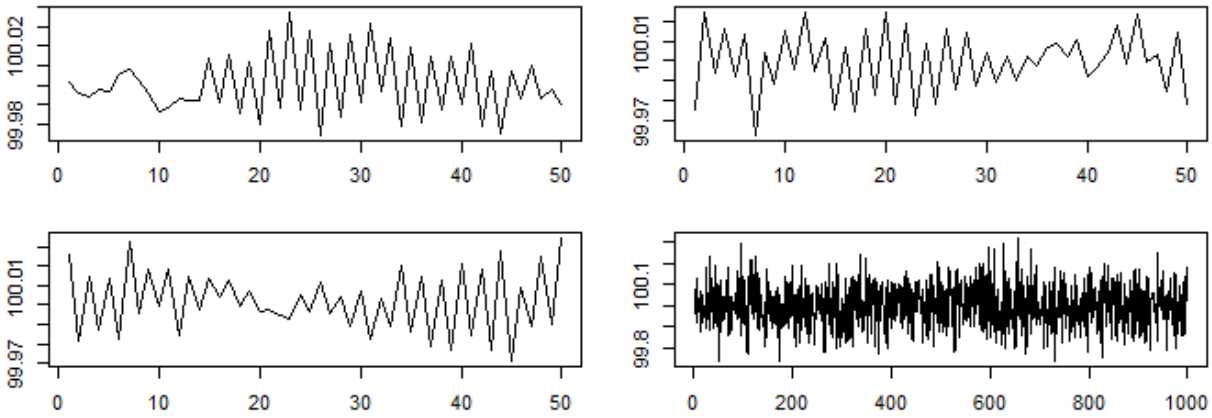


Add noise



Null Hypothesis

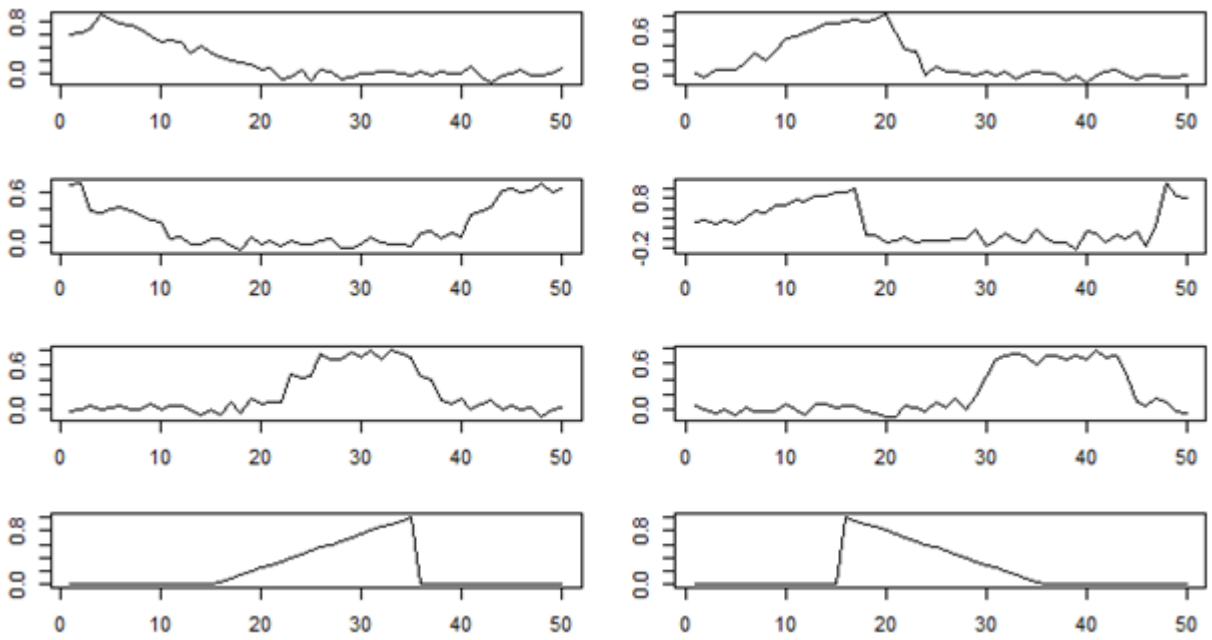
If the claim made in Keogh's paper is true that subsequence clustering produces the same outcome for any input, then it should produce the same outcome even when the input is just noise. This could not be confirmed. I created a series of 10,000 datapoints all of the same value. To this series I introduced just noise. I produce a subsequence of window size 50. Kmeans clustering was able to produce 3 clusters from the time series which look not like a sine wave:



