Mathematics for Data Sciences **Sep 20th, 2011**

Lecture 4. Diffusion Map, an introduction

Instructor: Xiuyuan Cheng, Princeton University *Scribe: Peng Luo, Wei Jin*

1 Review Of The Last Class And Some Hints Of The First Homework

In the last class, we introduced how to calculate the largest eigenvalue of matrix $\widehat{\Sigma}_n$ and the properties of the corresponding eigenvector v̂. First we say some points about last class.

Random vectors:$\{Y_i\}_{i=1}^n \sim N(0, \sigma_x^2 uu^T + \sigma_\varepsilon^2 I_p)$, where$\|u\|^2 = 1$. Define $R = SNR = \frac{\sigma_x^2}{\sigma_\varepsilon^2}$. Without of generality,we assume $\sigma_\varepsilon^2 = 1$.

The sample covariance matrix of Y is:$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n y_i y_i^t = \frac{1}{n}YY^T$, suppose one of its eigenvalue is λ and the corresponding unit eigenvector is $\hat{v}$, so $\hat{\Sigma}_n \hat{v} = \lambda\hat{v}$. After that, we relate the λ to the MP distribution by the trick:

$$Y_i = \Sigma^{\frac{1}{2}} Z_i \to Z_i \sim N(0, I_p), \text{where } \Sigma^{\frac{1}{2}} = \sigma_x^2 uu^T + \sigma_\varepsilon^2 I_p = Ruu^T + I_p \tag{1}$$

Then $S_n = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$ ~MP distribution.

Notice: $\hat{\Sigma}_n = \Sigma^{\frac{1}{2}} S_n \Sigma^{\frac{1}{2}}$ and λ $\hat{v}$ is eigenvalue and eigenvector of matrix $\hat{\Sigma}_n$. So

$$\Sigma^{\frac{1}{2}} S_n \Sigma^{\frac{1}{2}} \hat{v} = \lambda\hat{v} \text{ which implies } S_n\Sigma(\Sigma^{-\frac{1}{2}}\hat{v}) = \lambda(\Sigma^{-\frac{1}{2}}\hat{v}) \tag{2}$$

From the above equation, we find that λ and $\Sigma^{-\frac{1}{2}}\hat{v}$ is the eigenvalue and eigenvector of matrix $S_n\Sigma$. Suppose $c\Sigma^{-\frac{1}{2}}\hat{v} = v$ where the constant c makes v a unit eigenvector. So we have

$$c\hat{v} = \Sigma^{\frac{1}{2}} v \Rightarrow c^2 = c\hat{v}^T\hat{v} = v^T\Sigma v = v^T(\sigma_x^2 uu^T + \sigma_\varepsilon^2)v = R(u^Tv)^2 + 1 \tag{3}$$

In the last class, we computed the inner product of u and v(lecture03 equation22):

$$|u^Tv|^2 = \{\sigma_x^4 \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2)^2} d\mu^{MP}(t)\}^{-1} \tag{4}$$

$$= \{\frac{\sigma_x^4}{4\gamma}(-4\lambda + (a+b) + 2(\sqrt{(\lambda-a)(\lambda-b)}) + \frac{\lambda(2\lambda - (a+b))}{\sqrt{(\lambda-a)(\lambda-b)}})\}^{-1} \tag{5}$$

$$= \frac{1 - \frac{\gamma}{R^2}}{1 + \gamma + \frac{2\gamma}{R}} \tag{6}$$

where $R = SNR = \frac{\sigma_x^2}{\sigma_\varepsilon^2} = \sigma_x^2$,$\gamma = \sqrt{\frac{p}{n}}$. We can compute the inner product of u and $\hat{v}$ which we are really interested in from the above equation:

$$\begin{aligned} |u^T\hat{v}|^2 &= (\frac{1}{c}u^T\Sigma^{\frac{1}{2}}v)^2 = \frac{1}{c^2}((\Sigma^{\frac{1}{2}}u)^Tv)^2 = \frac{1}{c^2}(((Ruu^T + I_p)^{\frac{1}{2}}u)^Tv)^2 = \frac{1}{c^2}((\sqrt{(1+R)}u)^Tv)^2 \\ &= \frac{(1+R)(u^Tv)^2}{R(u^Tv)^2 + 1} = \frac{1 + R - \frac{\gamma}{R} - \frac{\gamma}{R^2}}{1 + R + \gamma + \frac{\gamma}{R}} = \frac{1 - \frac{\gamma}{R^2}}{1 + \frac{\gamma}{R}} \end{aligned}$$

