Fermat distance and its critical parameter

Matthieu Jonckheere, CNRS LAAS, Toulouse

with P. Groisman (UBA) and F. Sapienza (Berkeley)

and F. Chazal, L. Ferraris (INRIA), F. Pascal (Centralesupelec)

Stochastic Geometric Days, Dijon 2023

Overview

Motivation

Fermat distance

Some clustering results

The critical parameter

Motivation

Our original motivation

Problem

- Clustering of high dimensional chemical formulas

Data size

- 10^6 formulas
- Dimension $d \sim 4000$

Clustering in high-dimensional spaces is usually very difficult

Our original motivation

Problem

- Clustering of high dimensional chemical formulas

Data size

- 10^6 formulas
- Dimension $d \sim 4000$

Clustering in high-dimensional spaces is usually very difficult and Euclidian or ad-hoc distances might be misleading...

A curse of dimensionality

Bad news

Let $\omega_D(r) = \omega_D(1) r^D$ be the volume of the ball of radius r in $\mathbb{R}^D$.

$$\frac{\omega_D(1) - \omega_D(1-\varepsilon)}{\omega_D(1)} = 1 - (1-\varepsilon)^D \xrightarrow{D \to \infty} 1$$

A curse of dimensionality

Bad news

Let $\omega_D(r) = \omega_D(1)r^D$ be the volume of the ball of radius r in $\mathbb{R}^D$.

$$\frac{\omega_D(1) - \omega_D(1-\varepsilon)}{\omega_D(1)} = 1 - (1-\varepsilon)^D \xrightarrow{D\to\infty} 1$$

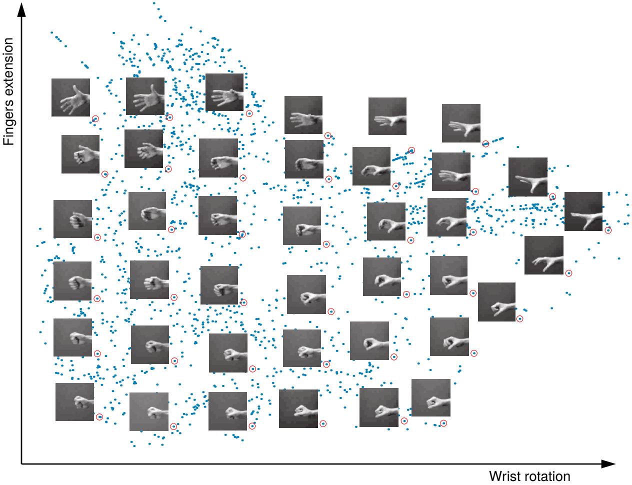
In high dimensional Euclidean spaces every two points of a typical large set are at similar distance.

Manifold hope

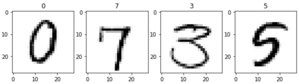
Good news: many structured data live in a manifold of dimension much lower than ambient space ($d \ll D$).

Manifold hope

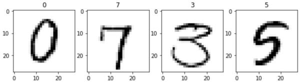
Good news: many structured data live in a manifold of dimension much lower than ambient space ($d \ll D$).

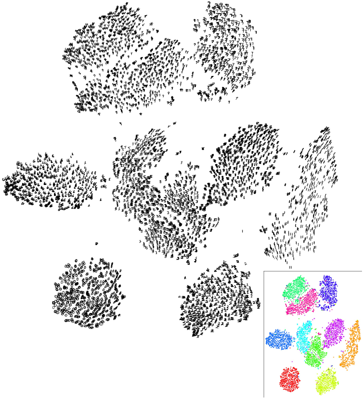


Motivation: MNIST Dataset



Motivation: MNIST Dataset





van der Maaten, L.J.P.; Hinton, G.E. (Nov 2008). Visualizing Data Using t-SNE. Journal of Machine Learning Research. 9: 2579–2605.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.
- The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.
- The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.
- Since the data lies in an (unknown) lower dimensional surface, this distance has to be inferred from the data itself.

Dimension reduction and distances

- In most unsupervised learning tasks, a notion of similarity between data points is both crucial and usually not directly available as an input.
- The efficiency of tasks like dimensionality reduction and clustering might crucially depend on the distance chosen.
- Since the data lies in an (unknown) lower dimensional surface, this distance has to be inferred from the data itself.
- Delicate game between dimensionality reduction, choice of the distance and clustering...

