Practical: Clustering applied to a toy example (moving-curve)

Lecturer: Diego Peluffo – Course: Machine Learning
https://sites.google.com/site/yachaycoursesdp/machine-learning
Yachay Tech University
September 9, 2019

This practical is aimed at verifying some basics and foundations on distance-based clustering while testing the dynamic effect of a curve moving in an arc. Its grading value is the 10 % of the overall course evaluation. The maximum grading value of each item is highlighted next to the itemization numbering. The maximum score of the whole practical is 10.0, being 20 % devoted to the technical and linguistic evaluation. MatLab (Octave) and/or Python scripts are highly desirable (otherwise, provide also software installers and indications). Every group must upload all the files (document and scripts) in the corresponding Google drive shared folder (`...\Practicals\Practical1`).

I. Practical description

A. Artificial data generation: Moving-curve

This toy example is introduced in [1] (Chapter 5, section 5.3.3, page 79).

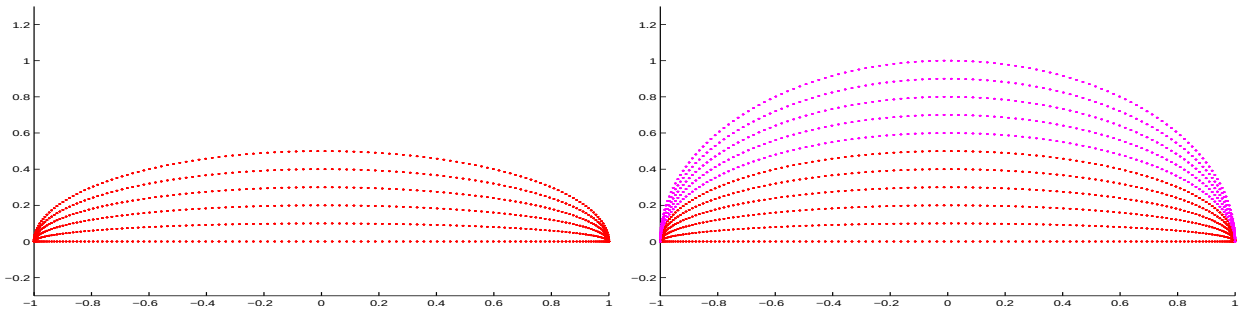
Let us consider the following toy example of a moving-curve. At time instance t, the effect of a 2-D curve moving in an arc from down up can be emulated by the XY coordinates:

$$\boldsymbol{x}^{(t)} = \begin{pmatrix} |\cos(2\pi\boldsymbol{\tau})|^{\top} \\ -|\cos(2\pi\boldsymbol{\tau})|^{\top} \end{pmatrix} \tag{1}$$

and

$$\boldsymbol{y}^{(t)} = \begin{pmatrix} |t\sin(2\pi\boldsymbol{\tau})|^{\top} \\ |t\sin(2\pi\boldsymbol{\tau})|^{\top} \end{pmatrix}, \tag{2}$$

where each entry of vector $\boldsymbol{\tau}$ is $\tau_n = n/N$ with $n \in \{1, \ldots, N/2\}$, being N the number of samples per frame. Then, we can form the corresponding data matrix $\boldsymbol{X}^{(t)} \in \mathbb{R}^{N\times 2}$ as $\boldsymbol{X}^{(t)} = [\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}]$ as well as the frame sequence $\{\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N_f)}\}$. Such a sequence can be arranged into the frame matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N_f N\times 2}$. Then the video effect until a certain frame T is done by keeping the previous frames to show the trace of path followed by the curve. Figure 1 depicts the arc moving effect, when considering $N = 100$, and $N_f = 10$. In this instance, the clustering can be done by using any approach, and its aim is to identify two natural movements or clusters ($\widetilde{K} = 2$).