In lecture03, we didn't compute two equations(see equation (17)and(22) in lecture03) in details. Here is my point to calculate them:

$$\int_a^b \frac{t}{\lambda - t}\mu^{MP}(t)dt := T(\lambda) \text{ (equation (17) in lecture03)} \tag{7}$$

From above equation, we can get:

$$\int_a^b \frac{t^2}{(\lambda - t)^2}\mu^{MP}(t)dt = -T(\lambda) - \lambda T^{'}(\lambda) \text{ (equation (22) in lecture03)} \tag{8}$$

So we just focus on $T(\lambda)$.

Define:

$$m(z) := \int_R \frac{1}{(z-t)}\mu^{MP}(t)dt,\ z \in C \tag{9}$$

$m(z)$ is called Stieltjes Transformation of density μ^{MP}.If z $\in R$, the transformation is called Hilbert Transformation. Further details can be found in Reference [Tao] (Topics on Random Matrix Theory), Sec. 2.4.3 (the end of page 169) for the definition of Stietljes transform of a density $p(t)dt$ on $\mathbb{R}$ (the book is using $s(z)$ instead of $m(z)$ in class).

$m(z)$ satisfies the equation:

$$\gamma z m(z)^2 + (z - (1-\gamma))m(z) + 1 = 0 \Longleftrightarrow z + \frac{1}{m(z)} = \frac{1}{1+\gamma m(z)} \tag{10}$$

From the equation, one can take derivative of z on both side to obtain $m'(z)$ in terms of m and z.

Notice:

$$1 + T(\lambda) = 1 + \int_a^b \frac{t}{\lambda - t}\mu^{MP}(t)dt = \int_a^b \frac{\lambda - t + t}{\lambda - t}\mu^{MP}(t)dt = \lambda m(\lambda) \tag{11}$$

So we can compute $T(\lambda)$ by $m(\lambda)$

In the last problem of first homework, we analyze Wigner Matrix $W = [w_{ij}]_{n\times n}, w_{ij} = w_{ji}, w_{ij} \sim N(0, \frac{\sigma}{\sqrt{n}})$. The answer is

eigenvalue is	$\lambda = R + \frac{1}{R}$
eigenvector satisfies	$(u^T\hat{v})^2 = 1 - \frac{1}{R^2}$

2 Introduction To The Diffusion Map

2.1 Manifold Learning Method

Here is the development of manifold learning method:

$$\text{PCA} \longrightarrow \text{LLE} \longrightarrow \begin{cases} \text{Laplacian Eigen Map} \\ \text{Hessian LLE} \\ \text{Diffusion MAp} \end{cases}$$

$$\text{MSE} \longrightarrow \text{ISOMAP}$$

Please read the Todd Wittman's slides for the comparison of different manifold learning method. You can find it in the website:http: www.math.pku.edu.cn/teachers/yaoy/Spring2011/. Lecture11.

Figure 1: Order the face

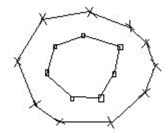

Figure 2: Two circles

2.2 Examples

The following three problems can be solved by diffusion map.

Ex1: order the face. How to put the photos of figure one in order?

Ex2:"3D". Fig. 2 of CoifmanLafon'06 paper.

Ex3:spectral clustering. How to separate the points in figure 2?

2.3 Method

In this section,we introduce the general diffusion map.