Dimensionality reduction and distance learning

There are many techniques to address dimensionality reduction and possibly finding distances in lower dimensional spaces:

- Principal components analysis (PCA),
- Multidimensional scaling (MDS),
- Embeddings (VAE, t-SNE,...)
- Isomap and variants.

Dimensionality reduction and distance learning

Dimensionality reduction

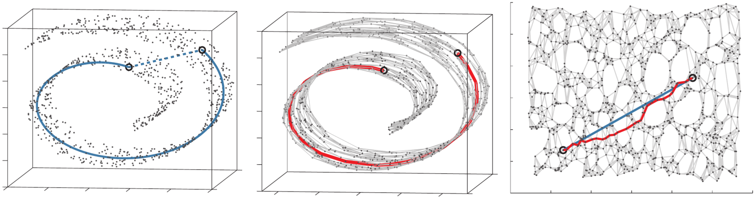
- Principal components analysis (PCA),
- Multidimensional scaling (MDS),
- Spectral embeddings
- Embeddings (VAE, t-SNE,...)

Distance learning

- Isomap and variants.

Dimensionality Reduction/distance learning: Isomap

Constructs the k-nn graph and finds the optimal path. The weight of an edge is given $|q_i - q_j|$.



Isomap

Theorem
Given $\varepsilon > 0$ and $\delta > 0$, for n large enough

$$\mathbb{P}\left(1 - \varepsilon \leq \frac{d_{\textsf{geodesic}}(x, y)}{d_{\textsf{graph}}(x, y)} \leq 1 + \varepsilon\right) > 1 - \delta.$$

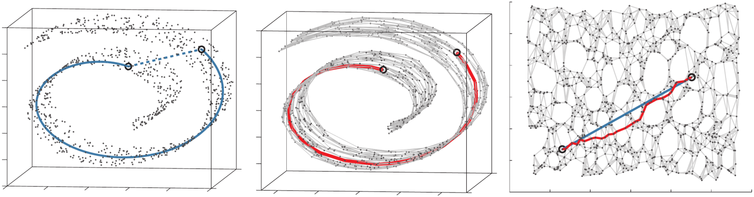
[Bernstein, de Silva, Langford, Tenenbaum (2000)].

Isomap

Theorem
Given $\varepsilon > 0$ and $\delta > 0$, for n large enough

$$\mathbb{P}\left(1-\varepsilon \leq \frac{d_{\textit{geodesic}}(x,y)}{d_{\textit{graph}}(x,y)} \leq 1+\varepsilon\right) > 1-\delta.$$

[Bernstein, de Silva, Langford, Tenenbaum (2000)].



Fermat distance

The Problem

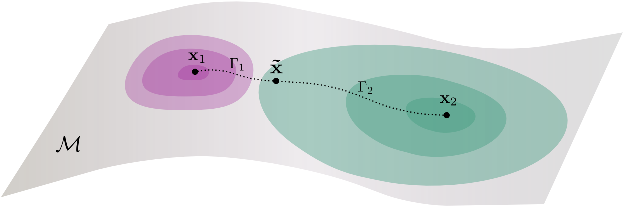
- Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).

The Problem

- Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).
- Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.

The Problem

- Let $\mathscr{M} \subseteq \mathbb{R}^D$ be a d-dimensional surface (we expect $d \ll D$).
- Consider n independent points on $\mathscr{M}$ with common density $f : \mathscr{M} \mapsto \mathbb{R}_{\geq 0}$.



Can we learn a better notion of distance between points (for say clustering)?

Objectives

We look for a distance that takes into account the underlying manifold $\mathscr{M}$ and the underlying density f.

Euclidean Percolation and Sample Fermat's distance

- $\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$.

