A Fuzzy Clustering Algorithm for the Mode-Seeking Framework

Thomas Bonis and Steve Oudot
DataShape Team
Inria Saclay

June 13, 2016

Abstract

In this paper, we propose a new fuzzy clustering algorithm based on the mode-seeking framework. Given a dataset in $\mathbb{R}^d$, we define regions of high density that we call cluster cores. We then consider a random walk on a neighborhood graph built on top of our data points which is designed to be attracted by high density regions. The strength of this attraction is controlled by a temperature parameter $\beta > 0$. The membership of a point to a given cluster is then the probability for the random walk to hit the corresponding cluster core before any other. While many properties of random walks (such as hitting times, commute distances, etc…) have been shown to enventually encode purely local information when the number of data points grows, we show that the regularization introduced by the use of cluster cores solves this issue. Empirically, we show how the choice of β influences the behavior of our algorithm: for small values of β the result is close to hard mode-seeking whereas when β is close to 1 the result is similar to the output of a (fuzzy) spectral clustering. Finally, we demonstrate the scalability of our approach by providing the fuzzy clustering of a protein configuration dataset containing a million data points in 30 dimensions.

1 Introduction

The analysis of large and possibly high-dimensional datasets is becoming ubiquitous in the sciences. The long-term objective is to gain insight into the structure of measurement or simulation data, for a better understanding of the underlying physical phenomena at work. Clustering is one of the simplest ways of gaining such insight, by finding a suitable decomposition of the data into clusters such that data points within a same cluster share common (and, if possible, exclusive) properties.

In this work, we are interested in the mode seeking approach to clustering. This approach assumes the data points to be drawn from some unknown probability distribution and defines the clusters as the basins of attraction of the maxima of the density, requiring a preliminary density estimation phase (Chen et al., 2014b; Chazal et al., 2013; Cheng, 1995; Cho and Lee, 2010; Comaniciu and Meer, 2002; Koontz et al., 1976). The theoretical analysis of this clustering framework has drawn increasing attention recently, see Chen et al. (2014a); Azizyan et al. (2015); Chen et al. (2015a,b); Arias-Castro et al. (2013). However, this (hard) clustering method provides a fairly limited knowledge on the structure of the data: while the partition into clusters is well understood, the interplay between clusters (respective locations, proximity relations, interactions) remains unknown. Identifying interfaces between clusters is the first step towards a higher-level understanding of the data, and it already plays a prominent role in some applications such as the study of the conformations space of a protein, where a fundamental question beyond the detection of metastable states is to understand when and how the protein can switch from one metastable state to another (Chodera et al., 2006). Hard clustering can be used in this context, for instance by defining the border between two clusters as the set of data points whose neighborhood (in the ambient space or in some neighborhood graph) intersects the two clusters, however this kind of information is by nature unstable with respect to perturbations of the data.

fuzzy clustering appears as the appropriate tool to deal with interfaces between clusters. Instead of assigning each data point to a single cluster, it computes a degree of membership to each cluster for each data point. The promise is that points close to the interface between two clusters will have similar degrees of membership to these clusters. Thus, fuzzy clustering uses a fuzzier notion of cluster membership in order to gain stability on the locations of the clusters boundaries.

Consider a smooth density f in $\mathbb{R}^d$. Under the mode seeking paradigm, clusters correspond to the modes of f. More precisely, considering the gradient flow induced by f:

$$y'(t) = \nabla f(u(t))$$

two points x and y are in the same cluster if the gradient flow started at x and the gradient flow started at y have the same limit which is a local maximum of f. A natural way to turn this approach into a fuzzy clustering algorithm is to follow a perturbed gradient flow instead, such as the diffusion process solution of

$$dY_t = \frac{1}{\beta}\nabla(\log f)dt + dB_t, \tag{1}$$

where B_t is a d-dimensional Brownian motion and β is a temperature parameter controlling the amount of noise introduced in the gradient flow. We use the gradient of the logarithm of f here as this quantity arises naturally in practice. Indeed, since we only have access to a discretization of the space through the sampled data points, we mimic this perturbed gradient flow by a random walk on the data points. Ting et al. (2010) proved an isotropic random walk on a neighborhood graph approximates the previous diffusion process for $\beta = 1$ while other values of β are obtained by putting weights on the edges of the graph. At this point, one could perform fuzzy clustering by considering the first local maximum of the density encountered by the random walk, an approach wich has been proposed by Chen et al. (2014b). However, as emphasized by Luxburg et al. (2010), the hitting time to a single point for a random walk on the graph converges to irrelevant quantities when the number of data points goes to infinity. We can thus expect the clustering to fail in that case. Indeed, if we apply this method to the fuzzy clustering of two different Gaussian measures (see Figure 1). The obtained fuzzy

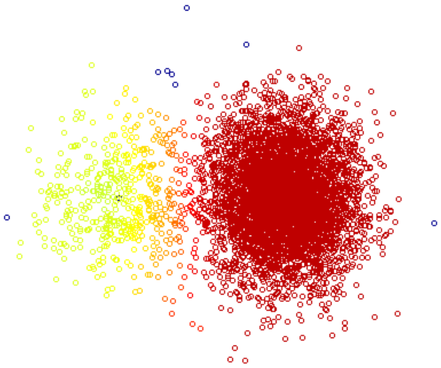


Figure 1: Fuzzy clustering output for an unbalanced mixture of gaussian. Red color corresponds to the right cluster, blue to the left one. Finally, green points have similar membership to both clusters.

memberships are unsatisfying. In order to circumvent this issue, we assign a zone of high density to each cluster, called *cluster core* and computed using the mode-seeking (hard) clustering algorithm ToMATo (Chazal et al., 2013). The fuzzy membership of a point to a given cluster is then given by the probability for the random walk started at this point to hit the corresponding cluster core first.

