Diffusion Maps

Mladen Victor WICKERHAUSER

Washington University in St. Louis, Missouri
victor@wustl.edu
http://www.math.wustl.edu/~victor

Dimensionality Reduction and Manifold Estimation
PMF — University of Zagreb
Winter, 2022

Goals for Diffusion Maps

Setup: $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subset \mathbf{R}^d$ is a finite data set.

- Expect both d and n to be large.
- Some sufficiently close pairs in $\mathcal{V}$ are related.
- Start with some relation S on those pairs, defined on a small subset $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.
- Normalize S for use in a diffusion process.
- Extend S by diffusion to all of $\mathcal{V} \times \mathcal{V}$.
- Use the diffusion time parameter to:
  - define diffusion distances,
  - decompose the geometry of $\mathcal{V}$ by scales.

Similarity from Distance

For pairs $\mathbf{x}, \mathbf{y} \in \mathcal{V}$:

- Small distance is good. EG: Norm $\|\mathbf{x} - \mathbf{y}\| < \epsilon$
- ... or use a more general metric: $\mathrm{d}(\mathbf{x}, \mathbf{y}) < \epsilon$

But there is no natural value (other than perhaps ∞) for initially unrelated points.

So use similarity, like adjacency or connectedness:

- Adjacency: $A(i, j) = 1$ iff $(\mathbf{v}_i, \mathbf{v}_j) \in \mathcal{E}$, otherwise zero.
- Connectedness: Markov transition probability $M(i, j)$.

Transform distance to similarity using a kernel.

Kernels and Affinity

Affinity is defined with a kernel $k : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$:

- $k(\mathbf{x}, \mathbf{y}) = k(\mathbf{y}, \mathbf{x})$ is symmetric,
- $k(\mathbf{x}, \mathbf{y}) \geq 0$, with $k(\mathbf{x}, \mathbf{x}) > 0$,

Restrict to finite $\mathcal{V} = \{\mathbf{v}_i\} \subset \mathbf{R}^d$ to get a kernel matrix:

$$K(i, j) \stackrel{\text{def}}{=} k(\mathbf{v}_i, \mathbf{v}_j)$$

Symmetry of k implies $K^T = K$ is symmetric.

Remark. K is a weighted adjacency matrix for the complete (undirected) graph on $\mathcal{V}$

Gaussian Kernel

This is an "adjustable" kernel wth parameter $\sigma > 0$:

$$k(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right)$$

- evidently symmetric and nonnegative
- fit $\sigma > 0$ to the data (how?)
- $k(\mathbf{x}, \mathbf{y}) \to 1$ as $\|\mathbf{x} - \mathbf{y}\| \to 0$
- $k(\mathbf{x}, \mathbf{y}) \to 0$ rapidly as $\|\mathbf{x} - \mathbf{y}\| \to \infty$

Gaussian kernel matrix is positive: for all i, j,

$$K(i, j) \stackrel{\text{def}}{=} k(\mathbf{v}_i, \mathbf{v}_j) > 0$$

Normalization to Row Stochastic

Degree matrix: $D(i,j) = 0$ if $i \neq j$, else

$$D(i,i) \stackrel{\text{def}}{=} \sum_j K(i,j) = \sum_j k(\mathbf{v}_i, \mathbf{v}_j) > 0$$

Transition matrix:

$$P \stackrel{\text{def}}{=} D^{-1}K.$$

Lemma

P is row stochastic. $\square$

Remark. P is not symmetric, in general.

Properties

Let P be the transition matrix obtained from the gaussian kernel matrix on $\mathcal{V}$. Then

- P is positive.
- P is ergodic, since positive matrices are irreducible and aperiodic.
- P can be made *almost band diagonal*, with $\sum_{|i-j|>b} P(i,j) < \epsilon$, for any fixed $b \geq 1$ and $\epsilon > 0$, by choosing $\sigma = \sigma(b,\epsilon) > 0$ small enough.

Row Stochastic Spectral Radius

Lemma

If P is row stochastic, then $\rho(P) = 1$.

Proof.

Let $\mathbf{1} = (1, \ldots, 1)$. Then $P\mathbf{1} = \mathbf{1}$, so $\rho(P) \geq 1$.

Now, $\|P\|_\infty = 1$ by definition, and likewise $\|P^k\|_\infty = 1$ for all k (exercise!). But if $\rho(P) > 1$, then $\lim_{k\to\infty} \|P^k\|_\infty = \infty$.