Euclidean Percolation and Sample Fermat's distance

- $\alpha \geq 1$ a parameter, $\mathbb{X} =$ a discrete set of points q, $x, y \in \mathbb{X}$.

$$\mathscr{D}_{\mathbb{X}}(\mathbf{p}, \mathbf{q}) = \inf\{\sum_{j=1}^{K-1} |\mathbf{y}_{i+1} - \mathbf{y}_i|^{\alpha} \colon K \geq 2,$$

$$\text{and } (\mathbf{y}_1, \ldots, \mathbf{y}_K) \text{ is a } \mathbb{X}\text{-path from } \mathbf{p} \text{ to } \mathbf{q}\}.$$

Visualisation

http://www.aristas.com.ar/fermat/index.html

Homogeneous Poisson Point Process : Shape theorem

We based our analysis on:

Theorem (Howard and Newman (1997))

Let $\mathbb{X}$ a PPP with intensity $\lambda = 1$. Then there exists $0 < \mu < \infty$ such that

$$\lim_{|\mathbf{q}|\to\infty} \frac{\mathscr{D}_{\mathbb{X}}(\mathbf{0},\mathbf{q})}{|\mathbf{q}|} = \mu, \qquad \textit{almost surely.}$$

They also give bounds on fluctuations.

Sample to Macroscopic Fermat's distance

Theorem (Groisman, J., Sapienza, '20)

Under mild assumptions on f, there exists $\mu > 0$, such that for $x, y \in \mathscr{M}$ and $\mathbb{X}_n$ i.i.d $\sim f$ we have

$$\lim_{n \to \infty} n^{\beta} D_{\mathbb{X}_n}(x, y) = \mu \mathscr{D}(x, y),$$

almost surely, with $\beta = (\alpha - 1)/d$.

$$\mathscr{D}(x, y) = \inf_{\Gamma} \int_{\Gamma} \frac{1}{f^{\beta}}.$$

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}, \quad \mathrm{n} = \text{refractive index}$$

Fermat's principle

In optics, the path taken between two points by a ray of light is an extreme of the functional

$$\Gamma \mapsto \int_\Gamma \mathrm{n}, \quad \mathrm{n} = \text{refractive index}$$

$$\mathscr{D}(x,y) = \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \qquad f^{-\beta} \sim \mathbf{n}$$

Fermat's principle

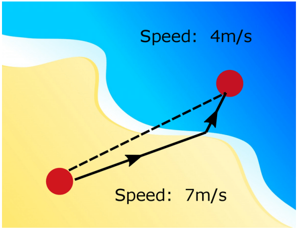
In optics, the path taken between two points by a ray of light is an extreme of the functional

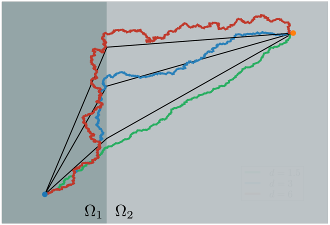
$$\Gamma \mapsto \int_\Gamma \mathrm{n}, \quad \mathrm{n} = \text{refractive index}$$

$$\mathscr{D}(x,y) = \inf_\Gamma \int_\Gamma \frac{1}{f^\beta} \qquad f^{-\beta} \sim \mathbf{n}$$

©S.Thorgerson - Pink Floyd, The Dark Side of the Moon (1973), Harvest, Capitol.

Snell's law, the lifeguard and Fermat's distance





Algorithmic considerations and generalizations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}_n}^{(\alpha,k)}(x,y) = \inf_{\substack{r = (q_1, \ldots, q_K) \\ q_{i+1} \in \mathcal{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1} - q_i|^{\alpha}.$$

Generalization of Isomap and Fermat's distance.

Algorithmic considerations and generalizations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}_n}^{(\alpha,k)}(x,y) = \inf_{\substack{r=(q_1,\dots,q_K) \\ q_{i+1}\in \mathscr{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Generalization of Isomap and Fermat's distance.

Proposition (Groisman, J., Sapienza, '20)

Given $\varepsilon > 0$, we can choose $k = \mathscr{O}(\log(n/\varepsilon))$ such that

$$\mathbb{P}\left(D_{\mathbb{X}_n}^{(k)}(x,y) = D_{\mathbb{X}_n}(x,y)\right) > 1-\varepsilon.$$

Algorithmic considerations and generalizations

Restricted Fermat's distance:

$$\mathbb{D}_{\mathbb{X}_n}^{(\alpha,k)}(x,y) = \inf_{\substack{r=(q_1,\ldots,q_K) \\ q_{i+1}\in\mathscr{N}_k(q_i)}} \sum_{k=1}^{K-1} |q_{i+1}-q_i|^{\alpha}.$$

Generalization of Isomap and Fermat's distance.

