Mathematics for Data Sciences **2011.10.11**

Lecture 7. Random Walk on Graphs: Perron-Frobenius Vector and PageRank

Instructor: Yuan Yao, Peking University *Scribe: Yuan Lu, Bowei Yan*

Introduction

In the past month we talked about two topics: one is the sample mean and sample covariance matrix (PCA) in high dimensional spaces. We have learned that when dimension p is large and sample size n is relatively small, in contrast to the traditional statistics where p is fixed and $n \to \infty$, both sample mean and PCA may have problems. In particular, Stein's phenomenon shows that in high dimensional space with independent Gaussian distributions, the sample mean is worse than a shrinkage estimator; moreover, random matrix theory sheds light on that in high dimensional space with sample size in a fixed ratio of dimension, the sample covariance matrix and PCA may not reflect the signal faithfully. These phenomena start a new philosophy in high dimensional data analysis that to overcome the curse of dimensionality, additional constraints has to be put that data never distribute in every corner in high dimensional spaces. Sparsity is a common assumption in modern high dimensional statistics. For example, data variation may only depend on a small number of variables; independence of Gaussian random fields leads to sparse covariance matrix; and the assumption of conditional independence can also lead to sparse inverse covariance matrix. In particular, an assumption that data concentrate around a low dimensional manifold in high dimensional spaces, leads to manifold learning or nonlinear dimensionality reduction, e.g. ISOMAP, LLE, and Diffusion Maps etc. This assumption often finds example in computer vision, graphics, and image processing.

We have talked about Diffusion Map as a model of Random walk (Markov Chain) on data graph. In all, we need to consider the prior distribution of the data in high dimensional space. Among other methods of Manifold Learning, Diffusion Map is the most powerful and successful method in that it combines both geometry and stochastic process. In the next few sections, we will talk about general theory of random walks (Markov chains) on graphs which are related to data analysis, including other applications related to diffusion maps. From this one can learn the origin of diffusion map ideas.

Random Walk on Graphs.

- Perron-Frobenius Vector and Google's PageRank: this is about Perron-Frobenius theory for non-negative matrices, which leads to the characterization of nonnegative primary eigenvectors, such as stationary distributions of Markov chains; application examples include Google's PageRank.
- Fiedler Vector, Cheeger's Inequality, and Spectral Bipartition: this is about the second eigenvector in a Markov chain, mostly reduced from graph Laplacians (Fiedler theory, Cheeger's Inequality), which is the basis for spectral partition.
- Lumpability/Metastability, piecewise const right eigenvector, and Multiple spectral clustering ("MN-cut" by Maila-Shi, 2001): this is about when to use multiple eigenvectors, whose relationship with lumpability or metastability of Markov chains, widely used in diffusion map, image segmentation, etc.
- Mean first passage time, commute time distance: the origins of diffusion distances.

Today we shall discuss the first part.

1 Perron-Frobenius Theory and Applications

Given $A_{n\times n}$, we define $A > 0$, *positive matrix*, iff $A_{ij} > 0\ \forall i, j$, and $A \geq 0$, *nonnegative matrix*, iff $A_{ij} \geq 0$ $\forall i, j$.
Note that this definition is different from positive definite:
$A \succ 0 \Leftrightarrow$ A is positive-definite $\Leftrightarrow x^T Ax > 0 \quad \forall x \neq 0$
$A \succeq 0 \Leftrightarrow$ A is semi-positive-definite $\Leftrightarrow x^T Ax \geq 0 \quad \forall x \neq 0$

Theorem 1.1 (Perron Theorem for Positive Matrix)**.** Assume that $A > 0$, *i.e.* a positive matrix. Then
1) $\exists \lambda^* > 0, \nu^* > 0, \|\nu^*\|_2 = 1, s.t.\ A\nu^* = \lambda^*\nu^*$, ν^* is a right eigenvector
($\exists \lambda^* > 0, \omega > 0, \|\omega\|_2 = 1, s.t.\ (\omega^T)A = \lambda^*\omega^T$, left eigenvector)
2) $\forall$ other eigenvalue λ of A, $|\lambda| < \lambda^*$
3) ν^* is unique up to rescaling or λ^* is simple
4) Collotz-Wielandt Formula

$$\lambda^* = \max_{x\geq 0, x\neq 0} \min_{x_i \neq 0} \frac{[Ax]_i}{x_i} = \min_{x>0} \max \frac{[Ax]_i}{x_i}.$$

Such eigenvectors will be called Perron vectors. This result can be extended to nonnegative matrices.

