Diffusion Geometry Review

Given Points $S = \{x_i\}_{i=1}^n$ in $\mathbb{R}^D$ we seek some parametrization $\phi : S \to \mathbb{R}^k$ (k should be small) that reveals the geometry (Low dimension structure, clustering).

Introduce a "kernel" $k(x, y) = \exp(-\frac{1}{\epsilon^2}\|x - y\|^2)$ where ϵ is a tuning parameter. Let L be the $n \times n$ matrix with entries $L(i, j) = k(x_i, x_j)$. Let $D(i, i) = \sum_{j=1}^n L(i, j)$. Set $M = LD^{-1}$ then M is a set of transition probabilities for a random walk on S.

For $t = 1, 2, 3, \ldots$ we are interested in the matrix M^t of transition probabilities for t steps of the random walk (t is another tuning parameter). Recall symmetrization "trick": set

$$\tilde{M} = D^{-\frac{1}{2}} M D^{\frac{1}{2}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}.$$

So $\tilde{M}$ is symmetric. Compute EVD of $\tilde{M}$. $\tilde{M} = V\Lambda V^*$. Then

$$M^t = D^{\frac{1}{2}} \tilde{M}^t D^{-\frac{1}{2}} = D^{\frac{1}{2}} V \Lambda^t V^* D^{-\frac{1}{2}}.$$

Assume the evals decaly, and pick a truncation parameter k. Then the (truncated) diffusion distance is

$$d_t(i, j) = \left(\sum_{p=1}^k \lambda_p^{2k} |v_p(i) - v_p(j)|^2 \right)^{\frac{1}{2}}.$$

So,

$$\begin{aligned} \Phi : S &\to \mathbb{R}^k \\ i &\mapsto \begin{bmatrix} \lambda_1^t v_1(i) \\ \vdots \\ \lambda_k^t v_k(t) \end{bmatrix} =: \mathbb{Z}_i \end{aligned}$$

Connection to heat conduction. Let $p \in \mathbb{R}^n$ be the vector of limiting probabilities $p = \lim_{t\to\infty} M^t p_0$. Recall $Mp = p \Rightarrow LD^{-1}p = p \Rightarrow (LD^{-1} - I)p = 0 \Rightarrow (L - D)D^{-1}p = 0$ where $(L - D)$ is graph Laplacian.

Ex. Square lattice in 2D. Consider heat conduction. Let $u \in \mathbb{R}^n$ be the vector of temperatures.

$$(u_w + u_e + u_n + u_s) - 4u_c = 0.$$

Standard 5-point stencil

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 4 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 8 & & & & \\ & 8 & & & \\ & & 8 & & \\ & & & 8 & \\ & & & & 8 \end{bmatrix}$$

and

$$A = L - D = \begin{bmatrix} 8 & & & & \\ & 8 & & & \\ 1 & 1 & -4 & 1 & 1 \\ & & & 8 & \\ & & & & 8 \end{bmatrix}$$

Graph Laplacian $Au = 0$. Heat conduction $\frac{\partial u}{\partial t} = Au$, solution $u = \exp(At)u_0$ where $\exp(At)$ is heat kernel and u_0 is initial value.

Recall n points $\{x_i\}_{i=1}^n$ in $\mathbb{R}^D$,

Computation issues: If n is large, e.g. $10^3 \le n \le 10^9$, D can be large! $D = 2, 3, \cdots 10^3$. Cost to assemble L is $O(Dn^2)$. Cost to compute top k evecs & evals of L is $O(kn^2)$. This is prohibitive when n is large.

Observe that many entries of L are very close to 0. Let us modify the kernel function. Pick a truncation distance δ and set

$$k(x, y) = \begin{cases} \exp(-\frac{1}{\epsilon^2}\|x-y\|^2, & \text{if } \|x-y\| \leq \delta, \\ 0 & \text{if } \|x-y\| > \delta. \end{cases}$$

This sparsifies L. On row i of L, the only non-zero entries $L(i, j)$ are the ones for which $\|x - y\| \leq \delta$. Then $\tilde{M}$ is sparse, and we can use e.g. Lanczos to compute the top k evals & evecs.

Problem: Finding the nearest neighbors can be costly. If done naivey, the cost is still Dn^2.

Solution - first try.

Say $D = 2$. Put down quad tree on domain. Assume points are distributed fairly uniformly. Cost to build the tree $\sim n$. Cost to search $\leq n$.

In 2D, the number of neighbors boxes $= 3^2 - 1 = 8$. In n-D, the number of neighbors boxes $= 3^n - 1$. This method scales abysmally with dimension.

Let us consider a non-uniform distribution. Build the tree adaptively. Split boxes only with "many" points in them. This still scales very badly with dimension. The search stage can get nasty.

"K-d trees": A technique to make tree searches work well for non-uniform distributions and for "sort of" high dimensions.

"Binary tree":