Proposition (Groisman, J., Sapienza, '20)

Given $\varepsilon > 0$, we can choose $k = \mathscr{O}(\log(n/\varepsilon))$ such that

$$\mathbb{P}\left(D_{\mathbb{X}_n}^{(k)}(x,y) = D_{\mathbb{X}_n}(x,y)\right) > 1-\varepsilon.$$

$\rightarrow$ We can reduce the running time from $\mathscr{O}(n^3)$ to $\mathscr{O}(n^2(\log n)^2)$.

Open theoretical questions

How to choose α, k ?

- k independent of n, $\alpha = 1$, f uniform $\Rightarrow$ Isomap.
- k scales like $\log(n)$, $\alpha > 1 \Rightarrow$ Fermat.

- General proof of convergence for k fixed, α?
- How to choose α, k ?

Other previous mathematical results

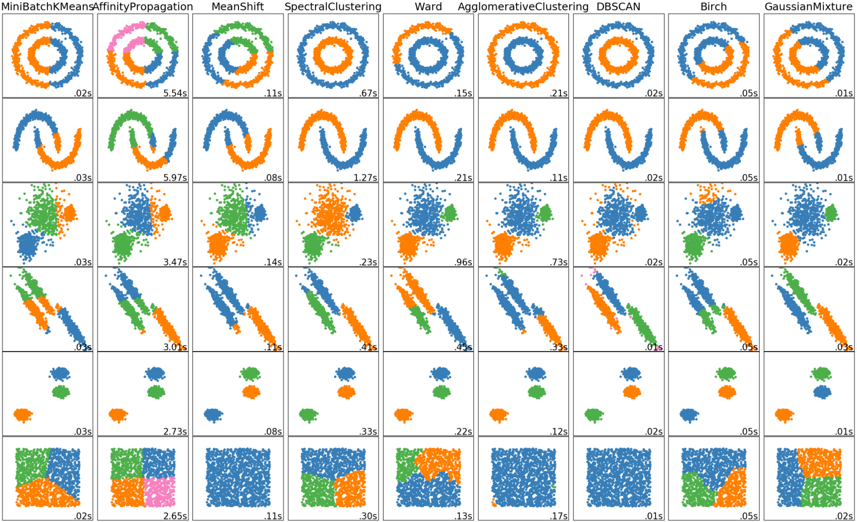
Sung Jin Hwang, Steven B. Damelin, Alfred O. Hero III,

Shortest Path through Random Points,

The Annals of Applied Probability, 2016, Vol. 26, No. 5, pp 2791-2823.

Some clustering results

Clustering

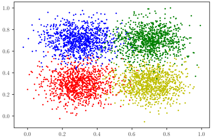


©scikit-learn developers

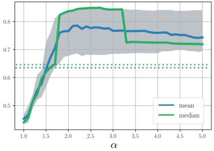
Clustering with Fermat (Simulation L. Ferraris)

FIGURE 5. An example of the Fermat K-medoids predictions for different α values. Each dataset is composed with 500 samples and 100 outliers.

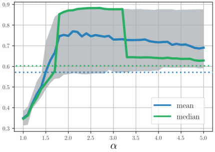
Clustering with Fermat — K-medoids in the Swiss roll



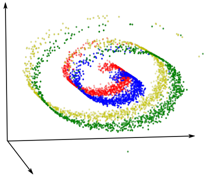
(a) 2D data



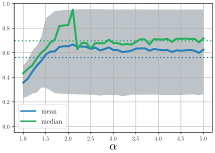
(c) Adjusted mutual information



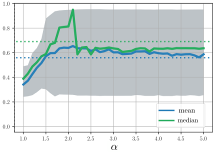
(e) Adjusted Rand index



(b) 3D data

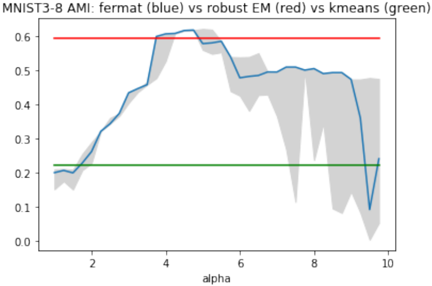


(d) Accuracy



(f) F1 score

MNIST



Performance of Fermat + k-medoids compared to state of the art robust clustering, Simulations Alfredo Umfurer.

