Lecture 3. Eigenmaps and Diffusion Maps

Juno Kim

Department of Mathematics & Statistics
Seoul National University

Manifold Learning, Spring 2022

Remark

The algorithms of the previous lecture follow a general principle:

Step 1. Construct a neighborhood graph G as a proxy of the unknown manifold $\mathcal{M}$.

Step 2. Compute a discrete aspect of G that approximates some geometrical structure on $\mathcal{M}$.

- geodesics, local weights, tangent space, etc

Step 3. Perform spectral embedding of G by optimizing the aspect via eigenanalysis.

In this lecture, we continue this philosophy with more complicated objects: Laplace operator, Hessian form, Markov process, etc.

Table of Contents

Laplacian

- The Laplace-Beltrami operator on a Riemannian manifold $\mathcal{M}$ with metric tensor g is defined as:

$$\Delta f := \nabla \cdot \nabla f = \frac{1}{\sqrt{\det(g_{ij})}} \partial_j \left(\sqrt{\det(g_{ij})}\, g^{ij} \partial_i f \right)$$

- Given a measure ν absolutely continuous w.r.t. the canonical measure $\mathrm{vol}_{\mathcal{M}}$ corresponding to the volume form, with p.d.f. $d\nu = P \cdot d\,\mathrm{vol}_{\mathcal{M}}$, we also define the weighted Laplacian:

$$\Delta_P f := \frac{1}{P} \nabla \cdot (P \nabla f)$$

The *Laplacian eigenmap* (LEM) algorithm works as follows.

Step 1. Given n points $V = \{x_1, \cdots, x_n\}$ on $\mathcal{M}$, construct a complete/KNN graph with weights given by the Gaussian kernel,

$$\mathbf{W}_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{4\rho}\right)$$

Step 2. Define the corresponding graph Laplacian matrix $\mathbf{L}$ as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \quad \mathbf{D} = \mathrm{diag}(\textstyle\sum_j \mathbf{W}_{ij})_i$$

Step 3. Find t-dimensional embedding coordinates $\mathbf{Y}$ subject to $\mathbf{Y}\mathbf{D}\mathbf{Y}^T = \mathbf{I}$ by minimizing

$$\sum_{i,j} \mathbf{W}_{ij} \|y_i - y_j\|^2 = \mathrm{tr}(\mathbf{Y}\mathbf{L}\mathbf{Y}^T)$$

The solution is obtained by eigenanalysis of $\mathbf{D}^{-1/2}\mathbf{Y}\mathbf{D}^{-1/2}$.

So why does the graph Laplacian serve as a discrete model for Δ?

- Consider $\mathbf{L}_n^\rho = \mathbf{L}$ as acting on functions $f : V \to \mathbb{R}$

$$\begin{pmatrix} \mathbf{L}_n^\rho f(x_1) \\ \vdots \end{pmatrix} = \mathbf{L}_n^\rho \begin{pmatrix} f(x_1) \\ \vdots \end{pmatrix}, \text{ that is,}$$

$$\mathbf{L}_n^\rho f(x_i) = f(x_i) \sum_j e^{-\frac{\|x_i - x_j\|^2}{4\rho}} - \sum_j f(x_j) e^{-\frac{\|x_i - x_j\|^2}{4\rho}}$$

- $\mathbf{L}_n^\rho$ naturally extends to an integral operator $\mathcal{L}_n^\rho$ on $C^\infty(\mathcal{M})$:

$$\mathcal{L}_n^\rho f(x) = \frac{1}{n} f(x) \sum_j e^{-\frac{\|x - x_j\|^2}{4\rho}} - \frac{1}{n} \sum_j f(x_j) e^{-\frac{\|x - x_j\|^2}{4\rho}}$$

- We may retain the Euclidean norm on $\mathbb{R}^r$ since $d^{\mathcal{M}}(x, y) = \|x - y\| + O(\|x - y\|^3)$.

Exercise. Prove the above statement. First consider curves in $\mathbb{R}^2$.