After a presentation of the algorithm, we provide convergence guarantees for its output. We then study the empirical behaviour of our algorithm on synthetic data and provide a method to select relevant values for the temperature β. Finally, we perform some quantitative experiments on a few UCI datasets.

2 The Algorithm

Our algorithm is a fuzzy generalization of the ToMATo algorithm which relies on the concept of *prominence*. Let G be a graph and f be a real valued function on the vertices of this graph. For any $\alpha \in \mathbb{R}$, let $F^\alpha = f^{-1}([\alpha, +\infty])$ be the α-superlevel-set of f. A new connected component C is born in F^α when α reaches a local maximum of f on G and we denote by $\alpha_{b,C}$ the corresponding value of α. This component then dies at $\alpha = \alpha_{d,C} < \alpha_{b,C}$ when it gets connected, in F^α, to another connected component C' such that $\alpha_{b,C'} > \alpha_{b,C}$. The prominence of C (and by extension, of the corresponding local maximum of f) is then simply $\alpha_{b,C} - \alpha_{d,C}$.

The algorithm takes as input a finite set of points $\mathcal{X} = \{X_1, \cdots, X_n\}$ together with pairwise distances $d(X_i, X_j)$. In practice only the distances are used, so there is no need for point coordinates. Additionally, the algorithm takes in the following set of parameters:

- a density estimator $\hat{f} : \mathcal{X} \to \mathbb{R}$,
- a kernel k, for example the Gaussian kernel,
- a window size $h > 0$,
- a prominence threshold $\tau > 0$,
- a temperature $\beta > 0$.

The first four parameters are in fact required by ToMATo for hard mode-seeking, upon which our algorithm relies. The last parameter is the one added in for fuzzy mode-seeking, as per Equation (1).
Given this input, our algorithm proceeds as follows:

1. It builds a weighted neighborhood graph G on top of the point cloud $\mathcal{X}$, adding an edge with weight

$$w_{i,j} = \left(1 + \frac{1-\beta}{\beta}\hat{f}(X_j)\right) k\left(\frac{d(X_i, X_j)}{h}\right) \tag{2}$$

   between each pair of points (X_i, X_j). Remark that it is possible to replace our kernel-based graph by a nearest neighbour graph.

2. It computes the cluster cores by running ToMATo with input $\mathcal{X}$, d, $\log(\hat{f})$, τ and the unweighted neighborhood graph $\bar{G}$ obtained from G by removing the edges with weights lower than $0.5\max(k)$. The output of ToMATo is a set of K clusters $C_1, \cdots, C_K$. Each cluster C_i corresponds to the basin of attraction of some peak of $\log(\hat{f})$ of prominence at least τ within $\bar{G}$. Up to a reordering of the data points, we can assume this peak to be X_i. The i-th cluster core $\mathcal{C}_i$ is then taken to be the *highest and most stable part* of C_i, defined formally as the connected component containing X_i within the subgraph of $\bar{G}$ spanned by those vertices X_j such that $\log(\hat{f})(X_j) > \log(\hat{f})(X_i) - \tau/2$.

3. It computes the fuzzy-membership values $\mu_1, \ldots, \mu_K$ by solving the linear system $A^T\mu = \mu$, where the matrix A is defined by:

$$A_{kl} = \begin{cases} \delta_{kl} & \text{if } X_k \text{ belongs to some cluster core} \\ K^h(X_k, X_l) & \text{otherwise,} \end{cases}$$

   where K^h is the transition kernel of the random walk on the graph, i.e.

$$K^h(X_i, X_j) = \frac{w_{i,j}}{\sum_z w_{i,z}}. \tag{3}$$

The output of the algorithm is the set of fuzzy-membership values $\mu_1, \ldots, \mu_K$ computed at step 3.

3 Parameters selection

3.1 Density estimator, window size, kernel and prominence threshold

These 4 parameters are tied to the classical hard mode-seeking framework. The density estimator can be linked to the window size in practice, as is done e.g. in Mean-Shift (Cheng, 1995) and its successors. For instance, one can consider the kernel density estimator associated to the kernel k. This not only reduces the number of parameters to tune in practice, but it also gives a way to select h using standard parameter selection techniques for density estimation, which is done for example in Chen et al. (2014b). Finally, the prominence threshold τ is used to distinguish between relevant and irrelevant peaks in the discrete setting. It can be selected by running ToMATo twice: once to get the distribution of prominences of the peaks of $\hat{f}$ within the neighborhood graph $\bar{G}$, from which τ can be inferred by looking for a gap in the distribution; then a second time, using the chosen value of τ, to get the final hard clustering. This procedure is detailed in Chazal et al. (2013).

3.2 Temperature parameter

This parameter is standard in fuzzy clustering. Outputs corresponding to large values of β will tend to have smooth interfaces between clusters, while small values of β will encourage quick transitions from one cluster to another. β can also be interpreted as a trade-off between the respective influence of the metric and of the density in the diffusion process: when β is small, the output of our algorithm is mostly guided by the density and therefore close to the output of mode seeking algorithms; by contrast, when β is large, the algorithm becomes oblivious to the density. In practice, one may get insights into the choice of β by looking at the evolution of a certain measure of *fuzziness* of the output clustering across a range of values of β. We elaborate on this in Section 5.

4 Convergence guarantees

In this section we provide guarantees to our fuzzy clustering scheme by exploiting the convergence of the random walk over the neighborhood graph to a continuous diffusion process.