Theorem 1.2 (Nonnegative Matrix, Perron)**.** Assume that $A \geq 0$, *i.e.* nonnegative. Then
1') $\exists \lambda^* > 0, \nu^* \geq 0, \|\nu^*\|_2 = 1, s.t.\ A\nu^* = \lambda^*\nu^*$ (similar to left eigenvector)
2') $\forall$ other eigenvalue λ of A, $|\lambda| \leq \lambda^*$
3') ν^* is NOT unique
4) Collotz-Wielandt Formula

$$\lambda^* = \max_{x\geq 0, x\neq 0} \min_{x_i \neq 0} \frac{[Ax]_i}{x_i} = \min_{x>0} \max \frac{[Ax]_i}{x_i}$$

Notice the changes in 1'), 2'), and 3'). Perron vectors are nonnegative rather than positive. In the nonnegative situation what we lose is the uniqueness in λ^* (2')and ν^* (3'). The next question is: can we add more conditions such that the loss can be remedied? Now recall the concept of irreducible and primitive matrices introduced before.

Irreducibility exactly describes the case that the induced graph from A is connected, *i.e.* every pair of nodes are connected by a path of arbitrary length. However primitivity strengths this condition to k-connected, *i.e.* every pair of nodes are connected by a path of length k.

Definition 1 (Irreducible)**.** The following definitions are equivalent:
1) For any $1 \leq i, j \leq n$, there is an integer $k \in \mathbb{Z}$, s.t. $A^k_{ij} > 0$; $\Leftrightarrow$
2) Graph $G = (V, E)$ ($V = \{1, \ldots, n\}$ and $\{i, j\} \in E$ iff $A_{ij} > 0$) is (path-) connected, *i.e.* $\forall \{i, j\} \in E$, there is a path $(x_0, x_1, \ldots, x_t) \in V^{n+1}$ where $i = x_0$ and $x_t = j$, connecting i and j.

Definition 2 (Primitive)**.** The following characterizations hold:
1) There is an integer $k \in \mathbb{Z}$, such that $\forall i, j$, $A^k_{ij} > 0$; $\Leftrightarrow$
2) Any node pair $\{i, j\} \in E$ are connected with a path of length no more than k; $\Leftrightarrow$
3) A has unique $\lambda^* = \max |\lambda|$; $\Leftarrow$
4) A is irreducible and $A_{ii} > 0$, for some i,

Note that condition 4) is sufficient for primitivity but not necessary; all the first three conditions are necessary and sufficient for primitivity.

When A is a primitive matrix, A^k becomes a positive matrix for some k, then we can recover 1), 2) and 3) for positivity and uniqueness. This leads to the following Perron-Frobenius theorem.

Theorem 1.3 (Nonnegative Matrix, Perron-Frobenius)**.** Assume that $A \geq 0$ and A is primitive. Then
1) $\exists \lambda^* > 0, \nu^* > 0, \|\nu^*\|_2 = 1, s.t.\ A\nu^* = \lambda^*\nu^*$ (right eigenvector)
and $\exists \omega > 0, \|\omega\|_2 = 1, s.t.\ (\omega^T)A = \lambda^*\omega^T$ (left eigenvector)
2) $\forall$ other eigenvalue λ of A, $|\lambda| < \lambda^*$
3) ν^* is unique
4) Collotz-Wielandt Formula

$$\lambda^* = \max_{x>0} \min \frac{[Ax]_i}{x_i} = \min_{x>0} \max \frac{[Ax]_i}{x_i}$$

Such eigenvectors and eigenvalue will be called as Perron-Frobenius or primary eigenvectors/eigenvalue.