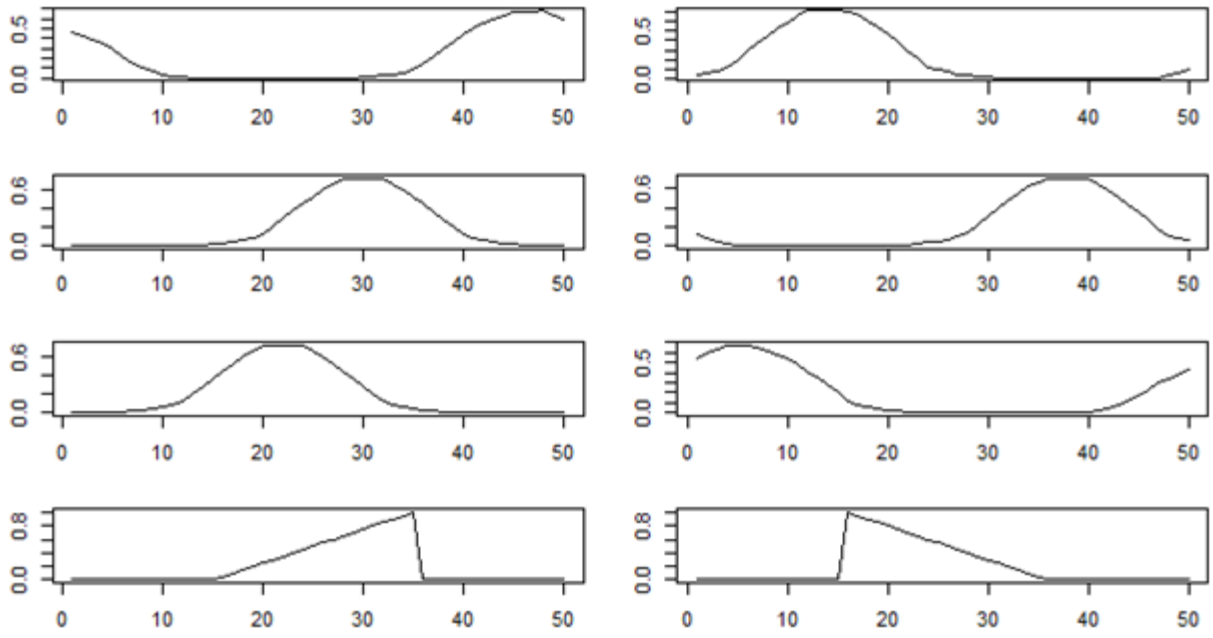
The last plot is the first 1000 datapoints in the series, the first 3 are the three clusters produced using subsequence clustering

Test 1 – Use K-means (6 clusters)

