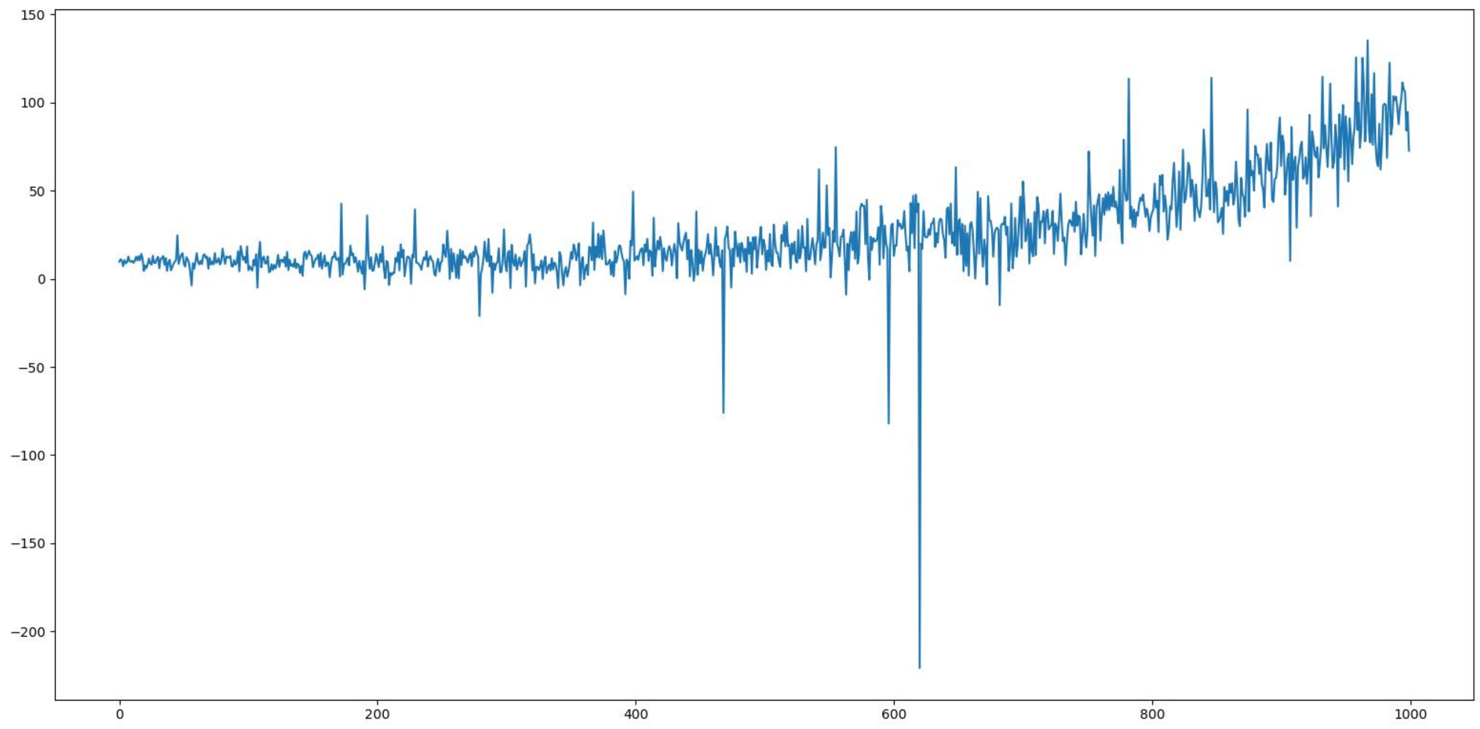
Report 5

Signal



Data + feature

1	2	3	4	-	-	-	-	20						40					60	-	-	-	998	999	1000

Number of segments = Data/20 +50%

	index	skew	max	rms	mean	median	Kur	std
Seg 1 →	1							
	2							
	3							
	…							
	48							
	49							
Seg 50 →	50							

Details

50% Train

Input of clustering algorithm is the Table of features

50% Test

Score : Silhouette Score

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Select the optimal number of clusters

Select the optimal number of clusters based on multiple clustering validation metrics like Gap Statistic, Silhouette Coefficient, Calinski-Harabasz Index etc.

Gap Statistic

Elbow Method: https://en.wikipedia.org/wiki/Elbow_method_(clustering)

Silhouette Coefficient

Calinski-Harabasz Index

Davies-Bouldin Index

Dendrogram: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html

Bayesian information criterion (BIC)

elbow method

It is the most popular method for determining the optimal number of clusters. The method is based on calculating the Within-Cluster-Sum of Squared Errors (WSS) for different number of clusters (k) and selecting the k for which change in WSS first starts to diminish.

The idea behind the elbow method is that the explained variation changes rapidly for a small number of clusters and then it slows down leading to an elbow formation in the curve.

The elbow point is the number of clusters we can use for our clustering algorithm.

