A Mathematical Introduction to Data Science

Yuan Yao

School of Mathematical Sciences, Peking University, Beijing, China 100871

E-mail address: yuany@math.pku.edu.cn

URL: http://www.math.pku.edu.cn/teachers/yaoy/Fall2012/lectures.pdf

This is a working draft last updated on
October 14, 2014

2000 *Mathematics Subject Classification.* Primary

Key words and phrases. keywords

Special thanks to Amit Singer, Weinan E, Xiuyuan Cheng, and the following students in PKU who help scribe lecture notes with various improvements: Hong Cheng, Chao Deng, Yanzhen Deng, Chendi Huang, Lei Huang, Shujiao Huang, Longlong Jiang, Yuwei Jiang, Wei Jin, Changcheng Li, Xiaoguang Li, Tengyuan Liang, Feng Lin, Yaning Liu, Peng Luo, Wulin Luo, Tangjie Lv, Yuan Lv, Hongyu Meng, Ping Qin, Jie Ren, Hu Sheng, Zhiming Wang, Yuting Wei, Jiechao Xiong, Jie Xu, Bowei Yan, Jun Yin, and Yue Zhao.

Abstract. This monograph aims to provide graduate students or senior graduates in applied mathematics, computer science and statistics an introduction to data science from a mathematical perspective. It is focused around a central topic in data analysis, Principal Component Analysis (PCA), with a divergence to some mathematical theories for deeper understanding, such as random matrix theory, convex optimization, random walks on graphs, geometric and topological perspectives in data analysis.

Contents

Preface

This book is used in a course instructed by Yuan Yao at Peking University, part of which is based on a similar course led by Amit Singer at Princeton University.

> *If knowledge comes from the impressions made upon us by natural objects, it is impossible to procure knowledge without the use of objects which impress the mind.* –John Dewey

> *It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.* –John W. Tukey

CHAPTER 1

Multidimensional Scaling and Principal Component Analysis

1. Classical MDS

Multidimensional Scaling (MDS) roots in psychology [YH41] which aims to recover Euclidean coordinates given pairwise distance metrics or dissimilarities. It is equivalent to PCA when pairwise distances are Euclidean. In the core of theoretical foundation of MDS lies the notion of positive definite functions [Sch37, Sch38a, Sch38b] (or see the survey [Bav11]) which has been the foundation of the kernel method in statistics [Wah90] and modern machine learning society (http://www.kernel-machines.org/).

In this section we study classical MDS, or metric Multidimensional scaling problem. The problem of classical MDS or isometric Euclidean embedding: given pairwise distances between data points, can we find a system of Euclidean coordinates for those points whose pairwise distances meet given constraints?

Consider a forward problem: given a set of points $x_1, x_2, ..., x_n \in \mathbb{R}^p$, let

$$X = [x_1, x_2, ..., x_n]^{p\times n}.$$

The distance between point x_i and x_j satisfies

$$d_{ij}^2 = \|x_i - x_j\|^2 = (x_i - x_j)^T (x_i - x_j) = {x_i}^T x_i + {x_j}^T x_j - 2{x_i}^T x_j.$$

Now we are considering the inverse problem: given d_{ij}, find a $\{x_i\}$ satisfying the relations above. Clearly the solutions are not unique as any Euclidean transform on $\{x_i\}$ gives another solution. General ideas of classic (metric) MDS is:

(1) transform squared distance matrix $D = [d_{ij}^2]$ to an inner product form;
(2) compute the eigen-decomposition for this inner product form.

Below we shall see how to do this given D.

Let K be the inner product matrix

$$K = X^T X,$$

with $k = \mathrm{diag}(K_{ii}) \in \mathbb{R}^n$. So

$$D = (d_{ij}^2) = k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K.$$

where $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^n$.

Define the mean and the centered data

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n} \cdot X \cdot \mathbf{1},$$

$$\widetilde{x}_i = x_i - \widehat{\mu}_n = x_i - \frac{1}{n} \cdot X \cdot \mathbf{1},$$

or

$$\widetilde{X} = X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T.$$

Thus,

$$\begin{aligned}
\tilde{K} &\triangleq \tilde{X}^T \tilde{X} \\
&= (X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T)^T (X - \frac{1}{n} X \cdot \mathbf{1} \cdot \mathbf{1}^T) \\
&= K - \frac{1}{n} K \cdot \mathbf{1} \cdot \mathbf{1}^T - \frac{1}{n} \mathbf{1} \cdot \mathbf{1}^T \cdot K + \frac{1}{n^2} \cdot \mathbf{1} \cdot \mathbf{1}^T \cdot K \cdot \mathbf{1} \ \cdot \mathbf{1}^T.
\end{aligned}$$

Let

$$B = -\frac{1}{2} H \cdot D \cdot H^T$$

where $H = I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T$. H is called as a *centering matrix.*

So

$$B = -\frac{1}{2} H \cdot (k \cdot \mathbf{1}^T + \mathbf{1} \cdot k^T - 2K) \cdot H^T$$

Since $k \cdot \mathbf{1}^T \cdot H^T = k \cdot \mathbf{1}(I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) = k \cdot \mathbf{1} - k(\frac{\mathbf{1}^T \cdot \mathbf{1}}{n}) \cdot \mathbf{1} = 0$, we have $H \cdot k \ \mathbf{1} \cdot H^T = H \cdot \mathbf{1} \cdot k^T \cdot H^T = 0$.

Therefore,

$$\begin{aligned}
B &= H \cdot K \cdot H^T = (I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) \cdot K \cdot (I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T) \\
&= K - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1} \cdot K - \frac{1}{n} \cdot K \cdot \mathbf{1} \cdot \mathbf{1}^T + \frac{1}{n^2} \cdot \mathbf{1}(\mathbf{1}^T \cdot K\mathbf{1}) \cdot \mathbf{1}^T \\
&= \tilde{K}.
\end{aligned}$$

That is,

$$B = -\frac{1}{2} H \cdot D \cdot H^T = \tilde{X}^T \tilde{X}.$$

Note that often we define the covariance matrix

$$\widehat{\Sigma}_n \triangleq \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \widehat{\mu}_n)(x_i - \widehat{\mu}_n)^T = \frac{1}{n-1} \widetilde{X} \widetilde{X}^T.$$

Above we have shown that given a squared distance matrix $D = (d_{ij}^2)$, we can convert it to an inner product matrix by $B = -\frac{1}{2} HDH^T$. Eigen-decomposition applied to B will give rise the Euclidean coordinates centered at the origin.

In practice, one often chooses top k nonzero eigenvectors of B for a k-dimensional Euclidean embedding of data.

Hence $\widetilde{X}_k$ gives k-dimensional Euclidean coordinations for the n points.

In Matlab, the command for computing MDS is "`cmdscale`", short for Classical Multidimensional Scaling. For non-metric MDS, you may choose "`mdscale`". Figure 1 shows an example of MDS.

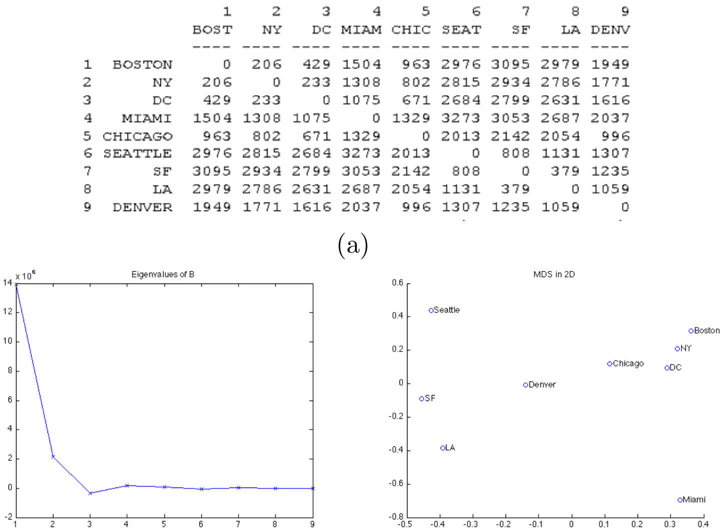
2. Theory of MDS (Young/Househölder/Schoenberg'1938)

Definition (Positive Semi-definite)**.** Suppose $A^{n \times n}$ is a real symmetric matrix,then: A is p.s.d.(positive semi-definite)$(A \succeq 0) \iff \forall v \in \mathbb{R}^n, v^T A v \geq 0 \iff A = Y^T Y$

Property. Suppose $A^{n \times n}$, $B^{n \times n}$ are real symmetric matrix, $A \succeq 0$, $B \succeq 0$. Then we have:

Algorithm 1: Classical MDS Algorithm

```
**Input**: A squared distance matrix $D^{n\times n}$ with $D_{ij} = d_{ij}^2$.
**Output**: Euclidean $k$-dimensional coordinates $\widetilde{X}_k \in \mathbb{R}^{k\times n}$ of data.
1 Compute $B = -\frac{1}{2} H \cdot D \cdot H^T$, where H is a centering matrix.
2 Compute Eigenvalue decomposition $B = U\Lambda U^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where
  $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;
3 Choose top $k$ nonzero eigenvalues and corresponding eigenvectors, $\widetilde{X}_k = U_k \Lambda_k{}^{\frac{1}{2}}$
  where
      $U_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n,$
      $\Lambda_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$
  with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$.
```



```
                1    2    3    4    5    6    7    8    9
             BOST   NY   DC MIAM CHIC SEAT   SF   LA DENV
             ---- ---- ---- ---- ---- ---- ---- ---- ----
1   BOSTON      0  206  429 1504  963 2976 3095 2979 1949
2       NY    206    0  233 1308  802 2815 2934 2786 1771
3       DC    429  233    0 1075  671 2684 2799 2631 1616
4    MIAMI   1504 1308 1075    0 1329 3273 3053 2687 2037
5  CHICAGO    963  802  671 1329    0 2013 2142 2054  996
6  SEATTLE   2976 2815 2684 3273 2013    0  808 1131 1307
7       SF   3095 2934 2799 3053 2142  808    0  379 1235
8       LA   2979 2786 2631 2687 2054 1131  379    0 1059
9   DENVER   1949 1771 1616 2037  996 1307 1235 1059    0
```

(a)

(b) (c)

FIGURE 1. MDS of nine cities in USA. (a) Pairwise distances between 9 cities; (b) Eigenvalues of $B = -\frac{1}{2} H \cdot D \cdot H^T$; (c) MDS embedding with top-2 eigenvectors.

(1) $A + B \succeq 0$;
(2) $A \circ B \succeq 0$;

where $A \circ B$ is called Hadamard product and $(A \circ B)_{i,j} := A_{i,j} \times B_{i,j}$.

Definition (Conditionally Negative Definite)**.** Let $A^{n\times n}$ be a real symmetric matrix. A is c.n.d.(conditionally negative definite) $\iff$ $\forall v \in \mathbb{R}^n$, such that $\mathbf{1}^T v = \sum_{i=1}^n v_i = 0$, there holds $v^T A v \leq 0$

Lemma 2.1 (Young/Househölder-Schoenberg '1938)**.** For any signed probability measure α ($\alpha \in \mathbb{R}^n, \sum_{i=1}^n \alpha_i = 1$),

$$B_\alpha = -\frac{1}{2} H_\alpha C H_\alpha^T \succeq 0 \iff C \text{ is c.n.d.}$$

where H_α is Householder centering matrix: $H_\alpha = \mathbf{I} - \mathbf{1} \cdot \alpha^T$.

PROOF. $\Leftarrow$ We are to show if C is c.n.d., then $B_\alpha \geq 0$. Taking an arbitrary $x \in \mathbb{R}^n$,

$$x^T B_\alpha x = -\frac{1}{2} x^T H_\alpha C H_\alpha^T x = -\frac{1}{2}(H_\alpha^T x)^T C (H_\alpha^T x).$$

Now we are going to show that $y = H_\alpha^T x$ satisfies $\mathbf{1}^T y = 0$. In fact,

$$\mathbf{1}^T \cdot H_\alpha^T x = \mathbf{1}^T \cdot (\mathbf{I} - \alpha \cdot \mathbf{1}^T) x = (1 - \mathbf{1}^T \cdot \alpha)\mathbf{1}^T \cdot x = 0$$

as $\mathbf{1}^T \cdot \alpha = 1$ for signed probability measure α. Therefore,

$$x^T B_\alpha x = -\frac{1}{2}(H_\alpha^T x)^T C (H_\alpha^T x) \geq 0,$$

as C is c.n.d.

$\Rightarrow$ Now it remains to show if $B_\alpha \geq 0$ then C is c.n.d. For $\forall x \in \mathbb{R}^n$ satisfying $\mathbf{1}^T \cdot x = 0$, we have

$$H_\alpha^T x = (\mathbf{I} - \alpha \cdot \mathbf{1}^T) x = x - \alpha \cdot \mathbf{1}^T x = x$$

Thus,

$$x^T C x = (H_\alpha^T x)^T C (H_\alpha^T x) = x^T H_\alpha C H_\alpha^T x = -2 x^T B_\alpha x \leq 0,$$

as desired.

This completes the proof. □

Theorem 2.2 (Classical MDS)**.** Let $D^{n \times n}$ a real symmetric matrix. $C = D - \dfrac{1}{2} d \cdot \mathbf{1}^T - \dfrac{1}{2} \mathbf{1} \cdot d^T$, $d = \mathrm{diag}(D)$. Then:

(1) $B_\alpha = -\frac{1}{2} H_\alpha D H_\alpha^T = -\frac{1}{2} H_\alpha C H_\alpha^T$ for $\forall \alpha$ signed probability measrue;
(2) $C_{i,j} = B_{i,i}(\alpha) + B_{j,j}(\alpha) - 2B_{i,j}(\alpha)$
(3) D c.n.d. $\iff$ C c.n.d.
(4) C c.n.d. $\Rightarrow$ C is a square distance matrix (i.e. $\exists Y^{n \times k}$ s.t. $C_{i,j} = \sum_{m=1}^{k} (y_{i,m} - y_{j,m})^2$)

PROOF. (1) $H_\alpha D H_\alpha^T - H_\alpha C H_\alpha^T = H_\alpha (D - C) H_\alpha^T = H_\alpha (\frac{1}{2} d \cdot \mathbf{1}^T + \frac{1}{2} \mathbf{1} \cdot d^T) H_\alpha^T$.
Since $H_\alpha \cdot \mathbf{1} = 0$, we have

$$H_\alpha D H_\alpha^T - H_\alpha C H_\alpha^T = 0$$

(2) $B_\alpha = -\frac{1}{2} H_\alpha C H_\alpha^T = -\frac{1}{2}(\mathbf{I} - \mathbf{1} \cdot \alpha^T) C (\mathbf{I} - \alpha \cdot \mathbf{1}^T) = -\frac{1}{2} C + \frac{1}{2} \mathbf{1} \cdot \alpha^T C + \frac{1}{2} C \alpha \cdot \mathbf{1}^T - \frac{1}{2} \mathbf{1} \cdot \alpha^T C \alpha \cdot \mathbf{1}^T$, so we have:

$$B_{i,j}(\alpha) = -\frac{1}{2} C_{i,j} + \frac{1}{2} c_i + \frac{1}{2} c_j - \frac{1}{2} c$$

where $c_i = (\alpha^T C)_i$, $c = \alpha^T C \alpha$. This implies

$$B_{i,i}(\alpha) + B_{j,j}(\alpha) - 2B_{i,j}(\alpha) = -\frac{1}{2} C_{ii} - \frac{1}{2} C_{jj} + C_{ij} = C_{ij},$$

where the last step is due to $C_{i,i} = 0$.

(3) According to Lemma 2.1 and the first part of Theorem 2.2: C c.n.d. $\iff$ B p.s.d $\iff$ D c.n.d.

(4) According to Lemma 2.1 and the second part of Theorem 2.2: C c.n.d. $\iff$ B p.s.d $\iff$ $\exists Y$ s.t. $B_\alpha = Y^T Y$ $\iff$ $B_{i,j}(\alpha) = \sum_k Y_{i,k}Y_{j,k} \Rightarrow C_{i,j} = \sum_k (Y_{i,k} - Y_{j,k})^2$

This completes the proof. □

Sometimes, we may want to transform a square distance matrix to another square distance matrix. The following theorem tells us the form of all the transformations between squared distance matrices.

Theorem 2.3 (Schoenberg Transform). Given D a squared distance matrix, $C_{i,j} = \Phi(D_{i,j})$. Then

$$C \text{ is a squared distance matrix} \iff \Phi \text{ is a Schoenberg Transform.}$$

A *Schoenberg Transform* Φ is a transform from $\mathbb{R}^+$ to $\mathbb{R}^+$, which takes d to

$$\Phi(d) = \int_0^\infty \frac{1 - \exp(-\lambda d)}{\lambda} g(\lambda) d\lambda,$$

where $g(\lambda)$ is some nonnegative measure on $[0, \infty)$ s.t

$$\int_0^\infty \frac{g(\lambda)}{\lambda} d\lambda < \infty.$$

Examples of Schoeberg transforms include

- $\phi_0(d) = d$ with $g_0(\lambda) = \delta(\lambda)$;
- $\phi_1(d) = \dfrac{1 - \exp(-ad)}{a}$ with $g_1(\lambda) = \delta(\lambda - a)$ $(a > 0)$;
- $\phi_2(d) = \ln(1 + d/a)$ with $g_2(\lambda) = \exp(-a\lambda)$;
- $\phi_3(d) = \dfrac{d}{a(a+d)}$ with $g_3(\lambda) = \lambda \exp(-a\lambda)$;
- $\phi_4(d) = d^p$ $(p \in (0,1))$ with $g_4(\lambda) = \dfrac{p}{\Gamma(1-p)}\lambda^{-p}$ (see more in [Bav11]).

