Lecture 4

Fuzzy clustering by PCCA+

CONTENTS

A. Crisp clustering

Clustering or coarse-graining is a technique used in data analysis to group similar data points or objects together based on certain features or characteristics they share. In the context of dynamical systems (deterministic and stochastic), clustering refers to the discretization of the state space into subsets containing states with similar static and kinetic properties (e.g. equilibrium distribution and rates). Clustering is useful because it allows to represent continuous dynamic as a discrete process. Consequently, continuous objects (operators and functions) can be represented by discrete objects (matrices and vectors) that are easier to use in practical applications.

For example, consider a dynamical system defined on the state space $\Gamma \subset \mathbb{R}^{N_d}$ discretized with a Voronoi tessellation of k disjoint cells (clusters) Γ_i such that $\Gamma = \cup_i^k \Gamma_i$, where each cell Γ_i is defined by the indicator function

$$\mathbf{1}_i(x) = \begin{cases} 1 & \text{if } \; x \in \Gamma_i \; , \\ 0 & \text{if } \; x \notin \Gamma_i \; . \end{cases} \tag{1}$$

The choice of the tessellation is arbitrary, it could be a tessellation made of either regular or irregular N_d-polytopes (polygons in 2D, polyhedra in 3D).

This kind of clustering is known as crisp clustering and is largely used. However, it has limitations:

- Although dynamics is Markovian, its cluster representation may lose this property.
- Crisp clustering is not robust to noise. Small perturbations in the dynamics could be amplified in its cluster representation.
- Crisp clustering methods struggle with identifying clusters of irregular shapes or clusters that are connected but separated by sparse regions.
- When analysing metastable regions of a dynamical system, it is not always possible to uniquely determine the boundaries of metastable regions.
- Crisp clustering requires the specification of the number of clusters k a priori. Increasing the resolution, i.e. the number of clusters, may alleviate some problems, but the initial dataset may not have enough data points. Furthermore, a high number of clusters may require more resources to perform calculations with the matrices involved.

To address some of these limitations, researchers often explore alternative clustering approaches, such as fuzzy clustering, hierarchical clustering, and density-based clustering, which offer more flexibility and improved performance in specific scenarios.

B. Fuzzy clustering

Instead of assigning each data point to a single cluster (as in crisp clustering), fuzzy clustering assigns a membership value to each data point for each cluster, indicating the degree (or the probability) to which the point belongs to that cluster. This provides a more nuanced representation of the inherent uncertainty or ambiguity in the data.

Consider a dataset that can be divided into n_c clusters, then we introduce the membership function

$$\chi_i(x) \in [0,1]\,, \quad \text{with}\, i = 0, \ldots, n_c - 1\,, \tag{2}$$

also known as almost characteristic function, that indicates the probability, or membership degree, that a state x belongs to the ith cluster. The membership functions fulfil the partition

of the unit

$$\sum_{i=0}^{n_c-1} \chi_i(x) = 1\,. \tag{3}$$

Note that, if we indicate χ without the sub-index i, we refer to the set $\chi = \{\chi_0, \chi_1, \ldots, \chi_{n_c-1}\}$ containing all the n_c membership functions organized in columns. From a geometrical point of view, the points of the χ functions lie on the "standard" (n_c-1)-simplex as illustrated in fig. 1. The term "standard" indicates that the vertices of the simplex are the unit vectors $e_1, e_2, \ldots, e_{n_c}$.

