# Curve Clustering Project Documentation

# Overview

This documentation will explain the logic and file structure of the project. Overall, the input of a project is an matrix where is the number of sample *types* (for instance, each voltage applied in an I(V)) measurement and is the number of samples taken (for instance, if we have 1000 I(V) curves measured).

We will use the fact that Python is object oriented, therefore we will have an *object* which will be modified. An *object* in Python can have *attributes* and *methods*. This will be clearer later. An object in python is realized calling a new *class*.

Thus, the first thing we will do is to define our object.

# Main class: data\_2\_cluster

## \_\_init\_\_\_

This is the first thing that happens when you create an object. To create an object, you need to specify a folder where all the samples reside. Then the object is created with 4 new attributes.

*Input:* the source folder

*Result:* 4 new attributes:

* data\_2\_cluster.X, data\_2\_cluster.X\_norm: Two matrices where is the number of samples *types* and is the number of samples taken.
* data\_2\_cluster.X\_df, data\_2\_cluster.X\_norm\_df: Same as above but the shape of pandas dataframe.

The normalization is done according to z-score normalization:

– the mean of the measurement , – the standard deviation.

## Data\_2\_cluster.plot()

This is the plotting method of the object, for plotting its’ attributes.

*Input:* The user is prompted to enter what he chooses to plot

*Output:* The data is plotted according to the user selections.

## PCA

This method calculates the principal components of the data.

First, the mean of each measurement *type* is calculated (the mean of every *row* in the data matrix). Then, the mean is subtracted. **Note:** this means that the original I(V) curves are being distorted, or, in other words, we treat each voltage measurement to be independent to one another.