Theorem (Belkin, Niyogi, 2008)

Let $\mathcal{M}^t$ be a compact Riemannian submanifold of $\mathbb{R}^r$ with a probability density P from which i.i.d. data points $x_1, \cdots, x_n$ are drawn. Let $\rho_n = n^{-1/(t+2+\epsilon)}$, $\epsilon > 0$. Then for $f \in C^\infty(\mathcal{M})$,

$$\lim_{n\to\infty} \frac{1}{\rho_n(4\pi\rho_n)^{t/2}} \mathcal{L}_n^{\rho_n} f(x) = P(x)\,\Delta_{P^2} f(x)$$

In particular, the LHS converges to $\frac{1}{\mathrm{vol}\,\mathcal{M}}\Delta f(x)$ if P is uniform.

By the Law of Large Numbers, the LHS approximates

$$\mathcal{L}^\rho f(x) = f(x)\int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4\rho}} d\nu(y) - \int_{\mathcal{M}} f(y) e^{-\frac{\|x-y\|^2}{4\rho}} d\nu(y)$$

This relationship with the heat kernel connects $\mathcal{L}^\rho$ with Δ, which governs diffusion on $\mathcal{M}$ via the *heat equation*.

Heat Equation

- Consider the heat equation

$$\frac{\partial}{\partial \rho} u(x,\rho) - \Delta u(x,\rho) = 0, \quad u(x,0) = f(x)$$

- For $\mathbb{R}^r$, the solution is given by convolution with the heat kernel $H^t(x,y) = (4\pi\rho)^{-t/2} e^{-\frac{\|x-y\|^2}{4\rho}}$. Then:

$$\Delta f(x) = -\frac{\partial}{\partial \rho} u(x,\rho)\bigg|_{\rho=0} = \lim_{\rho \to 0} \frac{1}{\rho}(f(x) - u(x,\rho))$$

$$= \lim_{\rho \to 0} \frac{(4\pi\rho)^{-t/2}}{\rho} \left(f(x) \int_{\mathbb{R}^r} e^{-\frac{\|x-y\|^2}{4\rho}} dy - \int_{\mathbb{R}^r} f(y) e^{-\frac{\|x-y\|^2}{4\rho}} dy \right)$$

- H^t is unknown for general manifolds - the full proof requires careful analysis of local behaviour

Table of Contents

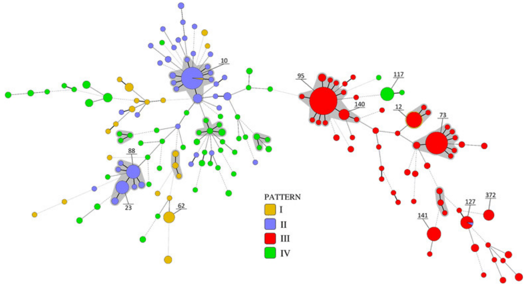

Figure: Minimum spanning tree of bionumeric data.

A robust version of LEM utilizing global information has been proposed (Ch.2). Certain minimal graphs contribute to Δ, on the grounds that their length functionals have asymptotic properties depending only on $\mathcal{M}$.

- Let $\mathbf{X}_n = \{x_1, \cdots, x_n\} \subset \mathbb{R}^t$. Its minimum spanning tree (MST) is the tree spanning $\mathbf{X}_n$ with shortest total length:

$$L_\gamma^t(\mathbf{X}_n) = \min_{T:span} \sum_{e \in E(T)} |e|^\gamma, \quad \text{where } \gamma \in (0, t)$$

- We may also use KNN graph, Traveling Salesman tour, etc.
- Let μ be a probability measure on $\mathbb{R}^t$ with compact support. Let f denote the density of the continuous component μ_c of μ w.r.t. Lebesgue decomposition.
- The Rényi α-entropy of the density f is defined as:

$$H_\alpha^t(f) = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^t} f(y)^\alpha dy$$

Exercise. Show that $\alpha \to 1$ gives Shannon entropy, $-\int f(y) \log f(y) dy$.

Theorem (Beardwood-Halton-Hammersley)

Let $t \geq 2$ and $\alpha = (t-\gamma)/t \in (0,1)$. For $x_1, \cdots, x_n$ sampled i.i.d. from μ on $\mathbb{R}^t$, the MST length functional satisfies:

$$\lim_{n\to\infty} \frac{1}{n^\alpha} L_\gamma^t(\mathbf{X}_n) = \beta_t \int_{\mathbb{R}^t} f(y)^\alpha dy \quad a.s.$$

where β_t are constants depending only on t.