As is usual in mode-seeking, we assume our input data points $X_1, ..., X_n$ to be i.i.d random variables drawn from some unknown probability density f over $\mathbb{R}^d$. We also assume that the metric d that equips the data points is the Euclidean norm, and that f satisfies the following technical conditions:

- f is Lipschitz continuous over $\mathbb{R}^d$ and C^1-continuous over the domain $\Omega = \{x \in \mathbb{R}^d \mid f(x) > 0\}$,
- $\lim\limits_{\|x\|_2 \to \infty} f(x) = 0$,

- The SDE 1 is well-posed.

Standard sufficient conditions ensuring the well-posedness (particularly the non-explosion) of the SDE 1 can be found in Albeverio et al. (2003) or in Krylov and Röckner (2005), for example one can assume $\nabla \log f$ to be Lipschitz continuous.

Our analysis connects random walks on graphs built on top of the input point cloud $\mathcal{X}$ using a density estimator $\hat{f}$ to the solution of Equation 1, for a fixed temperature parameter $\beta > 0$. Specifically, let $M^{x,h}$ denote the Markov Chain whose initial state is the closest neighbour of x in the point cloud $\mathcal{X}$ (break ties arbitrarily), and whose transition kernel K^h is given by Equation 3. Following the approach of Ting et al. (2010), we show that, under suitable conditions on the estimator $\hat{f}$, this graph-based random walk approximates the diffusion process in the continuous domain in the following sense: there exists s depending on h such that, as n tends to infinity, with high probability, $M^{x,h}_{\lfloor t/s \rfloor}$ converges weakly to the solution of Equation (1). From there, under standard conditions for mode estimation on the window size h and on the density estimator $\hat{f}$ (see Chen et al. (2015b); Arias-Castro et al. (2013)), we obtain the convergence of the fuzzy-membership values $\mu_i(x)$ computed by the algorithm to the membership defined from the underlying continuous diffusion process $\tilde{\mu}_i(x)$. Formally, letting $v_1, \ldots, v_K$ be the local maxima of f of prominence higher than τ, and $\tilde{\mathcal{C}_1}, \cdots, \tilde{\mathcal{C}_K}$, their associated cluster cores in the continuous domain (i.e. $\tilde{\mathcal{C}_i}$ is the connected component containing v_i in $\{x \in \mathbb{R}^d | \log f(x) \geq \log f(v_i) - \tau/2\}$), we define $\tilde{\mu}_i(x)$ as the probability for the diffusion process solution of (1) to hit $\tilde{\mathcal{C}_i}$ before any other $\tilde{\mathcal{C}_j}$.

Theorem 1. *Let $\beta > 0$ and assume $\|\nabla f\|$ is bounded from below on the boundary of the underlying cluster cores $\tilde{C}$. Let $h : \mathbb{N} \to \mathbb{R}^+$ be a decreasing window size such that $\lim\limits_{n\to\infty} h(n) = 0$ while $\lim\limits_{n\to\infty} \frac{h(n)^{d+2} n}{\log n} = \infty$. Suppose the density estimator $\hat{f}_n$ satisfies, for any compact set $U \subset \Omega$ and any $\epsilon > 0$,*

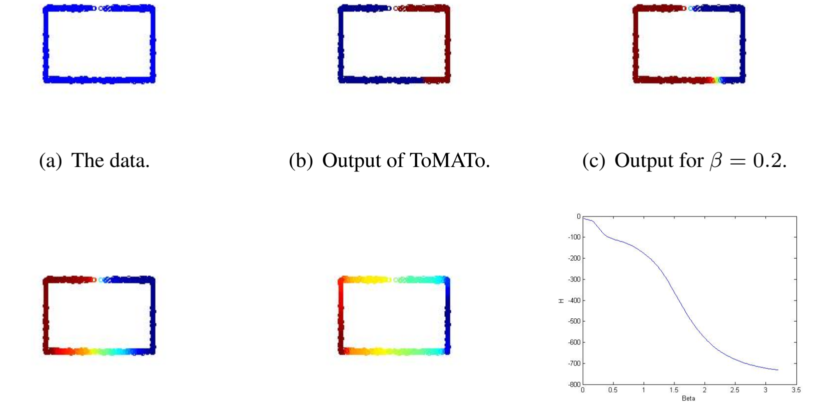
$$\lim_{n\to\infty} \mathbb{P}(\sup_{x \in U} |\nabla f(x) - \nabla \hat{f}_n(x)| \geq h(n)^2 \epsilon) = 0.$$

Then, for any compact set $U \subset \Omega$, any $\epsilon > 0$ and any i,

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{x \in U} |\mu_i(x) - \tilde{\mu}_i(x)| \geq \epsilon\right) = 0.$$

5 Experiments

We first illustrate the effect of the temperature parameter β on the clustering output using synthetic data. We then apply our method on a couple UCI repository datasets and on simulated protein conformations data. In all our experiments we use a k-nearest neighbor graph along with a distance to measure density estimator (Biau et al., 2011) computed using the $2k$ nearest-neighbors.



(a) The data.

(b) Output of ToMATo.

(c) Output for $\beta = 0.2$.

(d) Output for $\beta = 1$.

(e) Output for $\beta = 2$.

(f) Evolution of H with respect to β.

Figure 2: Output of our algorithm on a simple dataset composed of two overlapping clusters. For fuzzy clustering green corresponds to an equal membership to both clusters.