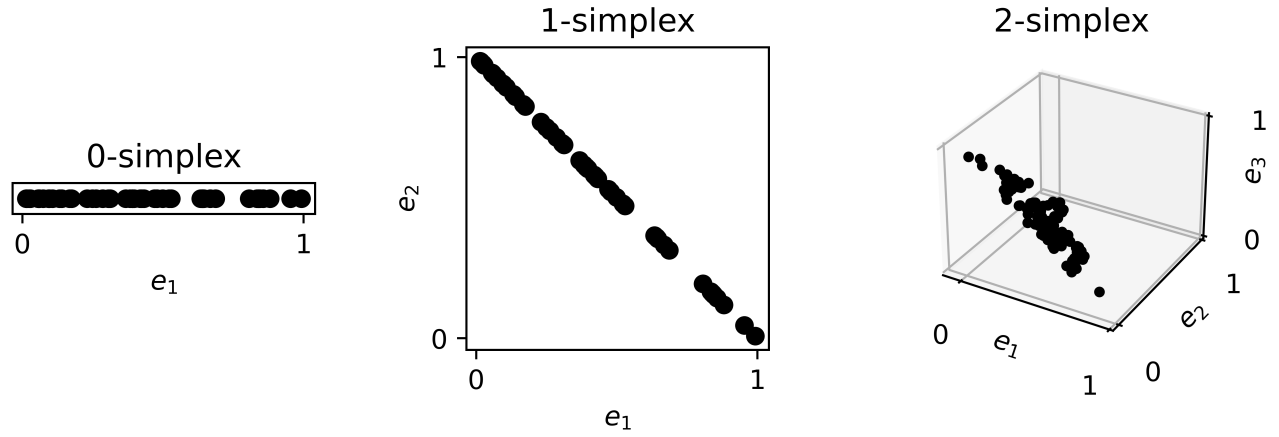


FIG. 1. Random points that fulfill the partition of unity.

C. Robust Perron Cluster Cluster Analysis

Consider now a dynamical system defined by a potential energy function $V(x) : \Gamma \subset \mathbb{R}^{N_d} \to \mathbb{R}$ with n_c minima separated by energy barriers higher than the thermal energy $k_B T$. Then, we state that the system is characterized by n_c metastable states, also referred to as metastable macro-state, or conformations. For example, given a double well potential: $n_c = 2$; for a triple-well potential $n_c = 3$. The dynamics can be described by the infinitesimal generator

$$\mathcal{Q}\psi_i = \kappa_i \psi_i\,, \tag{4}$$

or the associated Koopman operator

$$\mathcal{K}_\tau \psi_i = \lambda_{\tau,i} \psi_i\,, \tag{5}$$

where the first eigenfunction ψ_0 is equal to 1, while the other eigenfunctions have positive and negative values, and represent the dominant processes that constitute the dynamics of the system.

Similar to the χ functions, also the dominant n_c eigenfunctions form an (n_c-1)-simplex, the vertices of which, however, are not the unit vectors (see figs. E and F as examples). The idea underlying robust the Perron Cluster Cluster Analysis (PCCA+) algorithm [1–4] is then to find the linear transformation such that

$$\chi = \psi A\,, \tag{6}$$

where A is a matrix of size $n_c \times n_c$. In other words, PCCA+ determines the membership functions χ as a linear combination of the first n_c dominant eigenfunctions $\psi = \{\psi_0, \psi_1, \ldots, \psi_{n_c}\}$. The standard simplex, in the context of dynamical systems, has a physical interpretation: the vertices represent the conformations of the system, the points on the edges represent the transition regions. Additionally, the membership functions allow the direct Galerkin discretization of the infinitesimal generator

$$\mathbf{Q}_c = \langle \chi,\, \chi \rangle_\pi^{-1} \langle \chi,\, \mathcal{Q}\chi \rangle_\pi\,, \tag{7}$$

where $\mathbf{Q}_c$ is an $n_c \times n_c$ matrix whose entries expresses the transition rates between fuzzy sets.

D. Solution for $n_c = 2$

Unfortunately, determining the matrix A is not easy, as there are an infinite number of possible solutions, which can only be determined solving an optimization problem after appropriate objective functions have been defined [1]. However, for the sole case when $n_c = 2$, a unique solution can be determined.

If $n_c = 2$ the matrix A reads.

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11}\,. \end{pmatrix} \tag{8}$$

First, we pose the following three constraints on A:

1.

$$\chi = \psi A \rightarrow \chi_i(x) = \sum_{j=0}^{n_c-1} a_{ji}\psi_j(x)\,,$$

2.

$$\chi_i(x) \geq 0\,.$$

3.

$$\sum_{i=0}^{n_c-1} \chi_i(x) = 1\,,$$

Then, we rewrite the first condition as

$$\chi_i(x) = \sum_{j=0}^{n_c-1} a_{ji}\psi_j(x) \tag{9}$$

$$= a_{0i}\psi_0 + \sum_{j=1}^{n_c-1} a_{ji}\psi_j(x)\,, \tag{10}$$

and applying the second condition, we obtain an expression for a_{0i}:

$$a_{0i}\psi_0(x) + \sum_{j=1}^{n_c-1} a_{ji}\psi_j(x) \geq 0 \tag{11}$$

$$a_{0i}\psi_0(x) \geq -\sum_{j=1}^{n_c-1} a_{ji}\psi_j(x) \tag{12}$$

$$a_{0i} \geq -\sum_{j=1}^{n_c-1} a_{ji}\psi_j(x) \tag{13}$$

$$a_{0i} = \max_x \left(-\sum_{j=1}^{n_c-1} a_{ji}\psi_j(x) \right) \tag{14}$$

$$a_{0i} = -\min_x \sum_{j=1}^{n_c-1} a_{ji}\psi_j(x)\,, \tag{15}$$

where we used $\psi_0(x) = 1$ and $\max(-f) = -\min(f)$, for an arbitrary function f.