Example 1 (Markov Chain)**.** Given a graph $G = (V, E)$, consider a random walk on G with transition probability $P_{ij} = Prob(x_{t+1} = j | x_t = i) \geq 0$. Thus P is a row-stochastic or row-Markov matrix i.e. $P \cdot \overrightarrow{1} = \overrightarrow{1}$ where $\overrightarrow{1} \in \mathbb{R}^n$ is the vector with all elements being 1. From Perron theorem for nonnegative matrices, we know
$\nu^* = \overrightarrow{1} > 0$ is a right Perron eigenvector of P
$\lambda^* = 1$ is a Perron eigenvalue and all other eigenvalues $|\lambda| \leq 1 = \lambda^*$
$\exists$ left PF-eigenvector π such that $\pi^T P = \pi^T$ where $\pi \geq 0$, $1^T\pi = 1$; such π is called an invariant/equilibrium distribution
P is irreducible (G is connected) $\Rightarrow \pi$ unique
P is primitive (G connected by paths of length $\leq k$) $\Rightarrow |\lambda| = 1$ unique

$$\Leftrightarrow \lim_{t\to\infty} \pi_0^T P^k \to \pi^T \quad \forall \pi_0 \geq 0, 1^T\pi_0 = 1$$

This means when we take powers of P, *i.e.* P^k, all rows of P^k will converge to the stationary distribution π^T. Such a convergence only holds when P is primitive. If P is not primitive, *e.g.* $P = [0, 1; 1, 0]$ (whose eigenvalues are 1 and -1), P^k always oscillates and never converges.

What's the rate of the convergence? Let

$$\gamma = \max\{|\lambda_2|, \cdots, |\lambda_n|\}, \quad \lambda_1 = 1$$

and $\pi_t = (P^T)^t\pi_0$, roughly speaking we have

$$\|\pi_t - \pi\|_1 \sim O(e^{-\gamma t}).$$

This type of rates will be seen in various mixing time estimations.

A famous application of Markov chain in modern data analysis is Google's PageRank, although Google's current search engine only exploits that as one factor among many others. But you can still install Google Toolbar on your browser and inspect the PageRank scores of webpages.

Example 2 (Pagerank)**.** Consider a directed weighted graph $G = (V, E, W)$ whose weight matrix decodes the webpage link structure:

$$w_{ij} = \begin{cases} \#\{link: \quad i \mapsto j\}, & (i,j) \in E \\ 0, \quad otherwise \end{cases}$$

Define an out-degree vector $d_i^o = \sum_{j=1}^n w_{ij}$, which measures the number of out-links from i. A diagonal matrix $D = \text{diag}(d_i)$ and a row Markov matrix $P_1 = D^{-1}W$, assumed for simplicity that all nodes have non-empty out-degree. This P_1 accounts for a random walk according to the link structure of webpages. One would expect that stationary distributions of such random walks will disclose the importance of webpages: the more visits, the more important. However Perron-Frobenius above tells us that to obtain a unique stationary distribution, we need a primitive Markov matrix. For this purpose, Google's PageRank does the following trick.

Let $P_\alpha = \alpha P_1 + (1-\alpha)E$, where $E = \frac{1}{n}1 \cdot 1^T$ is a random surfer model, *i.e.* one can jump to any other webpage uniformly. So in the model P_α, a browser will play a dice: he will jump according to link structure with probability α or randomly surf with probability $1-\alpha$. With $1 > \alpha > 0$, the existence of random surfer model makes P a positive matrix, whence $\exists! \pi s.t. P_\alpha^T \pi = \pi$ (means 'there exists a unique π'). Google choose $\alpha = 0.85$ and in this case π gives PageRank scores.