(a) `Frames from` 1 `to` $N_f/2$ (b) `Frames from` $N_f/2+1$ `to` N_f

Fig. 1: 2-D moving-curve

To add noise to the moving-curve model, we consider an additive noise to be applied over the Y coordinate in the form $\nu\mathbf{n}$, where ν is the noise level and $\mathbf{n} \in \mathbb{R}^N$ is the introduced noise following a Gaussian distribution $\mathbf{n} \sim \mathcal{N}(0, 1)$, so:

$$\boldsymbol{y}_{\mathbf{n}}^{(t)} = \boldsymbol{y} + \nu\mathbf{n}. \tag{3}$$

Some examples of the tracking performance are shown in figures 2 and 3 when setting the number of frames to be $N_f = 100$ and the number of samples per frame to be $N = 100$, representing two experiments changing the number of groups and noise conditions. Figure 2 shows the behavior of tracking vector for clustering the frame matrix into 3 clusters ($\widetilde{K} = 3$) while Figure 3 is for $\widetilde{K} = 4$. Experiments can be run over different noise levels $\nu \in \{0.05, 0.1\}$. No noise case is also considered ($\nu = 0$).

For visualization purposes, some meaningful frames from the two experiments are selected. For the first experiment, selected frames are $t \in \{31, 71, 100\}$. Likewise, frames $t \in \{31, 50, 75, 100\}$ are selected for the second experiment.

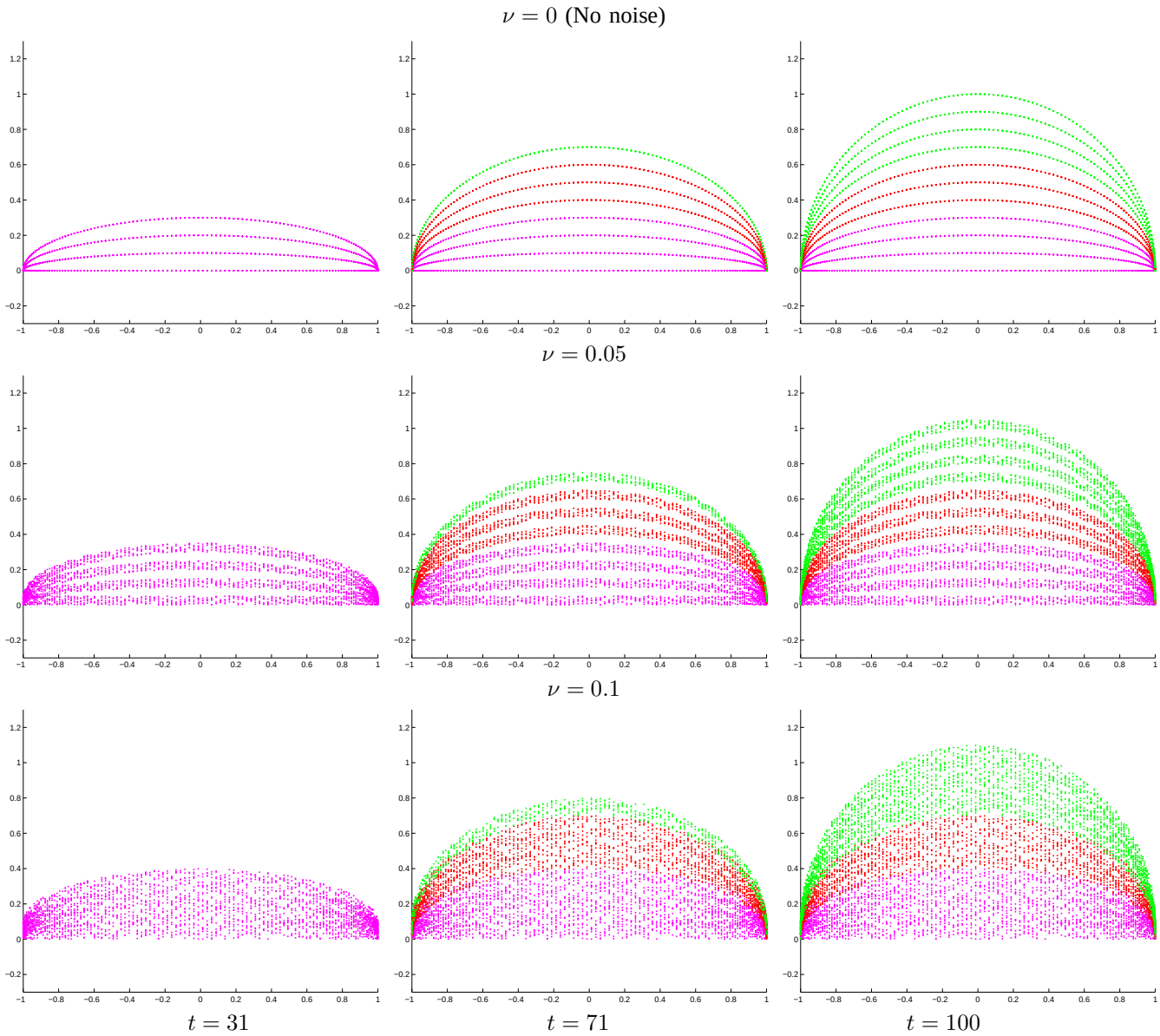


Fig. 2: Clustering of 2-D moving-curve into $\widetilde{K} = 3$ clusters with $N_f = 100$ frames and $N = 100$ samples per frame