As a corollary,

$$\hat{H}_\alpha^t(\mathbf{X}_n) := \frac{t}{\gamma}\left(\log L_\gamma^t(\mathbf{X}_n) - \alpha \log n - \log \beta_t\right)$$

is an asymptotically unbiased and consistent estimator of $H_\alpha^t(f)$.

Exercise. Estimate $\beta_2 = 0.71...$ by simulating uniformly sampled data from the unit square and computing MST or TSP length $L_1^2/\sqrt{n}$.

- Now consider data $\mathbf{X}_n$ sampled from a distribution μ on a compact Riemannian manifold $(\mathcal{M}^t, g)$ with volume form $\mathrm{vol}_{\mathcal{M}}$, embedded in $\mathbb{R}^r$ via an isometry φ. Assume $r \geq t \geq 2$.
- Since we are only given $\mathbf{Y}_n = \varphi(\mathbf{X}_n)$, the exact value of the *geodesic* length functional is unknown:

$$L_\gamma^{\mathcal{M}}(\varphi^{-1}(\mathbf{Y}_n)) = \min_{T:span} \sum_{(x_i,x_j)\in E(T)} d^{\mathcal{M}}(x_i, x_j)^\gamma$$

- However, we may estimate geodesic distances $d^{\mathcal{M}}$ with shortest path length $d^{\mathbf{Y}_n}$ as in Isomap, yielding an estimator $\hat{L}_\gamma^{\mathcal{M}}$ (computed from a potentially different MST).
- Then, $\hat{L}_\gamma^{\mathcal{M}}(\mathbf{Y}_n) = (1 + o(1))^\gamma \, L_\gamma^{\mathcal{M}}(\varphi^{-1}(\mathbf{Y}_n))$. For asymptotic considerations for Isomap, see Tenenbaum et al (2000).

Theorem (Costa, Hero, 2004)

Let data $\mathbf{X}_n$ be sampled i.i.d. from the probability measure μ on the compact Riemannian manifold $(\mathcal{M}^t, g)$. For an isometry $\varphi : \mathcal{M} \to \mathbb{R}^r$, $\mathbf{Y}_n = \varphi(\mathbf{X}_n)$, define the MST functional estimator:

$$\hat{L}_\gamma^{\mathcal{M}}(\mathbf{Y}_n) = \min_{T:span} \sum_{(y_i, y_j) \in E(T)} d^{\mathbf{Y}_n}(y_i, y_j)^\gamma$$

Let $\alpha = (r - \gamma)/r \in (0, 1)$ and $d\mu_c = f \cdot d\,\mathrm{vol}_{\mathcal{M}}$. Then,

$$\lim_{n\to\infty} \frac{1}{n^\alpha} \hat{L}_\gamma^{\mathcal{M}}(\mathbf{Y}_n) = \lim_{n\to\infty} \frac{1}{n^\alpha} L_\gamma^r(\varphi^{-1}(\mathbf{Y}_n))$$
$$= \beta_r \int_{\mathcal{M}} f(y)^\alpha \det(g_{ij}(y))^{(\alpha-1)/2} \mathrm{vol}_{\mathcal{M}}(y) \quad a.s.$$

Exercise. Prove the generalization. Construct estimators for the dimensionality constant α and manifold entropy $H_\alpha^{\mathcal{M}}(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{M}} f(y)^\alpha d\,\mathrm{vol}_{\mathcal{M}}$.

- The Global LEM algorithm uses MST as a 'backbone' to increase global stability.
- The *graph sum* of graphs sharing vertices is defined as:

$$G_i = (V, E_i, W_i) \quad \to \quad \bigoplus \lambda_i G_i := (V, \cup E_i, \sum \lambda_i W_i)$$

- Optimize w.r.t. the sum of the KNN and MST graphs with Gaussian weights, thus with combined Laplacian

$$\mathbf{L}(G_{KNN} \bigoplus \lambda G_{MST}) = \mathbf{L}(G_{KNN}) + \lambda \cdot \mathbf{L}(G_{MST})$$

- The asymptotic properties of minimal graphs and the parameters λ, K ensure neither graph dominates.