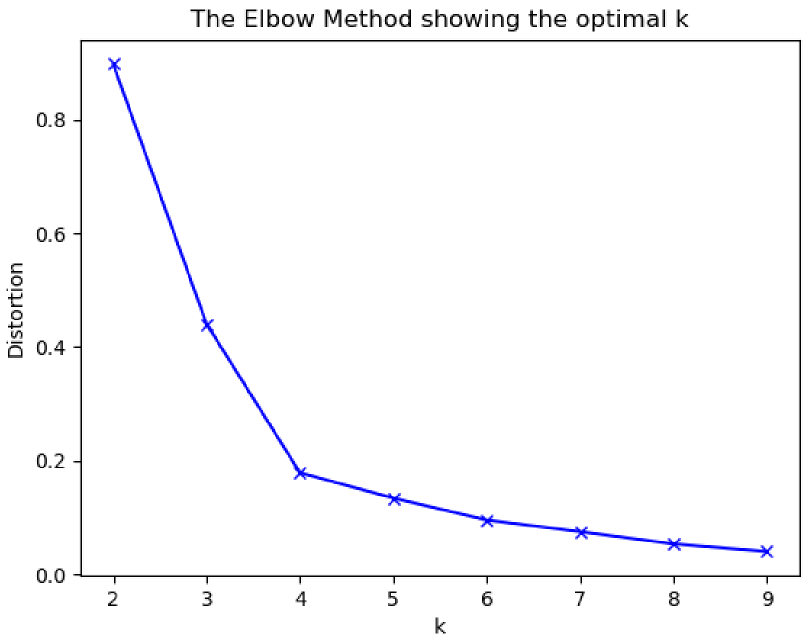
Further details on this method can be found in this paper by Chunhui Yuan and Haitao Yang.

using elbow method (use k-means clustering)

Explained variance. The "elbow" is indicated by the red circle. The number of clusters chosen should therefore be 3 or 4.
Distortion: Sum of squared distances of samples to their closest cluster center, weighted by the sample weights if provided. (the highest is better)

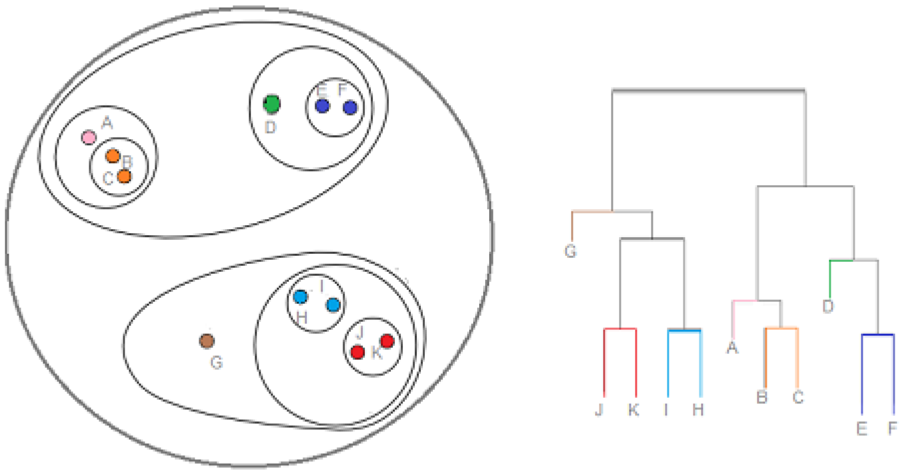
distortions - List (8 elements)

Index	Type	Size	Value
0	float	1	0.8977892495420725
1	float	1	0.4405725813228102
2	float	1	0.1790216944369348
3	float	1	0.13437550975121867
4	float	1	0.09576241887885968
5	float	1	0.07540434171185194
6	float	1	0.05383987117959501
7	float	1	0.0404616178462972



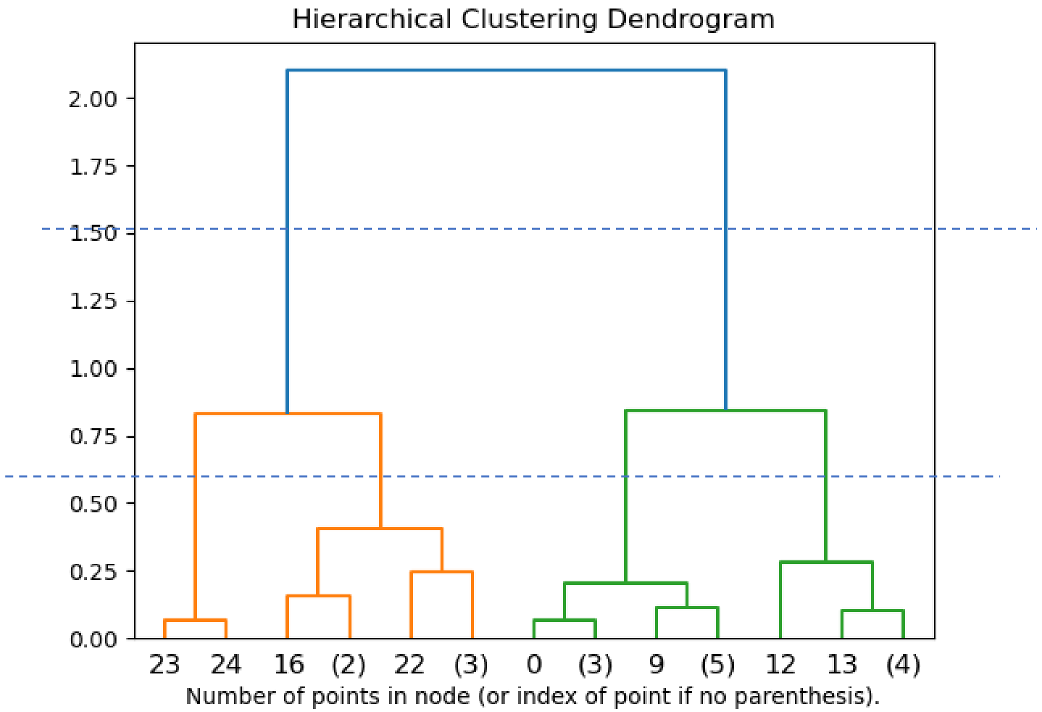
Dendrogram

- A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data
- This technique is specific to the agglomerative hierarchical method of clustering.
- The agglomerative hierarchical method of clustering starts by considering each point as a separate cluster and starts joining points to clusters in a hierarchical fashion based on their distances.
- To get the optimal number of clusters for hierarchical clustering, we make use a dendrogram which is tree-like chart that shows the sequences of merges or splits of clusters.