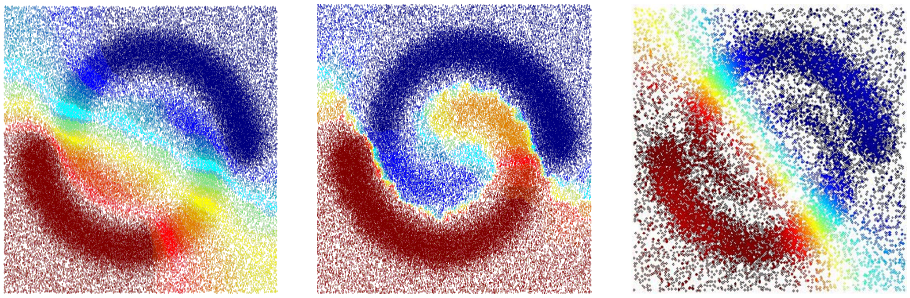
5.1 Synthetic data

The first dataset is presented in Figure 2(a) and is composed of two high-density clusters connected by two links. The bottom link is sampled from a uniform density while the top link is sampled from a density that has a gap inbetween the two clusters. Standard mode seeking algorithms will have a hard time clustering the bottom link as a density estimation can create many "noisy" local maxima: for instance, ToMATo miss-clusters most of the bottom link (see Figure 2(b)). We display the results of our algorithm for three values of β : $\beta = 0.2$ in Figure 2(c), $\beta = 1$ in Figure 2(d) and $\beta = 2$ in Figure 2(e). As we can see from the output of the algorithm, for small values of β, the amount of noise injected in our trajectory is not large enough to compensate for the influence of the noise in the density estimation, so the result obtained is really close to hard clustering. Large values of β do not give enough weight to the density function which leads to a smooth transition between the two clusters on the top link. Intermediate values of β seem to give more satisfying results. In order to gain intuition regarding which value of β one should use, it is possible to look at the evolution of a fuzziness value for the clustering. For example, one can consider a notion of clustering entropy:

$$H = \sum_i \sum_j \mu_j(X_i) \log(\mu_j(X_i)), \tag{4}$$

which gets lower when the fuzziness of the clustering increases. As we can see in Figure 2(f), the evolution of H with respect to β presents three distincts plateaus corresponding to the three behaviour highlighted earlier.

The second dataset we consider is composed of two interleaved spirals—see Fig-



(a) Output for $\beta = 1$.

(b) Output for $\beta = 0.28$.

(c) Fuzzy Spectral Clustering.

Figure 3: Experiments on a cluttered spirals dataset.

ure 3. An interesting property of this dataset is that the head of each spiral is close (in Euclidean distance) to the tail of the other spiral. Thus, the two clusters are well-separated by a density gap but not by the Euclidean metric. We use our algorithm with two different values of β: $\beta = 1$ and $\beta = 0.3$. We also run the spectral fuzzy-C means on a subsampling of this dataset. The first thing we want to emphasize is that the result of spectral clustering and our algorithm using $\beta = 1$ are similar, this is to be expected as both algorithms rely on properties of the same diffusion operator, this also means that other fuzzy clustering techniques based on spectral clustering will fail on this dataset. Moreover, we can see that for $\beta = 1$, the density gap between the two spirals is not strong enough to compensate for the proximity of the two clusters in the Euclidean metric. On the other hand, for $\beta \simeq 0.3$ we recover the two clusters as we give more weight to the density structure.

5.2 UCI datasets

In order to obtain quantitative results regarding our fuzzy clustering scheme, we evaluate it in a classification scenario on a few datasets from the UCI repository: the Pendigits dataset ($10,000$ points and 10 clusters), the Waveset dataset (5000 points and 3 clusters) and the Statlog dataset ($6,435$ points for 7 clusters). We preprocess each dataset by renormalizing the various coordinates so they have unit variance. Then, for each dataset, we run our algorithm with various values of the parameter β between 0.3 and 5, but a single value of k and τ (given by a prominence gap), along with the fuzzy C-means algorithm for fuzziness parameters between 1.2 and 5. We also consider the fuzzy clustering algorithm proposed by Chen et al. (2014b), for which the cluster cores are reduced to a single point. Let $X_1, \dots, X_n$ denote our sample points and $Y_1, \dots, Y_n$ their respective labels taking values in $\{1, \dots, K'\}$. In these datasets, there are only two plateaus, thus we choose β. Thus, we propose an automatic selection of β by computing the values of the clustering entropy H for multiple values of β and by selecting

$$\beta = arg\,max \frac{dH}{d\beta},$$

Algorithm / Data	Waveform	Pendigits	Statlog
Ours with optimal β	-1.1	-0.61	-0.51
Ours with automatic β selection	-1.1	-0.64	-0.55
FCM	-1.1	-1.35	-0.57
fuzzy mode-seeking of Chen et al. (2014b)	-3.2	-0.76	-0.58

Table 1: Entropic purity obtained by fuzzy clustering algorithms on UCI datasets.

in other words we take β inbetween the two plateaus by choosing the value of β maximizing the slope of H. In order to evaluate hard clustering algorithms, it is common to use the purity measure defined by

$$P = \max_{\pi} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} 1_{\tilde{\mu}_j(X_i)=1} 1_{Y_i=\pi(j)},$$

where π is a map from the set of clusters $\{1,\ldots,K\}$ to the set of labels $\{1,\ldots,K'\}$. As this measure is not adapted to fuzzy clustering, we define the *ϵ-entropic purity* as

$$HP_{\epsilon} = \max_{\pi} \frac{1}{n} \sum_{i} \log\left(\epsilon + \sum_{j,\pi(j)=Y_i} \tilde{\mu}_j(X_i)\right),$$

for some $\epsilon > 0$. The ϵ parameter is used to prevent the quantity from exploding due to possible outliers. This extension of the traditional purity can be useful to evaluate fuzzy clustering as it can be seen as an approximation of $\mathbb{E}[\log(\epsilon + \sum_{j,\pi(j)=Y} \tilde{\mu}_j(X))]$ which enjoys the following property.