Suppose $x_1, x_2, \ldots, x_n \in R^p$,we create a symmetric matrix $W_{n\times n} = \{w_{ij}\}$, such that $w_{ij} = k(x_i, x_j) = k(\|x_i - x_j\|^2_{R^p})$, where $k(x, y)$ is the similarity function. For example, we can choose

$$k(x, y) = \exp\{-\frac{\|x-y\|^2}{2\varepsilon}\} \text{ or } k(x, y) = I_{\{\|x_i - x_j\| < \delta\}} \tag{12}$$

Next, we create a $n \times n$ diagonal matrix D, where $D_{ii} = \sum_{j=1}^{n} W_{ij}$.

$A := D^{-1}W$, So

$$\sum_{j=1}^{n} A_{ij} = 1 \ \forall i \in \{1, 2, \cdots, n\} \ (A_{ij} \geq 0) \tag{13}$$

Based on matrix A, we can construct a discrete time Markov chain: $\{X_t\}_{t\in N}$ which satifies

$$P(X_{t+1} = x_j \mid X_t = x_i) = A_{ij} \tag{14}$$

$$S := D^{\frac{1}{2}} W D^{\frac{1}{2}} = V\Lambda V^T \text{where } VV^T = I_n, \Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_n) \tag{15}$$

So

$$A = D^{-1}W = D^{-1}(D^{-\frac{1}{2}} S D^{-\frac{1}{2}}) = D^{-\frac{1}{2}} S D^{\frac{1}{2}} = D^{-\frac{1}{2}} V\Lambda V^T D^{\frac{1}{2}} = \Phi\Lambda\Psi^T (\Phi = D^{-\frac{1}{2}} V, \Psi = V^T D^{\frac{1}{2}}) \tag{16}$$

Thus $\Phi\Psi^T = I_n$ and we can get $A\Phi = \Phi\Lambda, \Psi^T A = \Lambda\Psi^T$.

Suppose $\Phi = [\phi_0, \phi_1, \cdots, \phi_n]$, So $A[\phi_0, \phi_1, \cdots, \phi_n] = [\lambda_0\phi_0, \lambda_1\phi_1, \cdots, \lambda_n\phi_n]$, where $\lambda_0 = 1$, $\phi_0 = e_n$.

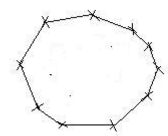

Figure 3: EX2 single circle

Define map:

$$\Phi_t(x_i) = [(\lambda_1)^t\phi_1(i), (\lambda_2)^t\phi_2(i), \cdots, (\lambda_{n-1})^t\phi_{n-1}(i)] \ (t > 0) \tag{17}$$

$\phi_k(i)$ is the i-th entry of ϕ_k.

Truncate the mapping where only those eigenvalues whose absolute value are larger than δ, some positive constant, are saved: suppose $\lambda_1, \lambda_2, \cdots, \lambda_m$ $s.t.$ $|\lambda_i| \geq \delta$

$$\Phi_t^\delta(x_i) = [(\lambda_1)^t\phi_1(i), (\lambda_2)^t\phi_2(i), \cdots, (\lambda_m)^t\phi_m(i)] \tag{18}$$

Diffusion distance:

$$D_t(x_i, x_j) := \| \Phi_t(x_i) - \Phi_t(x_j) \|^2 \tag{19}$$

2.4 Simple examples

Three examples about diffusion map:

EX1: two circles.

Suppose graph $G : (V, E)$. Matrix W satisfies $w_{ij} > 0$, if and only if $(i, j) \in E$. Choose $k(x, y) = I_{\|x-y\|<\delta}$. In this case,

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},$$

where A_1 is a $n_1 \times n_1$ matrix, A_2 is a $n_2 \times n_2$ matrix, $n_1 + n_2 = n$.