Sampling (20 samples of size 50 each) – The last 2 lots are the actual patterns



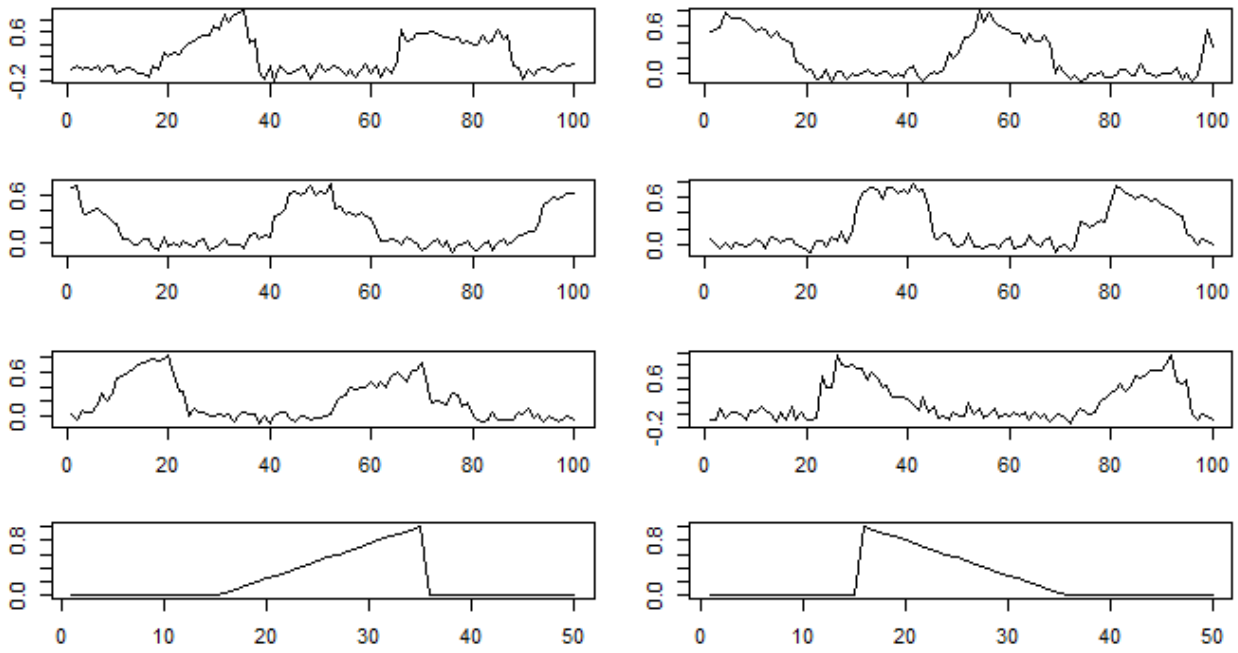
Subsequences – Window size 50 – Extreme smoothing



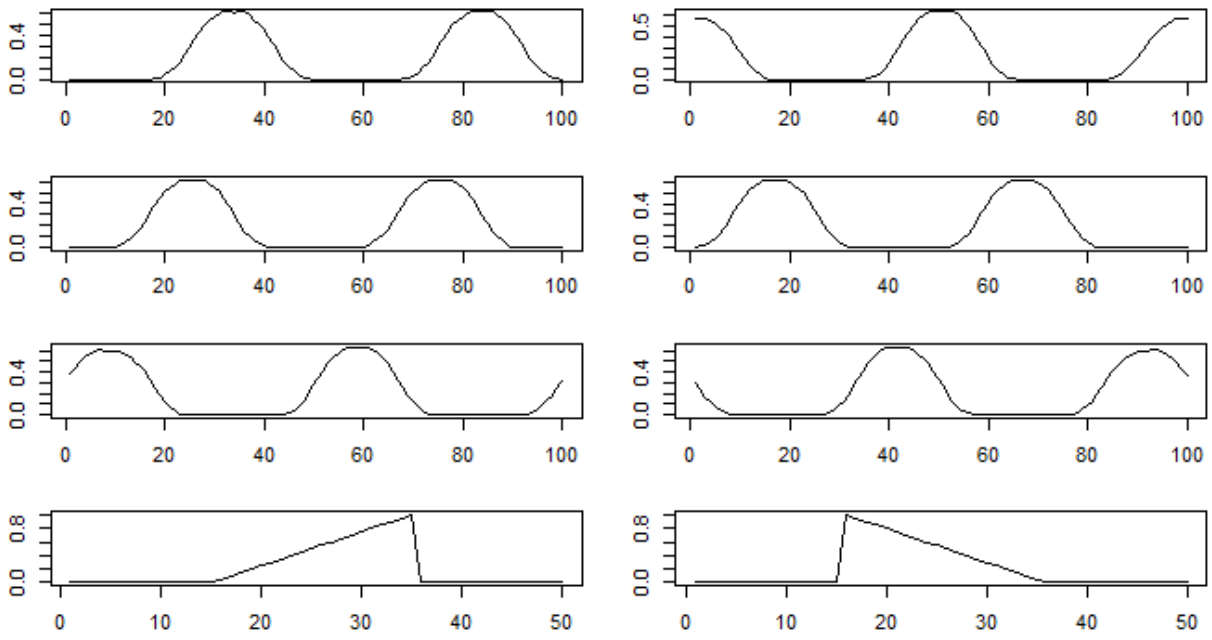
We know when the patterns are simple and obvious, sampling works better