Proposition 2. *Suppose that $Y \in \{1,\ldots,K\}$ and let $\epsilon > 0$, then*

$$arg\,max_{f\in\mathbb{R}^d\to\mathbb{R}^K,\|f\|_1=1}\mathbb{E}[\log(\epsilon + f(X))] = (1-\epsilon)^{-1}(\mathbb{P}(Y=j \mid X))_{1\le j\le K} - \epsilon.$$

Thus, for small values of ϵ, a fuzzy clustering minimizing the ϵ-entropic purity recovers the conditional probabilities of the labels with respect to the coordinates.

We provide the best 10^{-3}-entropic purity obtained by each algorithm on all datasets in Table 5.2. As we can see, our algorithm outperforms the other fuzzy clustering algorithms on these datasets. In particular we can see that the simple fuzzy mode-seeking algorithm of Chen et al. (2014b) fails on the Waveform dataset.

Alanine dipeptide conformations. We now turn to the problem of clustering protein conformations. We consider the case of the alanine-dipeptide molecule. Our dataset is composed of $1,420,738$ protein conformations, each one represented as a 30-dimensional vector. The metric used on this type of data is the root-mean-squared deviation (RMSD). The goal of fuzzy clustering in this case is twofold: first, to find the right number of clusters corresponding to metastable states of the molecule; second, to find the conformations lying at the border between different clusters, as these represent the transition phases between metasable states. It is well-known that the conformations of alanine-dipeptide only have two relevant degrees of freedom, so it is possible to project the data down to two dimensions (called a Ramachadran plot) to have a comfortable view of the clustering output. See Figure 4 for an illustration, and note that the clustering is performed in the original space. In order to highlight interfaces between clusters, we only display the second highest membership function. As we can see there are 5 clusters and 6 to 7 interfaces.

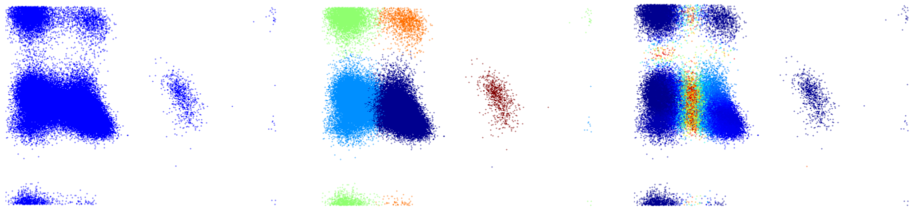


Figure 4: From left to right: (a) the dataset projected on the Ramachadran plot, (b) ToMATo output, (c) second highest membership obtained with our algorithm for $\beta = 0.2$

6 Proofs

6.1 Background on diffusion processes

Convergence of Markov chains to diffusion processes occurs in the *Skorokhod space* $D([0,T],\mathbb{R}^d)$, composed of the trajectories $[0,T] \to \mathbb{R}^d$ that are right-continuous and have left limits, for some fixed $T > 0$. It is equipped with the following metric:

$$d(f,g) = \inf_{\epsilon}\{\exists \lambda \in \Lambda, \|\lambda\| \leq \epsilon, \sup_t |f(t) - g(\lambda(t))| \leq \epsilon\},$$

where Λ denotes the space of strictly increasing automorphisms on the unit segment $[0,1]$, and where $\|\lambda\|$ is the quantity:

$$\|\lambda\| = \sup_{s \neq t} \left| \log \left(\frac{\lambda(t) - \lambda(s)}{t - s} \right) \right|.$$

In diffusion approximation, standard results prove the weak convergence of a Markov chain to a difussion process in $D([0,T],\mathbb{R}^d)$. A stochastic process M^s converges weakly to a diffusion process Y in $D([0,T],\mathbb{R}^d)$ as s tends to 0 if and only if

$$\lim_{s \to 0} \mathbb{P}(M^s \in B) = \mathbb{P}(Y \in B) \tag{5}$$

for any Borel set B such that $\mathbb{P}(Y \in \partial B) = 0$.

Let us state the convergence result when Y is the Solution of the Stochastic Differential Equation 1. For this case, $b = \frac{1}{\beta}\nabla \log f$ and $a = I_d$. Consider a family of Markov chains $(M^{x_0,s})$ defined on discrete state spaces $S_s \subset \Omega$, transition kernels K^s and initial states $M_0^{x_0,s} \in S_s$. For $x \in S_s$ and $\gamma > 0$, let

- $a^s(x) = \frac{1}{s}\sum_{y \in S_s} K^s(x,y)(y-x)(y-x)^T$;
- $b^s(x) = \frac{1}{s}\sum_{y \in S_s} K^s(x,y)(y-x)$;
- $\Delta_s^\gamma = \frac{1}{s}K^s(x, \mathcal{B}(x,\gamma)^c)$,

where $\mathcal{B}(x,\gamma)^c$ is the complementary of the ball of radius γ centered at x.

Proposition 3 (Adapted from Theorem 7.1 in Durrett (1996))**.** *Let U be a compact subset of Ω. Let also B be a Borel set in $D([0,T],\mathbb{R}^d)$ for some $T > 0$ such that $\mathbb{P}(Y^{x_0} \in \partial B) = 0$ for all $x_0 \in U$. For any $\epsilon > 0$, there exist parameters ν and γ such that*

$$\sup_{x_0 \in U} |\mathbb{P}(M^{x_0,s}_{\lfloor t/s \rfloor} \in B) - \mathbb{P}(Y_t^{x_0} \in B)| \leq \epsilon$$

whenever the following conditions are met:

(i) $\sup_{x \in S_s} \|a^s - a\|_\infty \leq \nu$;

(ii) $\sup_{x \in S_s} \|b^s - b\|_\infty \leq \nu$;

(iii) $\sup_{x \in S_s} \Delta_s^\gamma \leq \nu$;

(iv) $\sup_{x_0 \in S_s} \|M_0^{x_0,s} - x_0\|_\infty \leq \nu$.