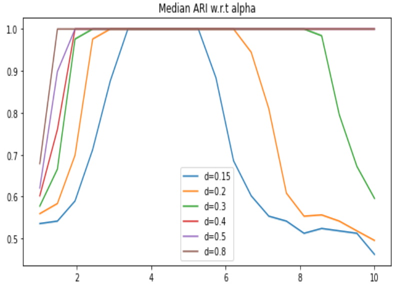
Application in genetics

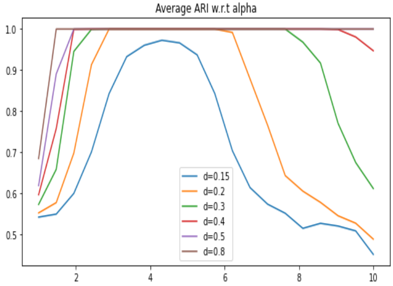
Fingerprints of cancer by persistent homology, 2019. A. Carpio, L. L. Bonilla, J. C. Mathews, A. R. Tannenbaum,

- They compute Fermat's distance between genes'expressions (dimension 77) (They choose $\alpha \sim 3$.)
- They study clusters based on the Fermat distance.
- These clusters make noticeable the relations between gene expressions in healthy samples and those in cancerous samples."

The critical parameter

The critical parameter





Performance of clustering in function of α for different scenarios

Generic Conjecture: There exists a window of critical parameters which maximizes the clustering performance.

How to find the critical window

- $\alpha > \alpha_0$

  Link to a macroscopic clustering problem:

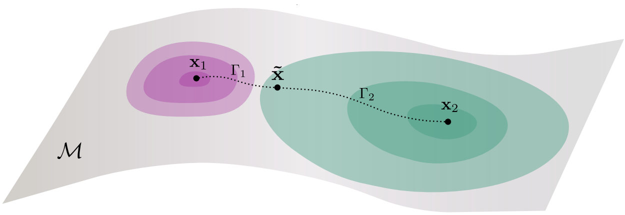
  Define a minimal α_0 such that in the limit all points are perfectly classified when it is possible.

How to find the critical window

- $\alpha > \alpha_0$

  Link to a macroscopic clustering problem:

  Define a minimal α_0 such that in the limit all points are perfectly classified when it is possible.



How to find the critical window

- $\alpha > \alpha_0$

  Link to a macroscopic clustering problem:

  Define a minimal α_0 such that in the limit all points are perfectly classified when it is possible.

The critical parameter through the macroscopic problem

Recall the macroscopic Fermat distance:

$$\mathscr{D}_\alpha(x,y) = \inf_{\gamma \subset \mathscr{M}} \int_\gamma \frac{1}{f^{(\alpha-1)/d}} \, d\gamma. \tag{4.1}$$

Definition (Strictly feasible macroscopic classification)

Given a family of clusters $(C_i)_{i \leq m}$ we say that a macroscopic clustering problem is strictly feasible if there exists $1 \leq \alpha < \infty$ and ϵ such that

$$\mathscr{D}_\alpha(x, c_i) \leq \mathscr{D}_\alpha(x, c_j) - \epsilon, \forall i, \forall x \in C_i, \forall j \neq i. \tag{4.2}$$

where c_i is "some" center for the set C_i.

"Nice" Geometry

Definition (Critical Parameter)

$\alpha_0 = \inf\{\alpha : \exists\epsilon \text{ such that } \mathscr{D}_\alpha(x, c_i) \leq \mathscr{D}_\alpha(x, c_j) - \epsilon,\ \forall x \in C_i, \forall j \neq i.\}$

Proposition

If the clusters are convex and the density of points is bigger than a_1 in the clusters and smaller than a_0 outside, then

$$\alpha \geq \alpha_0(d) = 1 + d\,\frac{\omega}{\log(a_1/a_0)}\,, \tag{4.3}$$

with d the intrinsic dimension, and ω a geometric constant.

More difficult Geometry

Proposition

If the clusters have a finite "reach" and the density of points is bigger than a_1 in the clusters and smaller than a_0 outside, then

$$\alpha \geq \alpha_0(d) = 1 + d^2 \, \frac{\tilde{\omega}}{\log(a_1/a_0)} \,, \tag{4.4}$$

with d the intrinsic dimension, and $\tilde{\omega}$ a geometric constant.