Dendrogram (use Agglomerative Clustering)

But we can use 4 cluster by seeing dendrogram diagram.



K=2

K=4

10 kind of clustering

List of Clustering + Reference

- ✓ spectral clustering (Spectral clustering is **a technique with roots in graph theory**, where the approach is used to identify communities of nodes in a graph based on the edges connecting them. The method is flexible and allows us to cluster non graph data as well)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html

https://en.wikipedia.org/wiki/Spectral_clustering

- ✓ birch clustering (**Balanced Iterative Reducing and Clustering** using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.Birch.html

https://en.wikipedia.org/wiki/BIRCH

- ✓ agglomerative clustering (The agglomerative clustering is **the most common type of hierarchical clustering used to group objects in clusters based on their similarity**. ... Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

https://en.wikipedia.org/wiki/Hierarchical_clustering

List of Clustering + Reference

- ✓ k-means clustering (K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

https://en.wikipedia.org/wiki/K-means_clustering

- ✓ mini-batch k-means clustering (The Mini-batch K-means clustering algorithm is **a version of the standard K-means algorithm in machine learning**. It uses small, random, fixed-size batches of data to store in memory, and then with each iteration, a random sample of the data is collected and used to update the clusters.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html

https://www.geeksforgeeks.org/ml-mini-batch-k-means-clustering-algorithm/

- ✓ affinity propagation clustering (Affinity propagation (AP) is **a graph based clustering algorithm** similar to k Means or K medoids, which does not require the estimation of the number of clusters before running the algorithm. Affinity propagation finds "exemplars" i.e. members of the input set that are representative of clusters.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AffinityPropagation.html

https://en.wikipedia.org/wiki/Affinity_propagation

- ✓ DB scan clustering (The principle of DBSCAN is **to find the neighborhoods of data points exceeds certain density threshold**. The density threshold is defined by two parameters: the radius of the neighborhood (eps) and the minimum number of neighbors/data points (minPts) within the radius of the neighborhood.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

https://en.wikipedia.org/wiki/DBSCAN

List of Clustering + Reference

- ✓ optics clustering (Ordering points to identify the clustering structure (OPTICS) is an **algorithm for finding density-based clusters in spatial data**. ... Its basic idea is similar to DBSCAN, but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html

https://en.wikipedia.org/wiki/OPTICS_algorithm

- ✓ mean shift clustering (Mean shift clustering using a flat kernel. Mean shift clustering aims to discover "blobs" in a smooth density of samples. It is a **centroid-based algorithm**, which works by updating candidates for centroids to be the mean of the points within a given region. ... If not set, the seeds are calculated by clustering.)

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html

https://en.wikipedia.org/wiki/Mean_shift

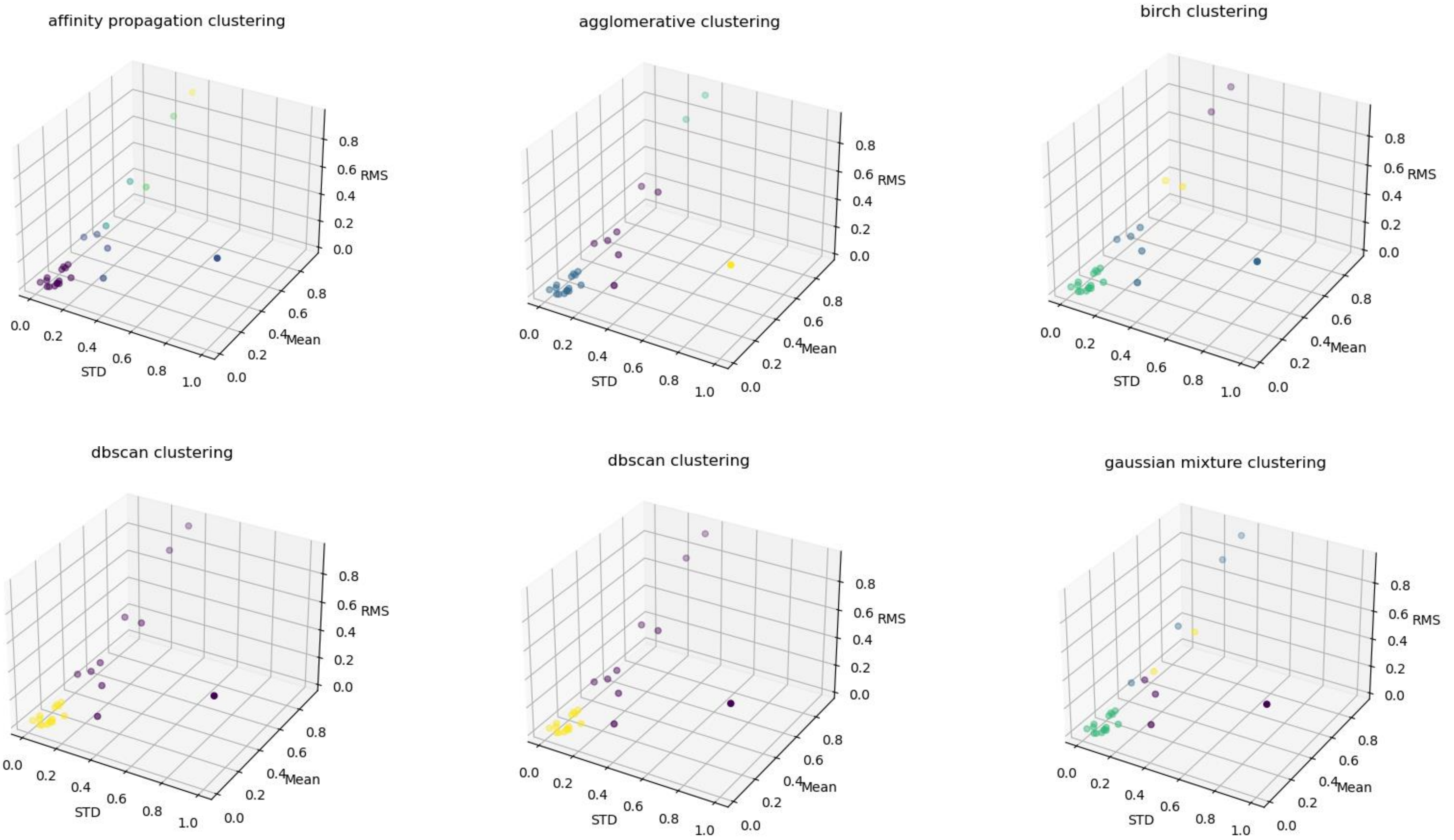
- ✓ gaussian mixture clustering (Gaussian mixture models (GMMs) are **often used for data clustering**. You can use GMMs to perform either hard clustering or soft clustering on query data. To perform hard clustering, the GMM assigns query data points to the multivariate normal components that maximize the component posterior probability, given the data.)