Now you probably can figure out how to cheat PageRank. If there are many cross links between a small set of nodes (for example, Wikipedia), those nodes must appear to be high in PageRank. This phenomenon actually has been exploited by spam webpages, and even scholar citations. After learning the nature of PageRank, we should be aware of such mis-behaviors.

Finally we discussed a bit on Kleinberg's HITS algorithm, which is based on singular value decomposition (SVD) of link matrix W. Above we have defined the out-degree d^o. Similarly we can define in-degree $d_k^i = \sum_j w_{jk}$. High out-degree webpages can be regarded as *hubs*, as they provide more links to others. On the other hand, high in-degree webpages are regarded as *authorities*, as they were cited by others intensively. Basically in/out-degrees can be used to rank webpages, which gives relative ranking as authorities/hubs. It turns out Kleinberg's HITS algorithm gives pretty similar results to in/out-degree ranking.

Definition 3 (HITS-authority)**.** This use primary right singular vector of W as scores to give the ranking. To understand this, define $L_a = W^T W$. Primary right singular vector of W is just a primary eigenvector of nonnegative symmetric matrix L_a. Since $L_a(i,j) = \sum_k W_{ki} W_{kj}$, thus it counts the number of references which cites both i and j, *i.e.* $\sum_k \#\{i \leftarrow k \rightarrow j\}$. The higher value of $L_a(i,j)$ the more references received on the pair of nodes. Therefore Perron vector tend to rank the webpages according to authority.

Definition 4 (HITS-hub)**.** This use primary left singular vector of W as scores to give the ranking. Define $L_h = WW^T$, whence primary left singular vector of W is just a primary eigenvector of nonnegative symmetric matrix L_h. Similarly $L_h(i,j) = \sum_k W_{ik} W_{jk}$, which counts the number of links from both i and j, hitting the same target, *i.e.* $\sum_k \#\{i \rightarrow k \leftarrow j\}$. Therefore the Perron vector L_h gives hub-ranking.

The last example is about Economic Growth model where the Debreu introduced nonnegative matrix into its study. Similar applications include population growth and exchange market, etc.

Example 3 (Economic Growth/Population/Exchange Market)**.** Consider a market consisting n sectors (or families, currencies) where A_{ij} represents for each unit investment on sector j, how much the outcome in sector i. The nonnegative constraint $A_{ij} \geq 0$ requires that i and j are not *mutually inhibitor*, which means that investment in sector j does not decrease products in sector i. We study the dynamics $x_{t+1} = Ax_t$ and its long term behavior as $t \to \infty$ which describes the economic growth.

Moreover in exchange market, an additional requirement is put as $A_{ij} = 1/A_{ji}$, which is called *reciprocal matrix*. Such matrices are also used for preference aggregation in decision theory by Saaty.

From Perron-Frobenius theory we get: $\exists \lambda* > 0 \quad \exists \nu^* \geq 0 \quad A\nu^* = \lambda^* \nu^*$ and $\exists \omega^* \geq 0 \quad A^T \omega^* = \lambda^* \omega^*$. When A is primitive, ($A^k > 0$, *i.e.* investment in one sector will increase the product in another sector in no more than k industrial periods), we have for all other eigenvalues λ, $|\lambda| < \lambda^*$ and ω^*, ν^* are unique. In this case one can check that the long term economic growth is governed by

$$A^t \to (\lambda^*)^t \nu^* \omega^{*T}$$

where
1) for all i, $\frac{(x_t)_i}{(x_{t-1})_i} \to \lambda^*$
2) distribution of resources $\to \nu^* / \sum_i \nu_i^*$, so the distribution is actually not balanced
3) ω_i^* gives the relative value of investment on sector i in long term

2 Proof of Perron Theorem for Positive Matrices

A complete proof can be found in Meyer's book, Chapter 8. Our proof below is based on optimization view, which is related to the Collotz-Wielandt Formula.