The first one gives the identity transform and the last one implies that for a distance function, $\sqrt{d}$ is also a distance function but d^2 is not. To see this, take three points on a line $x = 0, y = 1, z = 2$ where $d(x,y) = d(y,z) = 1$, then for $p > 1$ $d^p(x,z) = 2^p > d^p(x,y) + d^p(y,z) = 2$ which violates the triangle inequality. In fact, d^p $(p \in (0,1))$ is Euclidean distance function immediately implies the following triangle inequality

$$d^p(0, x+y) \leq d^p(0,x) + d^p(0,y).$$

Note that Schoenberg transform satisfies $\phi(0) = 0$,

$$\phi'(d) = \int_0^\infty exp(-\lambda d) g(\lambda) d\lambda \geq 0,$$

$$\phi''(d) = -\int_0^\infty \exp(-\lambda d)\lambda g(\lambda) d\lambda \leq 0,$$

and so on. In other words, ϕ is a *completely monotonic function* defined by $(-1)^n \phi^{(n)}(x) \geq 0$, with additional constraint $\phi(0) = 0$. Schoenberg showed in 1938 that a function ϕ is completely monotone on $[0, \infty)$ if and only if $\phi(d^2)$ is positive definite and radial on $\mathbb{R}^s$ for all s.

3. Hilbert Space Embedding and Reproducing Kernels

Schoenberg [Sch38b] shows that Euclidean embedding of finite points can be characterized completely by positive definite functions, which paves a way toward Hilbert space embedding. Later Aronzajn [Aro50] developed Reproducing Kernel Hilbert spaces based on positive definite functions which eventually leads to the kernel methods in statistics and machine learning [Vap98, BTA04, CST03].

Theorem 3.1 (Schoenberg 38)**.** *A separable space M with a metric function $d(x, y)$ can be isometrically imbedded in a Hilbert space H, if and only if the family of functions $e^{-\lambda d^2}$ are positive definite for all $\lambda > 0$ (in fact we just need it for a sequence of λ_i whose accumulate point is 0).*

Here a symmetric function $k(x, y) = k(y, x)$ is called *positive definite* if for all finite x_i, x_j,

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0, \quad \forall c_i, c_j$$

with equality $=$ holds iff $c_i = c_j = 0$. In other words the function k restricted on $\{(x_i, x_j) : i, j = 1, \ldots, n\}$ is a positive definite matrix.

Combined this with Schoenberg transform, one shows that if $d(x, y)$ is an Euclidean distance matrix, then $e^{-\lambda\Phi(d)^2}$ is positive definite for all $\lambda > 0$. Note that for homogeneous function $e^{-\lambda\Phi(tx)} = e^{-\lambda t^k \Phi(x)}$, it suffices to check positive definiteness for $\lambda = 1$.

Symmetric positive definite functions $k(x, y)$ are often called reproducing kernels [Aro50]. In fact the functions spanned by $k_x(\cdot) = k(x, \cdot)$ for $x \in X$ made up of a Hilbert space, where we can associate an inner product induced from $\langle k_x, k_y \rangle = k(x, y)$. The radial basis function $e^{-\lambda d^2} = e^{-\lambda \|x\|^2}$ is often called Gaussian kernel or heat kernel in literature and has been widely used in machine learning.

On the other hand, every Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ with bounded evaluation functional can be regarded as a reproducing kernel Hilbert space [Wah90]. By Riesz representation, for every $x \in \mathcal{X}$ there exists $E_x \in \mathcal{H}$ such that $f(x) = \langle f, E_x \rangle$. By boundedness of evaluation functional, $|f(x)| \leq \|f\|_H \|E_x\|$, one can define a reproducing kernel $k(x, y) = \langle E_x, E_y \rangle$ which is bounded, symmetric and positive definite. It is called 'reproducing' because we can reproduce the function value using $f(x) = \langle f, k_x \rangle$ where $k_x(\cdot) := k(x, \cdot)$ as a function in $\mathcal{H}$. Such an universal property makes RKHS a unified tool to study Hilbert function spaces in nonparametric statistics, including Sobolev spaces consisting of splines [Wah90].

4. Linear Dimensionality Reduction

We have seen that given a set of paired distances d_{ij}, how to find an Euclidean embedding $x_i \in \mathbb{R}^p$ such that $\|x_i - x_j\| = d_{ij}$. However the dimensionality of such an embedding p can be very large. For example, any $n + 1$ points can be isometrically embedded into $\mathbb{R}^n_\infty$ using $(d_{i1}, d_{i2}, \ldots, d_{in})$ and l_∞-metric: $d_\infty(x_j, x_k) = \max_{i=1,\ldots,n} |d_{ij} - d_{jk}| = d_{ik}$ due to triangle inequality. Moreover, via the heat kernel $e^{-\lambda t^2}$ they can be embedded into Hilbert spaces of infinite dimensions.

Therefore dimensionality reduction is desired when p is large, at the best preservation of pairwise distances.

Given a set of points $x_i \in \mathbb{R}^p$ $(i = 1, 2, \cdots, n)$; form a data Matrix $X^{p\times n} = [X_1, X_2 \cdots X_n]^T$, when p is large, especially in some cases larger than n, we want to find k-dimensional projection with which pairwise distances of the data point are preserved as well as possible. That is to say, if we know the original pairwise distance $d_{ij} = \|X_i - X_j\|$ or data distances with some disturbance $\tilde{d}_{ij} = \|X_i - X_j\| + \epsilon$, we want to find $Y_i \in \mathbb{R}^k$ s.t.:

$$\min_{Y_i \in \mathbb{R}^k} \sum_{i,j} (\|Y_i - Y_j\|^2 - d_{ij}^2)^2 \tag{1}$$

take the derivative w.r.t $Y_i \in \mathbb{R}^k$:

$$\sum_{i,j} (\|Y_i\|^2 + \|Y_j\|^2 - 2Y_i^T Y_j - d_{ij}^2)(Y_i - Y_j) = 0$$

which implies $\sum_i Y_i = \sum_j Y_j$. For simplicity set $\sum_i Y_i = 0$, *i.e.*putting the origin as data center.

Use a linear transformation to move the sample mean to be the origin of the coordinates, *i.e.* define a matrix $B_{ij} = -\frac{1}{2}HDH$ where $D = (d_{ij}^2)$, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then, the minimization (1) is equivalent to find $Y_i \in \mathbb{R}^k$:

$$\min_{Y \in \mathbb{R}^{k\times n}} \|Y^T Y - B\|_F^2$$

then the row vectors of matrix Y are the eigenvectors corresponding to k largest eigenvalues of $B = \widetilde{X}^T \widetilde{X}$, or equivalently the top k *right singular vectors* of $\widetilde{X} = USV^T$.

We have seen in the first section that the covariance matrix of data $\widehat{\Sigma}_n = \frac{1}{n-1}\widetilde{X}\widetilde{X}^T = \frac{1}{n}US^2U^T$, passing through the singular vector decomposition (SVD) of $\widetilde{X} = USV^T$. Taking top k *left singular vectors* as the embedding coordinates is often called *Principal Component Analysis* (PCA). In PCA, given (centralized) Euclidean coordinate $\widetilde{X}$, ususally one gets the inner product matrix as covariance matrix $\widehat{\Sigma}_n = \frac{1}{n-1}\widetilde{X} \cdot \widetilde{X}^T$ which is a $p \times p$ *positive semi-definite* matrix, then the top k eigenvectors of $\widehat{\Sigma}_n$ give rise to a k-dimensional embedding of data, as principal components. So both MDS and PCA are unified in SVD of centralized data matrix.

The following introduces PCA from another point of view as best k-dimensional affine space approximation of data.

5. Principal Component Analysis

Principal component analysis (PCA), invented by Pearson (1901) and Hotelling (1933), is perhaps the most ubiquitous method for dimensionality reduction with high dimensional Euclidean data, under various names in science and engineering such as Karhunen-Loève Transform, Empirical Orthogonal Functions, and Principal Orthogonal Decomposition, etc. In the following we will introduce PCA from its sampled version.

Let $X = [X_1|X_2|\cdots|X_n] \in \mathbb{R}^{p\times n}$. Now we are going to look for a k-dimensional affine space in $\mathbb{R}^p$ to best approximate these n examples. Assume that such an affine space can be parameterized by $\mu + U\beta$ such that $U = [u_1, \ldots, u_k]$ consists of k-columns of an orthonormal basis of the affine space. Then the best approximation in terms of Euclidean distance is given by the following optimization problem.

$$\min_{\beta,\mu,U} I := \sum_{i=1}^{n} \|X_i - (\mu + U\beta_i)\|^2 \tag{2}$$

where $U \in \mathbb{R}^{p\times k}$, $U^TU = I_p$, and $\sum_{i=1}^n \beta_i = 0$ (nonzero sum of β_i can be represented by μ). Taking the first order optimality conditions,

$$\frac{\partial I}{\partial \mu} = -2\sum_{i=1}^{n}(X_i - \mu - U\beta_i) = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\frac{\partial I}{\partial \beta_i} = (x_i - \mu - U\beta_i)^T U = 0 \Rightarrow \beta_i = U^T(X_i - \mu)$$

Plug in the expression of $\hat{\mu}_n$ and β_i

$$\begin{aligned} I &= \sum_{i=1}^{n} \|X_i - \hat{\mu}_n - UU^T(X_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|X_i - \hat{\mu}_n - P_k(X_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|Y_i - P_k(y_i)\|^2, \quad Y_i := X_i - \hat{\mu}_n \end{aligned}$$

where $P_k = UU^T$ is a projection operator satisfying the idempotent property $P_k^2 = P_k$.

Denote $Y = [Y_1|Y_2|\cdots|Y_n] \in \mathbb{R}^{p\times n}$, whence the original problem turns into

$$\begin{aligned} \min_U \sum_{i=1}^{n} \|Y_i - P_k(Y_i)\|^2 &= \min \operatorname{trace}[(Y - P_kY)^T(Y - P_kY)] \\ &= \min \operatorname{trace}[Y^T(I - P_k)(I - P_k)Y] \\ &= \min \operatorname{trace}[YY^T(I - P_k)^2] \\ &= \min \operatorname{trace}[YY^T(I - P_k)] \\ &= \min[\operatorname{trace}(YY^T) - \operatorname{trace}(YY^TUU^T)] \\ &= \min[\operatorname{trace}(YY^T) - \operatorname{trace}(U^TYY^TU)]. \end{aligned}$$

Above we use cyclic property of trace and idempotent property of projection.

Since Y does not depend on U, the problem above is equivalent to

$$\max_{UU^T=I_k} Var(U^TY) = \max_{UU^T=I_k} \frac{1}{n}\operatorname{trace}(U^TYY^TU) = \max_{UU^T=I_k} \operatorname{trace}(U^T\hat{\Sigma}_nU) \tag{3}$$

where $\hat{\Sigma}_n = \frac{1}{n}YY^T = \frac{1}{n}(X - \hat{\mu}_n\mathbf{1}^T)(X - \hat{\mu}_n\mathbf{1}^T)^T$ is the sample variance. Assume that the sample covariance matrix, which is positive semi-definite, has the eigenvalue decomposition $\hat{\Sigma}_n = \hat{U}\hat{\Lambda}\hat{U}^T$, where $\hat{U}^T\hat{U} = I$, $\Lambda = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$, and $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_n \geq 0$. Then

$$\max_{UU^T=I_k} \operatorname{trace}(U^T\hat{\Sigma}_nU) = \sum_{i=1}^{k} \hat{\lambda}_i$$

In fact when $k = 1$, the maximal covariance is given by the largest eigenvalue along the direction of its associated eigenvector,

$$\max_{\|u\|=1} u^T \hat{\Sigma}_n u =: \hat{\lambda}_1.$$

Restricted on the orthogonal subspace $u \perp \hat{u}_1$ will lead to

$$\max_{\|u\|=1, u^T \hat{u}_1 = 0} u^T \hat{\Sigma}_n u =: \hat{\lambda}_2,$$

and so on.

Here we conclude that the k-affine space can be discovered by eigenvector decomposition of $\hat{\Sigma}_n$. The sample principal components are defined as column vectors of $\hat{Q} = \hat{U}^T Y$, where the j-th observation has its projection on the k-th component as $\hat{q}_k(j) = \hat{u}_k^T y_j = \hat{u}_k^T (x_i - \hat{\mu}_n)$. Therefore, PCA takes the eigenvector decomposition of $\hat{\Sigma}_n = \hat{U} \hat{\Lambda} \hat{U}^T$ and studies the projection of centered data points on top k eigenvectors as the principle components. This is equivalent to the singular value decomposition (SVD) of $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}$ in the following sense,

$$Y = X - \frac{1}{n} \mathbf{1}\mathbf{1}^T X = \tilde{U} \tilde{S} \tilde{V}^T, \quad \mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$$

where top right singular vectors of centered data matrix Y gives the same principle components. From linear algebra, k-principal components thus gives the best rank-k approximation of centered data matrix Y.

Given a PCA, the following quantities are often used to measure the variances

- total variance:
$$\mathrm{trace}(\hat{\Sigma}_n) = \sum_{i=1}^{p} \hat{\lambda}_i;$$
- percentage of variance explained by top-k principal components:
$$\sum_{i=1}^{k} \hat{\lambda}_i / \mathrm{trace}(\hat{\Sigma}_n);$$
- generalized variance as total volume:
$$\det(\hat{\Sigma}_n) = \prod_{i=1}^{p} \hat{\lambda}_i.$$

Example. Take the dataset of hand written digit "3", $\hat{X} \in \mathbb{R}^{658 \times 256}$ contains 658 images, each of which is of 16-by-16 grayscale image as hand written digit 3. Figure 2 shows a random selection of 9 images, the sorted singular values divided by total sum of singular values, and an approximation of x_1 by top 3 principle components: $x_1 = \hat{\mu}_n - 2.5184\tilde{v}_1 - 0.6385\tilde{v}_2 + 2.0223\tilde{v}_3$.

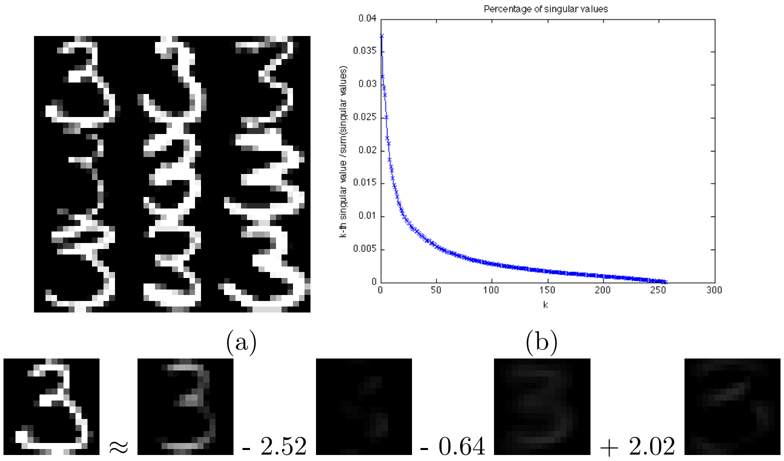
6. Dual Roles of MDS vs. PCA in SVD

Consider the data matrix

$$X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}.$$

Let the centered data admits a singular vector decomposition (SVD),

$$\widetilde{X} = X - \frac{1}{n} \mathbf{1}\mathbf{1}^T X = \widetilde{U} \widetilde{S} \widetilde{V}^T, \quad \mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n.$$



(a) (b)

(c)

Figure 2. (a) random 9 images. (b) percentage of singular values over total sum. (c) approximation of the first image by top 3 principle components (singular vectors).

We have seen that both MDS and PCA can be obtained from such a SVD of centered data matrix.

- MDS embedding is given by top k *left* singular vectors $Y_k^{MDS} = \widetilde{U}_k \widetilde{S_k}^{1/2} \in \mathbb{R}^{n\times k}$;
- PCA embedding is given by top k *right* singular vectors $Y_k^{PCA} = \widetilde{V}_k \widetilde{S_k}^{1/2} \in \mathbb{R}^{n\times k}$.

Altogether $\widetilde{U}_k \widetilde{S}_k \widetilde{V}_k^T$ gives best rank-k approximation of $\widetilde{X}$ in any unitary invariant norms.

CHAPTER 2

Random Projections and Almost Isometry

1. Introduction

For this class, we introduce Random Projection method which may reduce the dimensionality of n points in $\mathbb{R}^p$ to $k = O(c(\epsilon)\log n)$ at the cost of a uniform metric distortion of at most $\epsilon > 0$, with high probability. The theoretical basis of this method was given as a lemma by Johnson and Lindenstrauss [JL84] in the study of a Lipschitz extension problem. The result has a widespread application in mathematics and computer science. The main application of Johnson-Lindenstrauss Lemma in computer science is high dimensional data compression via random projections [Ach03]. In 2001, Sanjoy Dasgupta and Anupam Gupta [DG03a], gave a simple proof of this theorem using elementary probabilistic techniques in a four-page paper. Below we are going to present a brief proof of Johnson-Lindenstrauss Lemma based on the work of Sanjoy Dasgupta, Anupam Gupta [DG03a], and Dimitris Achlioptas [Ach03].

Recall the problem of MDS: given a set of points $x_i \in \mathbb{R}^p$ $(i = 1, 2, \cdots, n)$; form a data Matrix $X^{p\times n} = [X_1, X_2 \cdots X_n]^T$, when p is large, especially in some cases larger than n, we want to find k-dimensional projection with which pairwise distances of the data point are preserved as well as possible. That is to say, if we know the original pairwise distance $d_{ij} = \|X_i - X_j\|$ or data distances with some disturbance $\tilde{d}_{ij} = \|X_i - X_j\| + \epsilon_{ij}$, we want to find $Y_i \in \mathbb{R}^k$ s.t.:

$$\min \sum_{i,j} (\|Y_i - Y_j\|^2 - d_{ij}^2)^2 \tag{4}$$

take the derivative w.r.t $Y_i \in \mathbb{R}^k$:

$$\sum_{i,j} (\|Y_i\|^2 + \|Y_j\|^2 - 2Y_i^T Y_j - d_{ij}^2)(Y_i - Y_j) = 0$$

which implies $\sum_i Y_i = \sum_j Y_j$. For simplicity set $\sum_i Y_i = 0$, *i.e.*putting the origin as data center.

Use a linear transformation to move the sample mean to be the origin of the coordinates, *i.e.* define a matrix $K = -\frac{1}{2}HDH$ where $D = (d_{ij}^2)$, $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, then, the minimization (4) is equivalent to find $Y_i \in \mathbb{R}^k$:

$$\min_{Y \in \mathbb{R}^{k\times n}} \|Y^T Y - K\|_F^2 \tag{5}$$

then the row vectors of matrix Y are the eigenvectors (singular vectors) corresponding to k largest eigenvalues (singular values) of B.