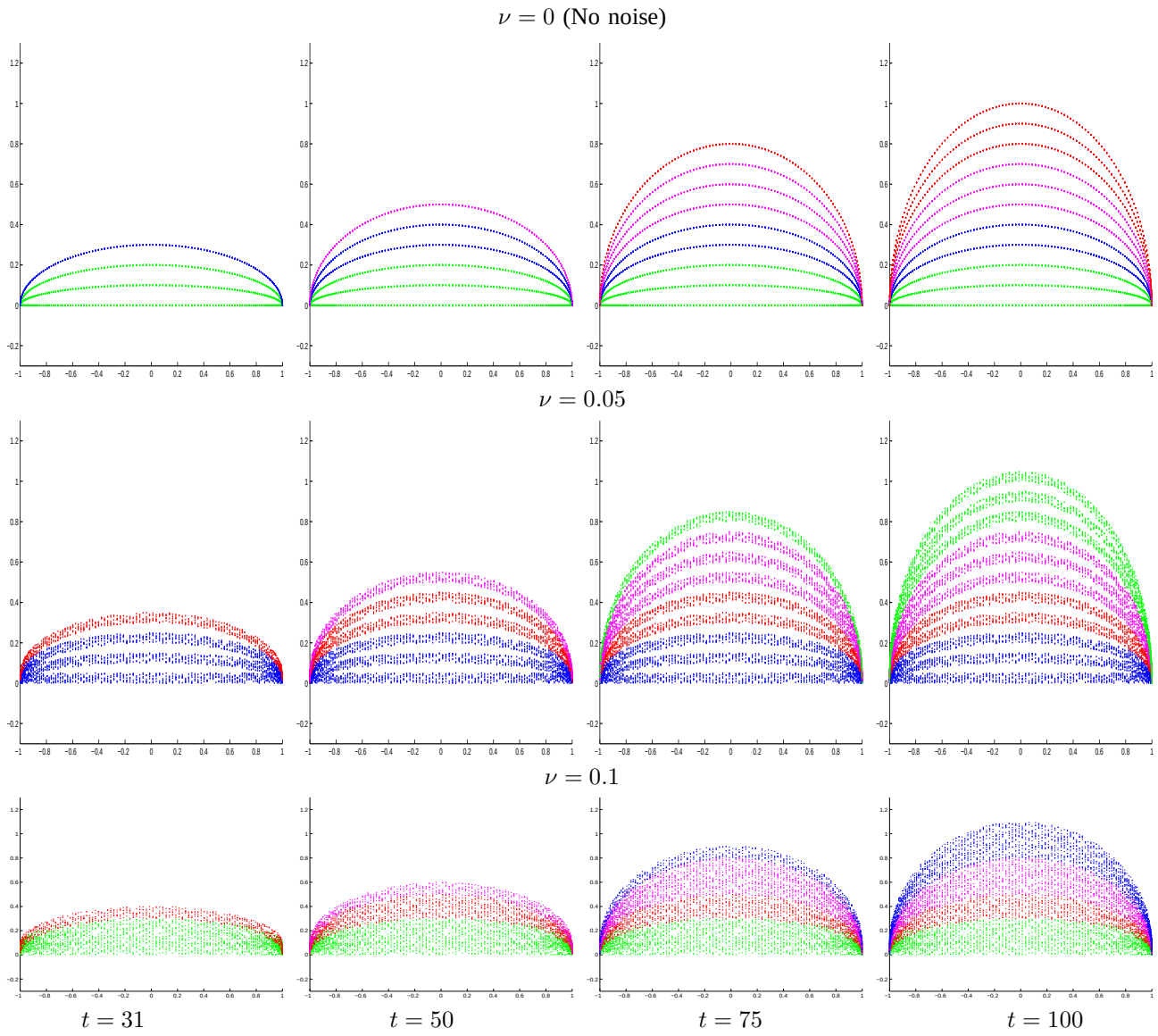


Fig. 3: Clustering of 2-D moving-curve into $\widetilde{K} = 4$ clusters with $N_f = 100$ frames and $N = 100$ samples per frame

More details and information in [1], [2].

