Mathematics for Data Sciences	**October 23,2011**

Lecture 9. Lumpability (Metastability) and MNcut

Instructor: Yuan Yao, Peking University *Scribe: Hong Cheng, Ping Qin*

1 Review of Diffusion Map

Recall $x_i \in \mathbb{R}^d, i = 1, 2, \cdots, n$,

$$W_{ij} = \exp\left(-\frac{d(x_i, x_j)^2}{2\varepsilon}\right),$$

W is a symmetrical $n \times n$ matrix.

Let $d_i = \sum_{j=1}^n W_{ij}$ and

$$D = \text{diag}(d_i), \quad P = D^{-1}W$$

and

$$S = D^{-1/2}WD^{-1/2} = I - \mathcal{L}, \quad \mathcal{L} = D^{-1/2}(D - W)D^{-1/2}.$$

Then

1) S is symmetrical, has n orthogonal eigenvectors $V = [v_1, v_2, \cdots, v_n]$,

$$S = V\Lambda V^T, \ \Lambda = \text{diag}(\lambda_i)^T \in \mathbb{R}^{n-1}, \ V^TV = I.$$

Here we assume that $1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \ldots \geq \lambda_{n-1}$ due to positivity of W.

2) $\Phi = D^{-1/2}V = [\phi_1, \phi_2, \cdots, \phi_n]$ are right eigenvectors of P, $P\Phi = \Phi\Lambda$.

3) $\Psi = D^{1/2}V = [\psi_1, \psi_2, \cdots, \psi_n]$ are left eigenvectors of P, $\Psi^T P = \Lambda\Psi^T$. Note that $\phi_0 = 1 \in \mathbb{R}^n$ and $\psi_0(i) = d_i / \sum_i d_i^2$. Thus ψ_0 is the same eigenvector as the stationary distribution $\pi(i) = d_i / \sum_i d_i$ $(\pi^T 1 = 1)$ up to a scaling factor.

Φ and Ψ are bi-orthogonal basis, i.e., $\phi_i^T D\psi_j = \delta_{ij}$ or simply $\Phi^T D\Psi = I$.

Define diffusion map

$$\Phi_t(x_i) = [\lambda_1^t \phi_1(i), \cdots, \lambda_{n-1}^t \phi_{n-1}(i)], \ t > 0.$$

A central question in this section is:

Why we choose right eigenvectors ϕ_i in diffusion map?

To answer this we will introduce the concept of *lumpability* in this lecture.

2 Lumpability of Markov Chain

P is row stochastic matrix on $V = \{1, 2, \cdots, n\}$. V has a partition Ω:

$$V = \cup_{i=1}^{k} \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \ i \neq j.$$

$$\Omega = \{\Omega_s : s = 1, \cdots, k\}.$$

Observe a sequence$\{x_0, x_1, \cdots, x_t\}$ sampled from a Markov chain whose transition matrix $\text{Prob}\{x_t = j : x_{t-1} = i\} = P_{ij}$. Relabel $x_t \mapsto y_t \in \{1, \cdots, k\}$ by

$$y_t = \sum_{s=1}^{k} s \mathcal{X}_{\Omega_s}(x_t).$$

Thus we obtain a sequence (y_t) which is a coarse-grained representation of original sequence.

Definition (Lumpability, Kemeny-Snell 1976)**.** P is lumpable with respect to partition Ω if the sequence $\{y_t\}$ is Markovian. In other words, the transition probabilities do not depend on the choice of initial distribution π_0 and history, *i.e.*

$$\text{Prob}_{\pi_0}\{y_t = k_t : y_{t-1} = k_{t-1}, \cdots, y_0 = k_0\} = \text{Prob}\{y_t = k_t : y_{t-1} = k_{t-1}\} \tag{1}$$

Theorem 2.1. **I.** (Kemeny-Snell 1976) P is lumpable with respect to partition $\Omega \Leftrightarrow \forall \Omega_s, \Omega_t \in \Omega$, $\forall i, j \in \Omega_s$, $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$, where $\hat{P}_{i\Omega_t} = \sum_{j \in \Omega_t} P_{ij}$.

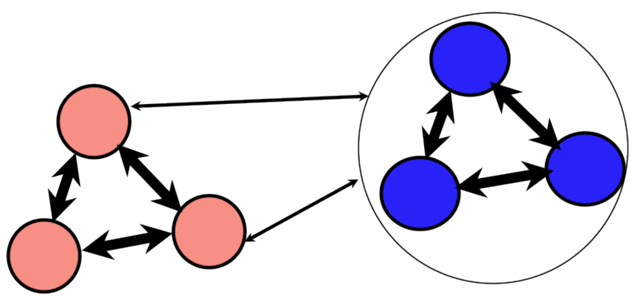

Figure 1: Lumpability condition $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$

II. (Meila-Shi 2001) P is lumpable with respect to partition Ω and $\hat{P}$ ($\hat{p}_{st} = \sum_{i \in \Omega_s, j \in \Omega_t} p_{ij}$) is nonsingular $\Leftrightarrow$ P has k independent piecewise constant right eigenvectors in $\text{span}\{\chi_{\Omega_s} : s = 1, \cdots, k\}$, χ is the characteristic function.

