Lecture 16: Diffusion maps

Last time we talked about spectral embeddings

$$F: X \to \mathbb{R}^J \qquad F(\underline{x}_i) = \left((\underline{v}_1)_i, (\underline{v}_2)_i, \dots, (\underline{v}_J)_i\right)^T$$

which formed the basis of the spectral clustering algorithm. At a high level, spectral clustering is nothing more than a two step procedure:

Step 1) Pre-process/embed data via F

Step 2) Apply standard clustering such as k-NN.

It is natural to ask whether our choice of F can be made more flexible? or more mathematically meaningful. For starters, observe that the kernel

$$K^J(\underline{x}, \underline{x}') = F(\underline{x})^T F(\underline{x}), \qquad F: X \to \mathbb{R}^J$$

is inconsistent, in the sense that $\lim_{J \to \infty} K^J$ does not exist.

Throughout this section we will consider the random walk graph Laplacian

$$L_{RW} = I - A, \qquad A = D^{-1} W$$

The material presented here is based on the seminal paper Coifman & Lafon, "Diffusion Maps" (2006).

Let $G = \{X, W\}$ be a weighted graph & for notational purposes we let $V = \{1, \dots, n\}$ be the set of vertices of G.

It is helpful to think of functions on the vertices of G: $u : V \longrightarrow \mathbb{R}$ $\quad u \equiv \underline{u} \in \mathbb{R}^n$

If $\underline{u}$ is such that $u(i) \geq 0$ & $\sum_{i=1}^{n} u(i) = 1$ then $\underline{u}$ represents a probability distribution on V. & we write $\underline{u} \in \mathbb{P}(V)$.

Observe that (assum's no lonely nodes)

$$A\mathbb{1} = D^{-1} W \mathbb{1} = D^{-1} \underline{d} = \mathbb{1}$$

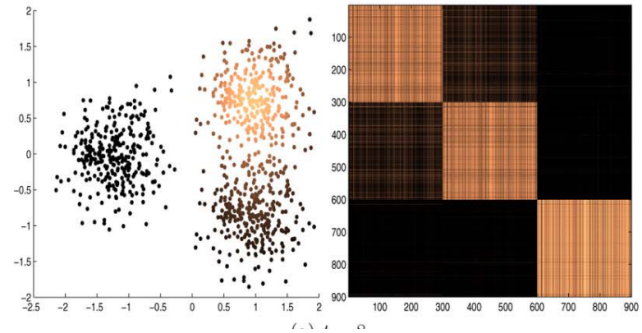
Thus, A is right stochastic & hence a transition matrix for a discrete Markov chain on V. Indeed the entry A_{ij} denotes the probability of going from vertex i to vertex j.

Thus, if $\underline{u} \in \mathbb{P}(V)$ denotes the initial state of the chain, the distribution at future time steps is given by

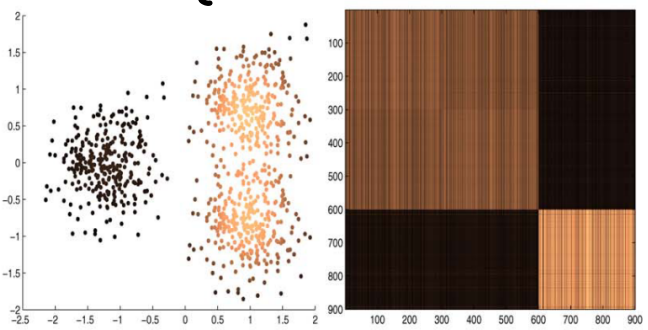
$$\underline{u}^t = \underline{u}^{t-1} A = \underline{u}^{t-2} A^2 = \underline{u}^0 A^t$$

The idea of Diffusion maps is to exploit the spectrum of A^t to reveal multi-scale geometric information in the data set.

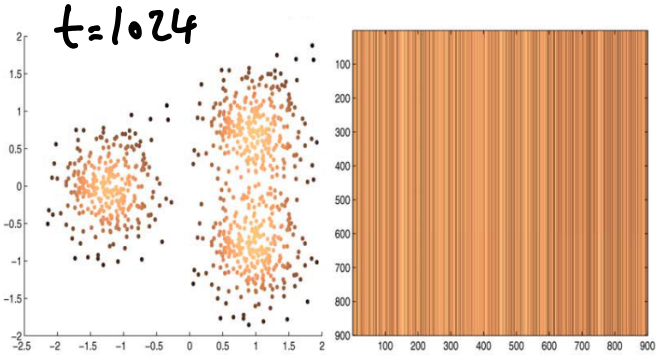
$t=8$



$t=64$



$t=1024$



- When t is small A^t reveals small scale clustering.
- When t is large we see the large scale clusters.