Table of Contents

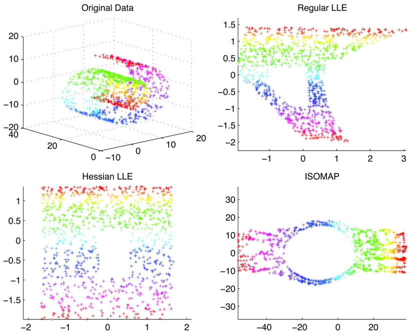

Figure: Embedding comparison of Swiss roll with hole.

HLLE (Donoho, Grimes, 2003) recovers parameter spaces of high-dimensional data such as articulated image libraries, where the convexity assumption is often violated.

- For a Riemannian manifold $\mathcal{M}$ with Levi-Civita connection ∇, $f \in C^\infty(\mathcal{M})$, the Hessian tensor $H_f \in \Gamma(T^*\mathcal{M} \otimes T^*\mathcal{M})$ (w.r.t. some local coordinate system $\{x^i\}$) is defined as:

$$H_f := \nabla df = \left(\frac{\partial^2 f}{\partial x^i \partial x^j} - \Gamma_{ij}^k \frac{\partial f}{\partial x^k} \right) dx^i \otimes dx^j$$

- Assumptions: $\mathcal{M} = \psi(\Theta)$ where the parameter space Θ is an open region of $\mathbb{R}^t$, and $\psi : \Theta \to \mathbb{R}^r$ is a locally isometric embedding. ($\Gamma_{ij}^k = 0$)
- Let μ be a probability measure on $\mathcal{M}$ with strictly positive density on $\mathcal{M} \setminus \partial\mathcal{M}$ from which the data is sampled.

$H_f(X, Y) = \nabla df(X, Y) = (\nabla_X df)(Y) = \nabla_X(df(Y)) - df(\nabla_X Y)$
$= \nabla_X(Y(f)) - \nabla_X Y(f) = X^i \nabla_i (Y^j \partial_j f) - X^i \nabla_i (Y^j \partial_j)(f)$
$= X^i \partial_i Y^j \partial_j f + X^i Y^j \partial_i \partial_j f \ - X^i \partial_i Y^j \partial_j f - X^i Y^j \Gamma_{ij}^k \partial_k f = X^i Y^j (\partial_i \partial_j - \Gamma_{ij}^k \partial_k) f$

- Let U_p be a neighborhood of $p \in \mathcal{M}$. Via ψ^{-1}, U_p inherits local *isometric* coordinates $\theta^i(q) = (x^i \circ \psi^{-1})(q)$ and the associated Hessian matrix $H_f^{iso}(p)$.
- Alternatively, smoothly identify U_p with a ball in $T_p\mathcal{M}$ via projection or the Riemannian exponential map.
- Viewing $T_p\mathcal{M}$ as an affine subspace of $\mathbb{R}^r$ gives orthonormal *tangent* / *geodesic* coordinates and matrices $H_f^{tan}(p)$ and $H_f^{geo}(p)$. Then:

$$\left\|H_f^{tan}(p)\right\|_F = \left\|H_f^{geo}(p)\right\|_F = \left\|H_f^{iso}(p)\right\|_F$$

Note only $H_f^{tan}(p)$ provides a tractable estimation scheme.

Exercise. Show the above quantities are well-defined w.r.t. coordinate transformations on $T_p\mathcal{M}$.

- Define the following quadratic form on the Sobolev space $W_2^2(\mathcal{M})$ which measures the average 'curviness' of f over $\mathcal{M}$:

$$\mathcal{H}(f) := \int_{\mathcal{M}} \left\| H_f^{tan}(p) \right\|_F^2 d\mu(p)$$

Theorem

The quadratic form $\mathcal{H}(\cdot)$ on $W_2^2(\mathcal{M})$ has a $(t+1)$-dimensional nullspace spanned by the constant function and the original isometric coordinates θ_i.

- Given data $\mathbf{X}$ on $\mathcal{M}$, our goal is to find an estimator $\hat{\mathcal{H}}$ of the linear operator form of $\mathcal{H}$, i.e. $\mathcal{H}(f) \simeq f(\mathbf{X})^T \cdot \hat{\mathcal{H}} \cdot f(\mathbf{X})$
- Eigenanalysis of $\hat{\mathcal{H}}$ retrieves the original coordinates.