The main features of MDS are the following.

- MDS looks for Euclidean embedding of data whose *total* or *average* metric distortion are minimized.

- MDS embedding basis is *adaptive* to the data, namely as a function of data via eigen-decomposition.

Note that distortion measure here amounts to a certain distance between the set of projected points and the original set of points B. Under the Frobenius norm the distortion equals the sum of the squared lengths of these vectors. It is clear that such vectors captures a significant global property, but it does not offer any local guarantees. Chances are that some points deviate greatly from the original if we only consider the total metric distortion minimization.

What if we want a *uniform* control on metric distortion at every data pair, say

$$(1-\epsilon)d_{ij} \leq \|Y_i - Y_j\| \leq (1+\epsilon)d_{ij}?$$

Such an embedding is an almost isometry or a Lipschitz mapping from metric space $\mathcal{X}$ to Euclidean space $\mathcal{Y}$. If $\mathcal{X}$ is an Euclidean space (or more generally Hilbert space), Johnson-Lindenstrauss Lemma tells us that one can take $\mathcal{Y}$ as a subspace of $\mathcal{X}$ of dimension $k = O(c(\epsilon)\log n)$ via random projections to obtain an almost isometry with high probability. As a contrast to MDS, the main features of this approach are the following.

- Almost isometry is achieved with a *uniform* metric distortion bound (*Lipschitz* bound), with high probability, rather than average metric distortion control;
- The mapping is *universal*, rather than being adaptive to the data.

2. The Johnson-Lindenstrauss Lemma

Theorem 2.1 (Johnson-Lindenstrauss Lemma)**.** For any $0 < \epsilon < 1$ and any integer n, let k be a positive integer such that

$$k \geq (4+2\alpha)(\epsilon^2/2 - \epsilon^3/3)^{-1}\ln n, \quad \alpha > 0.$$

Then for any set V of n points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$

$$(1-\epsilon) \parallel u - v \parallel^2 \leq \parallel f(u) - f(v) \parallel^2 \leq (1+\epsilon) \parallel u - v \parallel^2 \tag{6}$$

Such a f in fact can be found in randomized polynomial time. In fact, inequalities (6) holds with probability at least $1 - 1/n^\alpha$.

Remark. We have following facts.

(1) The embedding dimension $k = O(c(\epsilon)\log n)$ which is independent to ambient dimension d and logarithmic to the number of samples n. The independence to d in fact suggests that the Lemma can be generalized to the Hilbert spaces of infinite dimension.

(2) How to construct the map f? In fact we can use random projections:

$$Y^{n\times k} = X^{n\times d}R^{d\times k}$$

where the following random matrices R can cater our needs.

- $R = [r_1, \cdots, r_k] \quad r_i \in S^{d-1} \quad r_i = (a_1^i, \cdots, a_d^i)/ \parallel a^i \parallel \quad a_k^i \sim N(0,1)$
- $R = A/\sqrt{k} \quad A_{ij} \sim N(0,1)$
- $R = A/\sqrt{k} \quad A_{ij} = \begin{cases} 1 & p = 1/2 \\ -1 & p = 1/2 \end{cases}$

- $R = A/\sqrt{k/3} \quad A_{ij} = \begin{cases} 1 & p = 1/6 \\ 0 & p = 2/3 \\ -1 & p = 1/6 \end{cases}$

The proof below actually takes the first form of R as an illustration.

Now we are going to prove Johnson-Lindenstrauss Lemma using a random projection to k-subspace in $\mathbb{R}^d$. Notice that the distributions of the following two events are identical:

$$\begin{aligned} & \text{unit vector was randomly projected to } k\text{-subspace} \\ \Longleftrightarrow\ & \text{random vector on } S^{d-1} \text{ fixed top-}k \text{ coordinates.} \end{aligned}$$

Based on this observation, we change our target from random k-dimensional projection to random vector on sphere S^{d-1}.

If $x_i \sim N(0,1),\ (i = 1, \cdots, d),\ X = (x_1, \cdots, x_d)$, then $Y = X/\|x\| \in S^{d-1}$ is uniformly distributed. Fixing top-k coordinates, we get $z = (x_1, \cdots, x_k, 0, \cdots, 0)^T/\|x\| \in \mathbb{R}^d$. Let $L = \|Z\|^2$ and $\mu = \mathbb{E}[L] = k/d$.

The following lemma is crucial to reach the main theorem.

Lemma 2.2. let any $k < d$ then we have

(a) if $\beta < 1$ then

$$\text{Prob}[L \leq \beta\mu] \leq \beta^{k/2}\left(1 - \frac{(1-\beta)k}{d-k}\right)^{d-k/2} \leq \exp\left(\frac{k}{2}(1 - \beta + \ln\beta)\right)$$

(b) if $\beta > 1$ then

$$\text{Prob}[L \geq \beta\mu] \leq \beta^{k/2}\left(1 + \frac{(1-\beta)k}{d-k}\right)^{d-k/2} \leq \exp\left(\frac{k}{2}(1 - \beta + \ln\beta)\right)$$

Here $\mu = k/d$.

We first show how to use this lemma to prove the main theorem – Johnson-Lindenstrauss lemma.

Proof of Johnson-Lindenstrauss Lemma. If $d \leq k$,the theorem is trivial. Otherwise take a random k-dimensional subspace S, and let v_i' be the projection of point $v_i \in V$ into S, then setting $L = \|v_i' - v_j'\|^2$ and $\mu = (k/d)\|v_i - v_j\|^2$ and applying Lemma 2(a), we get that

$$\begin{aligned} \text{Prob}[L \leq (1-\epsilon)\mu] &\leq \exp(\frac{k}{2}(1 - (1-\epsilon) + \ln(1-\epsilon))) \\ &\leq \exp(\frac{k}{2}(\epsilon - (\epsilon + \frac{\epsilon^2}{2}))), \\ &\qquad \text{by } \ln(1-x) \leq -x - x^2/2 \text{ for } 0 \leq x < 1 \\ &= \exp(-\frac{k\epsilon^2}{4}) \\ &\leq \exp(-(2+\alpha)\ln n), \quad \text{for } k \geq 4(1+\alpha/2)(\epsilon^2/2)^{-1}\ln n \\ &= \frac{1}{n^{2+\alpha}} \end{aligned}$$

$$\begin{aligned}\text{Prob}[L \geq (1+\epsilon)\mu] &\leq \exp(\frac{k}{2}(1-(1+\epsilon)+\ln(1+\epsilon)))\\ &\leq \exp(\frac{k}{2}(-\epsilon+(\epsilon-\frac{\epsilon^2}{2}+\frac{\epsilon^3}{3}))),\\ &\qquad \text{by } \ln(1+x) \leq x - x^2/2 + x^3/3 \text{ for } x \geq 0\\ &= \exp(-\frac{k}{2}(\epsilon^2/2-\epsilon^3/3)),\\ &\leq \exp(-(2+\alpha)\ln n), \quad \text{for } k \geq 4(1+\alpha/2)(\epsilon^2/2-\epsilon^3/3)^{-1}\ln n\\ &= \frac{1}{n^{2+\alpha}}\end{aligned}$$

Now set the map $f(x) = \sqrt{\frac{d}{k}}x' = \sqrt{\frac{d}{k}}(x_1, \ldots, x_k, 0, \ldots, 0)$. By the above calculations, for some fixed pair i, j, the probability that the distortion

$$\frac{\|f(v_i)-f(v_j)\|^2}{\|v_i-v_j\|^2}$$

does not lie in the range $[(1-\epsilon),(1+\epsilon)]$ is at most $\frac{2}{n^{(2+\alpha)}}$. Using the trivial union bound with C_n^2 pairs, the chance that some pair of points suffers a large distortion is at most:

$$C_n^2\frac{2}{n^{(2+\alpha)}} = \frac{1}{n^\alpha}\left(1-\frac{1}{n}\right) \leq \frac{1}{n^\alpha}.$$

Hence f has the desired properties with probability at least $1-\frac{1}{n^\alpha}$. This gives us a randomized polynomial time algorithm. □

Now, it remains to Lemma 3.6.

Proof of Lemma 3.6.

$$\begin{aligned}\text{Prob}(L \leq \beta\mu) =&\text{Prob}(\sum_{i=1}^{k}(x_i^2) \leq \beta\mu(\sum_{i=1}^{d}(x_i^2)))\\ =&\text{Prob}(\beta\mu\sum_{i=1}^{d}(x_i^2) - \sum_{i=1}^{k}(x_i^2) \leq 0)\\ =&\text{Prob}[\exp(t\beta\mu\sum_{i=1}^{d}(x_i^2) - t\sum_{i=1}^{k}(x_i^2)) \leq 1] \quad (t>0)\\ \leq&\mathbb{E}[\exp(t\beta\mu\sum_{i=1}^{d}(x_i^2) - t\sum_{i=1}^{k}(x_i^2))] \quad (\textit{by Markov's inequality})\\ =&\Pi_{i=1}^{k}\mathbb{E}\exp(t(\beta\mu-1)x_i^2)\Pi_{i=k+1}^{d}\mathbb{E}exp(t(\beta\mu)x_i^2)\\ =&(\mathbb{E}\exp(t(\beta\mu-1)x^2))^k(\mathbb{E}\exp(t\beta\mu^2))^{d-k}\\ =&(1-2t(\beta\mu-1))^{-k/2}(1-2t\beta\mu)^{-(d-k)/2}\end{aligned}$$

We use the fact that if $X \sim N(0,1)$,then $\mathbb{E}[e^{sX^2}] = \frac{1}{\sqrt{(1-2s)}}$, for $-\infty < s < 1/2$.

Now we will refer to last expression as $g(t)$. The last line of derivation gives us the additional constraints that $t\beta\mu \leq 1/2$ and $t(\beta\mu - 1) \leq 1/2$, and so we have $0 < t < 1/(2\beta\mu)$. Now to minimize $g(t)$, which is equivalent to maximize

$$h(t) = 1/g(t) = (1 - 2t(\beta\mu - 1))^{k/2}(1 - 2t\beta\mu)^{(d-k)/2}$$

in the interval $0 < t < 1/(2\beta\mu)$. Setting the derivative $h'(t) = 0$, we get the maximum is achieved at

$$t_0 = \frac{1 - \beta}{2\beta(d - \beta k)}$$

Hence we have

$$h(t_0) = (\frac{d-k}{d-k\beta})^{(d-k)/2}(1/\beta)^{k/2}$$

And this is exactly what we need.

The proof of Lemma 3.6 (b) is almost exactly the same as that of Lemma 3.6 (a). □

2.1. Conclusion. As we can see, this proof of Lemma is both simple (using just some elementary probabilistic techniques) and elegant. And you may find in the field of machine learning, stochastic method always turns out to be really powerful. The random projection method we approaching today can be used in many fields especially huge dimensions of data is concerned. For one example, in the term document, you may find it really useful for compared with the number of words in the dictionary, the words included in a document is typically sparse (with a few thousands of words) while the dictionary is hugh. Random projections often provide us a useful tool to compress such data without losing much pairwise distance information.

3. Example: MDS in Human Genome Diversity Project

Now consider a SNPs (Single Nucleid Polymorphisms) dataset in Human Genome Diversity Project (HGDP, http://www.cephb.fr/en/hgdp_panel.php) which consists of a data matrix of n-by-p for $n = 1064$ individuals around the world and $p = 644258$ SNPs. Each entry in the matrix has 0, 1, 2, and 9, representing "AA", "AC", "CC", and "missing value", respectively. After removing 21 rows with all missing values, we are left with a matrix X of size 1043×644258.

Consider the projection of 1043 persons on the MDS (PCA) coordinates. Let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ be the centering matrix. Then define

$$K = HXX^TH = U\Lambda U^T$$

which is a positive semi-define matrix as centered Gram matrix whose eigenvalue decomposition is given by $U\Lambda U^T$. Taking the first two eigenvectors $\sqrt{\lambda_i}u_i$ ($i = 1, \ldots, 2$) as the projections of n individuals, Figure 1 gives the projection plot. It is interesting to note that the point cloud data exhibits a continuous trend of human migration in history: origins from Africa, then migrates to the Middle East, followed by one branch to Europe and another branch to Asia, finally spreading into America and Oceania.

One computational concern is that the high dimensionality caused by $p = 644,258$, which is much larger than the number of samples $n = 1043$. However random projections introduced above will provide us an efficient way to compute MDS (PCA) principal components with an almost isometry.

We randomly select (without replacement) $\{n_i, i = 1, \ldots, k\}$ from $1, \ldots, p$ with equal probability. Let $R \in \mathbb{R}^{k \times p}$ is a Bernoulli random matrix satisfying:

$$R_{ij} = \begin{cases} 1/k & j = n_i, \\ 0 & otherwise. \end{cases}$$

Now define

$$\widetilde{K} = H(XR^T)(RX^T)H$$

whose eigenvectors leads to new principal components of MDS. In the middle and right, Figure 1 plots the such approximate MDS principal components with $k = 5,000$, and $k = 100,000$, respectively. These plots are qualitatively equivalent to the original one.

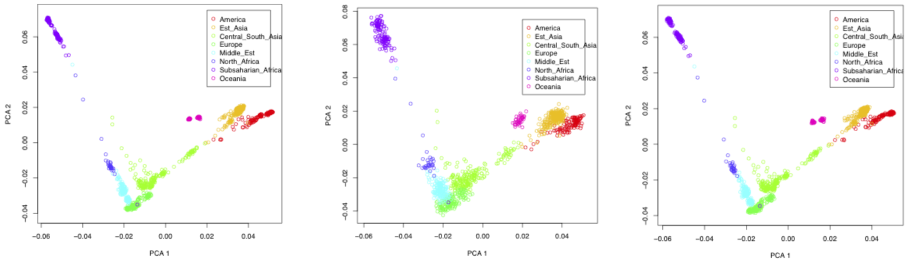


FIGURE 1. (Left) Projection of 1043 individuals on the top 2 MDS principal components. (Middle) MDS computed from 5,000 random projections. (Right) MDS computed from 100,000 random projections. Pictures are due to Qing Wang.

4. Random Projections and Compressed Sensing

There are wide applications of random projections in high dimensional data processing, e.g. [Vem04]. Here we particularly choose a special one, the compressed (or compressive) sensing (CS) where we will use the Johnson-Lindenstrauss Lemma to prove the Restricted Isometry Property (RIP), a crucial result in CS. A reference can be found at [BDDW08].

Compressive sensing can be traced back to 1950s in signal processing in geography. Its modern version appeared in LASSO [Tib96] and BPDN [CDS98], and achieved a highly noticeable status by [CT05, CRT06, CT06]. For a comprehensive literature on this topic, readers may refer to http://dsp.rice.edu/cs.

The basic problem of compressive sensing can be expressed by the following under-determined linear algebra problem. Assume that a signal $x^* \in \mathbb{R}^p$ is sparse with respect to some basis (measurement matrix) $\Phi \in \mathbb{R}^{n \times p}$ where $n < p$, given measurement $b = \Phi x^* \in \mathbb{R}^n$, how can one recover x^* by solving the linear equation system

$$\Phi x = b? \tag{7}$$

As $n < p$, it is an under-determined problem, whence without further constraint, the problem does not have an unique solution. To overcome this issue, one popular assumption is that the signal x^* is sparse, namely the number of nonzero components $\|x^*\|_0 := \#\{x^*_i \neq 0 : 1 \leq i \leq p\}$ is small compared to the total dimensionality p. Figure 2 gives an illustration of such sparse linear equation problem.

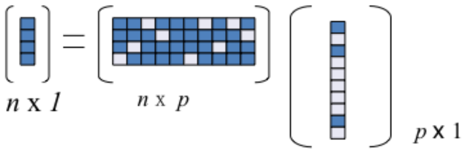


Figure 2. Illustration of Compressive Sensing (CS). Φ is a rectangular matrix with more columns than rows. The dark elements represent nonzero elements while the light ones are zeroes. The signal vector x^*, although high dimensional, is sparse.

With such a sparse assumption, we would like to find the sparsest solution satisfying the measurement equation.

$$\begin{aligned} (P_0) \quad & \min \quad \|x\|_0 \\ & s.t. \quad \Phi x = b. \end{aligned} \tag{8}$$

This is an NP-hard combinatorial optimization problem. A convex relaxation of (8) is called Basis Pursuit [CDS98],

$$\begin{aligned} (P_1) \quad & \min \quad \|x\|_1 := \sum |x_i| \\ & s.t. \quad \Phi x = b. \end{aligned} \tag{9}$$

This is a linear programming problem. Figure 3 shows different projections of a sparse vector x^* under l_0, l_1 and l_2, from which one can see in some cases the convex relaxation (9) does recover the sparse signal solution in (8). Now a natural problem arises, under what conditions the linear programming problem (P_1) has the solution exactly solves (P_0), i.e. exactly recovers the sparse signal x^*?

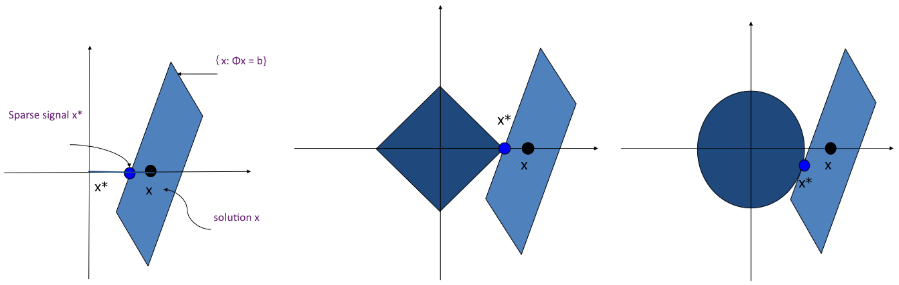


Figure 3. Comparison between different projections. Left: projection of x^* under $\|\cdot\|_0$; middle: projection under $\|\cdot\|_1$ which favors sparse solution; right: projection under Euclidean distance.