6.2 Weak-Convergence

In this section, we prove the following result.

Proposition 4. *Let Y be the diffusion process solution of the SDE 1. Let $h : \mathbb{N} \to \mathbb{R}^+$ be a decreasing function such that $\lim\limits_{n\to\infty} h(n) = 0$ and $\lim\limits_{n\to\infty} \frac{h(n)^{d+2}n}{\log n} = \infty$. Suppose our estimator $\hat{f}_n$ satisfies, for any compact set $U \subset \Omega$ and any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\sup_{x \in C} |f(x) - \hat{f}_n(x)| \geq h(n)^2\epsilon) = 0.$$

Then, for any $T, \epsilon > 0$, for any compact set $U \subset \Omega$, and for any Borel set B of $D([0,T],\mathbb{R}^d)$ such that $\mathbb{P}(Y^y \in \partial B) = 0$ for all $y \in U$, there exists a constant C depending on d *such that for $s(n) = Ch^2$, we have*

$$\lim_{n\to\infty} \mathbb{P}(\sup_{x \in U} |\mathbb{P}(M^{x,h(n)}_{\lfloor t/s(n) \rfloor} \in B) - \mathbb{P}(Y_t^x \in B)| \geq \epsilon) = 0.$$

The proof relies on Theorem 3 of Ting *et al.* (2010) along with a proper control of boundary effects. Let T and ϵ be strictly positive real numbers, throughout the course of the proof, the notation $M^{x,h(n)}$ stands for the continuous time process $M^{x,h(n)}_{\lfloor t/s(n) \rfloor}$. We denote by $\mathcal{X}_n = (X_1, ..., X_n)$ the i.i.d sampling which is also the state space of $M^{x,h(n)}$. For $\alpha > 0$, let $F^\alpha = \{x \in \mathbb{R}^d \mid f(x) \geq \alpha\}$ be the α superlevel-set of f and $B_\alpha = \{w \in D([0,T],\mathbb{R}^d) \mid \forall t, w(t) \in F^\alpha\}$ be trajectories staying in F^α up to time T. Since Y does not explode in finite time, there exists α such that, for any $x \in U$, $\mathbb{P}(Y^x \in B_\alpha) \geq 1 - \epsilon/4$. To obtain a good approximation of the trajectories of Y staying in F^α using $M^{x,h(n)}$, we only need to check assumptions (i)-(iv) of Proposition 3 on F^α. F^α is closed as f is continuous and it is also bounded as $\lim_{\|x\|_2 \to \infty} f(x) = 0$, it is therefore compact. Applying Theorem 3 from Ting *et al.* (2010) on the points of the compact set F^α, we have, with probability 1,

(i) $\lim\limits_{n\to\infty} \mathbb{P}(\sup_{y \in \mathcal{X}_n \cap F^\alpha} \|a^s - I_d\|_\infty \leq \nu) = 0$,

(ii) $\lim\limits_{n\to\infty} \mathbb{P}(\sup_{y \in \mathcal{X}_n \cap F^\alpha} \|b^s - \frac{\nabla f}{\beta f}\|_\infty \leq \nu) = 0$,

(iii) $\sup_{y \in \mathcal{X}_n \cap F^\alpha} \Delta_s^h = 0$,

(iv) $\lim\limits_{n\to\infty} \mathbb{P}(\|M_0^{x,h(n)} - x\|_\infty \leq \nu) = 0$.

Thus, the assumptions (i)-(iv) of Proposition 3 are verified on F^α.

Since f is continuous, B_α is an open set. Therefore, there exists $n_0 > 0$ such that for any $n > n_0$,

$$\sup_{x \in U} \mathbb{P}(M^{x,h(n)} \in B_\alpha) \geq \sup_{x \in U} \mathbb{P}(Y^x \in B_\alpha) - \epsilon/4 \geq 1 - \epsilon/2.$$

Therefore, for any Borel set B,

$$\sup_{x \in U} |\mathbb{P}(M^{x,h(n)} \in B) - \mathbb{P}(M^{x,h(n)} \in B \cap B_\alpha)| \leq \epsilon/2.$$

Thus, we only need to approximate trajectories that do not leave F^α to obtain a good approximation of $\mathbb{P}(M^{x,h(n)} \in B)$. So we can apply Corollary 3 on these trajectories with an accuracy of $\epsilon/2$ to obtain,

$$\sup_{x \in U} |\mathbb{P}(M^{x,h(n)} \in B) - \mathbb{P}(Y^x \in B)| \leq \epsilon.$$

Every step of the proof hold almost surely as n tends to infinity, thus the proof of Proposition 4 is complete.

6.3 Proof of Theorem 1

Let β and τ be strictly positive real numbers and let $U \subset \Omega$ be a compact set. Let $\mathcal{C}_1, \ldots, \mathcal{C}_K$ be the cluster cores used by the algorithm and computed with the density estimator $\hat{f}$. These cluster cores are approximations of the sets $\tilde{\mathcal{C}_1}, \ldots, \tilde{\mathcal{C}_K}$ obtained using the same computation with the true density f. Since, by assumptions, f is $\mathcal{C}^1$-continuous on Ω and $\|\nabla f\|$ is non-zero on the boundary of the $\tilde{\mathcal{C}}$, we have

(i) The $\tilde{\mathcal{C}_i}$ are compact sets of $\mathbb{R}^d$ that are well-separated (i.e. $\tilde{\mathcal{C}_j} \cap \tilde{\mathcal{C}_j} = \emptyset$ for all $i \neq j$).