We now use the second condition to rewrite the third as

$$\sum_{i=0}^{n_c-1} \chi_i(x) = \tag{16}$$

$$\sum_{i=0}^{n_c-1} \sum_{j=0}^{n_c-1} a_{ji}\psi_j(x) = \tag{17}$$

$$\sum_{j=0}^{n_c-1} \sum_{i=0}^{n_c-1} a_{ji}\psi_j(x) = 1\,. \tag{18}$$

using $\psi_0(x) = 1$, we observe that

$$\sum_{j=0}^{n_c-1} \delta_{j0}\psi_j(x) = 1\,, \tag{19}$$

where δ_{j0} is the Kronecker-delta, then we obtain the equality

$$\delta_{j0} = \sum_{i=0}^{n_c-1} a_{ji} \,, \tag{20}$$

because $\psi_0(x) = 1$. Then, we write

$$\sum_{i=0}^{n_c-1} a_{ji} = \delta_{j0} \tag{21}$$

$$a_{j0} + \sum_{i=1}^{n_c-1} a_{ji} = \delta_{j0} \tag{22}$$

$$a_{j0} = \delta_{j0} - \sum_{i=1}^{n_c-1} a_{ji} \,, \,. \tag{23}$$

Applying $i = 0, 1$ and $j = 0, 1$ to eqs. 15, 23 yields:

$$\begin{cases} a_{00} = -\min\limits_x a_{10}\psi_1(x) \,, \\ a_{01} = -\min\limits_x a_{11}\psi_1(x) \,, \\ a_{00} = 1 - a_{01} \,, \\ a_{10} = -a_{11} \,. \end{cases} \tag{24}$$

From the forth equality, the first one becomes

$$a_{01} = a_{10} \max_x \psi_1(x) \,. \tag{25}$$

Finally we have two equations for ψ_1:

$$\max_x \psi_1(x) = \frac{a_{01}}{a_{11}} \,, \tag{26}$$

and

$$-\min_x \psi_1(x) = \frac{a_{01}}{a_{11}} \,. \tag{27}$$

The sum eq. 26 and eq. 27 yields

$$\max_x \psi_1(x) - \min_x \psi_1(x) = \frac{a_{00} + a_{01}}{a_{11}} = \frac{1}{a_{11}} \,. \tag{28}$$