Notice that the eigenvalue $\lambda_0 = 1$ of A is of multiplicity 2, the two eigenvectors are $\phi_0 = 1_n$ and $\phi_0' = [c_1 1_{n1}^T, c_2 1_{n2}^T]^T$ $c_1 \neq c_2$.

$$\text{Diffusion Map} : \begin{cases} \Phi_t^{1D}(x_1), \cdots, \Phi_t^{1D}(x_{n_1}) = c_1 \\ \Phi_t^{1D}(x_{n_1+1}), \cdots, \Phi_t^{1D}(x_n) = c_2 \end{cases}$$

EX2: ring graph. "single circle"

In this case, W is a circulant matrix

$$W = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 1 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

The eigenvalue of W is $\lambda_k = \cos\frac{2\pi k}{n}$ $k = 0, 1, \cdots, \frac{n}{2}$ and the corresponding eigenvector is $(u_k)_j = e^{i\frac{2\pi}{n}kj}$ $j = 1, \cdots, n$. So we can get $\Phi_t^{2D}(x_i) = (\cos\frac{2\pi kj}{n}, \sin\frac{2\pi kj}{n})c^t$

EX3: order the face.

$$L := A - I = D^{-1}W - I$$

$$L_\varepsilon := \frac{1}{\varepsilon}(A_\varepsilon - I) \overset{\varepsilon\to 0}{\longrightarrow} \text{backward Kolmogorov operator}$$

$$L_\varepsilon f = \frac{1}{2}\triangle_M f - \nabla f \cdot \nabla v \Rightarrow L_\varepsilon = \lambda\phi \Rightarrow \begin{cases} \frac{1}{2}\phi^{''}(s) - \phi^{'}(s)V^{'}(s) = \lambda\phi(s) \\ \phi^{'}(0) = \phi^{'}(1) = 0 \end{cases}$$

Where $V(s)$ is the Gibbs weight and $p(s) = e^{-V(x)}$ is the density of data points along the curve. $\triangle_M$ is Laplace-Beltrami Operator. If $p(x) = const$, we can get

$$V(s) = const \Rightarrow \phi^{''}(s) = 2\lambda\phi(s) \Rightarrow \phi_k(s) = \cos(k\pi s), 2\lambda_k = -k^2\pi^2 \tag{20}$$

On the other hand $p(s) \neq const$, one can show [1] that $\phi_1(s)$ is monotonic for arbitrary $p(s)$. As a result, the faces can still be ordered by using $\phi_1(s)$.

2.5 Properties of Transition Matrix of Markov Chain

Suppose A is a Markov Chain Transition Matrix.

1 $\lambda(A) \subset [-1, 1]$.

proof: assume λ and v are the eigenvalue and eigenvector of A, so$Av = \lambda v$. Find j_0 $s.t.$ $|v_{j_0}| \geq |v_j|, \forall j \neq j_0$ where v_j is the j-th entry of v. Then:

$$\lambda v_{j_0} = (Av)_{j_0} = \sum_{j=1}^{n} A_{j_0 j} v_j$$

So:

$$|\lambda||v_{j_0}| = |\sum_{j=1}^{n} A_{j_0 j} v_j| \leq \sum_{j=1}^{n} A_{j_0 j}|v_j| \leq |v_{j_0}|$$

2 Define: A is irreducible, if and only if $\forall(i,j)$ $\exists t$ $s.t.$ $(A^t)_{ij} > 0 \Leftrightarrow$ Graph is connected

fact:A is irreducible $\Rightarrow \lambda = 1$

3 Define: A is primitive, if and only if $\exists t > 0$ $s.t. \forall(i,j)$ $(A^t)_{ij} > 0$

fact: A is primitive $\Rightarrow -1 \notin \lambda(A)$

fact: A is irreducible and $A_{ii} > 0$ $\forall i \Rightarrow$ A is primitive

4 Theory(Perron-Frobenius):if $A_{ij} > 0$, then:

$$\exists r > 0, s.t.\ r \in \lambda(A)\ and\ \forall\lambda \in \lambda(A), \lambda \neq r, |\lambda| < r$$

5 Fact: If $k(x,y)$ is heat kernel $\Rightarrow \lambda(A) \geq 0$

[1] by changing to polar coordinate $p(s)\phi'(s) = r(s)\cos\theta(s)$, $\phi(s) = r(s)\sin\theta(s)$ (the so-called 'Prufer Transform') and then try to show that $\phi'(s)$ is never zero on $(0, 1)$.