Convergence in the microscopic setting

We say that a (microscopic) classification/clustering problem is strictly feasible if there exists $1 \leq \alpha < \infty$ and ϵ such that

$$\mathscr{D}_{\mathbb{X}_n,\alpha}(x, \hat{c}_i) \leq \mathscr{D}_{\mathbb{X}_n,\alpha}(x, \hat{c}_j) - \epsilon, \forall i, \forall x \in C_i, \forall j \neq i, \tag{4.5}$$

where $\hat{c}_i$ are estimations of the the center of clusters C_i.

Definition (Microscopic critical Parameter)

$$\alpha_0^n = \inf \Big\{ \alpha \geq 1 : \exists \epsilon \text{ such that:} \\ \mathscr{D}_{\mathbb{X}_n,\alpha}(x, \hat{c}_i) \leq \mathscr{D}_{\mathbb{X}_n,\alpha}(x, \hat{c}_j) - \epsilon,\ \forall i,\ \forall x \in C_i,\ \forall j \neq i. \Big\}$$

Convergence in the microscopic setting

Proposition

Assume consistency on the empirical means, then there exists some constant $C, c, \gamma > 0$ *such that:*

$$\mathbb{P}(\alpha_0^n > \alpha_0) \leq Cne^{-cn^\gamma} + \epsilon_n.$$

Conversely, if $\alpha < \alpha_0$*, then the microscopic clustering problem is not strictly feasible with overwhelming probability.*

Influence of the noise

Define the coefficient of variation

$$CV_n = \sqrt{\frac{\mathrm{Var}(\mathscr{D}_{\mathbb{X}_n,\alpha})}{\mathbb{E}[\mathscr{D}_{\mathbb{X}_n,\alpha}]^2}}$$

Proposition

If $d = 1$,

$$CV_n \underset{n\to\infty}{\sim} \frac{1}{\sqrt{n}} \frac{\sqrt{\Gamma(2\alpha+1)}}{\Gamma(\alpha+1)} \sim_\alpha 2^\alpha/\sqrt{n}.$$

Influence of the noise

Define the coefficient of variation

$$CV_n = \sqrt{\frac{\mathrm{Var}(\mathscr{D}_{\mathbb{X}_n,\alpha})}{\mathbb{E}[\mathscr{D}_{\mathbb{X}_n,\alpha}]^2}}$$

Conjecture

There exists $\psi, c > 0$ *such that when* n *large and fixed,* α *large:*

$$CV_n \underset{n\to\infty}{=\sim} c^{\alpha/d}/n^{\psi}.$$

Hence, clustering should be fine if the geometry is not too rough (α_0 scales as d)...

Applications

- Clustering
- Dimension reduction
- Density estimation
- Regression
- Any learning task that requires a notion of distance (not necessarily in Euclidean space) as an input.

Download

A prototype implementation is available at

▸ http://www.aristas.com.ar/fermat/index.html

- *Weighted Geodesic Distance Following Fermat's Principle* (2018); F. Sapienza, P. Groisman, M. Jonckheere; 6th International Conference on Learning Representations (ICRL) 2018.

- *Nonhomegeneous First Passage Percolation and Distance Learning*; P. Groisman, M. Jonckheere, F. Sapienza; Bernouilli 2021.

Thanks!

Homogeneous Poisson Point Process : Shape theorem

We based our analysis on:

Theorem (Howard and Newman (1997))

Let $\mathbb{X}$ a PPP with intensity $\lambda = 1$. Then there exists $0 < \mu < \infty$ such that

$$\lim_{|\mathbf{q}|\to\infty} \frac{\mathscr{D}_{\mathbb{X}}(\mathbf{0},\mathbf{q})}{|\mathbf{q}|} = \mu, \qquad \textit{almost surely.}$$

They also give bounds on fluctuations.

Uniform distribution on compact sets

$\mathbb{X}_N \sim \mathsf{PPP}(C, n)$ on a convex set $C \subset \mathbb{R}^D$ (with strictly positive volume).

Uniform distribution on compact sets

$\mathbb{X}_N \sim \mathsf{PPP}(C, n)$ on a convex set $C \subset \mathbb{R}^D$ (with strictly positive volume).