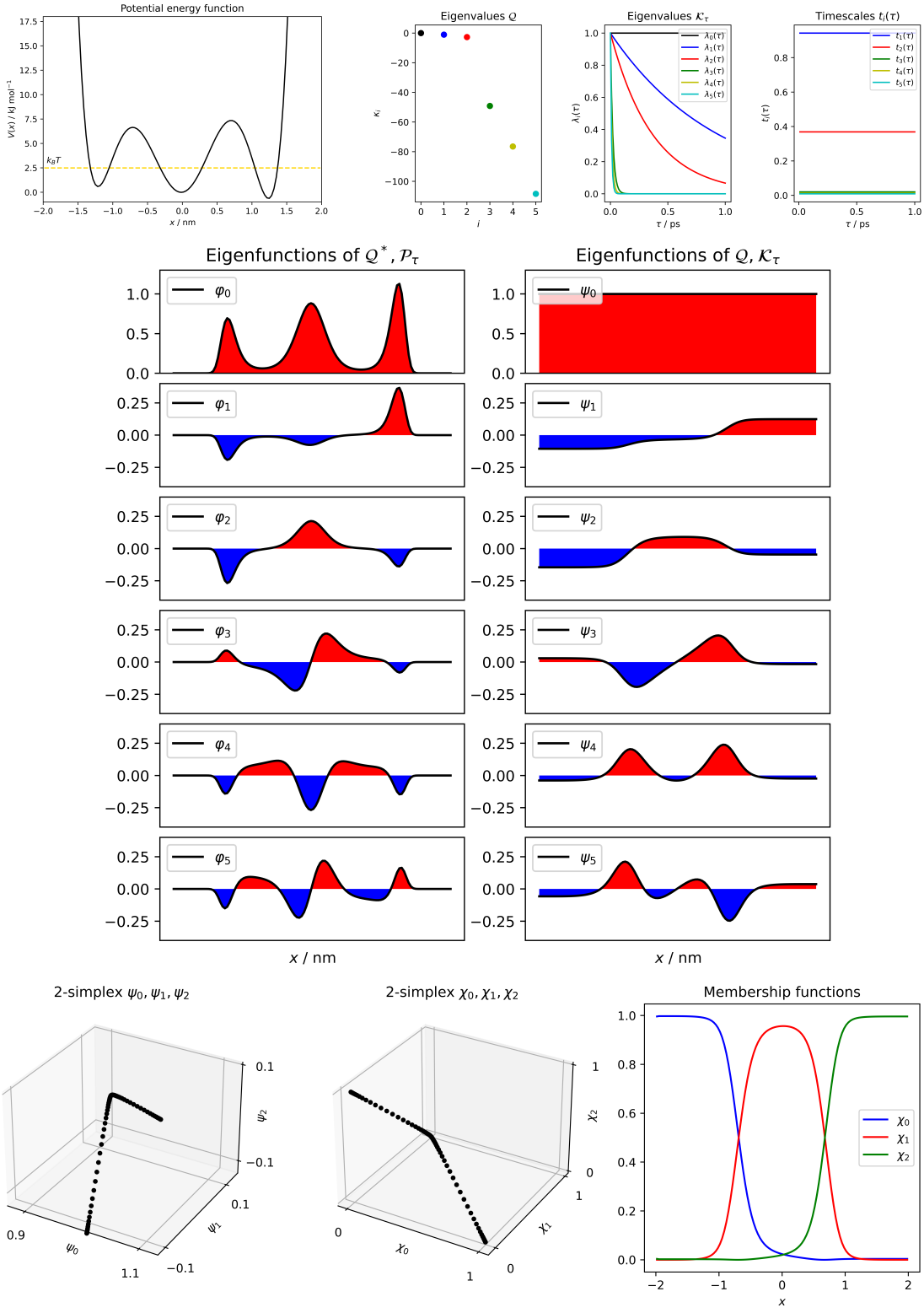
Thus one obtains an expression for each entry of the matrix A:

$$\begin{cases} a_{00} = \dfrac{\max_x \psi_1(x)}{\max_x \psi_1(x) - \min_x \psi_1(x)} \\ a_{01} = -\dfrac{\min_x \psi_1(x)}{\max_x \psi_1(x) - \min_x \psi_1(x)} \\ a_{10} = -\dfrac{1}{\max_x \psi_1(x) - \min_x \psi_1(x)} \\ a_{11} = \dfrac{1}{\max_x \psi_1(x) - \min_x \psi_1(x)} \,, \end{cases} \tag{29}$$

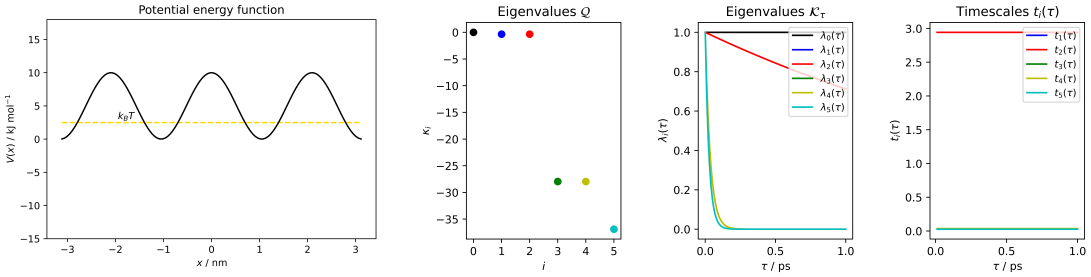
Finally, the two membership functions for the case $n_c = 2$ read