Assume that $A > 0$. Consider the following optimization problem:

$$\begin{aligned} & \max \delta \\ s.t. \quad & Ax \geq \delta x \\ & x \geq 0 \\ & x \neq 0 \end{aligned}$$

Without loss of generality, assume that $1^T x = 1$. Let $y = Ax$ and consider the growth factor $\frac{y_i}{x_i}$, for $x_i \neq 0$. Our purpose above is to maximize the minimal growth factor δ ($y_i/x_i \geq \delta$).

Let λ^* be optimal value with $\nu^* \geq 0, \quad 1^T \nu^* = 1$, and $A\nu^* \geq \lambda^* \nu^*$. Our purpose is to show
1) $A\nu^* = \lambda^* \nu^*$
2) $\nu^* > 0$
3) ν^* and λ^* are unique.
4) For other eigenvalue $\lambda \quad (\lambda z = Az \quad when \quad z \neq 0)$, $|\lambda| < \lambda^*$.

Sketchy Proof of Perron Theorem. 1) If $A\nu^* \neq \lambda^* \nu^*$, then for some i, $[A\nu^*]_i > \lambda^* \nu_i^*$. Below we will find an increase of λ^*, which is thus not optimal. Define $\tilde{\nu} = \nu^* + \epsilon e_i$ with $\epsilon > 0$ and e_i denotes the vector which is one on the i^{th} component and zero otherwise.

For those $j \neq i$,

$$(A\tilde{\nu})_j = (A\nu^*)_j + \epsilon (Ae_i)_j = \lambda^* \nu_j^* + \epsilon A_{ji} > \lambda^* \nu_j^* = \lambda^* \tilde{\nu}_j$$

where the last inequality is due to $A > 0$.

For those $j = i$,

$$(A\tilde{\nu})_i = (A\nu^*)_i + \epsilon (Ae_i)_i > \lambda^* \nu_i^* + \epsilon A_{ii}.$$

Since $\lambda^* \tilde{\nu}_i = \lambda^* \nu_i^* + \epsilon \lambda^*$, we have

$$(A\tilde{\nu})_i - (\lambda^* \tilde{\nu})_i + \epsilon (A_{ii} - \lambda^*) = (A\nu^*)_i - (\lambda^* \nu_i^*) - \epsilon (\lambda^* - A_{ii}) > 0,$$

where the last inequality holds for small enough $\epsilon > 0$. That means, for some small $\epsilon > 0$, $(A\tilde{\nu}) > \lambda^* \tilde{\nu}$. Thus λ^* is not optimal, which leads to a contradiction.

2) Assume on the contrary, for some k, $\nu_k^* = 0$, then $(A\nu^*)_k = \lambda^* \nu_k^* = 0$. But $A > 0$, $\nu^* \geq 0$ and $\nu^* \neq 0$, so there $\exists i$, $\nu_i^* > 0$, which implies that $A\nu^* > 0$. That contradicts to the previous conclusion. So $\nu^* > 0$, which followed by $\lambda^* > 0$ (otherwise $A\nu^* > 0 = \lambda^* \nu^* = A\nu^*$).

3) We are going to show that for every $\nu \geq 0$, $A\nu = \mu\nu \Rightarrow \mu = \lambda^*$. Following the same reasoning above, A must have a left Perron vector $\omega^* > 0$, s.t. $A^T \omega^* = \lambda^* \omega^*$. Then $\lambda^* (\omega^{*T} \nu) = \omega^{*T} A\nu = \mu (\omega^{*T} \nu)$. Since $\omega^{*T} \nu > 0$ ($\omega^* > 0, \quad \nu \geq 0$), there must be $\lambda^* = \mu$, i.e. λ^* is unique, and ν^* is unique.

4) For any other eigenvalue $Az = \lambda z$, $A|z| \geq |Az| = |\lambda||z|$, so $|\lambda| \leq \lambda^*$. Then we prove that $|\lambda| < \lambda^*$. Before proceeding, we need the following lemma.