II. Questionnaire

1) **(2.0)** Following equations (1) and (2), write a script to generate an artificial moving curve with an adjustable number of frames as well as number of samples (size).(Suggestion: Consider a function in the form: `output = toy_data_generation (input_arg, param)`, such that `input_arg = 'Moving curve'`, `param.frames =` N_f, `param.size =` N), and `output.data =` $\boldsymbol{\mathcal{X}}$. Explain the general idea of the script (a pseudocode is highly advisable).

2) **(2.0)** Clustering procedure can be carried out over the frame matrix $\boldsymbol{\mathcal{X}}$, whose purpose is to split $\boldsymbol{\mathcal{X}}$ into $\widetilde{K}$ disjoint clusters. Write a script to perform a distance-based clustering (K-means [3] is highly recommendable) approach over the frame matrix. Vary the number of clusters, analyze and write the observations. (Suggestion: Use a machine learning or statistical toolbox that includes the K-means algorithm).

3) **(4.0)** To assess the robustness of the clustering, performance can be measured over different noise levels (No noise case can bs ($\nu = 0$). Fisher's criterion is an advisable performance measure: **Fisher's Criterion** (J): This measure quantifies how well each data point is grouped into clusters regarding the Euclidean distance-based compactness, which is estimated

as follows:

$$J = \frac{\text{tr}\left(\sum_{k=1}^{\widetilde{K}} \left(\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{\mathcal{X}}}\right)\left(\overline{\boldsymbol{x}}_k - \overline{\boldsymbol{\mathcal{X}}}\right)^\top\right)}{\text{tr}\left(\sum_{k=1}^{\widetilde{K}} \boldsymbol{S}_k\right)}, \; J \in \mathbb{R}^+ \tag{4}$$

where $\overline{\boldsymbol{x}}_k \in \mathbb{R}^d$ is the mean of the k-th cluster, $\overline{\boldsymbol{\mathcal{X}}}$ is the mean of the whole data matrix $\boldsymbol{\mathcal{X}}$, and $\boldsymbol{S}_k$ is the covariance matrix associated to cluster k.

According to above, write a script to run clustering over both original and noisy frames, which can register the clustering performance (value of J) per every single value of $\widetilde{K}$. Plot some frames to depict the clustering effect. To do so, select meaningful frames for the considered experiments. Plot the value of J vs different values of $\widetilde{K}$. For each considered level noise, obtain the plotting and the value of J of the clustered sequence at time instance t. Vary the number of clusters, analyze and write the observations.

4) **(2.0)** Technical quality of the report and proper English language usage.

5) **(Extra +3.0)** Write a script to properly initialize the centers (not at random) and estimate the number of clusters for K-means. (Suggestion: [1], [4], [5]).

REFERENCES

[1] Diego Hernán Ordóñez-Peluffo. *Dynamic Spectral Clustering based on Kernels*. PhD thesis, 2013.

[2] O. R. Oña-Rocha, O. T. Sánchez-Manosalvas, A. C. Umaquinga-Criollo, P. D. Rosero-Montalvo, L. E. Suárez-Zambrano, J. L. Rodríguez-Sotelo, and D. H. Peluffo-Ordóñez. Automatic motion segmentation via a cumulative kernel representation and spectral clustering. In Hujun Yin, Yang Gao, Songcan Chen, Yimin Wen, Guoyong Cai, Tianlong Gu, Junping Du, Antonio J. Tallón-Ballesteros, and Minling Zhang, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2017*, pages 406–414, Cham, 2017. Springer International Publishing.

[3] Aristidis Likas, Nikos Vlassis, and Jakob J Verbeek. The global k-means clustering algorithm. *Pattern recognition*, 36(2):451–461, 2003.

[4] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734, 2000.

[5] Diego Hernán Peluffo Ordoñez et al. Estudio comparativo de métodos de agrupamiento no supervisado de latidos de señales ecg. Master's thesis, Universidad Nacional de Colombia-Sede Manizales, 2009.