Exercise. Prove the theorem. First show for C^∞ functions on Euclidean space, then use the natural pullback from $\mathcal{M}$ to Θ and $\left\| H_f^{tan} \right\|_F = \left\| H_f^{iso} \right\|_F$.

Estimation

- By abuse of notation, write N_i for the usual KNN set, the neighborhood U_{p_i}, and the identified ball in $T_{p_i}\mathcal{M}$.
- The tangent space at each point p_i is estimated by local PCA. The first t eigenvectors $u_1^{(i)}, \cdots, u_t^{(i)}$ of the Gram matrix give tangent coordinates for N_i.
- At each p_i, we find a $({}_t\mathrm{H}_2 \times K)$-matrix $\mathbf{H}^{(i)}$ that estimates the Hessian in the sense that for any $f \in C^2(\mathcal{M})$,

$$\mathbf{H}^{(i)} f^{(i)}, \quad \text{where } f^{(i)} = (\cdots, f(p_j), \cdots)^T, \quad j \in N_i$$

is a $t(t+1)/2-$vector whose entries approximate each

$$(H_f(p_i))_{\alpha\beta} = \frac{\partial^2 f}{\partial x^\alpha \partial x^\beta}(p_i), \quad \alpha \le \beta \le t$$

Our estimation scheme is: for each row corresponding to pairs α, β,

$$\frac{\partial^2 f}{\partial x^\alpha \partial x^\beta}(p_i) \simeq \sum_j \mathbf{H}^{(i)}_{(\alpha\beta),j} f(p_j) \quad \forall f \in C^\infty(\mathcal{M})$$

Write $\epsilon^{(i)}_j := p_j - p_i$. Substituting $f(p_j)$ by its 2nd order Taylor expansion $f(p_i) + \sum_k \frac{\partial f}{\partial x^k}(p_i)\,\epsilon^{(i)}_{j,k} + \frac{1}{2}\frac{\partial^2 f}{\partial x^k \partial x^\ell}(p_i)\,\epsilon^{(i)}_{j,k}\epsilon^{(i)}_{j,\ell}$ gives:

$$\frac{\partial^2 f}{\partial x^\alpha \partial x^\beta}(p_i) \simeq \left(\sum_j \mathbf{H}^{(i)}_{(\alpha\beta),j}\right) f(p_i) + \sum_k \frac{\partial f}{\partial x^k}(p_i) \left(\sum_j \mathbf{H}^{(i)}_{(\alpha\beta),j}\epsilon^{(i)}_{j,k}\right)$$
$$+ \frac{1}{2}\sum_{k,\ell} \frac{\partial^2 f}{\partial x^k \partial x^\ell}(p_i) \left(\sum_j \mathbf{H}^{(i)}_{(\alpha\beta),j}\epsilon^{(i)}_{j,k}\epsilon^{(i)}_{j,\ell}\right)$$

- Recalling $\epsilon^{(i)}_{*,k} = u^{(i)}_k$, we have the relations:

$$\mathbf{H}^{(i)\ T}_{(\alpha\beta)} 1_K = 0, \quad \mathbf{H}^{(i)\ T}_{(\alpha\beta)} u^{(i)}_k = 0, \quad \mathbf{H}^{(i)\ T}_{(\alpha\beta)} \left(u^{(i)}_k * u^{(i)}_\ell \right) = 2\delta^{\alpha,\beta}_{k,\ell}$$

where $*$ denotes entrywise multiplication.

- Solve by performing Gram-Schmidt orthogonalization on the following $K \times (1 + t + t(t+1)/2)$-matrix:

$$\left(1_K \cdots u^{(i)}_k \cdots u^{(i)}_k * u^{(i)}_\ell \cdots \right)$$

- Finally, $\hat{\mathcal{H}}$ is given by a form of contraction:

$$\hat{\mathcal{H}}_{jm} = \sum_i \sum_{\alpha,\beta} \mathbf{H}^{(i)}_{(\alpha\beta),j} \mathbf{H}^{(i)}_{(\alpha\beta),m}$$

Exercise. Show that $\mathcal{H}(f) \simeq f(\mathbf{X})^T \cdot \hat{\mathcal{H}} \cdot f(\mathbf{X})$ as desired. Calculate the time complexity of HLLE.

