Tutorial MATLAB Mapper

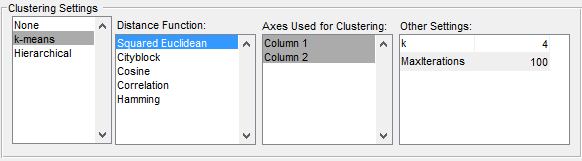
**Definition:** Mapper takes as input a set of points, a filtration function and a clustering function, and outputs an abstract simplicial complex (which in lower dimensions is equivalent to a graph) which illustrates the connectivity and topological structure of the data.

Topological analysis analyses the connectivity between points without implying a universal metric on the space. It only needs a similarity metric between any two points. This provides a way to analyse local paths

In contrast, in geometric space there is a vector associated with the path from one point to another which is often used to extrapolate further information about the structure. Geometric analysis typically embeds points in a metric space.

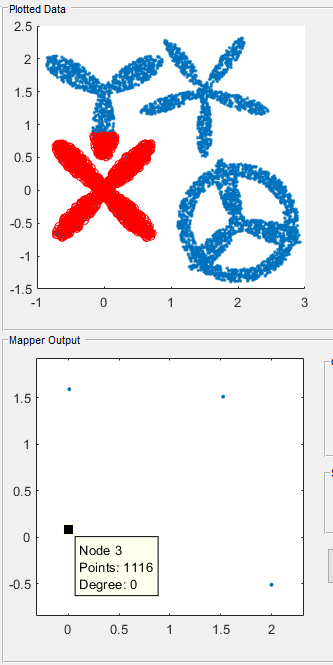
**Aim:** Any implementation of Mapper requires the two components, filtering and clustering. The following tutorial introduces filtering and clustering for the MATLAB Mapper.

1. Clustering – k-means, hierarchical
   1. Aims
      1. To ensure familiarity with clustering as part of the mathematical basis for the mapper algorithm;
      2. To introduce k-means and hierarchical clustering, as well as when they are appropriate to use with different data sets (as examples of common clustering algorithms in data science);
      3. To introduce the MATLAB Mapper interface: loading, visualising, and interacting with data
   2. Refresher material
      1. MATLAB; pointer to MATLAB tute
      2. Video of k-means (3blue 1brown)
      3. Video of hierarchical clustering (agglomerative)
   3. Activity (10-20 mins, see video walk-through [DF to make])
      1. Download MatMapper from <github>; unzip (there should be X files, folders, …)
      2. Launch MATLAB, navigate to the unzipped folder, enter MatMapperGuide on the command line; (the GUI will start and no further MATLAB commands are required in this tutorial)
      3. Load the *Combo* dataset by selecting it in the drop-down menu at the top left.   
         What you will see: The *Loaded Data* section in the top centre shows the generated points represented as a table, while the Plotted Data section at the top-right plots the imported data on selected axes.
      4. In this tutorial, you are only applying clustering, so you will not apply a filter. To deselect the default filter, Ctrl-Click any selected items in the Select Filters box. (the Filter Settings box should now be empty).
      5. In Clustering Settings, Clustering Algorithm set to k-means and other options should be automatically set as in the below screenshot (ie Distance Function set to Squared Euclidean; Axes used for Clustering set to both Columns; k set to 4, MaxIterations set to 100) {JW note: Settings and selections}

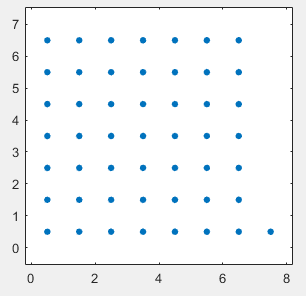
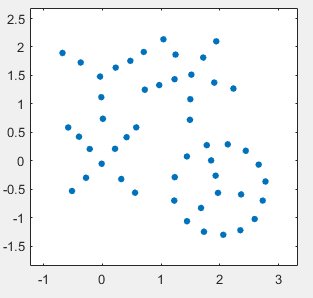
 [redo the image with k-means selected in blue, label clustering algorithm as first column]

* + 1. To apply the clustering algorithm, click the “Do Mapper” button at the bottom right. The Mapper Output window should display four points, one corresponding to each of the four clusters that are generated.
    2. To show the correspondence between the original plot and the clusters, click on the Data Tips icon in the red box in the below screenshot; while selected, it should appear inside a blue box. Then, click on one of the dots in the Mapper Output window. This will highlight the nodes in the Plotted Data window that are a member of the selected cluster.

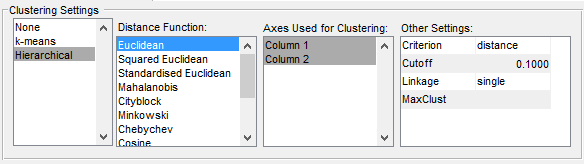
 



* + 1. [message] Try increasing the value of k to 50 or higher and pressing the Do Mapper button again. This should give a diagram similar to the one to the left below in the Mapper Output window. While this does illustrate the original structure, we are positioning the clusters according to their centroids from the plotted version. Click the “Force Directed” option in the Graph Layout box to visualise the same data in a different way, as shown on the right hand side – simply a list of points, with no information about their position in a relative or absolute sense. Click the “Geometric Mean” option to return to the previous setting.