Example 1. Consider a linear chain with $2n$ nodes (Figure 2) whose adjacency matrix and degree matrix are given by

$$A = \begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & & 1 & 0 \end{bmatrix}, \quad D = \text{diag}\{1, 2, \cdots, 2, 1\}$$

So the transition matrix is $P = D^{-1}A$ which is illustrated in Figure 2. The spectrum of P includes two eigenvalues of magnitude 1, *i.e.* $\lambda_0 = 1$ and $\lambda_{n-1} = -1$. Although P is not a *primitive* matrix here, it is *lumpable*. Let $\Omega_1 = \{\text{odd nodes}\}$, $\Omega_2 = \{\text{even nodes}\}$. We can check that I and II are satisfied.

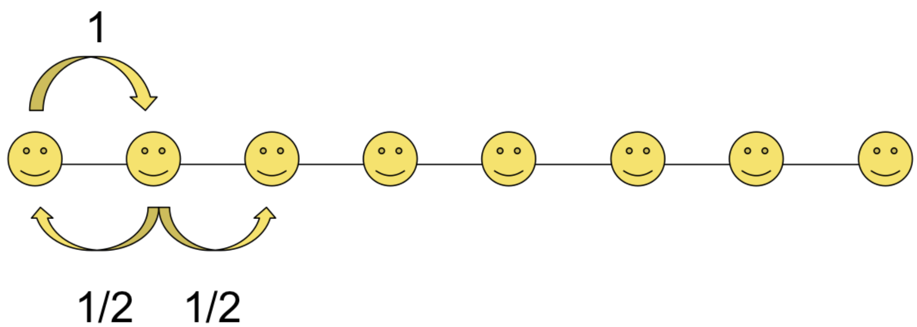

Figure 2: A linear chain of $2n$ nodes with a random walk.

To see I, note that for any two even nodes, say $i = 2$ and $j = 4$, $\hat{P}_{i\Omega_2} = \hat{P}_{j\Omega_2} = 1$ as their neighbors are all odd nodes, whence I is satisfied. To see II, note that ϕ_0 (associated with $\lambda_0 = 1$) is a constant vector while ϕ_1 (associated with $\lambda_{n-1} = -1$) is constant on even nodes and odd nodes respectively. Figure 3 shows the lumpable states when $n = 4$ in the left.

Note that lumpable states might not be optimal bi-partitions in $NCUT = Cut(S)/\min(vol(S), vol(\bar{S}))$. In this example, the optimal bi-partition by Ncut is given by $S = \{1, \ldots, n\}$, shown in the right of Figure 3. In fact the second largest eigenvalue $\lambda_1 = 0.9010$ with eigenvector

$$v_1 = [0.4714, 0.4247, 0.2939, 0.1049, -0.1049, -0.2939, -0.4247, -0.4714],$$

give the optimal bi-partition.

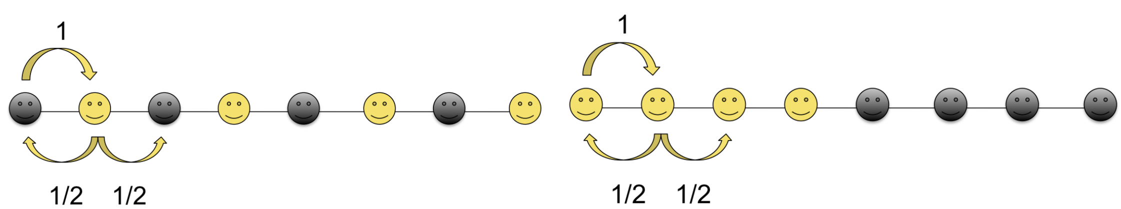

Figure 3: Left: two lumpable states; Right: optimal-bipartition of Ncut.

Example 2. Uncoupled Markov chains are lumpable, e.g.

$$P_0 = \begin{bmatrix} \Omega_1 & & \\ & \Omega_2 & \\ & & \Omega_3 \end{bmatrix}, \ \hat{P}_{it} = \hat{P}_{jt} = 0.$$

A markov chain $\tilde{P} = P_0 + O(\epsilon)$ is called nearly uncoupled Markov chain. Such Markov chains can be approximately represented as uncoupled Markov chains with *metastable states*, $\{\Omega_s\}$, where within metastable state transitions are fast while cross metastable states transitions are slow. Such a separation of scale in dynamics often appears in many phenomena in real lives, such as protein folding, your life transitions *primary schools* $\mapsto$ *middle schools* $\mapsto$ *high schools* $\mapsto$ *college/university* $\mapsto$ *work unit*, etc.