https://en.wikipedia.org/wiki/EM_algorithm_and_GMM_model

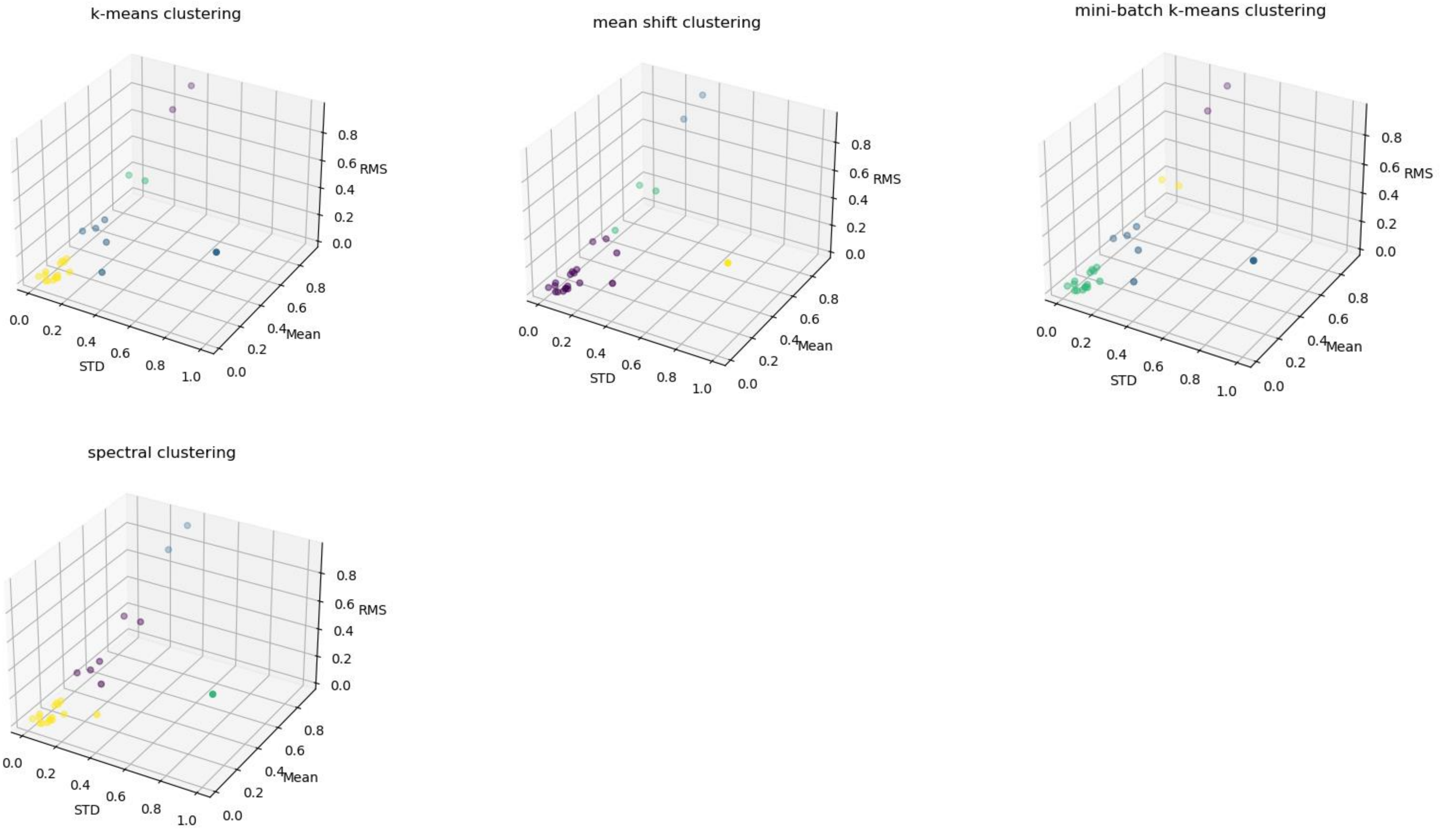
https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