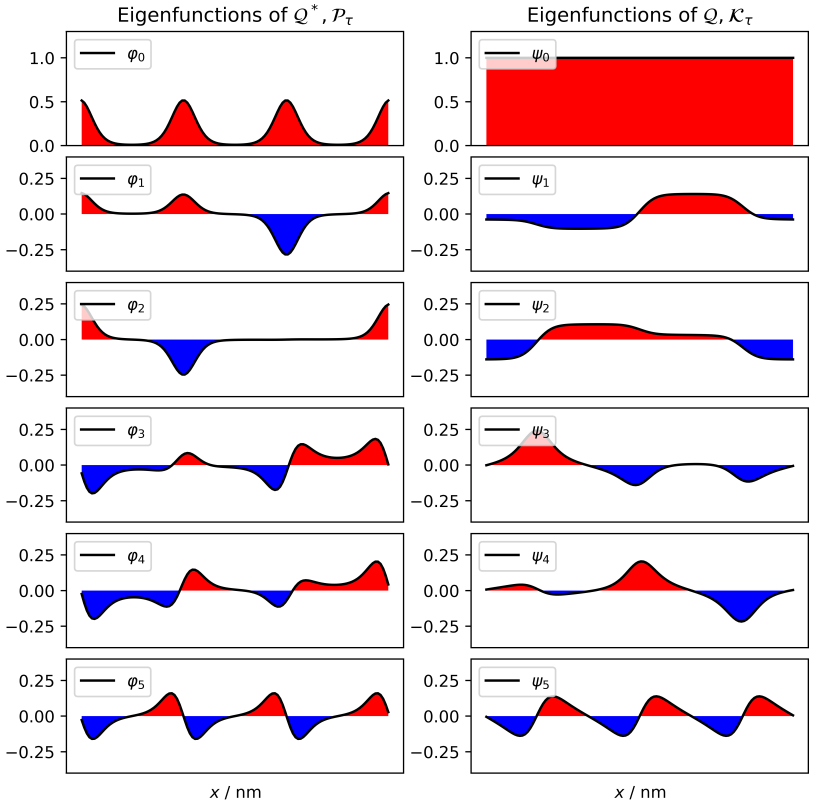
$$\begin{cases} \chi_0(x) = \dfrac{\max_x \psi_1(x) - \psi_1}{\max_x \psi_1(x) - \min_x \psi_1(x)} \,, \\ \chi_1(x) = \dfrac{\psi_1 - \min_x \psi_1(x)}{\max_x \psi_1(x) - \min_x \psi_1(x)} = 1 - \chi_0(x) \,. \end{cases} \tag{30}$$

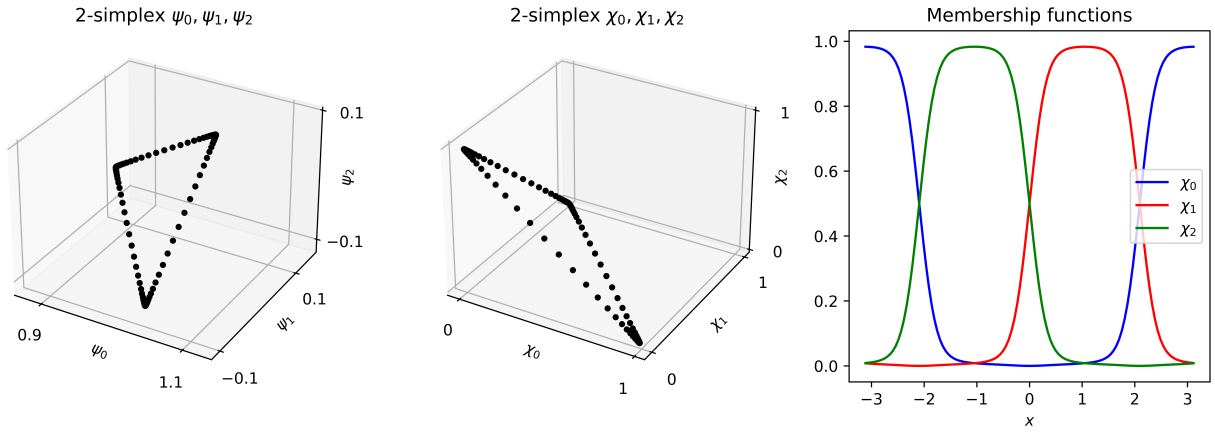
E. Example: triple-well potential



F. Example: periodic triple-well potential







[1] P. Deuflhard and M. Weber, Robust perron cluster analysis in conformation dynamics, Linear Algebra Appl. **398**, 161 (2004).

[2] S. Kube and M. Weber, A coarse graining method for the identification of transition rates between molecular conformations, J. Chem. Phys. **126**, 024103 (2007).

[3] S. Röblitz and M. Weber, Fuzzy spectral clustering by pcca+: Application to markov state models and data classification, Adv. Data Anal. Classif. **7** (2013).

[4] M. Weber, Implications of pcca+ in molecular simulation, Computation **6** (2018).