Conclude that $\rho(P) = 1$. $\square$

By the Perron-Frobenius theorem, such P has a maximal eigenvalue $\rho(P) = 1$ of multiplicity 1, with all other eigenvalues satisfying $|\lambda| < 1$.

The dual principal eigenvector $\mathbf{v}$ (which solves $\mathbf{v}P = \mathbf{v}$), normalized to be a pdf, is called a *stationary distribution*.

Principal Eigenvectors

Lemma

P has a stationary distribution $\pi P = \pi$ given by

$$\pi(j) = \frac{D(j,j)}{\sum_i D(i,i)}.$$

Proof.

Write $P = D^{-1}K$. Since $\pi D^{-1} = \frac{1}{\sum_i D(i,i)}\mathbf{1}$ and $K = K^T$, compute

$$\pi P = \frac{1}{\sum_i D(i,i)}\mathbf{1}K = \frac{1}{\sum_i D(i,i)}\mathbf{1}K^T = \pi,$$

since $\mathbf{1}K^T$ is the vector of row sums of K which are just the degrees $\{D(i,i)\}$ of the vertices. $\square$

Reversibility

Lemma

The Markov chain with transition matrix P is reversible.

Proof.

For any indices i, j, by the symmetry of K in $P = D^{-1}K$,

$$\begin{aligned}
\pi(i)P(i,j) &= \frac{D(i,i)}{\sum_l D(l,l)} \frac{1}{D(i,i)} K(i,j) = \frac{K(i,j)}{\sum_l D(l,l)} = \frac{K(j,i)}{\sum_l D(l,l)} \\
&= \frac{D(j,j)}{\sum_l D(l,l)} \frac{1}{D(j,j)} K(j,i) = \pi(j)P(j,i).
\end{aligned}$$

This is exactly the detailed balance equation. $\square$

Diffusion Distances

Remark. For any power k, the (row stochastic) matrix P^k also has π as a stationary distribution.

Given P and its stationary distribution π, define a distance function on $\mathcal{V} = \{\mathbf{v}_i\}$ for every power $k > 0$ of P:

$$d_k(i,j)^2 = d_k(\mathbf{v}_i, \mathbf{v}_j)^2 \stackrel{\text{def}}{=} \sum_l \frac{[P^k(i,l) - P^k(j,l)]^2}{\pi(l)}$$

Idea: get a multiscale geometric analysis of $\mathcal{V}$ from dyadic distances

$$d_1, d_2, d_4, d_8, \ldots,$$

with the limit $d_\infty(\mathbf{v}_i, \mathbf{v}_j) = 0$ giving the largest scale, at which all points in $\mathcal{V}$ are identical.

Interpretations

- Rows of P^k are the posterior distributions after k steps, from elementary intial distributions.
- $d_k(i, j)$ is a weighted L^2 distance between distributions $u \mapsto P^k(i, u)$ and $u \mapsto P^k(j, u)$.
- d_k is a likelihood summed over all paths of length k.

Singular Vector Expansion

Suppose that $P = USV^T$ is a singular value decomposition of the $n \times n$ matrix P, where U, V are orthogonal, and S is diagonal:

$$U = \begin{pmatrix} \vdots & & \vdots \\ \mathbf{u}_1 & \dots & \mathbf{u}_n \\ \vdots & & \vdots \end{pmatrix}; \quad V = \begin{pmatrix} \vdots & & \vdots \\ \mathbf{v}_1 & \dots & \mathbf{v}_n \\ \vdots & & \vdots \end{pmatrix}; \quad S = \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_n \end{pmatrix},$$

with $0 \leq s_1 \leq s_2 \leq \cdots \leq s_n$. Then

$$P = \sum_l s_l \mathbf{u}_l \otimes \mathbf{v}_l \quad \text{meaning} \quad P(i,j) = \sum_l s_l \mathbf{u}_l(i) \mathbf{v}_l(j).$$

Exercise: prove this.

Remark. Need more, like $U = V$, to represent P^k for $k > 1$.

Symmetrizing

K is symmetric but P is not, so introduce an intermediate:

$$A = \Pi^{1/2} P \Pi^{-1/2}, \qquad \text{where } \Pi = \begin{pmatrix} \pi(1) & & \\ & \ddots & \\ & & \pi(n). \end{pmatrix},$$

This A is symmetric:

$$A(i,j) = \frac{\sqrt{\pi(i)}}{\sqrt{\pi(j)}} P(i,j) = \frac{K(i,j)}{\sqrt{\pi(i)}\sqrt{\pi(j)}},$$

so its eigenvectors form an orthonormal basis $\Theta = (\theta_l)$, with corresponding eigenvalues $\{\lambda_l\}$. Then

$$A = \sum_l \lambda_l \theta_l \otimes \theta_l \quad \text{meaning} \quad A(i,j) = \sum_l \lambda_l \theta_l(i) \theta_l(j).$$

Note: It may be assumed that $\lambda_1 = 1$ with $\theta_1 = \sqrt{\pi}$.