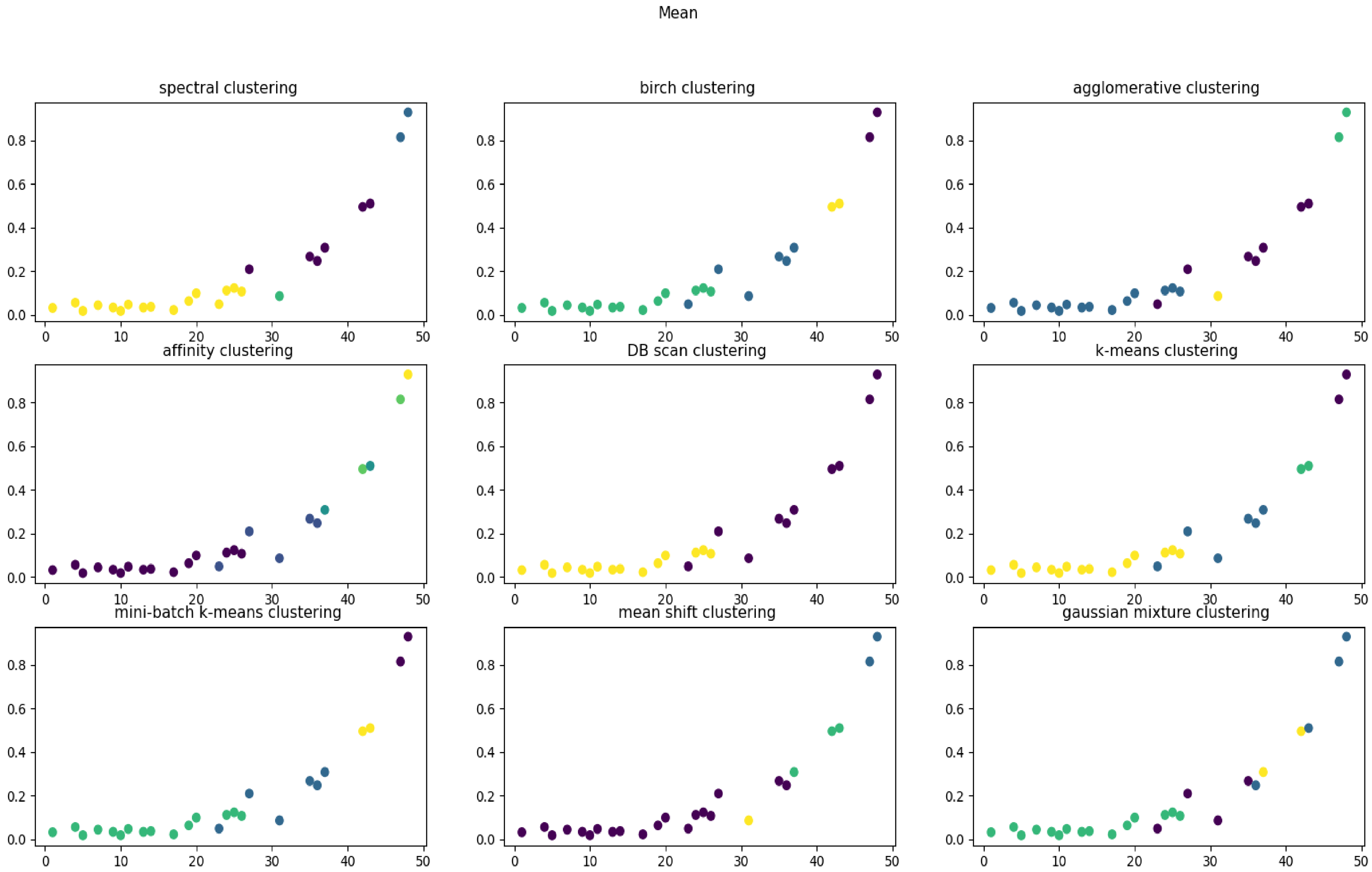
Clustering K=4

Result One	Result Two
✓ spectral clustering for k: 4 Score=61.311 % ✓ birch clustering for k: 4 Score=59.352 % ✓ agglomerative clustering for k: 4 Score=60.408 % ✓ affinity propagation clustering Score=40.297 % ✓ DB scan clustering Score=40.297 % ✓ k-means clustering for k: 4 Score=59.352 % ✓ mini-batch k-means clustering for k: 4 Score=59.352 % ✓ mean shift clustering Score=59.312 % ✓ gaussian mixture clustering Score=44.045 %	❖ spectral clustering for k: 4 Score=56.134 % ❖ birch clustering for k: 4 Score=56.209 % ❖ agglomerative clustering for k: 4 Score=60.862 % ❖ affinity propagation clustering Score=60.862 % ❖ DB scan clustering Score=60.862 % ❖ k-means clustering for k: 4 Score=56.209 % ❖ mini-batch k-means clustering for k: 4 Score=56.209 % ❖ mean shift clustering Score=59.500 % ❖ gaussian mixture clustering Score=54.964 %