Changing window size and sample width to 100

Samples

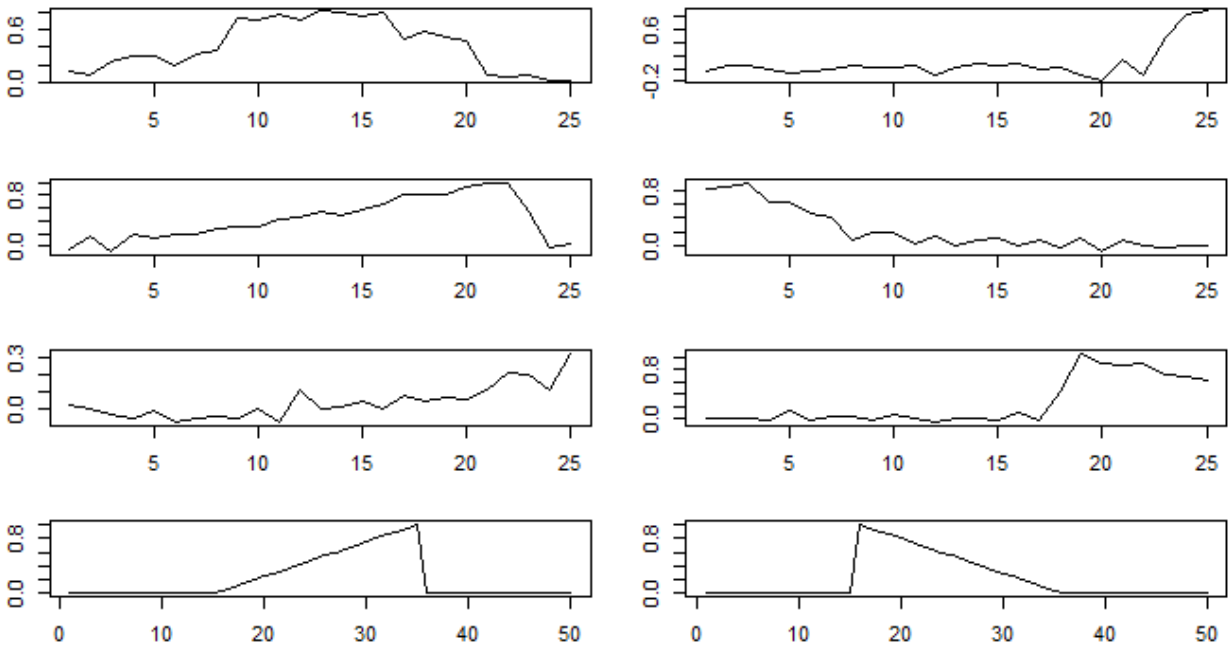


Subsequences

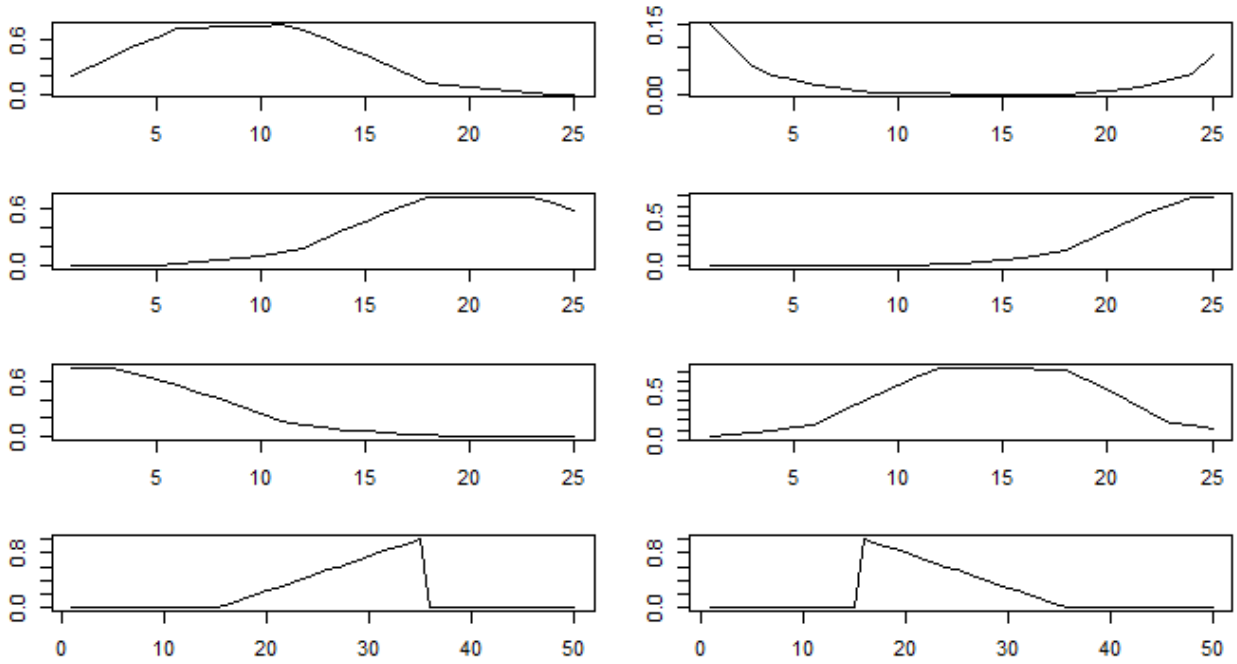


Changing window size to 25

Samples

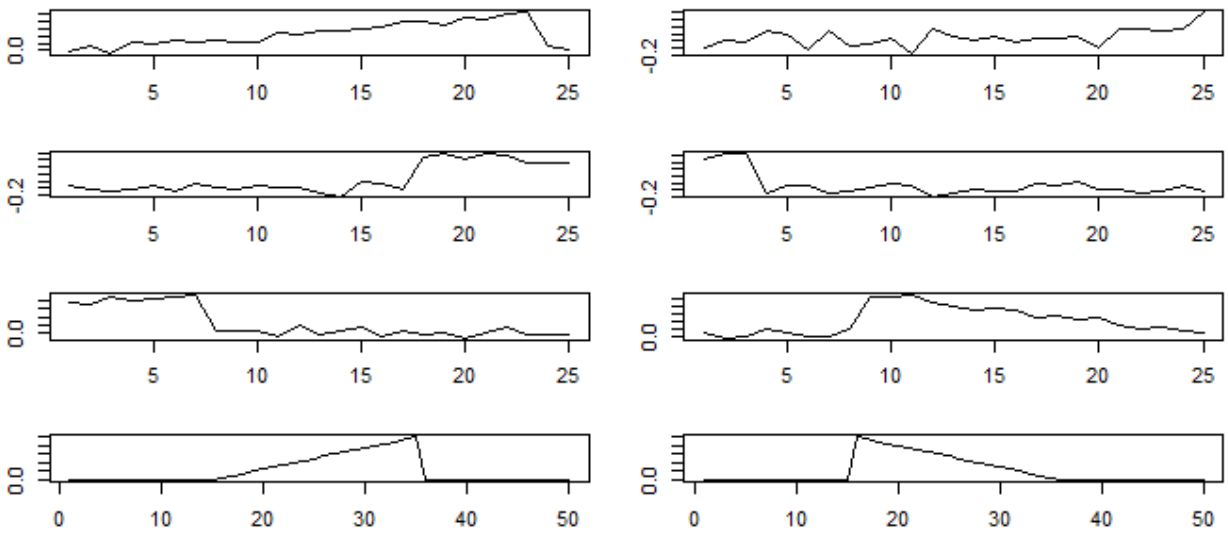


Subsequences

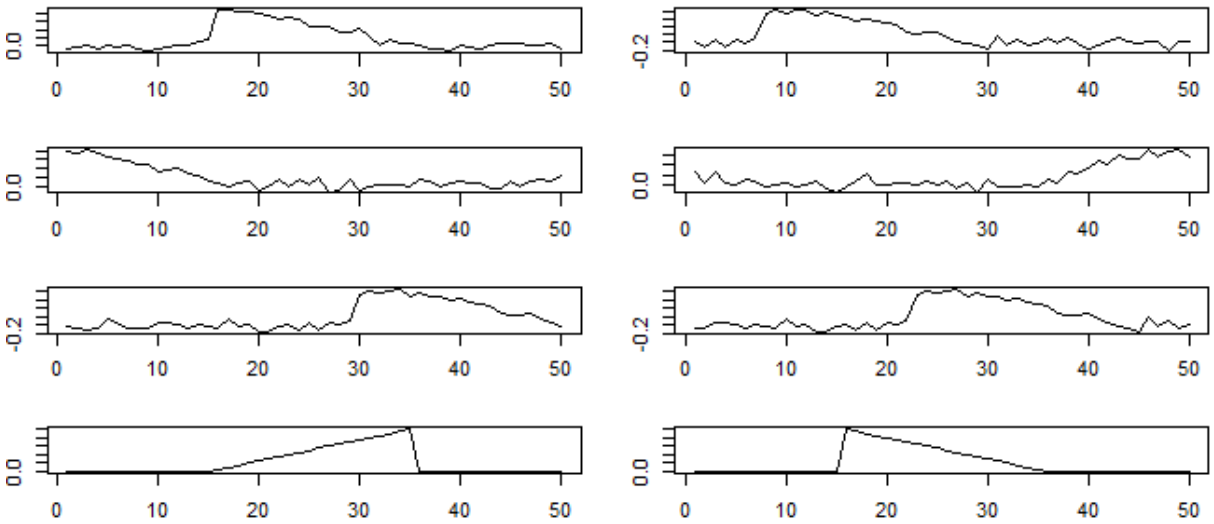


Test 2 - Using Medoids (pam) and window size =50

Samples