Lemma 2.1. $Az = \lambda z, |\lambda| = \lambda^*, z \neq 0 \quad \Rightarrow \quad A|z| = \lambda^*|z|. \quad \lambda^* = \max_i |\lambda_i(A)|$

Proof of Lemma. Since $|\lambda| = \lambda^*$,

$$A|z| = |A||z| \geq |Az| = |\lambda||z| = \lambda^*|z|$$

Assume that $\exists k, \quad \frac{1}{\lambda^*}A|z_k| > |z_k|$. Denote $Y = \frac{1}{\lambda^*}A|z| - |z| \geq 0$, then $Y_k > 0$. Using that $A > 0, x \geq 0, x \neq 0, \Rightarrow Ax > 0$, we can get

$$\Rightarrow \frac{1}{\lambda^*}AY > 0, \quad \frac{1}{\lambda^*}A|z| > 0$$

$$\Rightarrow \exists \epsilon > 0, \quad \frac{A}{\lambda^*}Y > \epsilon\frac{A}{\lambda^*}|z|$$

$$\Rightarrow \bar{A}Y > \epsilon\bar{A}|z|, \qquad \bar{A} = \frac{A}{\lambda^*}$$

$$\Rightarrow \bar{A}^2|z| - \bar{A}|z| > \epsilon\bar{A}|z|$$

$$\Rightarrow \frac{\bar{A}^2}{1+\epsilon}|z| > \bar{A}|z|$$

$$\Rightarrow B = \frac{\bar{A}}{1+\epsilon}, \quad 0 = \lim_{m\to\infty} B^m\bar{A}|z| \geq \bar{A}|z|$$

$$\Rightarrow \bar{A}|z| = 0 \quad \Rightarrow \quad |z| = 0 \quad \Rightarrow \quad Y = 0 \quad \Rightarrow \quad \bar{A}|z| = \lambda^*|z|$$

□

Equipped with this lemma, assume that we have $Az = \lambda z$ ($z \neq 0$) with $|\lambda| = \lambda^*$, then

$$A|z| = \lambda^*|z| = |\lambda||z| = |Az| \quad \Rightarrow \quad |\sum_j \bar{a}_{ij}z_j| = \sum_j \bar{a}_{ij}|z_j|, \quad \bar{A} = \frac{A}{\lambda^*}$$

which implies that z_j has the same sign, *i.e.* $z_j \geq 0$ or $z_j \leq 0$ ($\forall j$). In both cases $|z|$ ($z \neq 0$) is a nonnegative eigenvector $A|z| = \lambda|z|$ which implies $\lambda = \lambda^*$ by 3). □

3 Perron-Frobenius theory for Nonnegative Tensors

Some researchers, e.g. Liqun Qi (Polytechnic University of Hong Kong), Lek-Heng Lim (U Chicago) and Kung-Ching Chang (PKU) et al. recently generalize Perron-Frobenius theory to nonnegative tensors, which may open a field toward *PageRank* for hypergraphs and array or tensor data. For example, $A(i, j, k)$ is a 3-tensor of dimension n, representing for each object $1 \leq i \leq n$, which object of j and k are closer to i.

A tensor of order-m and dimension-n means an array of n^m real numbers:

$$A = (a_{i_1,\ldots,i_m}), \qquad 1 \leq i_1, \ldots, i_m \leq n$$

An n-vector $\nu = (\nu_1, \ldots, \nu_n)^T$ is called an *eigenvector*, if

$$A\nu^{[m-1]} = \lambda\nu^{m-1}$$

for some $\lambda \in \mathbb{R}$, where

$$A\nu^{[m-1]} := \sum_{i_2,\ldots,i_m=1}^{n} a_{ki_2\ldots i_m}\nu_{i_2}\cdots\nu_{i_m}, \quad \nu^{m-1} := (\nu_1^{m-1},\ldots,\nu_n^{m-1})^T.$$

Chang-Pearson-Zhang [2008] extends Perron-Frobenius theorem to show the existence of $\lambda^* > 0$ and $\nu^* > 0$ when $A > 0$ is irreducible.