Eigenvalue Expansion

The relation between A and P gives

$$P(i,j) = \sum_l \lambda_l \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}} \theta_l(i)\theta_l(j) \stackrel{\text{def}}{=} \sum_l \lambda_l \psi_l(i)\phi_l(j),$$

where $\psi_l(i) = \theta_l(i)/\sqrt{\pi(i)}$ and $\phi_l(j) = \theta_l(j)\sqrt{\pi(j)}$.

Bases $\Psi = (\psi_l) = \Pi^{-1/2}\Theta$ and $\Phi = (\phi_l) = \Pi^{1/2}\Theta$ are *biorthogonal duals:*

$$\Psi^T\Phi = I = \Phi^T\Psi \qquad \text{meaning} \quad \langle \psi_p, \phi_q \rangle = \begin{cases} 1, & p = q, \\ 0, & p \neq q. \end{cases}$$

(This is because $\Theta^T\Theta = I$ by construction.)

Biorthogonal Functional Calculus

Lemma

$$P = \Psi\Lambda\Phi^T, \text{ with } \Lambda = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n. \end{pmatrix}.$$ □

Corollary

ψ_l is an eigenvector of P with eigenvalue λ_l. □

Corollary

$P^k = \Psi\Lambda^k\Phi^T$. □

Exercise: Perform the computations to prove these results.

Eigenvector Expansion of Powers

Lemma

Let ψ_l be an eigenvector of the eigenvalue λ_l of P. Then

$$d_k(i,j) = \left(\sum_l \lambda_l^{2k} \left[\psi_l(i) - \psi_l(j) \right]^2 \right)^{1/2}$$

Proof.

Recognize $d_k(i,j)^2 = \sum_u \left[\frac{P^k(i,u)}{\sqrt{\pi(u)}} - \frac{P^k(j,u)}{\sqrt{\pi(u)}} \right]^2$ as the squared L^2 norm of the difference two functions (of u) with orthonormal expansions (in $\{\theta_l\}$):

- $u \mapsto P^k(i,u)/\sqrt{\pi(u)} = \sum_l \lambda_l^k \psi_l(i) \theta_l(u)$, and
- $u \mapsto P^k(j,u)/\sqrt{\pi(u)} = \sum_l \lambda_l^k \psi_l(j) \theta_l(u)$.

Apply Parseval's formula to get the result. □

Diffusion Maps

For $k = 1, 2, \ldots,$ define the mapping $\Psi_k : \mathcal{V} \to \mathbf{R}^n$ by

$$\Psi_k(\mathbf{v}_i) = \Psi_k(i) = \begin{pmatrix} \lambda_1^k \psi_1(i) & \ldots & \lambda_n^k \psi_n(i) \end{pmatrix},$$

which is the ith row of $\Psi\Lambda^k$.

Theorem

Ψ_k is an injection from $\mathcal{V} \subset \mathbf{R}^d$ into $\mathbf{R}^n$ that maps diffusion distance to Euclidean distance:

$$d_k(\mathbf{v}_i, \mathbf{v}_j) = \|\Psi_k(\mathbf{v}_i) - \Psi_k(\mathbf{v}_j)\|.$$

Remark. It follows that d_k is a metric on $\mathcal{V}$, for every $k = 1, 2, \ldots$

Numerical Rank

If $1 = \lambda_1 > |\lambda_2| \geq \cdots$ are chosen in decreasing order, then truncating Ψ_k to the first m coordinates gives the least-L^2-distortion approximation in $\mathbf{R}^m$ to the full data set $\mathcal{V}$.

Fix $\epsilon > 0$ and define the *numerical rank* of the matrix P^k to be

$$n_\epsilon \stackrel{\text{def}}{=} \#\{j : |\lambda_j|^k \geq \epsilon\}$$

Then $n_\epsilon \to 1$ as $k \to \infty$ since $|\lambda_j| < 1$ for all $j > 1$.

Thus for large k, the diffusion map Ψ_k injects $\mathcal{V}$ into $\mathbf{R}^1$.