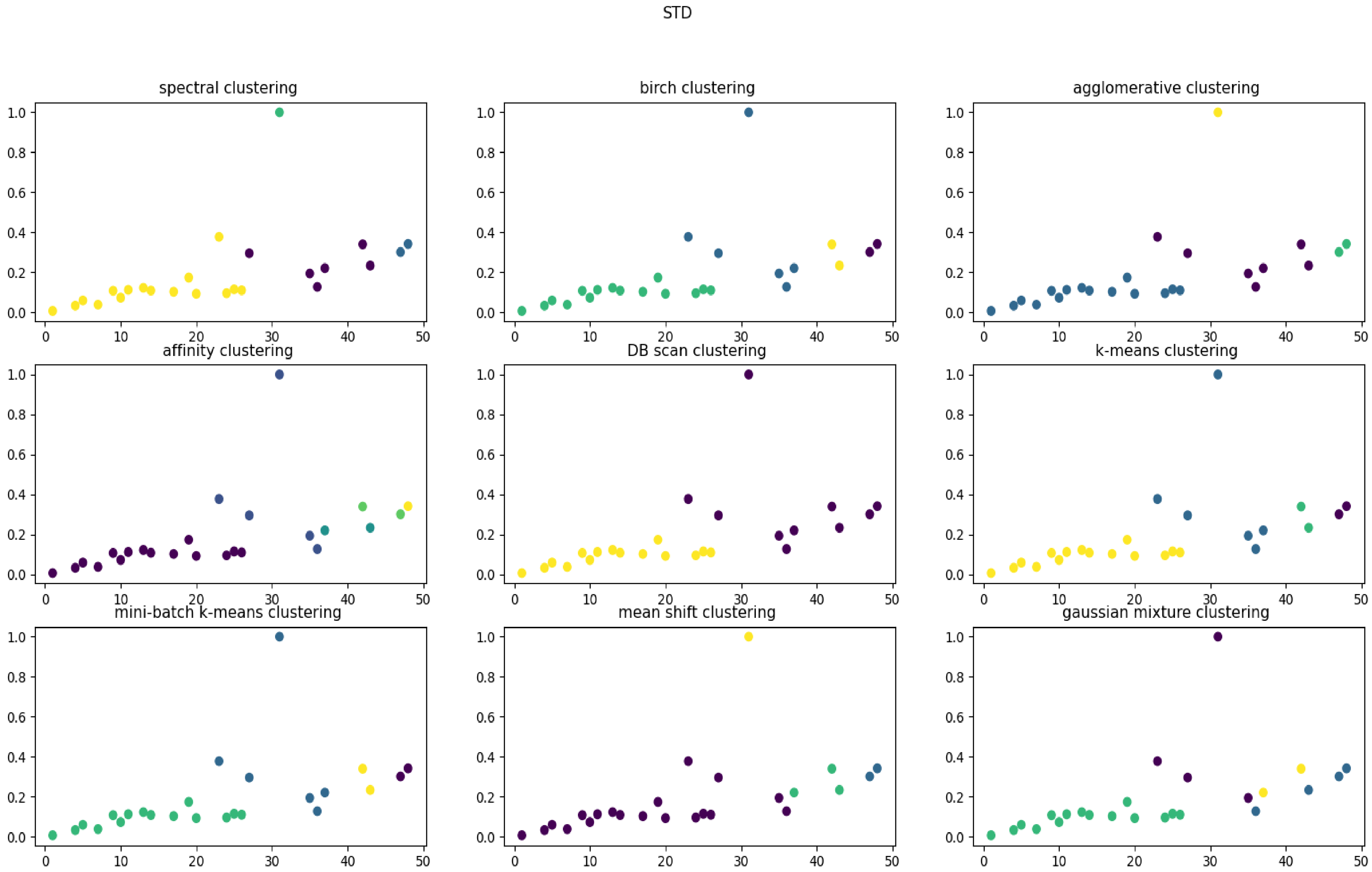


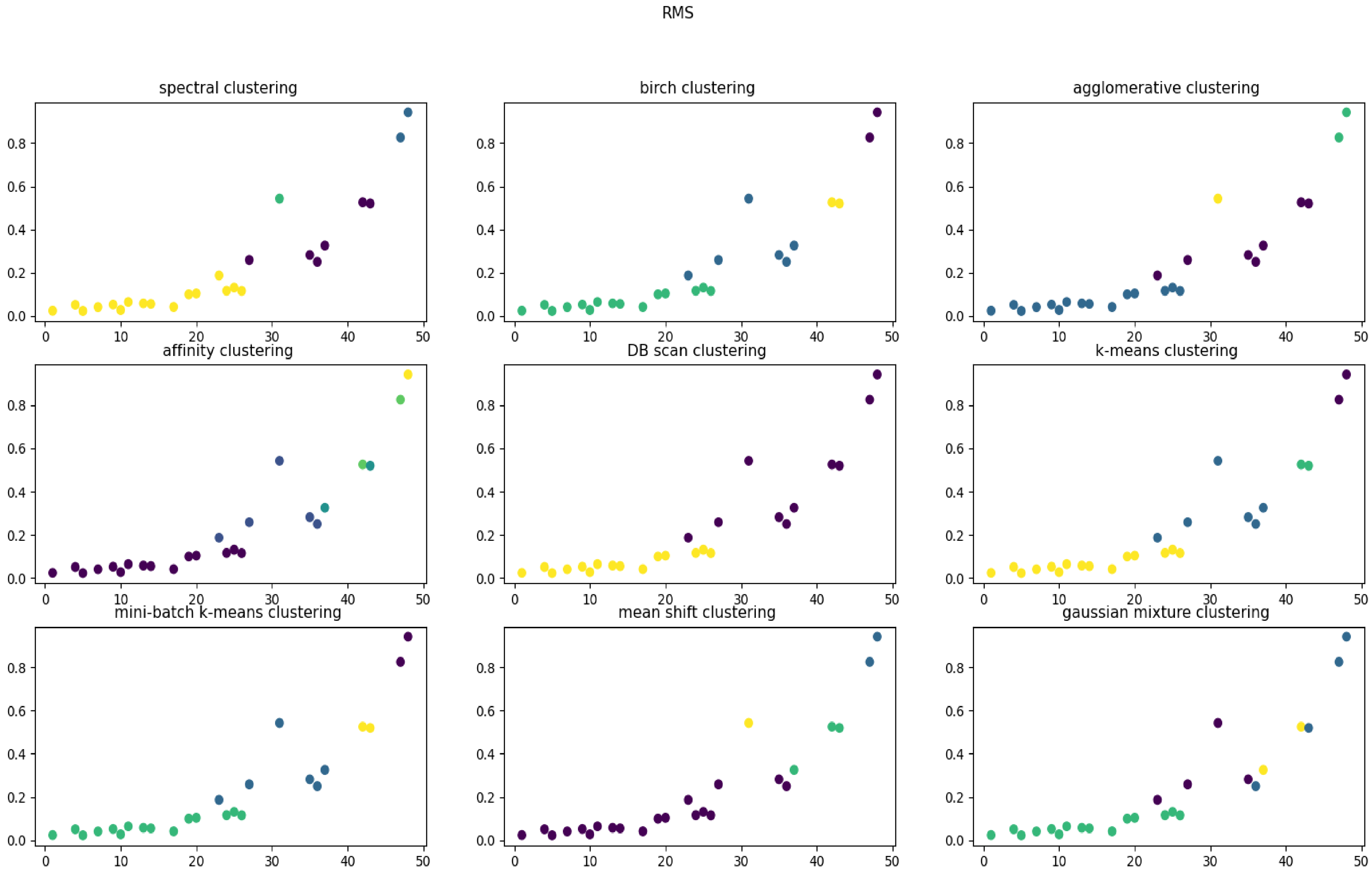
affinity propagation clustering
agglomerative clustering
birch clustering
dbscan clustering
dbscan clustering
gaussian mixture clustering
STD
Mean
RMS