To understand the equivalence between (P_0) and (P_1), one asks the question when the true signal x^* is the unique solution of P_0 and P_1. In such cases, P_1 is equivalent to P_0. For the uniqueness of P_1, one turns to the duality of Linear Programming via the Karush-Kuhn-Tucker (KKT) conditions. Take the Lagrangian of (P_1),

$$L(x;\lambda) = \|x\|_1 + \lambda^T(\Phi x - b), \quad \lambda \in \mathbb{R}^n.$$

Assume the support of x^* as $T \subseteq \{1, \ldots, p\}$, *i.e.* $T = \{1 \leq i \leq p : x_i \neq 0\}$, and denote its complement by T^c. x^* is an optimal solution of P_1 if

$$0 \in \partial L(x^*, \lambda)$$

which implies that $\mathrm{sign}(x^*_T) = \Phi_T^T \lambda$ and $|\Phi_{T^c}^T \lambda| \leq 1$. How to ensure that there are no other solutions than x^*? The following condition is used in [CT05] and other related works.

Lemma 4.1. Assume that Φ_T is of full rank. If there exists $\lambda \in \mathbb{R}^n$ such that:

(1) For each $i \in T$,

$$\Phi_i^T \lambda = \mathrm{sign}(x_i^*); \tag{10}$$

(2) For each $i \in T^c$,

$$|\Phi_i^T \lambda| < 1. \tag{11}$$

Then P_1 has a unique solution x^*.

These two conditions just ensure a special dual variable λ exists, under which any optimal solution of P_1 must have the same support T as x^* (strictly complementary condition in (2)). Since Φ_T is of full rank, then P_1 must have a unique solution x^*. In this case solving P_1 is equivalent to P_0. If these conditions fail, then there exists a problem instance (Φ, b) such that P_1 has a solution different to x^*. In this sense, these conditions are necessary and sufficient for the equivalence between P_1 and P_0.

Various sufficient conditions have been proposed in literature to meet the KKT conditions above. For example, these includes the *mutual incoherence* by Donoho-Huo (1999) [DH01], Elad-Bruckstein (2001) [EB01] and the Exact Recovery Condition by Tropp [Tro04] or Irrepresentative condition (IRR) by Zhao-Yu [ZY06] (see also [MY09]). The former condition essentially requires Φ to be a nearly orthogonal matrix,

$$\mu(\Phi) = \max_{i \neq j} |\phi_i^T \phi_j|,$$

where $\Phi = [\phi_1, \ldots, \phi_p]$ and $\|\phi_i\|_2 = 1$, under which [DH01] shows that as long as sparsity of x^* satisfies

$$\|x^*\|_0 = |T| < \frac{1 + \frac{1}{\mu(\Phi)}}{2}$$

which is later improved by [EB01] to be

$$\|x^*\|_0 = |T| < \frac{\sqrt{2} - \frac{1}{2}}{\mu(\Phi)},$$

then P_1 recovers x^*. The latter assumes that the dual variable λ lies in the column space of A_T, *i.e.* $\lambda = \Phi_T \alpha$. Then we solve λ explicitly in equation (10) and plugs in the solution to the inequality (11)

$$\|\Phi_{T^c}^T \Phi_T (\Phi_T^T \Phi_T)^{-1} \mathrm{sign}(x^*|_T)\|_\infty < 1$$

or simply

$$\|\Phi_{T^c}^T \Phi_T (\Phi_T^T \Phi_T)^{-1}\|_\infty < 1.$$

If for *every* k-sparse signal x^* with support T, conditions above are satisfied, then P_1 recovers x^*.

The most popular condition is proposed by [CRT06], called *Restricted Isometry Property* (RIP).

Definition. Define the isometry constant δ_k of a matrix Φ to be the smallest nonnegative number such that

$$(1-\delta_k)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1+\delta_k)\|x\|_2^2$$

holds for all k-sparse vectors $x \in \mathbb{R}^p$. A vector x is called k-sparse if it has at most k nonzero elements.

[AC09] shows that incoherence conditions implies RIP, whence RIP is a weaker condition. Under RIP condition, uniqueness of P_0 and P_1 can be guaranteed for all k-sparse signals, often called *uniform exact recovery*[Can08].

Theorem 4.2. The following holds for all k-sparse x^* satisfying $\Phi x^* = b$.

(1) If $\delta_{2k} < 1$, then problem P_0 has a unique solution x^*;
(2) If $\delta_{2k} < \sqrt{2}-1$, then the solution of P_1 (9) has a unique solution x^*, *i.e.* recovers the original sparse signal x^*.

The first condition is nothing but every $2k$-columns of Φ are linearly dependent. To see the first condition, assume by contradiction that there is another k-sparse solution of P_0, x'. Then by $\Phi y = 0$ and $y = x^* - x'$ is $2k$-sparse. If $y \neq 0$, it violates $\delta_{2k} < 1$ such that $0 = \|\Phi y\| \geq (1-\delta_{2k})\|y\| > 0$. Hence one must have $y = 0$, *i.e.* $x^* = x'$ which proves the uniqueness of P_0. The first condition is also necessary for the uniqueness of P_0's solutions. In fact, if $\delta_{2k} = 1$, this implies that there is a $2k$-subset $2T$ such that columns of Φ_{2T} are linearly dependent, *i.e.* $\Phi_{2T} z = 0$ for some $2k$-vector z. One can define x_1 to collect first k nonzero elements of z with zero otherwise, and x_2 to collect the second half nonzero entries of z but zero otherwise. Hence $\Phi_{2T}(x_1 + x_2) = 0 \Rightarrow \Phi_{T_1} x_1 = 0 = \Phi_{T_2} x_2$ with T_1 and T_2 consisting the first and second k columns of Φ_{2T} respectively, which violates the uniqueness of P_0 solutions. The proof of the second condition can be found in [Can08].

When measurement noise exists, *e.g.* $b = \Phi x + e$ with bound $\|e\|_2$, the following Basis Pursuit De-Noising (BPDN) [CDS98] or LASSO [Tib96] are used instead

$$(12) \qquad (BPDN) \quad \begin{array}{ll} \min & \|x\|_1 \\ s.t. & \|\Phi x - b\|_2 \leq \epsilon. \end{array}$$

$$(13) \qquad (LASSO) \quad \min_{x\in\mathbb{R}^p} \|\Phi x - b\|_2 + \lambda\|x\|_1$$

For bounded $\|e\|_\infty$, the following formulation is used in network analysis [JYLG12]

$$(14) \qquad \begin{array}{ll} \min & \|x\|_1 \\ s.t. & \|\Phi x - b\|_\infty \leq \epsilon \end{array}$$

RIP conditions also lead to upper bounds between solutions above and the true sparse signal x^*. For example, in the case of BPDN the follwoing result holds [Can08].

Theorem 4.3. Suppose that $\|e\|_2 \leq \epsilon$. If $\delta_{2k} < \sqrt{2} - 1$, then

$$\|\hat{x} - x^*\|_2 \leq C_1 k^{-1/2} \sigma_k^1(x^*) + C_2 \epsilon,$$

where $\hat{x}$ is the solution of BPDN and

$$\sigma_k^1(x^*) = \min_{\text{supp}(y) \leq k} \|x^* - y\|_1$$

is the best k-term approximation error in l_1 of x^*.

How to find matrices satisfying RIP? Equipped with Johnson-Lindenstrauss Lemma, one can construct such matrices by random projections with high probability [BDDW08].

Recall that in the Johnson-Lindenstrauss Lemma, a random matrix $\Phi \in \mathbb{R}^{n \times p}$ with each element is i.i.d. according to some distribution satisfying certain bounded moment conditions, e.g. $\Phi_{ij} \sim \mathcal{N}(0,1)$. The key step to establish Johnson-Lindenstrauss Lemma is the following fact

$$\Pr\left(\|\Phi x\|_2^2 - \|x\|_2^2 \geq \epsilon \|x\|_2^2\right) \leq 2e^{-nc_0(\epsilon)}. \tag{15}$$

With this one can establish a bound on the action of Φ on k-sparse x by an union bound via covering numbers of k-sparse signals.

Lemma 4.4. Let $\Phi \in \mathbb{R}^{n \times p}$ be a random matrix satisfying the concentration inequality (15). Then for any $\delta \in (0,1)$ and any set all T with $|T| = k < n$, the following holds

$$(1-\delta)\|x\|_2 \leq \|\Phi x\|_2 \leq (1+\delta)\|x\|_2 \tag{16}$$

for all x whose support is contained in T, with probability at least

$$1 - 2\left(\frac{12}{\delta}\right)^2 e^{-c_0(\delta/2)n}. \tag{17}$$

Proof. It suffices to prove the results when $\|x\|_2 = 1$ as Φ is linear. Let $X_T := \{x : \text{supp}(x) = T, \|x\|_2 = 1\}$. We first choose Q_T, a $\delta/4$-cover of X_T, such that for every $x \in X_T$ there exists $q \in Q_T$ satisfying $\|q - x\|_2 \leq \delta/4$. Since X_T has dimension at most k, it is well-known from covering numbers that the capacity $\#(Q_T) \leq (12/\delta)^k$. Now we are going to apply the union bound of (15) to the set Q_T with $\epsilon = \delta/2$. For each $q \in Q_T$, with probability at most $2e^{-c_0(\delta/2)n}$, $|\Phi q\|_2^2 - \|q\|_2^2 \geq \delta/2\|q\|_2^2$. Hence for all $q \in Q_T$, the same bound holds with probability at most

$$2\#(Q_T)e^{-c_0(\delta/2)n} = 2\left(\frac{12}{\delta}\right)^2 e^{-c_0(\delta/2)n}.$$

Now we define α to be the smallest constant such that

$$\|\Phi x\|_2 \leq (1+\alpha)\|x\|_2, \quad \text{for all } x \in X_T.$$

We can show that $\alpha \leq \delta$ with the same probability. For this, pick up a $q \in Q_T$ such that $\|q - x\|_2 \leq \delta/4$, whence by the triangle inequality

$$\|\Phi x\|_2 \leq \|\Phi q\|_2 + \|\Phi(x-q)\|_2 \leq 1 + \delta/2 + (1+\alpha)\delta/4.$$

This implies that $\alpha \leq \delta/2 + (1+\alpha)\delta/4$, whence $\alpha \leq 3\delta/4/(1-\delta/4) \leq \delta$. This gives the upper bound. The lower bound also follows this since

$$\|\Phi x\|_2 \geq \|\Phi q\|_2 - \|\Phi(x-q)\|_2 \geq 1 - \delta/2 - (1+\delta)\delta/4 \geq 1 - \delta,$$

which completes the proof. □

With this lemma, note that there are at most $\binom{p}{k}$ subspaces of k-sparse, an union bound leads to the following result for RIP.

Theorem 4.5. Let $\Phi \in \mathbb{R}^{n\times p}$ be a random matrix satisfying the concentration inequality (15) and $\delta \in (0,1)$. There exists $c_1, c_2 > 0$ such that if

$$k \le c_1 \frac{n}{\log(p/k)}$$

the following RIP holds

$$(1-\delta_k)\|x\|_2^2 \le \|\Phi x\|_2^2 \le (1+\delta_k)\|x\|_2^2$$

with probability at least $1 - 2e^{-c_2 n}$.

Proof. For each of k-sparse signal (X_T), RIP fails with probability at most

$$2\left(\frac{12}{\delta}\right)^2 e^{-c_0(\delta/2)n}.$$

There are $\binom{p}{k} \le (ep/k)^k$ such subspaces. Hence, RIP fails with probability at most

$$2\left(\frac{ep}{k}\right)^k \left(\frac{12}{\delta}\right)^2 e^{-c_0(\delta/2)n} = 2e^{-c_0(\delta/2)n + k[\log(ep/k)+\log(12/\delta)]}.$$

Thus for a fixed $c_1 > 0$, whenever $k \le c_1 n/\log(p/k)$, the exponent above will be $\le -c_2 n$ provided that $c_2 \le c_0(\delta/2) - c_1(1 + (1+\log(12/\delta))/\log(p/k)$. c_2 can be always chosen to be > 0 if $c_1 > 0$ is small enough. This leads to the results. □

Another use of random projections (random matrices) can be found in Robust Principal Component Analysis (RPCA) in the next chapter.

CHAPTER 3

High Dimensional Statistics: Mean and Covariance in Noise

In this very first lecture, we talk about data representation as vectors, matrices (*esp.* graphs, networks), and tensors, *etc.* Data are mappings of real world based on sensory measurements, whence the real world puts constraints on the variations of data. Data science is the study of laws in real world which shapes the data.

We start the first topic on sample mean and variance in high dimensional Euclidean spaces $\mathbb{R}^p$, as the maximal likelihood estimators based on multivariate Gaussian assumption. Principle Component Analysis (PCA) is the projection of high dimensional data on its top singular vectors. In classical statistics with the Law of Large Numbers, for fixed p when sample size $n \to \infty$, we know such sample mean and variance will converge, so as to PCA. Although sample mean $\hat{\mu}_n$ and sample covariance $\hat{\Sigma}_n$ are the most commonly used statistics in multivariate data analysis, they may suffer some problems in high dimensional settings, *e.g.* for large p and small n scenario. In 1956, Stein [**Ste56**] shows that the sample mean is not the best estimator in terms of the mean square error, for $p > 2$; moreover in 2006, Jonestone [**Joh06**] shows by random matrix theory that PCA might be overwhelmed by random noise for fixed ratio p/n when $n \to \infty$. Among other works, these two pieces of excellent works inspired a long pursuit toward modern high dimensional statistics with a large unexplored field ahead.

1. Maximum Likelihood Estimation

Consider the statistical model $f(X|\theta)$ as a conditional probability function on $\mathbb{R}^p$ with parameter space $\theta \in \Theta$. Let $X_1, ..., X_n \in \mathbb{R}^p$ are independently and identically distributed (i.i.d.) sampled according to $f(X|\theta_0)$ on $\mathbb{R}^p$ for some $\theta_0 \in \Theta$. The likelihood function is defined as the probability of observing the given data as a function of θ,

$$L(\theta) = \prod_{i=1}^n f(X_i|\theta),$$

and a maximum likelihood estimator is defined as

$$\hat{\theta}_n^{MLE} \in \arg\max_{\theta\in\Theta} L(\theta) = \arg\max_{\theta\in\Theta} \prod_{i=1}^n f(X_i|\theta)$$

which is equivalent to

$$\arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta).$$

Under some regularity conditions, the maximum likelihood estimator $\hat{\theta}_n^{MLE}$ has the following nice *limiting* properties:

A. (Consistency) $\hat{\theta}_n^{MLE} \to \theta_0$, in probability and almost surely.

B. (Asymptotic Normality) $\sqrt{n}(\hat{\theta}_n^{MLE} - \theta_0) \to \mathcal{N}(0, I_0^{-1})$ in distribution, where I_0 is the Fisher Information matrix

$$I(\theta_0) := \mathbb{E}[(\frac{\partial}{\partial\theta}\log f(X|\theta_0))^2] = -\mathbb{E}[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)].$$

C. (Asymptotic Efficiency) $\lim_{n\to\infty}\operatorname{cov}(\hat{\theta}_n^{MLE}) = I^{-1}(\theta_0)$. Hence $\hat{\theta}_n^{MLE}$ is the Uniformly Minimum-Variance Unbiased Estimator, i.e. the estimator with the least variance among the class of unbiased estimators, for any unbiased estimator $\hat{\theta}_n$, $\lim_{n\to\infty}\operatorname{var}(\hat{\theta}_n^{MLE}) \leq \lim_{n\to\infty}\operatorname{var}(\hat{\theta}_n^{MLE})$.

However in *finite sample* case, there are better estimators than MLEs, which include some *bias* in further reduction of *variance*.

1.1. Example: Multivariate Normal Distribution. For example, consider the normal distribution $\mathcal{N}(\mu, \Sigma)$,

$$f(X|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right],$$

where $|\Sigma|$ is the determinant of covariance matrix Σ.

To get the MLE of normal distribution, we need to

$$\max_{\mu,\Sigma} P(X_1,...,X_n|\mu,\Sigma) = \max_{\mu,\Sigma}\prod_{i=1}^n \frac{1}{\sqrt{2\pi|\Sigma|}}\exp[-(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)]$$

It is equivalent to maximize the log-likelihood

$$I = \log P(X_1,...,X_n|\mu,\Sigma) = -\frac{1}{2}\sum_{i=1}^n (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\log|\Sigma| + C$$

Let μ^* is the MLE of μ, we have

$$0 = \frac{\partial I}{\partial \mu^*} = -\sum_{i=1}^n \Sigma^{-1}(X_i - \mu^*)$$

$$\Rightarrow \mu^* = \frac{1}{n}\sum_{i=1}^n X_i = \hat{\mu}_n$$

To get the estimation of Σ, we need to maximize

$$I(\Sigma) = \operatorname{trace}(I) = -\frac{1}{2}\operatorname{trace}\sum_{i=1}^n (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) - \frac{n}{2}\operatorname{trace}\log|\Sigma| + C$$

$$\begin{aligned}
-\frac{1}{2}\operatorname{trace}\sum_{i=1}^n (X_i-\mu)^T\Sigma^{-1}(X_i-\mu) &= -\frac{1}{2}\sum_{i=1}^n \operatorname{trace}[\Sigma^{-1}(X_i-\mu)(X_i-\mu)^T] \\
&= -\frac{1}{2}(\operatorname{trace}\Sigma^{-1}\hat{\Sigma}_n)(n-1) \\
&= -\frac{n-1}{2}\operatorname{trace}(\Sigma^{-1}\hat{\Sigma}_n^{\frac{1}{2}}\hat{\Sigma}_n^{\frac{1}{2}}) \\
&= -\frac{n-1}{2}\operatorname{trace}(\hat{\Sigma}_n^{\frac{1}{2}}\Sigma^{-1}\hat{\Sigma}_n^{\frac{1}{2}}) \\
&= -\frac{n-1}{2}\operatorname{trace}(S)
\end{aligned}$$

where

$$\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T,$$

$S = \hat{\Sigma}_n^{\frac{1}{2}}\Sigma^{-1}\hat{\Sigma}_n^{\frac{1}{2}}$ is symmetric and positive definite. Above we repeatedly use cyclic property of trace:

- trace$(AB) =$ trace(BA), or more generally
- (invariance under cyclic permutation group) trace$(ABCD) =$ trace$(BCDA) =$ trace$(CDAB) =$ trace$(DABC)$.