The idea of diffusion maps:

Let $P_t(i,\cdot)$ denote the distribution of the chain initiated at vertex $i \in V$ at time $t \geq 0$. Then the diffusion dist. between $\underline{x}_i, \underline{x}_j \in X$ is given by

$$D_t(\underline{x}_i, \underline{x}_j) := \left(\sum_{l=1}^{n} \frac{1}{d_l} |P_t(i,l) - P_t(j,l)|^2 \right)^{1/2}$$

$$= \| A^t_{i:} - A^t_{j:} \|_{D^{-1}}$$ ← i,j th rows of A^t

Indeed, defining the maps

$$F : X \longrightarrow \mathbb{R}^n \qquad F(\underline{x}_i) = \begin{bmatrix} P_t(i,1) \\ P_t(i,2) \\ \vdots \\ P_t(i,n) \end{bmatrix}$$

we see that

$$D_t^2(\underline{x}_i, \underline{x}_j) = \| F(\underline{x}_i) - F(\underline{x}_j) \|^2_{D^{-1}}$$

The above definition is very intuitive but it is unclear whether it can be computed efficiently. Turns out a better formulation, related to the spectral embedding is possible which we now derive.

Let $A = D^{-1}W$ & define

$$B = D^{1/2} A D^{-1/2} = D^{-1/2} W D^{-1/2}$$

observe B is symm. infact $I - B = L_{\text{normalized}}$

(recall $L_{Norm} = D^{-1/2}(D-W)D^{-1/2}$)

write $B = V\Lambda V^T$ (eigen decomp)

& intun $A = D^{-1/2} B D^{1/2}$

$= D^{-1/2}(V\Lambda V^T)D^{1/2}$

(since V is orth) $= (D^{-1/2}V)\Lambda(V^{-1}D^{1/2})$

$= Q\Lambda Q^{-1}$

In otherwords, A has an eigen decomp. It has the same eig. values as B & Colums of $Q = D^{-1/2}V$ are the right eig. vectors of A & rows of $Q^{-1} = V^T D^{1/2}$ are its left eigen vectors. In other words,

$$A = \sum_{j=1}^{n} \lambda_j \, \underline{q}_j \, \underline{r}_j^T, \qquad \underline{q}_j = D^{-1/2}\underline{v}_j, \quad \underline{r}_j = D^{1/2}\underline{v}_j$$

Finally, we have also that

$$A^t = \sum_{j=1}^{n} \lambda_j^t \, \underline{q}_j \, \underline{r}_j^T$$

Now observe that

$$\langle \underline{r}_i, \underline{r}_j \rangle_{D^{-1}} = \underline{r}_i^T D^{-1} \underline{r}_j = \underline{v}_i^T D^{1/2} D^{-1} D D^{1/2} \underline{v}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

so that the $\underline{r}_j$ form an orthonormal basis of $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_{D^{-1}})$. Using this observation we can see that (see Sec 8, la Porte et al 2008)

$$D_t(\underline{x}_i, \underline{x}_j)^2 = \| A^t_{i:} - A^t_{j:} \|^2_{D^{-1}}$$

$$= \sum_{l=1}^{n} \lambda_l^{2t} \left| (\underline{q}_l)_i - (\underline{q}_l)_j \right|^2$$

$$= \| \mathcal{F}(\underline{x}_i) - \mathcal{F}(\underline{x}_j) \|_2^2$$

where

$$\mathcal{F} : X \to \mathbb{R}^n, \quad \mathcal{F}(\underline{x}_i) = \begin{bmatrix} \lambda_1^t q_1(i) \\ \lambda_2^t q_2(i) \\ \vdots \\ \lambda_n^t q_n(i) \end{bmatrix}$$

(diffusion map)

Since $q_1 \propto \mathbb{1}$ & $\lambda_1 = 1$, some practitioners often choose to leave this eigen pair out of their algorithms.

Observe: The map $\mathcal{F}$ is very similar to the spectral embedding

$$F(x_i) = \begin{bmatrix} (v_1)_i \\ (v_2)_i \\ \vdots \\ (v_J)_i \end{bmatrix} \leftarrow v_i \text{ eig. vec. of D-W}$$

The construction of the diff. maps suggests families of embeddings of the form

$$\mathcal{F}_t(x_i) = \begin{bmatrix} \lambda_1^t \, q_1(i) \\ \lambda_2^t \, q_2(i) \\ \vdots \\ \lambda_n^t \, q_n(i) \end{bmatrix} \qquad t \geq 0$$

where (λ_j, q_j) are eigen pairs of a Laplacian operator (normalized, unnormalized, etc.

The map $\Psi(\underline{x}) \longrightarrow \mathbb{R}^n$ defines a kernel $K(\underline{x}, \underline{x}') = \Psi(\underline{x})^T \Psi(\underline{x}')$ which we call the diffusion kernel.

Under appropriate assumptions one can show that this kernel coincides with the heat kernel on the manifold on which the data X lies.

(see Coifman & Lafon 2008 for discussion)