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8

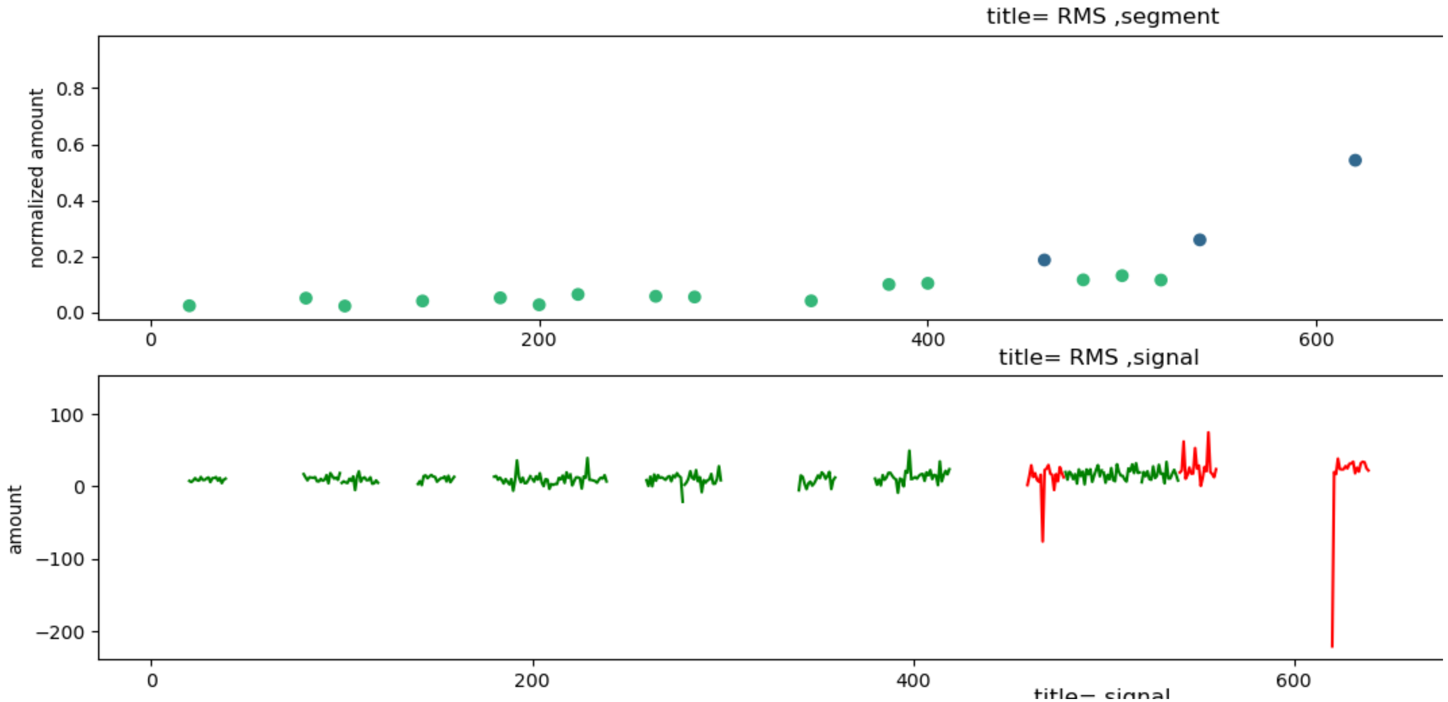

k-means clustering
mean shift clustering
mini-batch k-means clustering
spectral clustering
RMS
0.0
0.2
0.4
0.6
0.8
Mean
0.0
0.2
0.4
0.6
0.8
STD
0.0
0.2
0.4
0.6
0.8
1.0