Then we have

$$\Sigma = \hat{\Sigma}_n^{-\frac{1}{2}}S^{-1}\hat{\Sigma}_n^{-\frac{1}{2}}$$
$$-\frac{n}{2}\log|\Sigma| = \frac{n}{2}\log|S| + \frac{n}{2}\log|\hat{\Sigma}_n| = f(\hat{\Sigma}_n)$$

Therefore,

$$\max I(\Sigma) \Leftrightarrow \min \frac{n-1}{2}\mathrm{trace}(S) - \frac{n}{2}\log|S| + Const(\hat{\Sigma}_n, 1)$$

Suppose $S = U\Lambda U$ is the eigenvalue decomposition of S, $\Lambda = \mathrm{diag}(\lambda_i)$

$$J = \frac{n-1}{2}\sum_{i=1}^{p}\lambda_i - \frac{n}{2}\sum_{i=1}^{p}\log(\lambda_i) + Const$$
$$\frac{\partial J}{\partial \lambda_i} = \frac{n-1}{2} - \frac{n}{2}\frac{1}{\lambda_i} \Rightarrow \lambda_i = \frac{n}{n-1}$$
$$S = \frac{n}{n-1}I_p$$

This gives the MLE solution

$$\Sigma^* = \frac{n-1}{n}\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T,$$

which differs to $\hat{\Sigma}_n$ only in that the denominator $(n-1)$ is replaced by n. In covariance matrix, $(n-1)$ is used because for a single sample $n = 1$, there is no variance at all.

Fixed p, when $n \to \infty$, MLE satisfies $\hat{\mu}_n \to \mu$ and $\hat{\Sigma}_n \to \Sigma$. However as we can see in the following classes, they are not the best estimators when the dimension of the data p gets large, with finite sample n.

2. Bias-Variance Decomposition of Mean Square Error

Consider multivariate Gaussian model: let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \Sigma)$, $X_i \in \mathbb{R}^p (i = 1 \ldots n)$, then the maximum likelihood estimators (MLE) of the parameters (μ and Σ) are as follows:

$$\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n}X_i, \qquad \hat{\Sigma}_n^{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)(X_i - \hat{\mu}_n)^T.$$

For simplicity, take a coordinate transform (PCA) $Y_i = U^T X_i$ where $\Sigma = U\Lambda U^T$ is an eigen-decomposition. Assume that $\Lambda = \sigma^2 I_p$ and $n = 1$, then it suffices to consider $Y \sim \mathcal{N}(\mu, \sigma^2 I_p)$ in the sequel. In this case $\hat{\mu}^{MLE} = Y$.

To measure the performance of an estimator $\hat{\mu}_n$, one may look at the following so-called *risk*,

$$R(\hat{\mu}_n, \mu) = \mathbb{E}L(\hat{\mu}_n, \mu)$$

where the loss function takes the square loss here

$$L(\hat{\mu}_n, \mu) = \|\hat{\mu}_n - \mu\|^2.$$

The mean square error (MSE) to measure the risk enjoys the following *bias-variance decomposition*, from the Pythagorean theorem.

$$\begin{aligned} R(\hat{\mu}_n, \mu) &= \mathbb{E}\|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n] + \mathbb{E}[\hat{\mu}_n] - \mu\|^2 \\ &= \mathbb{E}\|\hat{\mu}_n - \mathbb{E}[\hat{\mu}_n]\|^2 + \|\mathbb{E}[\hat{\mu}_n] - \mu\|^2 \\ &=: Var(\hat{\mu}_n) + Bias(\hat{\mu}_n)^2 \end{aligned}$$

Example 1. For the simple case $Y_i \sim \mathcal{N}(\mu, \sigma^2 I_p)$ $(i = 1, \ldots, n)$, the MLE estimator satisfies

$$Bias(\hat{\mu}_n^{MLE}) = 0$$

and

$$Var(\hat{\mu}_n^{MLE}) = \frac{p}{n}\sigma^2$$

In particular for $n = 1$, $Var(\hat{\mu}^{MLE}) = \sigma^2 p$ for $\hat{\mu}^{MLE} = Y$.

Example 2. MSE of Linear Estimators. Consider $Y \sim \mathcal{N}(\mu, \sigma^2 I_p)$ and linear estimator $\hat{\mu}_C = CY$. Then we have

$$Bias(\hat{\mu}_C) = \|(I - C)\mu\|^2$$

and

$$Var(\hat{\mu}_C) = \mathbb{E}[(CY - C\mu)^T (CY - C\mu)] = \mathbb{E}[\text{trace}((Y - \mu)^T C^T C(Y - \mu))] = \sigma^2 \text{trace}(C^T C).$$

In applications, one often consider the diagonal linear estimators $C = \text{diag}(c_i)$, e.g. in Ridge regression

$$\min_{\mu} \frac{1}{2}\|Y - X\beta\|^2 + \frac{\lambda}{2}\|\beta\|^2.$$

For diagonal linear estimators, the risk

$$R(\hat{\mu}_C, \mu) = \sigma^2 \sum_{i=1}^{p} c_i^2 + \sum_{i=1}^{p} (1 - c_i)^2 \mu_i^2.$$

In this case, it is simple to find minimax risk over the hyper-rectangular model class $|\mu_i| \leq \tau_i$,

$$\inf_{c_i} \sup_{|\mu_i| \leq \tau_i} R(\hat{\mu}_C, \mu) = \sum_{i=1}^{p} \frac{\sigma^2 \tau_i^2}{\sigma^2 + \tau_i^2}.$$

From here one can see that for those sparse model classes such that $\#\{i : \tau_i = O(\sigma)\} = k \ll p$, it is possible to get smaller risk using linear estimators than MLE.

In general, is it possible to introduce some *biased* estimators which significantly reduces the *variance* such that the total risk is smaller than MLE uniformly for all μ? This is the notion of inadmissibility introduced by Charles Stein in 1956 and he find the answer is YES by presenting the James-Stein estimators, as the shrinkage of sample means.

3. Stein's Phenomenon and Shrinkage of Sample Mean

Definition (Inadmissible)**.** An estimator $\hat{\mu}_n$ of the parameter μ is called **inadmissible** on $\mathbb{R}^p$ with respect to the squared risk if there exists another estimator μ_n^* such that

$$\mathbb{E}\|\mu_n^* - \mu\|^2 \leq \mathbb{E}\|\hat{\mu}_n - \mu\|^2 \qquad \text{for all } \mu \in \mathbb{R}^p,$$

and there exist $\mu_0 \in \mathbb{R}^p$ such that

$$\mathbb{E}\|\mu_n^* - \mu_0\|^2 < \mathbb{E}\|\hat{\mu}_n - \mu_0\|^2.$$

In this case, we also call that μ_n^* **dominates** $\hat{\mu}_n$. Otherwise, the estimator $\hat{\mu}_n$ is called **admissible**.

The notion of inadmissibility or dominance introduces a partial order on the set of estimators where admissible estimators are local optima in this partial order.

Stein (1956) [**Ste56**] found that if $p \geq 3$, then the MLE estimator $\hat{\mu}_n$ is inadmissible. This property is known as ***Stein's phenomenon***. This phenomenon can be described like:

For $p \geq 3$, there exists $\hat{\mu}$ such that $\forall \mu$,

$$R(\hat{\mu}, \mu) < R(\hat{\mu}^{\text{MLE}}, \mu)$$

which makes MLE inadmissible.

A typical choice is the *James-Stein estimator* given by James-Stein (1961),

$$\tilde{\mu}_n^{JS} = \left(1 - \frac{\sigma^2(p-2)}{\|\hat{\mu}_n^{MLE}\|}\right)\hat{\mu}_n^{MLE}, \qquad \sigma = \varepsilon.$$

Theorem 3.1. Suppose $Y \sim \mathcal{N}_p(\mu, I)$. Then $\hat{\mu}^{\text{MLE}} = Y$. $R(\hat{\mu}, \mu) = \mathbb{E}_\mu\|\hat{\mu} - \mu\|^2$, and define

$$\hat{\mu}^{\text{JS}} = \left(1 - \frac{p-2}{\|Y\|^2}\right)Y$$

then

$$R(\hat{\mu}^{\text{JS}}, \mu) < R(\hat{\mu}^{\text{MLE}}, \mu)$$

We'll prove a useful lemma first.

3.1. Stein's Unbiased Risk Estimates (SURE).

Discussions below are all under the assumption that $Y \sim \mathcal{N}_p(\mu, I)$.

Lemma 3.2. (Stein's Unbiased Risk Estimates (SURE)) Suppose $\hat{\mu} = Y + g(Y)$, g satisfies [1]

(1) g is weakly differentiable.
(2) $\sum_{i=1}^p \int |\partial_i g_i(x)| \mathrm{d}x < \infty$

then

$$R(\hat{\mu}, \mu) = \mathbb{E}_\mu(p + 2\nabla^T g(Y) + \|g(Y)\|^2) \tag{18}$$

where $\nabla^T g(Y) := \sum_{i=1}^p \frac{\partial}{\partial y_i} g_i(Y)$.

[1] cf. p38, Prop 2.4 [GE]

Examples of $g(x)$: For James-Stein estimator

$$g(x) = -\frac{p-2}{\|Y\|^2} Y$$

and for soft-thresholding, each component

$$g_i(x) = \left\{ \begin{array}{cc} -\lambda & x_i > \lambda \\ -x_i & |x_i| \leq \lambda \\ \lambda & x_i < -\lambda \end{array} \right.$$

Both of them are weakly differentiable. But Hard-Thresholding:

$$g_i(x) = \left\{ \begin{array}{cc} 0 & |x_i| > \lambda \\ -x_i & |x_i| \leq \lambda \end{array} \right.$$

which is not weakly differentiable!

Proof. Let $\phi(y)$ be the density function of standard Normal distribution $\mathcal{N}_p(0, I)$.

$$\begin{aligned} R(\hat{\mu}, \mu) &= \mathbb{E}_\mu \|Y + g(Y) - \mu\|^2 \\ &= \mathbb{E}_\mu \left(p + 2(Y-\mu)^T g(Y) + \|g(Y)\|^2 \right) \end{aligned}$$

$$\begin{aligned} \mathbb{E}_\mu (Y-\mu)^T g(Y) &= \sum_{i=1}^p \int_{-\infty}^{\infty} (y_i - \mu_i) g_i(Y) \phi(Y-\mu) \mathrm{d}Y \\ &= \sum_{i=1}^p \int_{-\infty}^{\infty} -g_i(Y) \frac{\partial}{\partial y_i} \phi(Y-\mu) \mathrm{d}Y, \quad \text{derivative of Gaussian function} \\ &= \sum_{i=1}^p \int_{-\infty}^{\infty} \frac{\partial}{\partial y_i} g_i(Y) \phi(Y-\mu) \mathrm{d}Y, \quad \text{Integration by parts} \\ &= \mathbb{E}_\mu \nabla^T g(Y) \end{aligned}$$

□

Thus, we define

$$U(Y) := p + 2\nabla^T g(Y) + \|g(Y)\|^2 \tag{19}$$

for convenience, and $R(\hat{\mu}, \mu) = \mathbb{E}_\mu U(Y)$.

This lemma is in fact called the Stein's lemma in Tsybakov's book [Tsy09] (page 157∼158).

3.2. Risk of Linear Estimator.

$$\hat{\mu}_C(Y) = Cy$$

$$g(Y) = (C - I)Y$$

$$\nabla^T g(Y) = -\sum_i \frac{\partial}{\partial y_i} \left((C-I)Y\right) = \text{trace}(C) - p$$

$$\begin{aligned} U(Y) &= p + 2\nabla^T g(Y) + \|g(Y)\|^2 \\ &= p + 2(\text{trace}(C) - p) + \|(I-C)Y\|^2 \\ &= -p + 2\text{trace}(C) + \|(I-C)Y\|^2 \end{aligned}$$

In applications, $C = C(\lambda)$ often depends on some regularization parameter λ (e.g. ridge regression). So one could find optimal λ^* by minimizing the MSE over λ. Suppose $Y \sim \mathcal{N}(\mu, \sigma^2 I)$,

$$R(\hat{\mu}_C, \mu) = \|(I - C(\lambda))Y\|^2 - p\sigma^2 + 2\sigma^2 \text{trace}(C(\lambda)).$$

3.3. Risk of James-Stein Estimator. Recall

$$g(Y) = -\frac{p-2}{\|Y\|^2} Y$$

$$U(Y) = p + 2\nabla^T g(Y) + \|g(Y)\|^2$$

$$\|g(Y)\|^2 = \frac{(p-2)^2}{\|Y\|^2}$$

$$\nabla^T g(Y) = -\sum_i \frac{\partial}{\partial y_i} \left(\frac{p-2}{\|Y\|^2} Y \right) = -\frac{(p-2)^2}{\|Y\|^2}$$

we have

$$R(\hat{\mu}^{\text{JS}}, \mu) = \mathbb{E}U(Y) = p - \mathbb{E}_\mu \frac{(p-2)^2}{\|Y\|^2} < p = R(\hat{\mu}^{\text{MLE}}, \mu)$$

when $p \geq 3$.

Problem. What's wrong when $p = 1$? Does SURE still hold?

Remark. Indeed, we have the following theorem

Theorem 3.3 (Lemma 2.8 in Johnstone's book (GE))**.** *$Y \sim N(\mu, I)$, $\forall \hat{\mu} = CY$, $\hat{\mu}$ is admissable iff*

1. *C is symmetric.*
2. *$0 \leq \rho_i(C) \leq 1$ (eigenvalue).*
3. *$\rho_i(C) = 1$ for at most two i.*

To find an upper bound of the risk of James-Stein estimator, notice that $\|Y\|^2 \sim \chi^2(\|\mu\|^2, p)$ and [2]

$$\chi^2(\|\mu\|^2, p) \overset{d}{=} \chi^2(0, p + 2N), \quad N \sim \text{Poisson}\left(\frac{\|\mu\|^2}{2} \right)$$

we have

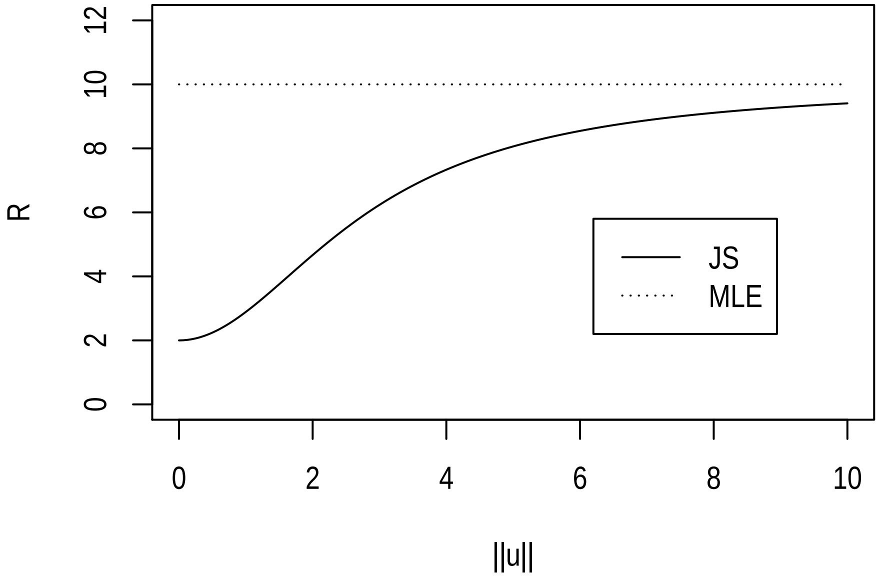
$$\begin{aligned}
\mathbb{E}_\mu \left(\frac{1}{\|Y\|^2} \right) &= \mathbb{E}\mathbb{E}_\mu \left[\frac{1}{\|Y\|^2} \Big| N \right] \\
&= \mathbb{E} \frac{1}{p + 2N - 2} \\
&\geq \frac{1}{p + 2\mathbb{E}N - 2} \text{ (Jensen's Inequality)} \\
&= \frac{1}{p + \|\mu\|^2 - 2}
\end{aligned}$$

that is

[2] This is a homework.

Proposition 3.4 (Upper bound of MSE for the James-Stein Estimator)**.** $Y \sim \mathcal{N}(\mu, I_p)$,

$$R(\hat{\mu}^{\mathrm{JS}}, \mu) \leq p - \frac{(p-2)^2}{p-2+\|\mu\|^2} = 2 + \frac{(p-2)\|\mu\|^2}{p-2+\|\mu\|^2}$$



3.4. Risk of Soft-thresholding.

Using Stein's unbiased risk estimate, we have soft-thresholding in the form of

$$\hat{\mu}(x) = x + g(x). \quad \frac{\partial}{\partial i} g_i(x) = -I(|x_i| \leq \lambda)$$

We then have

$$\begin{aligned}\mathbb{E}_\mu \|\hat{\mu}_\lambda - \mu\|^2 &= \mathbb{E}_\mu \left(p - 2\sum_{i=1}^p I(|x_i| \leq \lambda) + \sum_{i=1}^p x_i^2 \wedge \lambda^2 \right) \\ &\leq 1 + (2\log p + 1) \sum_{i=1}^p \mu_i^2 \wedge 1 \quad \text{if we take } \lambda = \sqrt{2\log p}\end{aligned}$$

By using the inequality

$$\frac{1}{2} a \wedge b \leq \frac{ab}{a+b} \leq a \wedge b$$

we can compare the risk of soft-thresholding and James-Stein estimator as

$$1 + (2\log p + 1) \sum_{i=1}^p (\mu_i^2 \wedge 1) \quad \lesseqgtr \quad 2 + c\left(\left(\sum_{i=1}^p \mu_i^2 \right) \wedge p \right) \quad c \in (1/2, 1)$$

In LHS, the risk for each μ_i is bounded by 1 so if μ is sparse ($s = \#\{i : \mu_i \neq 0\}$) but large in magnitudes (s.t. $\|\mu\|_2^2 \geq p$), we may expect LHS $= O(s \log p) < O(p) =$ RHS. [3]

In addition to L_1 penalty in LASSO, there are also other penalty functions like

- $\lambda\|\beta\|_0$ This leads to *hard-thresholding* when $X = I$. Solving this problem is normally NP-hard.
- $\lambda\|\beta\|_p$, $0 < p < 1$. Non-convex, also NP-hard.
- $\lambda \sum \rho(\beta_i)$. such that
  (1) $\rho'(0)$ singular (for sparsity in variable selection)
  (2) $\rho'(\infty) = 0$ (for unbiasedness in parameter estimation)
  Such ρ must be non-convex essentially (Jianqing Fan and Runze Li, 2001).