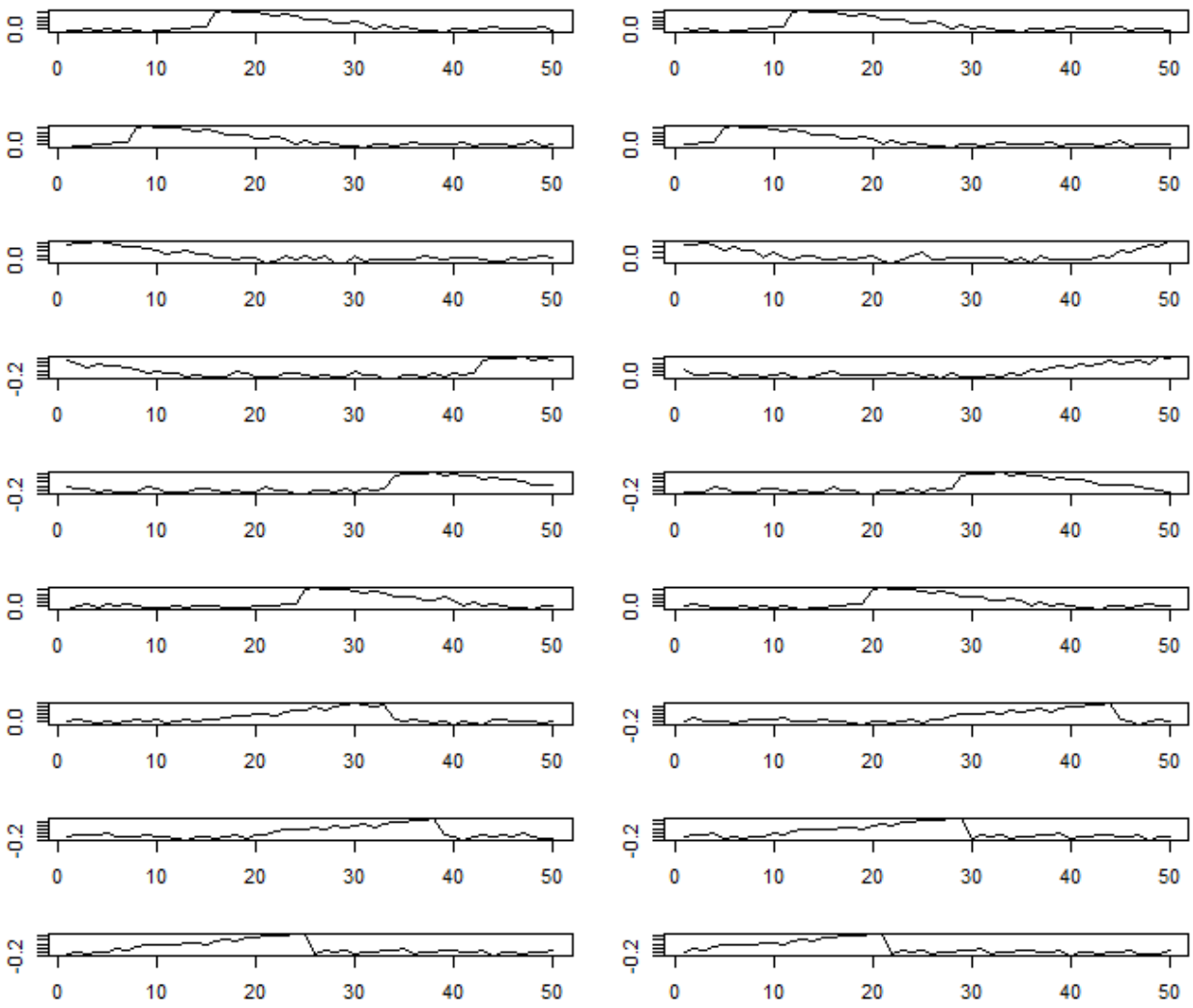
Subsequences



It is curious that only one of the patterns is detected by subsequences by samples finds both

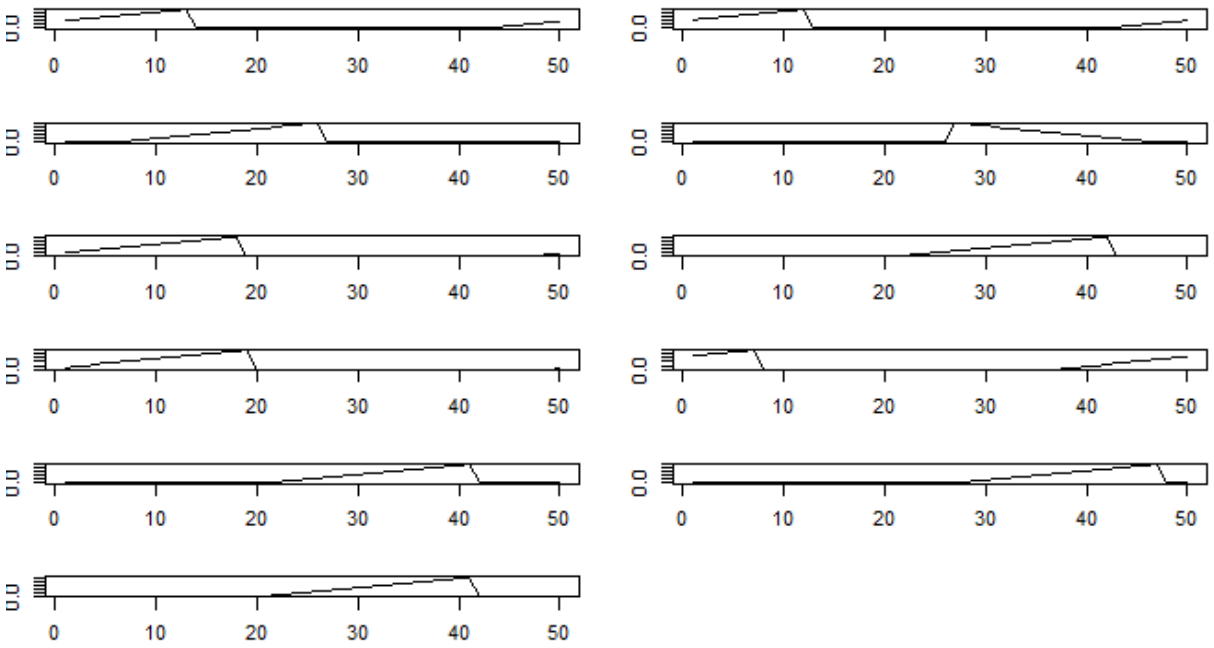
We have established that for simple patterns that occur frequently, sampling out-performs subsequences and that using kmeans for subsequences produces severely smoothed patterns while using pam's medoid finds the patterns.

Test 4 – One of the patterns occurred more frequently than the other, can subsequences find it? To check, I used 20 clusters:



Test 3 – Can subsequences find rare patterns?

To try I inserted one of the shapes only 1% of the times. I create a new sequence with one of the two patterns inserted 1% of the times (at points 50 and 150 in the 200 repeated patterns). I use subsequencing with window-size 50 and I us spam with 100 clusters. For the sample, I take 200 samples with no replacement, the sample width is 50 and so I know the entire series is represented. Samples does not find the rare pattern but subsequencing does.



Samples