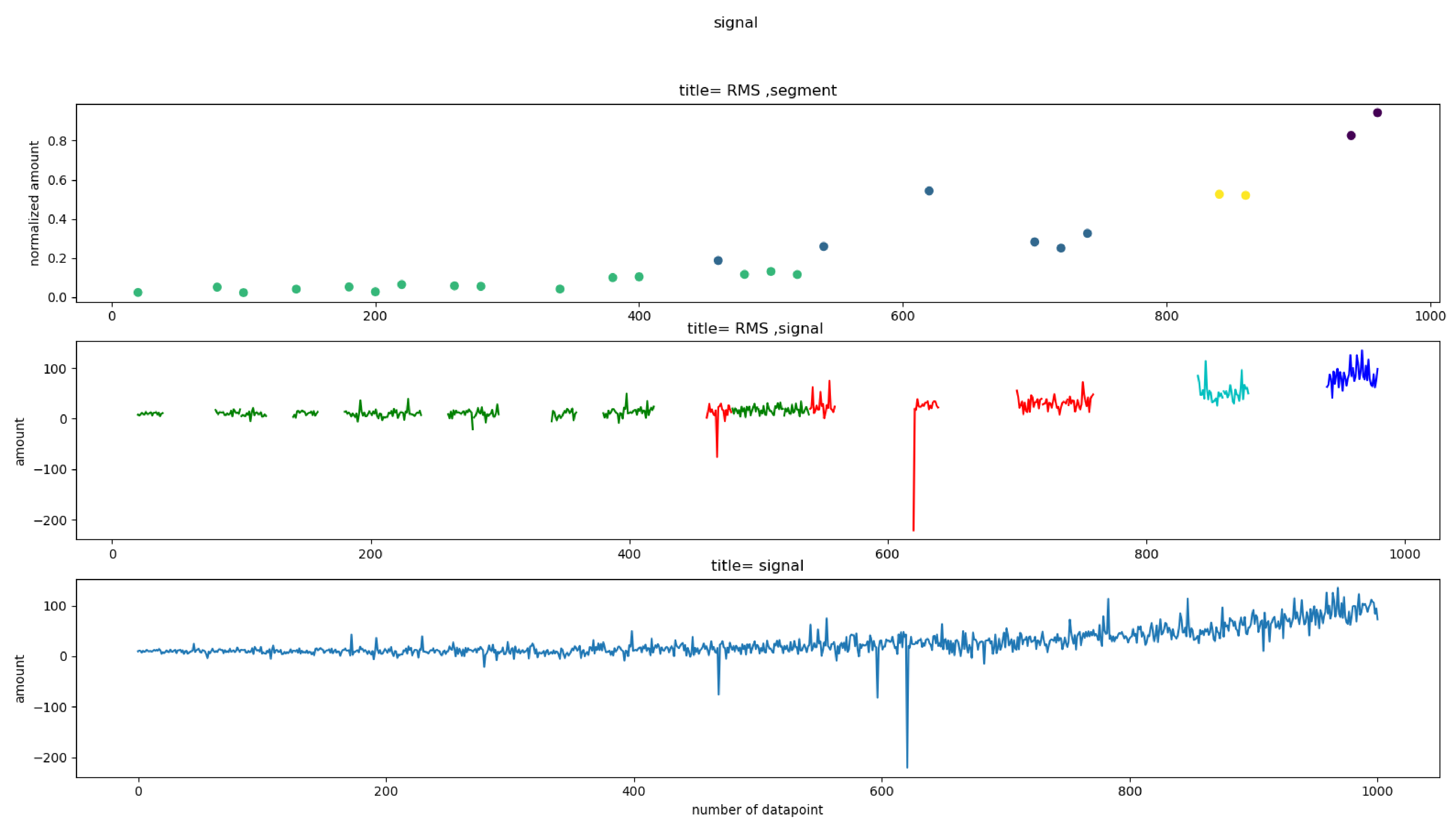


Mean
spectral clustering
birch clustering
agglomerative clustering
affinity clustering
DB scan clustering
k-means clustering
mini-batch k-means clustering
mean shift clustering
gaussian mixture clustering
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50


STD
spectral clustering
birch clustering
agglomerative clustering
affinity clustering
DB scan clustering
k-means clustering
mini-batch k-means clustering
mean shift clustering
gaussian mixture clustering
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50


RMS
spectral clustering
birch clustering
agglomerative clustering
affinity clustering
DB scan clustering
k-means clustering
mini-batch k-means clustering
mean shift clustering
gaussian mixture clustering
0.0
0.2
0.4
0.6
0.8
0
10
20
30
40
50


title= RMS ,segment
normalized amount
0.8
0.6
0.4
0.2
0.0
0
200
400
600
title= RMS ,signal
amount
100
0
−100
−200
0
200
400
600
title= signal


signal
title= RMS ,segment
normalized amount
0.0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000
title= RMS ,signal
amount
100
0
−100
−200
0
200
400
600
800
1000
title= signal
amount
100
0
−100
−200
0
200
400
600
800
1000
number of datapoint