3.5. How to Optimize the Constants in James-Stein Estimator? Now, let us look for a function g such that the risk of the estimator $\tilde{\mu}_n(Y) = (1 - g(Y))Y$ is smaller than the MLE of $Y \sim \mathcal{N}(\mu, \varepsilon^2 I_p)$. We have

$$\begin{aligned}\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 &= \sum_{i=1}^p \mathbb{E}[((1 - g(y))y_i - \mu_i)^2] \\ &= \sum_{i=1}^p \{\mathbb{E}[(y_i - \mu_i)^2] + 2\mathbb{E}[(\mu_i - y_i)g(y)y_i] \\ &\qquad + \mathbb{E}[y_i^2 g(y)^2]\}.\end{aligned}$$

Suppose now that the function g is such that the assumptions of Stein's Lemma 3.5 hold (page 157~158 in Tsybakov's book [Tsy09]), i.e. weakly differentiable.

Lemma 3.5 (Stein's lemma)**.** Suppose that a function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies:

(i) $f(u_1, \ldots, u_p)$ is absolutely continuous in each coordinate u_i for almost all values (with respect to the Lebesgue measure on $\mathbb{R}^{p-1}$) of other coordinates $(u_j, j \neq i)$

(ii)
$$\mathbb{E}\left|\frac{\partial f(y)}{\partial y_i}\right| < \infty, \qquad i = 1, \ldots, p.$$

then
$$\mathbb{E}[(\mu_i - y_i)f(y)] = -\varepsilon^2 \mathbb{E}\left[\frac{\partial f}{\partial y_i}(y)\right], \qquad i = 1, \ldots, p.$$

With Stein's Lemma, therefore
$$\mathbb{E}[(\mu_i - y_i)(1 - g(y))y_i] = -\varepsilon^2 \mathbb{E}\left[g(y) + y_i \frac{\partial g}{\partial y_i}(y)\right],$$
with
$$\mathbb{E}[(y_i - \mu_i)^2] = \varepsilon^2 = \sigma^2,$$
we have
$$\mathbb{E}[(\tilde{\mu}_{n,i} - \mu_i)]^2 = \varepsilon^2 - 2\varepsilon^2 \mathbb{E}\left[g(y) + y_i \frac{\partial g}{\partial y_i}(y)\right] + \mathbb{E}[y_i^2 g(y)^2].$$
Summing over i gives
$$\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 = p\varepsilon^2 + \mathbb{E}[W(y)] = \mathbb{E}\|\hat{\mu}_n - \mu\|^2 + \mathbb{E}[W(y)]$$

[3] also cf. p43 [GE]

with

$$W(y) = -2p\varepsilon^2 g(y) + 2\varepsilon^2 \sum_{i=1}^{p} y_i \frac{\partial g}{\partial y_i}(y) + \|y\|^2 g(y)^2.$$

The risk of $\tilde{\mu}_n$ is smaller than that of $\hat{\mu}_n$ if we choose g such that

$$\mathbb{E}[W(y)] < 0.$$

In order to satisfy this inequality, we can search for g among the functions of the form

$$g(y) = \frac{b}{a + \|y\|^2}$$

with an appropriately chosen constants $a \geq 0$, $b > 0$. Therefore, $W(y)$ can be written as

$$\begin{aligned} W(y) &= -2p\varepsilon^2 \frac{b}{a+\|y\|^2} + 2\varepsilon^2 \sum_{i=1}^{p} \frac{2by_i^2}{(a+\|y\|^2)^2} + \frac{b^2\|y\|^2}{(a+\|y\|^2)^2} \\ &= \frac{1}{a+\|y\|^2}\left(-2pb\varepsilon^2 + \frac{4b\varepsilon^2\|y\|^2}{a+\|y\|^2} + \frac{b^2\|y\|^2}{(a+\|y\|^2)^2}\right) \\ &\leq \frac{1}{a+\|y\|^2}(-2pb\varepsilon^2 + 4b\varepsilon^2 + b^2) \qquad \|y\|^2 \leq a + \|y\|^2 \text{ for } a \geq 0 \\ &= \frac{Q(b)}{a+\|y\|^2}, \qquad Q(b) = b^2 - 2pb\varepsilon^2 + 4b\varepsilon^2. \end{aligned}$$

The minimizer in b of quadratic function $Q(b)$ is equal to

$$b_{opt} = \varepsilon^2(p-2),$$

where the minimum of $W(y)$ satisfies

$$W_{min}(y) \leq -\frac{b_{opt}^2}{a+\|y\|^2} = -\frac{\varepsilon^4(p-2)^2}{a+\|y\|^2} < 0.$$

Note that when $b \in (b_1, b_2)$, *i.e.* between the two roots of $Q(b)$

$$b_1 = 0, \qquad b_2 = 2\varepsilon^2(p-2)$$

we have $W(y) < 0$, which may lead to other estimators having smaller mean square errors than MLE estimator.

When $a = 0$, the function g and the estimator $\tilde{\mu}_n = (1 - g(y))y$ associated to this choice of g are given by

$$g(y) = \frac{\varepsilon^2(p-2)}{\|y\|^2},$$

and

$$\tilde{\mu}_n = \left(1 - \frac{\varepsilon^2(p-2)}{\|y\|^2}\right) y =: \tilde{\mu}_{JS},$$

respectively. $\tilde{\mu}_{JS}$ is called **James-Stein estimator**. If dimension $p \geq 3$ and the norm $\|y\|^2$ is sufficiently large, multiplication of y by $g(y)$ shrinks the value of y to 0. This is called the ***Stein shrinkage***. If $b = b_{opt}$, then

$$W_{min}(y) = -\frac{\varepsilon^4(p-2)^2}{\|y\|^2}.$$

Lemma 3.6. Let $p \geq 3$. Then, for all $\mu \in \mathbb{R}^p$,

$$0 < \mathbb{E}\left(\frac{1}{\|y\|^2}\right) < \infty.$$

The proof of Lemma 3.6 can be found on Tsybakov's book [Tsy09] (page 158∼159). For the function W, Lemma 3.6 implies $-\infty < \mathbb{E}[W(y)] < 0$, provided that $p \geq 3$. Therefore, if $p \geq 3$, the risk of the estimator $\tilde{\mu}_n$ satisfies

$$\mathbb{E}\|\tilde{\mu}_n - \mu\|^2 = p\varepsilon^2 - \mathbb{E}\left(\frac{\varepsilon^4(p-2)^2}{\|y\|^2}\right) < \mathbb{E}\|\hat{\mu}_n - \mu\|^2$$

for all $\mu \in \mathbb{R}^p$.

Besides James-Stein estimator, there are other estimators having smaller mean square errors than MLE $\hat{mu}_n$.

- *Stein estimator*: $a = 0, b = \varepsilon^2 p$,
$$\tilde{\mu}_S := \left(1 - \frac{\varepsilon^2 p}{\|y\|^2}\right) y$$
- *James-Stein estimator*: $c \in (0, 2(p-2))$
$$\tilde{\mu}^c_{JS} := \left(1 - \frac{\varepsilon^2 c}{\|y\|^2}\right) y$$
- *Positive part James-Stein estimator*:
$$\tilde{\mu}_{JS+} := \left(1 - \frac{\varepsilon^2 (p-2)}{\|y\|^2}\right)_+ y$$
- *Positive part Stein estimator*:
$$\tilde{\mu}_{S+} := \left(1 - \frac{\varepsilon^2 p}{\|y\|^2}\right)_+ y$$

where $(x)_+ = \min(0, x)$. Denote the mean square error by $MSE(\tilde{\mu}) = \mathbb{E}\|\tilde{\mu} - \mu\|^2$, then we have

$$MSE(\tilde{\mu}_{JS+}) < MSE(\tilde{\mu}_{JS}) < MSE(\hat{\mu}_n), \qquad MSE(\tilde{\mu}_{S+}) < MSE(\tilde{\mu}_S) < MSE(\hat{\mu}_n).$$

See Efron's Book, Chap 1, Table 1.1.

Another dimension of variation is Shrinkage toward *any vector* rather than the origin.

$$\tilde{\mu}_{\mu_0} = \mu_0 + \left(1 - \frac{\varepsilon^2 c}{\|y\|^2}\right)(y - \mu_0), \qquad c \in (0, 2(p-2)).$$

In particular, one may choose $\mu_0 = \bar{y}$ where $\bar{y} = \sum_{i=1}^p y_i / p$.

3.6. Discussion. Stein's phenomenon firstly shows that in high dimensional estimation, shrinkage may lead to better performance than MLE, the sample mean. This opens a new era for modern high dimensional statistics. In fact discussions above study independent random variables in p-dimensional space, concentration of measure tells us some priori knowledge about the estimator distribution – samples are concentrating around certain point. Shrinkage toward such point may naturally lead to better performance.

However, after Stein's phenomenon firstly proposed in 1956, for many years researchers have not found the expected revolution in practice. Mostly because Stein's type estimators are too complicated in real applications and very small gain can be achieved in many cases. Researchers struggle to show real application examples where one can benefit greatly from Stein's estimators. For example, Efron-Morris (1974) showed three examples that JS-estimator significantly improves the multivariate estimation. On other other hand, deeper understanding on Shrinkage-type estimators has been pursued from various aspects in statistics.

The situation changes dramatically when LASSO-type estimators by Tibshirani, also called Basis Pursuit by Donoho et al. are studied around 1996. This brings sparsity and L1-regularization into the central theme of high dimensional statistics and leads to a new type of shrinkage estimator, thresholding. For example,

$$\min_{\tilde{\mu}} I = \min_{\tilde{\mu}} \frac{1}{2}\|\tilde{\mu} - \mu\|^2 + \lambda\|\tilde{\mu}\|_1$$

Subgradients of I over $\tilde{\mu}$ leads to

$$0 \in \partial_{\tilde{\mu}_j} I = (\tilde{\mu}_j - \mu_j) + \lambda \text{sign}(\tilde{\mu}_j) \Rightarrow \tilde{\mu}_j = \text{sign}(\mu_j)(|\mu_j| - \lambda)_+$$

where the set-valued map $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$, and $\text{sign}(x) = [-1, 1]$ if $x = 0$, is the subgradient of absolute function $|x|$. Under this new framework shrinkage estimators achieves a new peak with an ubiquitous spread in data analysis with high dimensionality.

4. Random Matrix Theory and Phase Transitions in PCA

In PCA, one often looks at the eigenvalue plot in an decreasing order as percentage or variations. A large gap in the eigenvalue drops may indicate those top eigenvectors reflect major variation directions, where those small eigenvalues indicate directions due to noise which will vanish when $n \to \infty$. Is this true in all situations? The answer is yes in classical setting $p << n$. Unfortunately, in high dimensional statistics even with fixed ratio $p/n = \gamma$, top eigenvectors of sample covariance matrices might not reflect the subspace of signals. In the following we consider one particularly simple example: rank-1 signal (spike) model, where random matrix theory will tell us when PCA fails to capture the signal subspace.

First of all, let's introduce some basic results in random matrix theory which will be used later.

4.1. Marčenko-Pastur Law of Sample Covariance Matrix.

Let $\mathrm{X} \in \mathbb{R}^{p*n}$, $X_i \sim \mathcal{N}(0, I_p)$.

When p fixed and $n \to \infty$, the classical Law of Large Numbers tells us

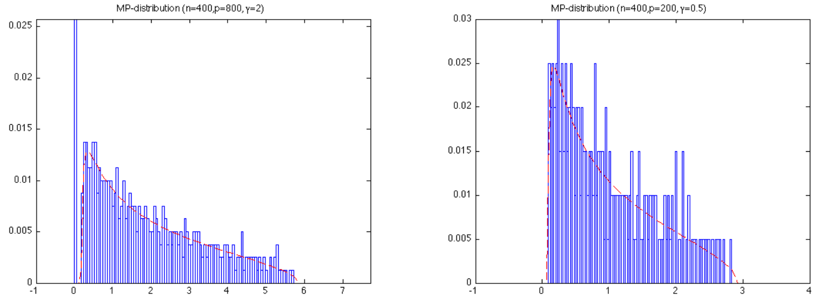
$$\widehat{\Sigma}_n = \frac{1}{n} X X' \to \mathrm{I_p}. \tag{20}$$

Such a random matrix $\widehat{\Sigma}_n$ is called Wishart matrix.

But when $\frac{p}{n} \to \gamma \neq 0$, the distribution of the eigenvalues of $\widehat{\Sigma}_n$ follows [BS10] (Chapter 3), if $\gamma \leq 1$,

$$\mu^{MP}(t) = \begin{cases} 0 & t \notin [a, b] \\ \frac{\sqrt{(b-t)(t-a)}}{2\pi\gamma t} dt & t \in [a, b] \end{cases} \tag{21}$$

and has an additional point mass $1 - 1/\gamma$ at the origin if $\gamma > 1$. Note that $a = (1 - \sqrt{\gamma})^2, b = (1 + \sqrt{\gamma})^2$. Figure 1 illustrates the MP-distribution by MATLAB simulations whose codes can be found below.



(a) (b)

FIGURE 1. (a) Marčenko-Pastur distribution with $\gamma = 2$. (b) Marčenko-Pastur distribution with $\gamma = 0.5$.

```
 %Wishart matrix
% S = 1/n*X*X.', X is p-by-n, X_ij i.i.d N(0,1),
% ESD_S converge to M.P. with parameter y = p/n

y = 2;

a = (1-sqrt(y))^2;
b = (1+sqrt(y))^2;

f_MP = @(t) sqrt(max(b-t, 0).*max(t-a, 0) )./(2*pi*y*t); %MP Distribution


%non-zero eigenvalue part
n = 400;
p = n*y;

X = randn(p,n);
S = 1/n*(X*X.');
evals = sort( eig(S), 'descend');

nbin = 100;
[nout, xout] = hist(evals, nbin);
hx = xout(2) - xout(1); % step size, used to compute frequency below
x1 = evals(end) -1;
x2 = evals(1) + 1; % two end points
xx = x1+hx/2:  hx:  x2;
fre = f_MP(xx)*hx;

figure,
h = bar(xout, nout/p);
set(h, 'BarWidth', 1, 'FaceColor', 'w', 'EdgeColor', 'b');
hold on;
plot(xx, fre, '--r');
```

```
if y > 1 % there are (1-1/y)*p zero eigenvalues
axis([-1 x2+1 0 max(fre)*2]);
end
```

In the following, we are going to show that if dimension is relatively large compared to sample size, *i.e.* $p/n \to \gamma > 0$, PCA may fail to identify signals from noise even the signal lies in a low-dimensional subspace. In fact, there is a phase transition for signal identifiability by PCA: below a threshold of signal-noise ratio, PCA will fail with high probability and above that threshold of signal-noise ratio, PCA will approximate the signal subspace with high probability. This will be illustrated by the following simplest rank-1 model.

4.2. Phase Transitions of PCA in Rank-1 Model. Consider the following rank-1 signal-noise model

$$Y = X + \varepsilon,$$

where signal lies in an one-dimensional subspace $X = \alpha u$ with $\alpha \sim \mathcal{N}(0, \sigma_X^2)$ and noise $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_\mathrm{p})$ is i.i.d. Gaussian. For multi-rank models, please see [**KN08**]. Therefore $Y \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = \sigma_X^2 uu' + \sigma_\varepsilon^2 I_p.$$

The whole question in the remaining part of this section is to ask, *can we recover signal direction u from principal component analysis on noisy measurements Y*?

Define the signal-noise ratio $SNR = R = \frac{\sigma_X^2}{\sigma_\varepsilon^2}$, where for simplicity $\sigma_\varepsilon^2 = 1$. We aim to show how SNR affect the result of PCA when p is large. A fundamental result by Johnstone in 2006 [**Joh06**], or see [**NBG10**], shows that the primary (largest) eigenvalue of sample covariance matrix satisfies

$$\lambda_{max}(\widehat{\Sigma}_n) \to \begin{cases} (1+\sqrt{\gamma})^2 = b, & \sigma_X^2 \le \sqrt{\gamma} \\ (1+\sigma_X^2)(1+\frac{\gamma}{\sigma_X^2}), & \sigma_X^2 > \sqrt{\gamma} \end{cases} \tag{22}$$

which implies that if signal energy is small, top eigenvalue of sample covariance matrix never pops up from random matrix ones; only if the signal energy is beyond the phase transition threshold $\sqrt{\gamma}$, top eigenvalue can be separated from random matrix eigenvalues. However, even in the latter case it is a biased estimation.

Moreover, the primary eigenvector associated with the largest eigenvalue (principal component) converges to

$$|\langle u, v_{max}\rangle|^2 \to \begin{cases} 0 & \sigma_X^2 \le \sqrt{\gamma} \\ \frac{1-\frac{\gamma}{\sigma_X^4}}{1+\frac{\gamma}{\sigma_X^2}}, & \sigma_X^2 > \sqrt{\gamma} \end{cases} \tag{23}$$

which means the same phase transition phenomenon: if signal is of low energy, PCA will tell us nothing about the true signal and the estimated top eigenvector is orthogonal to the true direction u; if the signal is of high energy, PCA will return a biased estimation which lies in a cone whose angle with the true signal is no more than

$$\frac{1-\frac{\gamma}{\sigma_X^4}}{1+\frac{\gamma}{\sigma_X^2}}.$$

Below we are going to show such results.

4.3. Stieltjes Transform. The following Stieltjes Transformation of MP-density will be useful in the next part. Define the Stieltjes Transformation of MP-density μ^{MP} to be

$$s(z) := \int_R \frac{1}{t-z} d\mu^{MP}(t), \ \ z \in C \tag{24}$$

If $z \in \mathbb{R}$, the transformation is called Hilbert Transformation. Further details can be found in Terry Tao's textbook, Topics on Random Matrix Theory [Tao11], Sec. 2.4.3 (the end of page 169) for the definition of Stieltjes transform of a density $p(t)dt$ on $\mathbb{R}$.