Corollary
Let $\beta = (\alpha - 1)/d$. For all $\mathbf{p}, \mathbf{q} \in C^o$ we have

$$\lim_{N\to\infty} n^\beta \mathscr{D}_{\mathbb{X}_N}(\mathbf{p}, \mathbf{q}) = \mu|\mathbf{p} - \mathbf{q}|, \qquad \textit{a.s.}$$

Moreover, given $\delta > 0$ there exist positive constants c_1, c_2, c_3, c_4, with c_2 depending on δ, such that if $|x - y| > \delta$ then

$$\mathbb{P}\left(\left|n^\beta D_{\mathbb{X}_N}(\mathbf{p}, \mathbf{q}) - \mu|\mathbf{p} - \mathbf{q}|\right| \geq c_4 n^{-1/3d}\right) \leq c_1 \exp\left(-c_2 n^{c_3}\right).$$

Some proof ideas I

For the case f constant and C convex, we saw that

$$\lim_{n\to\infty} n^{\beta}\mathscr{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mu\frac{1}{f^{\beta}}|\mathbf{p}-\mathbf{q}|, \qquad \text{a.s.}$$

Some proof ideas I

For the case f constant and C convex, we saw that

$$\lim_{n\to\infty} n^{\beta}\mathscr{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mu\frac{1}{f^{\beta}}|\mathbf{p}-\mathbf{q}|, \qquad \text{a.s.}$$

Locally, we can construct $\mathbb{X}_n^-, \mathbb{X}_n^+, \mathbb{X}_n$, where $\mathbb{X}_n^- \sim$ PPP$(f_{min}n)$ y $\mathbb{X}_n^+ \sim$ PPP$(f_{max}n)$, so that $\mathbb{X}_n^- \subset \mathbb{X}_n \subset \mathbb{X}_n^+$.

Some proof ideas I

For the case f constant and C convex, we saw that

$$\lim_{n\to\infty} n^{\beta}\mathscr{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) = \mu\frac{1}{f^{\beta}}|\mathbf{p}-\mathbf{q}|, \qquad \text{a.s.}$$

Locally, we can construct $\mathbb{X}_n^-, \mathbb{X}_n^+, \mathbb{X}_n$, where $\mathbb{X}_n^- \sim$ PPP$(f_{min}n)$ y $\mathbb{X}_n^+ \sim$ PPP$(f_{max}n)$, so that $\mathbb{X}_n^- \subset \mathbb{X}_n \subset \mathbb{X}_n^+$.

Lemma (Bounds)

$$\mu f_{max}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| \leq \liminf_{n\to\infty} n^{\beta}\mathbb{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}),$$

$$\mu f_{min}^{-\beta}|\,\mathbf{p}-\mathbf{q}\,| \geq \limsup_{n\to\infty} n^{\beta}\mathbb{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}), \qquad \textit{with overwhelming probability}$$

Some proof ideas II

Lemma (Restriction to a neighborhood)

$$\mathbb{P}\left(\mathbb{D}_{\mathbb{X}_n}(\mathbf{p},\mathbf{q}) \neq \mathbb{D}_{\mathbb{X}_n \cap B(\mathbf{p}, a|\,\mathbf{p}\,\mathbf{q}\,|)}(\mathbf{p},\mathbf{q})\right) < c_1 e^{-c_2 n^{c_3}}$$

Some proof ideas III

An important issue is to prove that optimal paths have bounded lenght.

Lemma

Let C *a connected set and* $p, q \in C$*. Sea* $(\mathbf{y}_1^*, \ldots, \mathbf{y}_K^*)$ *the* $\mathbb{X}_n$*-path that realizes* $\mathbb{D}_{\mathbb{X}_n}(\mathbf{p}, \mathbf{q})$ *with arc-length:*

$$L_n = \sum_{i=1}^{K-1} |\mathbf{y}_{i+1}^* - \mathbf{y}_i^*|.$$

then there exists $\ell_{\max} < \infty$ *such that*

$$\limsup_{n \to \infty} L_n < \ell_{\max} \quad \textit{a.s.}$$

Consequences

- Proving that Fermat's distance empirical geodesics converge to the macroscopic ones.

Other previous mathematical results

Sung Jin Hwang, Steven B. Damelin, Alfred O. Hero III,

Shortest Path through Random Points,

The Annals of Applied Probability, 2016, Vol. 26, No. 5, pp 2791-2823.