Before the proof of the theorem, we note that condition I is in fact equivalent to

$$VUPV = PV, \tag{2}$$

where U is a k-by-n matrix where each row is a uniform probability that

$$U_{is}^{k\times n} = \frac{1}{|\Omega_s|}\chi_{\Omega_s}(i), \quad i \in V,\ s \in \Omega,$$

and V is a n-by-k matrix where each column is a characteristic function on Ω_s,

$$V_{sj}^{n\times k} = \chi_{\Omega_s}(j).$$

With this we have $\hat{P} = UPV$ and $UV = I$. Such a matrix representation will be useful in the derivation of condition II. Now we give the proof of the main theorem.

Proof. **I.** "$\Rightarrow$" To see the necessity, P is lumpable w.r.t. partition Ω, then it is necessary that

$$\mathrm{Prob}_{\pi_0}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \mathrm{Prob}_{\pi_0}\{y_1 = t : y_0 = s\} = \hat{p}_{st}$$

which does not depend on π_0. Now assume there are two different initial distribution such that $\pi_0^{(1)}(i) = 1$ and $\pi_0^{(2)}(j) = 1$ for $\forall i, j \in \Omega_s$. Thus

$$\hat{p}_{i\Omega_t} = \mathrm{Prob}_{\pi_0^{(1)}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \hat{p}_{st} = \mathrm{Prob}_{\pi_0^{(2)}}\{x_1 \in \Omega_t : x_0 \in \Omega_s\} = \hat{p}_{j\Omega_t}.$$

"$\Leftarrow$" To show the sufficiency, we are going to show that if the condition is satisfied, then the probability

$$\mathrm{Prob}_{\pi_0}\{y_t = t : y_{t-1} = s, \cdots, y_0 = k_0\}$$

depends only on $\Omega_s, \Omega_t \in \Omega$. Probability above can be written as $\mathrm{Prob}_{\pi_{t-1}}(y_t = t)$ where π_{t-1} is a distribution with support only on Ω_s which depends on π_0 and history up to $t-1$. But since $\mathrm{Prob}_i(y_t = t) = \hat{p}_{i\Omega_t} \equiv \hat{p}_{st}$ for all $i \in \Omega_s$, then $\mathrm{Prob}_{\pi_{t-1}}(y_t = t) = \sum_{i\in\Omega_s} \pi_{t-1}\hat{p}_{i\Omega_t} = \hat{p}_{st}$ which only depends on Ω_s and Ω_t.

II.

"$\Rightarrow$"

Since $\hat{P}$ is nonsingular, let $\{\psi_i,\ i = 1, \cdots, k\}$ are independent right eigenvectors of $\hat{P}$, i.e., $\hat{P}\psi_i = \lambda_i\psi_i$. Define $\phi_i = V\psi_i$, then ϕ_i are independent piecewise constant vectors in $\mathrm{span}\{\chi_{\Omega_i}, i = 1, \cdots, k\}$. We have

$$P\phi_i = PV\psi_i = VUPV\psi_i = V\hat{P}\psi_i = \lambda_i V\psi_i = \lambda_i\phi_i,$$

i.e. ϕ_i are right eigenvectors of P.

"$\Leftarrow$"

Let $\{\phi_i, i = 1, \cdots, k\}$ be k independent piecewise constant right eigenvectors of P in $\mathrm{span}\{\mathcal{X}_{\Omega_i}, i = 1, \cdots, k\}$. There must be k independent vectors $\psi_i \in \mathbb{R}^k$ that satisfied $\phi_i = V\psi_i$. Then

$$P\phi_i = \lambda_i\phi_i \Rightarrow PV\psi_i = \lambda_i V\psi_i,$$

Multiplying VU to the left on both sides of the equation, we have

$$VUPV\psi_i = \lambda_i VUV\psi_i = \lambda_i V\psi_i = PV\psi_i,\ (UV = I),$$

which implies

$$(VUPV - PV)\Psi = 0, \quad \Psi = [\psi_1, \ldots, \psi_k].$$

Since Ψ is nonsingular due to independence of ψ_i, whence we must have $VUPV = PV$. □

3 Algorithm of Multiple Spectral Clustering

Meila-Shi (2001) calls the following algorithm as MNcut, standing for *modified Ncut*. Due to the theory above, perhaps we'd better to call it *multiple spectral clustering*.

1) Find top k right eigenvectors $P\Phi_i = \lambda_i \Phi_i$, $i = 1, \cdots, k$, $\lambda_i = 1 - o(\epsilon)$.
2) Embedding $Y^{n \times k} = [\phi_1, \cdots, \phi_k] \to$ diffusion map when $\lambda_i \approx 1$.
3) k-means (or other suitable clustering methods) on Y to k-clusters.