In [BS10], Lemma 3.11 on page 52 gives the following characterization of $s(z)$ (note that the book contains a typo that $4y\sigma^4$ in numerator should be replaced by $4yz\sigma^2$):

$$s(z) = \frac{(1-\gamma) - z + \sqrt{(z-1-\gamma)^2 - 4\gamma z}}{2\gamma z}, \tag{25}$$

which is the largest root of the quadratic equation,

$$\gamma z s(z)^2 + (z - (1-\gamma))s(z) + 1 = 0 \Longleftrightarrow z + \frac{1}{s(z)} = \frac{1}{1+\gamma s(z)}. \tag{26}$$

From the equation (25), one can take derivative of z on both side to obtain $s'(z)$ in terms of s and z. Using $s(z)$ one can compute the following basic integrals.

Lemma 4.1. (1)

$$\int_a^b \frac{t}{\lambda - t} \mu^{MP}(t)dt = -\lambda s(\lambda) - 1;$$

(2)

$$\int_a^b \frac{t^2}{(\lambda-t)^2} \mu^{MP}(t)dt = \lambda^2 s'(\lambda) + 2\lambda s(\lambda) + 1$$

Proof. For convenience, define

$$T(\lambda) := \int_a^b \frac{t}{\lambda - t} \mu^{MP}(t)dt. \tag{27}$$

Note that

$$1 + T(\lambda) = 1 + \int_a^b \frac{t}{\lambda - t}\mu^{MP}(t)dt = \int_a^b \frac{\lambda - t + t}{\lambda - t}\mu^{MP}(t)dt = -\lambda s(\lambda) \tag{28}$$

which give the first result.

4.4. Characterization of Phase Transitions with RMT. First of all, we give an overview of this part. Following the rank-1 model, consider random vectors $\{Y_i\}_{i=1}^n \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \sigma_x^2 uu^T + \sigma_\varepsilon^2 I_p$ and $\|u\|^2 = 1$. This covariance matrix Σ thus has a structure that low-rank plus sparse matrix. Define the Signal-Noise-Ratio (SNR) $R = \frac{\sigma_x^2}{\sigma_\varepsilon^2}$. Without of generality,we assume $\sigma_\varepsilon^2 = 1$.

The sample covariance matrix of Y is $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n Y_i Y_i^t = \frac{1}{n} YY^T$ where $Y = [Y_1, \ldots, Y_n] \in \mathbb{R}^{p\times n}$. 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:

$$Y_i = \Sigma^{\frac{1}{2}} Z_i \to Z_i \sim N(0, I_p), \text{where } \Sigma^{\frac{1}{2}} = \sigma_x^2 uu^T + \sigma_\varepsilon^2 I_p = Ruu^T + I_p \tag{29}$$

Then $S_n = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T$ is a Wishart random matrix whose eigenvalues follow the MP distribution.

Notice that $\hat{\Sigma}_n = \Sigma^{\frac{1}{2}} S_n \Sigma^{\frac{1}{2}}$ and $(\lambda, \hat{v})$ is eigenvalue-eigenvector pair of matrix $\hat{\Sigma}_n$. Therefore

$$\Sigma^{\frac{1}{2}} S_n \Sigma^{\frac{1}{2}} \hat{v} = \lambda \hat{v} \ \Rightarrow \ S_n \Sigma(\Sigma^{-\frac{1}{2}}\hat{v}) = \lambda(\Sigma^{-\frac{1}{2}}\hat{v}) \tag{30}$$

In other words, λ and $\Sigma^{-\frac{1}{2}}\hat{v}$ are 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 and thus satisfies,

$$c^2 = c\hat{v}^T\hat{v} = v^T\Sigma v = v^T(\sigma_x^2 uu^T + \sigma_\varepsilon^2)v = \sigma_x^2(u^Tv)^2 + \sigma_\varepsilon^2) = \mathbb{R}(u^Tv)^2 + 1. \tag{31}$$

With the aid of Stieltjes transform, we can calculate the largest eigenvalue of matrix $\hat{\Sigma}_n$ and the properties of the corresponding eigenvector $\hat{v}$.

In fact, the eigenvalue λ satisfies

$$1 = \sigma_X^2 \cdot \frac{1}{p}\sum_{i=1}^p \frac{\lambda_i}{\lambda - \sigma_\varepsilon^2\lambda_i} \sim \sigma_X^2 \cdot \int_a^b \frac{t}{\lambda - \sigma_\varepsilon^2 t} d\mu^{MP}(t), \tag{32}$$

and the inner product of u and v satisfies

$$\begin{aligned} & |u^Tv|^2 \\ = \ & \{\sigma_x^4 \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2)^2} d\mu^{MP}(t)\}^{-1} \\ = \ & \{\frac{\sigma_x^4}{4\gamma}(-4\lambda + (a+b) + 2(\sqrt{(\lambda-a)(\lambda-b)}) + \frac{\lambda(2\lambda-(a+b))}{\sqrt{(\lambda-a)(\lambda-b)}})\}^{-1} \\ = \ & \frac{1 - \frac{\gamma}{R^2}}{1 + \gamma + \frac{2\gamma}{R}} \end{aligned} \tag{33}$$

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:

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

Now we are going to present the details.

First of all, from

$$S_n\Sigma v = \lambda v, \tag{34}$$

we obtain the following by plugging in the expression of Σ

$$S_n(\sigma_X^2 uu' + \sigma_\varepsilon^2 I_p)v = \lambda v \tag{35}$$

Rearrange the term with u to one side, we got

$$(\lambda I_p - \sigma_\varepsilon^2 S_n)v = \sigma_X^2 S_n uu'v \tag{36}$$

Assuming that $\lambda I_p - \sigma_\varepsilon^2 S_n$ is invertable, then multiple its reversion at both sides of the equality, we get,

$$v = \sigma_X^2 \cdot (\lambda I_p - \sigma_\varepsilon^2 S_n)^{-1} \cdot S_n u(u'v). \tag{37}$$

4.4.1. *Primary Eigenvalue.* Multiply (37) by u' at both side,

$$u'v = \sigma_X^2 \cdot u'(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-1} S_n u \cdot (u'v) \tag{38}$$

that is, if $u'v \neq 0$,

$$1 = \sigma_X^2 \cdot u'(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-1} S_n u \tag{39}$$

For SVD $S_n = W \Lambda W'$, where Λ is diagonal, $W \cdot W' = W' \cdot W = I_p$, $W = [W_1, W_2, \cdots, W_n] \in \mathbb{R}^{p\times p}, \alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n] \in \mathbb{R}^{p\times 1}$, in which W_i is the corresponding eigenvector, then $u = \sum_{i=1}^p \alpha_i W_i = W \cdot \alpha$, then, $\alpha = W'u$, and,

$$1 = \sigma_X^2 \cdot u'[W(\lambda I_p - \sigma_\varepsilon^2 \Lambda)^{-1} W'][W\Lambda W']u = \sigma_X^2 \cdot (u'W)(\lambda I_p - \sigma_\varepsilon^2 \Lambda)^{-1} \Lambda (W'u) \tag{40}$$

Replace $W'u = \alpha$, then,

$$1 = \sigma_X^2 \cdot \sum_{i=1}^p \frac{\lambda_i}{\lambda - \sigma_\varepsilon^2 \lambda_i} \alpha_i^2 \tag{41}$$

where $\sum_{i=1}^p \alpha_i^2 = 1$. Since W is a random orthogonal basis on a sphere, α_i will concentrate on its mean $\alpha_i = \frac{1}{\sqrt{q}}$. According to the fact that p is large enough($\sim \infty$), due to Law of Large Numbers(LLN) and $\lambda \sim \mu^{MP}$(λ_i can be thought sampled from the μ^{MP}), the equation (12) can be thought of as the Expected Value (Monte-Carlo Integration), then equation (12) can be written as,

$$1 = \sigma_X^2 \cdot \frac{1}{p} \sum_{i=1}^p \frac{\lambda_i}{\lambda - \sigma_\varepsilon^2 \lambda_i} \sim \sigma_X^2 \cdot \int_a^b \frac{t}{\lambda - \sigma_\varepsilon^2 t} d\mu^{MP}(t) \tag{42}$$

For convenience, assume without loss of generosity that $\sigma_\varepsilon^2 = 1$, that is the noise volatility is 1. Now we unveil the story of the ratio γ, do the integration in equation (13), we got,

$$1 = \sigma_X^2 \cdot \int_a^b \frac{t}{\lambda - t} \frac{\sqrt{(b-t)(t-a)}}{2\pi\gamma t} dt = \frac{\sigma_X^2}{4\gamma}[2\lambda - (a+b) - 2\sqrt{|(\lambda - a)(b - \lambda)|}] \tag{43}$$

where the last step can be computed via Stieltjes transform introduced above.

From the definition of $T(\lambda)$, we have

$$\int_a^b \frac{t^2}{(\lambda - t)^2} \mu^{MP}(t) dt = -T(\lambda) - \lambda T'(\lambda). \tag{44}$$

Combined with the first result, we reach the second one. □

If we suppose $\sigma_\varepsilon^2 = 1$ as in the script. Then $R = \sigma_X^2$. Note that when $R \geq \sqrt{\frac{p}{n}}$, $\lambda \geq b$. Solve the equation:

$$\because \qquad 1 = \frac{\sigma_X^2}{4\gamma}[2\lambda - (a+b) - 2\sqrt{(\lambda - a)(\lambda - b)}$$

$$\therefore \qquad \lambda = \sigma_X^2 + \frac{\gamma}{\sigma_X^2} + 1 + \gamma = (1 + \sigma_X^2)(1 + \frac{\gamma}{\sigma_X^2})$$

Loose this assumption. Then all the equations above is true, except that all the λ will be replaced by $\frac{\lambda}{\sigma_\varepsilon^2}$ and λ_0 by SNR. Then we get:

$$\lambda = (1 + SNR)(1 + \frac{\gamma}{SNR})\sigma_\varepsilon^2$$

Here we observe the following phase transitions for primary eigenvalue:

- If $\lambda \in [a, b]$, then $\widehat{\Sigma}_n$ has eigenvalue λ within $\mathrm{supp}(\mu^{MP})$, so it is undistinguishable from the noise S_n.
- If $\lambda \geq b$, PCA will pick up the top eigenvalue as non-noise. So $\lambda = b$ is the phase transition where PCA works to pop up correct eigenvalue. Then plug in $\lambda = b$ in equation (14), we get,

$$1 = \sigma_X^2 \cdot \frac{1}{4\gamma}[2b - (a+b)] = \frac{\sigma_X^2}{\sqrt{\gamma}} \Leftrightarrow \sigma_X^2 = \sqrt{\frac{p}{n}} \tag{45}$$

So, in order to make PCA works, we need to let $SNR \geq \sqrt{\frac{p}{n}}$.

We know that if PCA works good and noise doesn't dominate the effect, the inner-product $|u'\hat{v}|$ should be close to 1. On the other hand, from RMT we know that if the top eigenvalue λ is merged in the M. P. distribution, then the top eigenvector computed is purely random and $|u'\hat{v}| = 0$, which means that from $\hat{v}$ we can know nothing about the signal u.

4.4.2. *Primary Eigenvector.* We now study the phase transition of top-eigenvector.

It is convenient to study $|u'v|^2$ first and then translate back to $|u'\hat{v}|^2$. Using the equation (37),

$$1 = |v'v| = \sigma_X^4 \cdot v'uu'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n uu'v = \sigma_X^4 \cdot (|v'u|)[u'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n u](|u'v|) \tag{46}$$

$$|u'v|^{-2} = \sigma_X^4 [u'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n u] \tag{47}$$

Using the same trick as the equation (39),

$$|u'v|^{-2} = \sigma_X^4 [u'S_n(\lambda I_p - \sigma_\varepsilon^2 S_n)^{-2} S_n u] \sim \sigma_X^4 \cdot \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2 t)^2} d\mu^{MP}(t) \tag{48}$$

and assume that $\lambda > b$, from Stieltjes transform introduced later one can compute the integral as

$$|u'v|^{-2} = \sigma_X^4 \cdot \int_a^b \frac{t^2}{(\lambda - \sigma_\varepsilon^2 t)^2} d\mu^{MP}(t) = \frac{\sigma_X^4}{4\gamma}(-4\lambda + (a+b) + 2\sqrt{(\lambda-a)(\lambda-b)} + \frac{\lambda(2\lambda - (a+b))}{\sqrt{(\lambda-a)(\lambda-b)}} \tag{49}$$

from which it can be computed that (using $\lambda = (1+R)(1+\frac{\gamma}{R})$ obtained above, where $R = SNR = \frac{\sigma_X^2}{\sigma_\epsilon^2}$)

$$|u'v|^2 = \frac{1 - \frac{\gamma}{R^2}}{1 + \gamma + \frac{2\gamma}{R}}.$$

Using the relation

$$u'\hat{v} = u'\left(\frac{1}{c}\Sigma^{1/2} v\right) = \frac{\sqrt{1+R}}{c}(u'v)$$

where the second equality uses $\Sigma^{1/2}u = \sqrt{1+R}u$, and with the formula for c^2 above, we can compute

$$(u'\hat{v})^2 = \frac{1+R}{1+R(u'v)^2}(u'v)^2$$

in terms of R. Note that this number holds under the condition that $R > \sqrt{\gamma}$.

4.5. Further Comments. When $\frac{log(p)}{n} \to 0$, we need to add more restrictions on $\widehat{\Sigma}_n$ in order to estimate it faithfully. There are typically three kinds of restrictions.

- Σ sparse
- Σ^{-1} sparse, also called–Precision Matrix
- banded structures (e.g. Toeplitz) on Σ or Σ^{-1}

Recent developments can be found by Bickel, Tony Cai, Tsybakov, Wainwright et al.

For spectral study on random kernel matrices, see El Karoui, Tiefeng Jiang, Xiuyuan Cheng, and Amit Singer et al.

CHAPTER 4

Generalized PCA/MDS via SDP Relaxations

1. Introduction of SDP with a Comparison to LP

Here we will give a short note on Semidefinite Programming (SDP) formulation of Robust PCA, Sparse PCA, MDS with uncertainty, and Maximal Variance Unfolding, etc. First of all, we give a short introduction to SDP based on a parallel comparison with LP.

Semi-definite programming (SDP) involves linear objective functions and linear (in)equalities constraint with respect to variables as positive semi-definite matrices. SDP is a generalization of linear programming (LP) by replacing nonnegative variables with positive semi-definite matrices. We will give a brief introduction of SDP through a comparison with LP.

LP (Linear Programming): for $x \in \mathbb{R}^n$ and $c \in \mathbb{R}^n$,

$$\begin{aligned} \min \quad & c^T x \\ s.t. \quad & Ax = b \\ & x \geq 0 \end{aligned} \tag{50}$$

This is the primal linear programming problem.

In SDP, the inner product between vectors $c^T x$ in LP will change to Hadamard inner product (denoted by $\bullet$) between matrices.

SDP (Semi-definite Programming): for $X, C \in \mathbb{R}^{n \times n}$

$$\begin{aligned} \min \quad & C \bullet X = \sum_{i,j} c_{ij} X_{ij} \\ s.t. \quad & A_i \bullet X = b_i, \quad \text{for } i = 1, \cdots, m \\ & X \succeq 0 \end{aligned} \tag{51}$$

Linear programming has a dual problem via the Lagrangian. The Lagrangian of the primal problem is

$$\max_{\mu \geq 0, y} \min_x L_{x;y,\mu} = c^T x + y^T (b - Ax) - \mu^T x$$

which implies that

$$\frac{\partial L}{\partial x} = c - A^T y - \mu = 0$$

$$\Longleftrightarrow c - A^T y = \mu \geq 0$$

$$\Longrightarrow \max_{\mu \geq 0, y} L = -y^T b$$

which leads to the following dual problem.

LD (Dual Linear Programming):

$$\begin{aligned} \min \quad & b^T y \\ s.t. \quad & \mu = c - A^T y \geq 0 \end{aligned} \tag{52}$$

In a similar manner, for SDP's dual form, we have the following.

SDD (Dual Semi-definite Programming):

$$\begin{aligned} \max \quad & -b^T y \\ s.t. \quad & S = C - \sum_{i=1}^{m} A_i y_i \succeq 0 =: C - A^T \otimes y \end{aligned} \tag{53}$$

where

$$A = \begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix}$$

and

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

1.1. Duality of SDP. Define the feasible set of primal and dual problems are $\mathbb{F}_p = \{X \succeq 0; A_i \bullet X = b_i\}$ and $\mathbb{F}_d = \{(y, S) : S = C - \sum_i y_i A_i \succeq 0\}$, respectively. Similar to linear programming, semi-definite programming also has properties of week and strong duality. The week duality says that the primal value is always an upper bound of dual value. The strong duality says that the existence of an interior point ensures the vanishing duality gap between primal value and dual value, as well as the complementary conditions. In this case, to check the optimality of a primal variable, it suffices to find a dual variable which meets the complementary condition with the primal. This is often called the *witness* method.

For more reference on duality of SDP, see e.g. [Ali95].

Theorem 1.1 (Weak Duality of SDP)**.** If $\mathbb{F}_p \neq \emptyset, \mathbb{F}_d \neq \emptyset$, We have $C \bullet X \geq b^T y$, for $\forall X \in \mathbb{F}_p$ and $\forall (y, S) \in \mathbb{F}_d$.

Theorem 1.2 (Strong Duality SDP)**.** Assume the following hold,

(1) $\mathbb{F}_p \neq \emptyset, \mathbb{F}_d \neq \emptyset$;
(2) At least one feasible set has an interior.

Then X^* is optimal iff

(1) $X^* \in \mathbb{F}_p$
(2) $\exists (y^*, S^*) \in \mathbb{F}_d$

s.t. $C \bullet X^* = b^T y^*$ or $X^* S^* = 0$ (note: in matrix product)

In other words, the existence of an interior solution implies the complementary condition of optimal solutions. Under the complementary condition, we have

$$\operatorname{rank}(X^*) + \operatorname{rank}(S^*) \leq n$$

for every optimal primal X^* and dual S^*.

2. Robust PCA

Let $X \in \mathbb{R}^{p\times n}$ be a data matrix. Classical PCA tries to find

$$\begin{aligned} \min \quad & \|X - L\| \\ s.t. \quad & \operatorname{rank}(L) \leq k \end{aligned} \tag{54}$$

where the norm here is matrix-norm or Frobenius norm. SVD provides a solution with $L = \sum_{i\leq k} \sigma_i u_i v_i^T$ where $X = \sum_i \sigma_i u_i v_i^T$ $(\sigma_1 \geq \sigma_2 \geq \ldots)$. In other words, classical PCA looks for decomposition

$$X = L + E$$

where the error matrix E has small matrix/Frobenius norm. However, it is well-known that classical PCA is sensitive to outliers which are sparse and lie far from the major population.