(ii) For each i, the boundary of $\tilde{\mathcal{C}_i}$ is smooth.

By our assumptions on the convergence of $\hat{f}$ along with Theorem 10.1 of Chazal et al. (2013),

$$\forall \delta > 0, \lim_{n \to \infty} \mathbb{P}(\tilde{\mathcal{C}}_i^{-\delta} \subset \mathcal{C}_i \subset \tilde{\mathcal{C}}_i^{\delta}) = 1, \tag{6}$$

where $\tilde{\mathcal{C}}_i^{\delta} = \cup_{x \in \tilde{\mathcal{C}}_i} \mathcal{B}(x, \delta)$ and $\tilde{\mathcal{C}}_i^{-\delta} = \tilde{\mathcal{C}}^i \setminus \cup_{x \notin \tilde{\mathcal{C}}_i} \mathcal{B}(x, \delta)$.

Without loss of generality, we can assume that Ω has a single connected component. Let ϵ be a strictly positive real and consider $x \in \Omega$, we let

- $\mu_{i,\delta}^{+}(x)$ be the probability that Y^x hits $\tilde{\mathcal{C}}_i^{\delta}$ before any other $\tilde{\mathcal{C}}_j^{-\delta}$,
- $\mu_{i,\delta}^{-}(x)$ be the probability that Y^x hits $\tilde{\mathcal{C}}_i^{-\delta}$ before any other $\tilde{\mathcal{C}}_j^{\delta}$.

Let us show that, for any i, a trajectory entering $\tilde{\mathcal{C}}_i^{\delta}$ has a high probability to enter $\tilde{\mathcal{C}}_i$ if δ is small enough. Since the $\tilde{\mathcal{C}}_i$ are closed and disjoint there exists $\delta_0 > 0$ such that the $\tilde{\mathcal{C}}_i^{\delta_0}$ are disjoints. Moreover, since the $\tilde{\mathcal{C}}_i$ have smooth boundaries, there exists $\delta_i^{+} > 0$ such that if $d(x, \tilde{\mathcal{C}}_i) \leq \delta_i^{+}$ then, the probability for Y^x to hit $\tilde{\mathcal{C}}_i$ before exiting $\tilde{\mathcal{C}}_i^{\delta_0}$ is at least $1 - \epsilon/8$.

Similarly, if a trajectory enters $\tilde{\mathcal{C}}_i$, then it enters $\tilde{\mathcal{C}}_i^{-\delta}$ with high probability. More precisely there exists δ_i^{-} such that if a trajectory hits $\tilde{\mathcal{C}}_i$, then it hits $\tilde{\mathcal{C}}_i^{-\delta}$ with probability at least $1 - \epsilon/8$.

Let $\delta = \min(\delta_j^{+}, \delta_j^{-})$, by combining our results and using the strong Markov property of Y^x we obtain that:

- $\mu_{i,\delta}^{+}(x) - \mu_i(x) \leq \epsilon/4$,
- $\mu_i(x) - \mu_{i,\delta}^{-}(x) \leq \epsilon/4$.

The next step is to show that the approximation of $\mu_{i,\delta}^{+}$ provided by the Markov chain is correct. For $T > 0$, let

$$B = \{w \in D([0, \infty], \mathbb{R}^d \mid \exists \tau \text{ such that } w(\tau) \in \tilde{\mathcal{C}}_i^{\delta} \\ \text{and } \forall t < \tau \text{ we have } w(t) \in \Omega \setminus \cup_j \tilde{\mathcal{C}}_j^{-\delta}\},$$

$$B_T = \{w \in D([0, T], \mathbb{R}^d \mid \exists \tau \text{ such that } w(\tau) \in \tilde{\mathcal{C}}_i^{\delta} \\ \text{and } \forall t < \tau \text{ we have } w(t) \in \Omega \setminus \cup_j \tilde{\mathcal{C}}_j^{-\delta}\}.$$

We define the stopping time

$$\tau(Y) = \inf_t Y \in \tilde{\mathcal{C}}_i^{\delta} \cup_{j \in \{1, \ldots, K\}, j \neq i} \tilde{\mathcal{C}}_j^{-\delta}.$$

Since $\mathcal{C}_i \subset \Omega$ and Ω has a single connected component, we have that $\mathbb{P}(\tau(Y_t^x) < \infty) = 1$, in particular that means that there exists T_0 such that for any $T \geq T_0$, $\mathbb{P}(\tau(Y^x) \leq T) \geq 1 - \epsilon/6$. Using Proposition 4, we have that, almost surely

$$\mathbb{P}(\mathbb{P}(\tau(M^{x,h(n)}) \leq T) \geq \mathbb{P}(\tau(Y^x) \leq T)) - \epsilon/6 \geq 1 - \frac{1}{3}\epsilon$$

Hence, we have

$$\mathbb{P}(M^{x,h(n)} \in B \setminus B_T) + \mathbb{P}(Y^x \in B \setminus B_T) \\ \leq \mathbb{P}(\tau(M^{x,h(n)}) > T) + \mathbb{P}(\tau(Y^x) > T) \leq \epsilon/2$$