Table of Contents

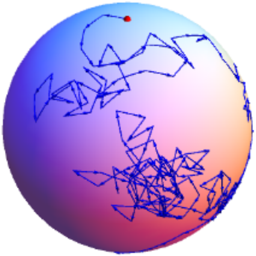

Figure: Random walk on the 2-sphere.

- Diffusion maps (Lafon et al, 2005) is based on the idea that two graphs are similar if they conduct analogous patterns of information propagation.

- Construct the KNN graph G of the data $\mathbf{X}$ with edge weights $\mathbf{W}_{ij}$ given by some kernel $k(x_i, x_j)$.
- Define a discrete-time Markov process (random walk) $M(t)$ on G with transition probability

$$\mathbf{P}_{ij} = P(M(t+1) = x_j | M(t) = x_i) = \mathbf{W}_{ij} / \sum_k \mathbf{W}_{ik}$$

- $\mathbf{P}$ has eigenvalues $1 = \lambda_0 \geq \cdots \geq \lambda_{n-1} \geq 0$ and a set of left and right eigenvectors

$$\phi_j^T \mathbf{P} = \lambda_j \phi_j^T, \quad \mathbf{P}\psi_j = \lambda_j \psi_j$$

  i.e. $\mathbf{P}$ has spectral decomposition $\Psi \Lambda \Phi^T$.
- $\lambda_1 < 1$ by the Perron-Frobenius theorem, $\psi_0 = 1_n$, and ϕ_0 is the unique stationary distribution for M.

- The m-step transition probabilities are:

$$\begin{aligned} p_m(x_j|x_i) &:= P(M(t+m) = x_j | M(t) = x_i) \\ &= (\mathbf{P}^m)_{ij} = \phi_0(x_j) + \textstyle\sum_{k=1}^{n-1} \lambda_k^m \psi_k(x_i)\phi_k(x_j) \end{aligned}$$

- Running M forward in time explores the geometry of $\mathbf{X}$ at larger scales. Thus m acts as both time and scale parameter.
- Define the *diffusion distance* as:

$$D_m(x_i, x_j)^2 = \sum_y \frac{1}{\phi_0(y)} (p_m(y|x_i) - p_m(y|x_j))^2$$

Note:

- Points are closer if they are connected by many short paths.
- D_m is robust to noise since it considers all possible routes.

Theorem (Spectral decay)

The diffusion distance satisfies:

$$D_m(x_i, x_j)^2 = \sum_{k=1}^{n-1} \lambda_k^{2m}(\psi_k(x_i) - \psi_k(x_j))^2 \quad \forall i, j$$

Exercise. Prove the theorem. Hint:

- Write $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D} = \mathrm{diag}(\sum_i \mathbf{W}_{ij})$
- Show that the symmetric operator $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ has eigenvalues λ_i, so that $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^T$ for some $\mathbf{U} \in O(n)$
- Compare with $\mathbf{P} = \Psi\Lambda\Psi^{-1}$ to show $\Psi = \mathbf{E}\mathbf{U}$, $\Phi = \mathbf{E}^{-1}\mathbf{U}$ for some diagonal matrix $\mathbf{E}$
- Prove that $\{\phi_k\}$ form a basis of $L^2(\mathbf{X}, \delta(\mathbf{X})/\phi_0)$.

- Truncating the sum at $k = t$ gives

$$D_m(x_i, x_j)^2 \simeq \left\| \Gamma_m^t(x_i) - \Gamma_m^t(x_j) \right\|^2$$

  where $\Gamma_m^t : \mathbf{X} \to \mathbb{R}^t$ is the *diffusion map*:

$$\Gamma_m^t(x) = (\lambda_1^m \psi_1(x), \cdots, \lambda_t^m \psi_t(x))^T$$

- Thus $\hat{y}_i = \Gamma_m^t(x_i)$ yields a nonlinear dimensionality reduction scheme which approximates diffusion distance by the embedded Euclidean distance.
- The algorithm depends on K, the kernel $\mathbf{W}$, step number m, and threshold t (decided by the spectral decay rate of $\mathbf{P}$).