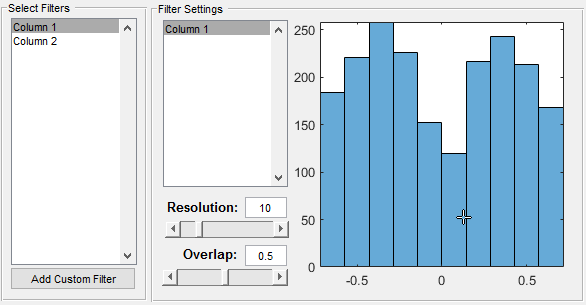


* + 1. Change the clustering settings to use Hierarchical clustering instead of k-means, using the settings shown in the diagram below. Click the “Do Mapper” button again and use Data Tips to see how this changed the position of Observe how this changes what is included

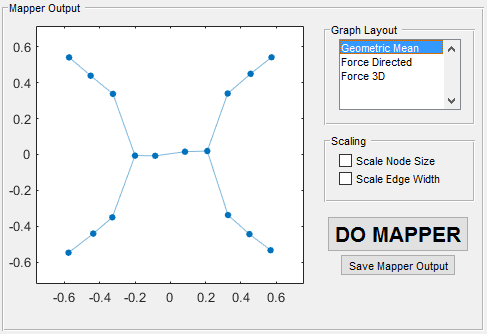


* + 1. Explore settings (k for k-means; cut-off level for hierarchical)
    2. Repeat this for the Pronged dataset and the Annular dataset.
  1. Questions they should be able to answer
     1. What is the effect of varying k?
     2. What’s the effect of varying the cutoff in h…?
     3. Why do the clusters sometimes change when repeating k-means, but not Hierarchical?
     4. The Annular dataset is clearly separated into two clusters. Why does k-means fail at separating them when k=2? Why does hierarchical succeed?
     5. Applying k-means with k=4 to the pronged dataset clusters each arm together.
  2. Wow factor
     1. Only takes 5 mins if you already know clustering
     2. Ease of use
     3. I can do it myself
     4. Elements of visualisation –

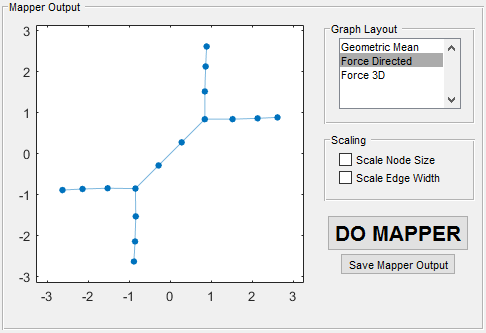
1. Mapper algorithm + basic Filtering
   1. Aims
      1. Understand filtering
      2. Use the interface to use Mapper
   2. Refresher material
      1. Level sets
      2. Video explaining mapper algorithm
      3. Pointer to classic paper
      4. Flares and loops
   3. Activity (1-2 hours)
      1. Load the Pronged dataset as in the first tutorial, and select the default Hierarchical clustering algorithm settings as before.
      2. In the Select Filters box, select Column 1; after this, in the Filter Settings box, again select Column 1. That section of the interface should look like as demonstrated below. The histogram to the right shows the distribution of the points along the selected axis.



* + 1. Press the Do Mapper button and observe the output, which should be similar to the screenshot below. Unlike in the previous tutorial, where we only applied clustering, we are now applying the Mapper algorithm, which provides connectivity information between clusters.



* + 1. Select some nodes using the Data Tips function from the previous tutorial, and observe the correspondence between the plotted data and the nodes in the graph. Note the overlap between the points contributing to any pair of connected nodes.
    2. Select the “Force Directed” option in the Graph Layout list. Unlike in the first tutorial, the nodes communicate information about the structure of the data through their connection to each other. It is important to recognise that neither of these is the “correct” or “true” representation of the output graph, simply a way of visualising it. The graph structure itself has no geometric information associated with it – no node has any intrinsic position, only a list of other nodes that it is connected to. Select the Geometric Mean option again before continuing.



* + 1. Investigate changing Resolution and Overlap
    2. Repeat with Column 2 and see how the output changes
    3. Repeat with a second data set
  1. Questions
     1. How many flares / holes? (make them read the analysis)
  2. Wow factor
     1. Separate out the flares and loops that the geometric approaches cannot do
     2. Elements of visualisation – select components in one rep and seeing it in the other
     3. Seeing the correspondence between the two representations

1. Filtering – advanced
   1. Aims
      1. To know how to use this on your own data
   2. Activity
      1. Import external data set (need format of table described)
2. Advanced – simplicial complex analysis (exporting)