Since $\mathbb{P}(Y^x \in \partial B_T) = \mathbb{P}(Y^x \in \partial B) = 0$, we can apply Proposition 4 on the set B_T, and obtain

$$\|\mathbb{P}(M^{x,h(n)} \in B_T) - \mathbb{P}(Y^x \in B_T)\|_{\infty,U} \leq \epsilon/4$$

Combined with our previous result, we obtain:

$$\|\mathbb{P}(M^{x,h(n)} \in B) - \mathbb{P}(Y^x \in B)\|_{\infty,U} \leq 3\epsilon/4$$

Using our assumption on $\hat{C}_{i,n}$, we have $\mathbb{P}(M^{x,h(n)} \in B) \geq \hat{\mu}_{i,h,n}$. Therefore, using our previous bound between μ and μ^+:

$$\hat{\mu}_{i,h,n}(x) - \mu_i(x) \leq \epsilon.$$

Similarly,

$$\mu_i(x) - \hat{\mu}_{i,h,n}(x) \leq \epsilon,$$

concluding the proof.

7 Conclusion

We have provided a fuzzy clustering algorithm based on the mode-seeking framework relying on the approximation of a diffusion process through the use of a random walk. Despite the convergence issues of random-walk-based quantities for large data highlighted by Luxburg et al. (2010), we have shown that our algorithm does converge to meaningful values. Our thereotical result is backed up by encouraging experiments. The main question still open regarding our algorithm is the choice of the temperature parameter β, while we have shown that the evolution of a quantification of the fuzziness of the clustering through the clustering entropy can give some hint about a correct choice for this parameter, it is not clear whether this can be done in all cases and for more complicated datasets.

Acknowledgements. The authors wish to thank Cecilia Clementi and her student Wenwei Zheng for providing the alanine-dipeptide conformation data used in Figure 4. This work was supported by the French Délégation Générale de l'Armement (DGA), by ANR project TopData ANR-13-BS01-0008 and by ERC grant Gudhi (ERC-2013-ADG-339025).

References

Albeverio, S., Kondratiev, Y., Röckner, M., 2003. Strong feller properties for distorted brownian motion and applications to finite particle systems with singular interactions, in: Finite and infinite dimensional analysis in honor of Leonard Gross (New Orleans, LA, 2001). Amer. Math. Soc., Providence, RI. volume 317 of *Contemp. Math.*, pp. 15–35.

Arias-Castro, E., Mason, D., Pelletier, B., 2013. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Unpublished .

Azizyan, M., Chen, Y.C., Singh, A., Wasserman, L., 2015. Risk Bounds For Mode Clustering. ArXiv e-prints `arXiv:1505.00482`.

Biau, G., Chazal, F., Cohen-Steiner, D., Devroye, L., Rodriguez, C., 2011. A weighted k-nearest neighbor density estimate for geometric inference. Electronic Journal of Statistics 5, 204–237. URL: `https://hal.archives-ouvertes.fr/hal-00606482`. http://imstat.org/ejs/.

Chazal, F., Guibas, L.J., Oudot, S.Y., Skraba, P., 2013. Persistence-based clustering in riemannian manifolds. J. ACM 60, 41. URL: `http://dblp.uni-trier.de/db/journals/jacm/jacm60.html#ChazalGOS13`.

Chen, Y.C., Genovese, C.R., Tibshirani, R.J., Wasserman, L., 2014a. Nonparametric Modal Regression. ArXiv e-prints, to appears in Annals of Statistics `arXiv:1412.1716`.

Chen, Y.C., Genovese, C.R., Wasserman, L., 2014b. A Comprehensive Approach to Mode Clustering. ArXiv e-prints, to appears in Electronic Journal of Statistics `arXiv:1406.1780`.

Chen, Y.C., Genovese, C.R., Wasserman, L., 2015a. Density Level Sets: Asymptotics, Inference, and Visualization. ArXiv e-prints `arXiv:1504.05438`.

Chen, Y.C., Genovese, C.R., Wasserman, L., 2015b. Statistical Inference using the Morse-Smale Complex. ArXiv e-prints `arXiv:1506.08826`.

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Trans. Pattern Anal. Mach. Intell. 17, 790–799. URL: `http://dx.doi.org/10.1109/34.400568`, doi:`10.1109/34.400568`.

Cho, M., Lee, K.M., 2010. Authority-shift clustering: Hierarchical clustering by authority seeking on graphs., in: CVPR, IEEE. pp. 3193–3200. URL: `http://dblp.uni-trier.de/db/conf/cvpr/cvpr2010.html#ChoL10`.

Chodera, J.D., Swope, W.C., Pitera, J.W., Dill, K.A., 2006. Long-time protein folding dynamics from short-time molecular dynamics simulations. Multiscale Modeling & Simulation 5, 1214–1226. URL: `http://link.aip.org/link/?MMS/5/1214/1`, doi:`10.1137/06065146X`.

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 603–619.

Durrett, R., 1996. Stochastic calculus : a practical introduction. Probability and stochastics series, CRC Press.

Koontz, W., Narendra, P., Fukunaga, K., 1976. A graph-theoretic approach to nonparametric cluster analysis. IEEE Transactions on Computers 25, 936–944. doi:http://doi.ieeecomputersociety.org/10.1109/TC.1976.1674719.

Krylov, N., Röckner, M., 2005. Strong solutions of stochastic equations with singular time dependent drift. Probab. Theory Relat. Fields 131, 154–196.

Luxburg, U.V., Radl, A., Hein, M., 2010. Getting lost in space: Large sample analysis of the resistance distance, in: Advances in Neural Information Processing Systems 23, pp. 2622–2630.

Ting, D., Huang, L., Jordan, M.I., 2010. An analysis of the convergence of graph laplacians., in: ICML.