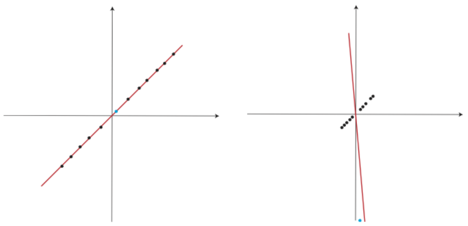


FIGURE 1. Classical PCA is sensitive to outliers

Robust PCA looks for the following decomposition instead

$$X = L + S$$

where

- L is a low rank matrix;
- S is a sparse matrix.

Example. Let $X = [x_1, \ldots, x_p]^T \sim \mathcal{N}(0, \Sigma)$ be multivariate Gaussian random variables. The following characterization [CPW12] holds

$$x_i \text{ and } x_j \text{ are conditionally independent given other variables}$$

$$\Leftrightarrow (\Sigma^{-1})_{ij} = 0$$

We denote it by $x_i \perp x_j | x_k (k \not\in \{i,j\})$. Let $G = (V, E)$ be a undirected graph where V represent p random variables and $(i,j) \in E \Leftrightarrow x_i \perp x_j | x_k (k \not\in \{i,j\})$. G is called a (Gaussian) graphical model of X.

Divide the random variables into observed and hidden (a few) variables $X = (X_o, X_h)^T$ (in semi-supervised learning, unlabeled and labeled, respectively) and

$$\Sigma = \begin{bmatrix} \Sigma_{oo} & \Sigma_{oh} \\ \Sigma_{ho} & \Sigma_{hh} \end{bmatrix} \quad \text{and} \quad Q = \Sigma^{-1} = \begin{bmatrix} Q_{oo} & Q_{oh} \\ Q_{ho} & Q_{hh} \end{bmatrix}$$

The following Schur Complement equation holds for covariance matrix of observed variables

$$\Sigma_{oo}^{-1} = Q_{oo} + Q_{oh} Q_{hh}^{-1} Q_{ho}.$$

Note that

- Observable variables are often conditional independent given hidden variables, so Q_{oo} is expected to be *sparse*;
- Hidden variables are of small number, so $Q_{oh}Q_{hh}^{-1}Q_{ho}$ is of *low-rank*.

In semi-supervised learning, the labeled points are of small number, and the unlabeled points should be as much conditional independent as possible to each other given labeled points. This implies that the labels should be placed on those most "influential" points.

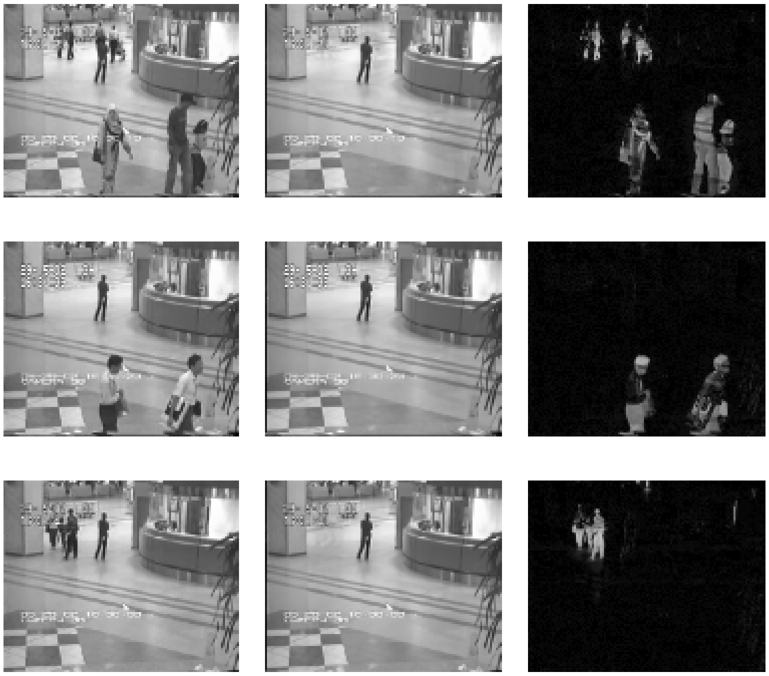


FIGURE 2. Surveilliance video as low rank plus sparse matrices: Left = low rank (middle) + sparse (right) [CLMW09]

Example (Surveilliance Video Decomposition)**.** Figure 2 gives an example of low rank vs. sparse decomposition in surveilliance video. On the left column, surveilliance video of a movie theatre records a great amount of images with the same background and the various walking customers. If we vectorize these images (each image as a vector) to form a matrix, the background image leads to a rank-1 part and the occasional walking customers contribute to the sparse part.

More examples can be found at [CLMW09, CSPW11, CPW12].

In Robust PCA the purpose is to solve

$$\begin{aligned} \min \quad & \|X - L\|_0 \\ s.t. \quad & \mathrm{rank}(L) \leq k \end{aligned} \tag{55}$$

where $\|A\|_0 = \#\{A_{ij} \neq 0\}$. However both the objective function and the constraint are non-convex, whence it is NP-hard to solve in general.

The simplest convexification leads to a Semi-definite relaxation:

$$\|S\|_0 := \#\{S_{ij} \neq 0\} \Rightarrow \|S\|_1$$

$$\mathrm{rank}(L) := \#\{\sigma_i(L) \neq 0\} \Rightarrow \|L\|_* = \sum_i \sigma_i(L),$$

where $\|L\|_*$ is called the *nuclear norm* of L, which has a semi-definite representation

$$\begin{aligned} \|L\|_* = \quad & \min \quad \frac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2)) \\ & s.t. \quad \begin{bmatrix} W_1 & L \\ L^T & W_2 \end{bmatrix} \succeq 0. \end{aligned}$$

With these, the relaxed Robust PCA problem can be solved by the following semi-definite programming (SDP).

$$\begin{aligned} \min \quad & \frac{1}{2}(\mathrm{trace}(W_1) + \mathrm{trace}(W_2)) + \lambda\|S\|_1 \\ s.t. \quad & L_{ij} + S_{ij} = X_{ij}, \quad (i,j) \in E \\ & \begin{bmatrix} W_1 & L \\ L^T & W_2 \end{bmatrix} \succeq 0 \end{aligned} \tag{56}$$

The following Matlab codes realized the SDP algorithm above by CVX (http://cvxr.com/cvx).

```
 % Construct a random 20-by-20 Gaussian matrix and construct a rank-1
% matrix using its top-1 singular vectors
R = randn(20,20);
[U,S,V] = svds(R,3);
A = U(:,1)*V(:,1)’;

% Construct a 90% uniformly sparse matrix
E0 = rand(20);
E = 1*abs(E0>0.9);

X = A + E;

% Choose the regularization parameter
lambda = 0.25;

% Solve the SDP by calling cvx toolbox
if exist(’cvx_setup.m’,’file’),
    cd /matlab_tools/cvx/
    cvx_setup
end

cvx_begin
    variable L(20,20);
    variable S(20,20);
    variable W1(20,20);
    variable W2(20,20);
    variable Y(40,40) symmetric;
    Y == semidefinite(40);
    minimize(.5*trace(W1)+0.5*trace(W2)+lambda*sum(sum(abs(S))));
    subject to
        L + S >= X-1e-5;
        L + S <= X + 1e-5;
        Y == [W1, L';L W2];
cvx_end

% The difference between sparse solution S and E
disp('$\—S-E\—_\infty$:')
norm(S-E,'inf')

% The difference between the low rank solution L and A
disp('\—A-L\—')
norm(A-L)
```

Typically CVX only solves SDP problem of small sizes (say matrices of size less than 100). Specific matlab tools have been developed to solve large scale RPCA, which can be found at http://perception.csl.uiuc.edu/matrix-rank/.

3. Probabilistic Exact Recovery Conditions for RPCA

A fundamental question about Robust PCA is: given $X = L_0 + S_0$ with low-rank L and sparse S, under what conditions that one can recover X by solving SDP in (56)?

It is necessary to assume that

- the low-rank matrix L_0 can not be sparse;
- the sparse matrix S_0 can not be of low-rank.

The first assumption is called incoherence condition. Assume that $L_0 \in \mathbb{R}^{n\times n} = U\Sigma V^T$ and $r = \mathrm{rank}(L_0)$.

Incoherence condition [CR09]: there exists a $\mu \geq 1$ such that for all $e_i = (0,\ldots,0,1,0,\ldots,0)^T$,

$$\|U^T e_i\|^2 \leq \frac{\mu r}{n}, \quad \|V^T e_i\|^2 \leq \frac{\mu r}{n},$$

and

$$|UV^T|_{ij}^2 \leq \frac{\mu r}{n^2}.$$

These conditions, roughly speaking, ensure that the singular vectors are not sparse, i.e. well-spread over all coordinates and won't concentrate on some coordinates. The incoherence condition holds if $|U_{ij}|^2 \vee |V_{ij}|^2 \leq \mu/n$. In fact, if U represent random projections to r-dimensional subspaces with $r \geq \log n$, we have $\max_i \|U^T e_i\|^2 \asymp r/n$.

To meet the second condition, we simply assume that the sparsity pattern of S_0 is uniformly random.

Theorem 3.1. Assume the following holds,

(1) L_0 is n-by-n with $\mathrm{rank}(L_0) \leq \rho_r n \mu^{-1} (\log n)^{-2}$,
(2) S_0 is uniformly sparse of cardinality $m \leq \rho_s n^2$.

Then with probability $1 - O(n^{-10})$, (56) with $\lambda = 1/\sqrt{n}$ is exact, *i.e.* its solution $\hat{L} = L_0$ and $\hat{S} = S_0$.

Note that if L_0 is a rectangular matrix of $n_1 \times n_2$, the same holds with $\lambda = 1/\sqrt{(\max n_1, n_2)}$. The result can be generalized to $1 - O(n^{-\beta})$ for $\beta > 0$. Extensions and improvements of these results to incomplete measurements can be found in [CT10, Gro11] etc., which solves the following SDP problem.

$$\begin{aligned} \text{(57)} \qquad & \min \quad \|L\|_* + \lambda \|S\|_1 \\ & s.t. \quad L_{ij} + S_{ij} = X_{ij}, \quad (i,j) \in \Omega_{obs}. \end{aligned}$$

Theorem 3.2. Assume the following holds,

(1) L_0 is n-by-n with $\mathrm{rank}(L_0) \leq \rho_r n \mu^{-1} (\log n)^{-2}$,
(2) Ω_{obs} is a uniform random set of size $m = 0.1n^2$,
(3) each observed entry is corrupted with probability $\tau \leq \tau_s$.

Then with probability $1 - O(n^{-10})$, (56) with $\lambda = 1/\sqrt{0.1n}$ is exact, *i.e.* its solution $\hat{L} = L_0$. The same conclusion holds for rectangular matrices with $\lambda = 1/\sqrt{\max dim}$.

All these results hold irrespective to the magnitudes of L_0 and S_0.

When there are no sparse perturbation in optimization problem (57), the problem becomes the classical Matrix Completion problem with uniformly random sampling:

$$\begin{aligned} \text{(58)} \qquad & \min \quad \|L\|_* \\ & s.t. \quad L_{ij} = L^0_{ij}, \quad (i,j) \in \Omega_{obs}. \end{aligned}$$

Assumed the same condition as before,[CT10] gives the following result: solution to SDP (58) is exact with probability at least $1 - n^{-10}$ if $m \geq \mu n r \log^a n$ where $a \leq 6$, which can be improved by [Gro11] to be near-optimal

$$m \geq \mu n r \log^2 n.$$

Another theory based on geometry can be found in [CSPW11, CRPW12].

3.1. Phase Transitions. Take $L_0 = UV^T$ as a product of $n \times r$ i.i.d. $\mathcal{N}(0,1)$ random matrices. Figure 3 shows the phase transitions of successful recovery probability over sparsity ratio $\rho_s = m/n^2$ and low rank ratio r/n. White color indicates the probability equals to 1 and black color corresponds to the probability being 0. A sharp phase transition curve can be seen in the pictures. (a) and (b) respectively use random signs and coherent signs in sparse perturbation, where (c) is purely matrix completion with no perturbation. Increasing successful recovery can be seen from (a) to (c).

4. Sparse PCA

Sparse PCA is firstly proposed by [ZHT06] which tries to locate sparse principal components, which also has a SDP relaxation.

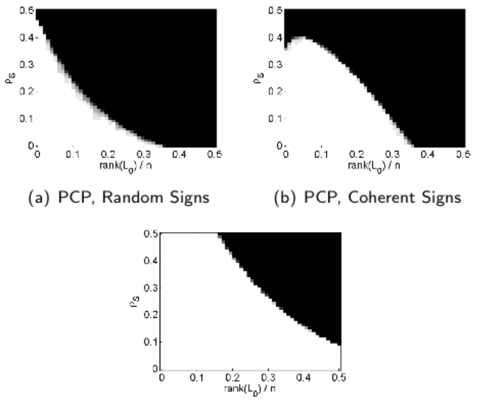
Recall that classical PCA is to solve

$$\begin{aligned} \max \quad & x^T \Sigma x \\ s.t. \quad & \|x\|_2 = 1 \end{aligned}$$

which gives the maximal variation direction of covariance matrix Σ.

Note that $x^T \Sigma x = \mathrm{trace}(\Sigma(xx^T))$. Classical PCA can thus be written as

$$\begin{aligned} \max \quad & \mathrm{trace}(\Sigma X) \\ s.t. \quad & \mathrm{trace}(X) = 1 \\ & X \succeq 0 \end{aligned}$$



(a) PCP, Random Signs (b) PCP, Coherent Signs

(c) Matrix Completion

FIGURE 3. Phase Transitions in Probability of Successful Recovery

The optimal solution gives a rank-1 X along the first principal component. A recursive application of the algorithm may lead to top k principal components. That is, one first to find a rank-1 approximation of Σ and extract it from $\Sigma_0 = \Sigma$ to get $\Sigma_1 = \Sigma - X$, then pursue the rank-1 approximation of Σ_1, and so on.

Now we are looking for sparse principal components, i.e. $\#\{X_{ij} \neq 0\}$ are small. Using 1-norm convexification, we have the following SDP formulation [dGJL07] for Sparse PCA

$$\begin{aligned} \max \quad & \mathrm{trace}(\Sigma X) - \lambda \|X\|_1 \\ s.t. \quad & \mathrm{trace}(X) = 1 \\ & X \succeq 0 \end{aligned}$$

The following Matlab codes realized the SDP algorithm above by CVX (http://cvxr.com/cvx).

```
 % Construct a 10-by-20 Gaussian random matrix and form a 20-by-20 correlation
% (inner product) matrix R
X0 = randn(10,20);
R = X0'*X0;

d = 20;
e = ones(d,1);

% Call CVX to solve the SPCA given R
if exist('cvx_setup.m','file'),
    cd /matlab_tools/cvx/
    cvx_setup
end

lambda = 0.5;
k = 10;

cvx_begin
```

```
    variable X(d,d) symmetric;
    X == semidefinite(d);
    minimize(-trace(R*X)+lambda*(e'*abs(X)*e));
    subject to
        trace(X)==1;
cvx_end
```

5. MDS with Uncertainty

In this lecture, we introduce Semi-Definite Programming (SDP) approach to solve some generalized Multi-dimensional Scaling (MDS) problems with uncertainty. Recall that in classical MDS, given pairwise distances $d_{ij} = \|x_i - x_j\|^2$ among a set of points $x_i \in \mathbb{R}^p$ ($i = 1, 2, \cdots, n$) whose coordinates are unknown, our purpose is to find $y_i \in \mathbb{R}^k (k \leq p)$ such that

$$\min \sum_{i,j=1}^{n} \left(\|y_i - y_j\|^2 - d_{ij} \right)^2 . \tag{59}$$

In classical MDS (Section 1 in Chapter 1) an eigen-decomposition approach is pursued to find a solution when all pairwise distances d_{ij}'s are known and noise-free. In case that d_{ij}'s are not from pairwise distances, we often use gradient descend method to solve it. However there is no guarantee that gradient descent will converge to the global optimal solution. In this section we will introduce a method based on convex relaxation, in particular the semi-definite relaxation, which will guarantee us to find optimal solutions in the following scenarios.

- Noisy perturbations: $d_{ij} \to \widetilde{d_{ij}} = d_{ij} + \epsilon_{ij}$
- Incomplete measurments: only partial pairwise distance measurements are available on an edge set of graph, *i.e.* $G = (V, E)$ and d_{ij} is given when $(i, j) \in E$ (*e.g.* x_i and x_j in a neighborhood).
- Anchors: sometimes we may fixed the locations of some points called *anchors*, *e.g.* in sensor network localization (SNL) problem.

In other words, we are looking for MDS on graphs with partial and noisy information.

5.1. SD Relaxation of MDS.

Like PCA, classical MDS has a semi-definite relaxation. In the following we shall introduce how the constraint

$$\|y_i - y_j\|^2 = d_{ij}, \tag{60}$$

can be relaxed into linear matrix inequality system with positive semidefinite variables.

Denote $Y = [y_1, \cdots, y_n]^{k \times n}$ where $y_i \in \mathbb{R}^k$, and

$$e_i = (0, 0, \cdots, 1, 0, \cdots, 0) \in \mathbb{R}^n.$$

Then we have

$$\|y_i - y_j\|^2 = (y_i - y_j)^T (y_i - y_j) = (e_i - e_j)^T Y^T Y (e_i - e_j)$$

Set $X = Y^T Y$, which is symmetric and positive semi-definite. Then

$$\|Y_i - Y_j\|^2 = (e_i - e_j)(e_i - e_j)^T \bullet X.$$

So

$$\|Y_i - Y_j\|^2 = d_{ij}^2 \Leftrightarrow (e_i - e_j)(e_i - e_j)^T \bullet X = d_{ij}^2$$

which is linear with respect to X.

Now we relax the constrain $X = Y^T Y$ to

$$X \succeq Y^T Y \Longleftrightarrow X - Y^T Y \succeq 0.$$

Through Schur Complement Lemma we know

$$X - Y^T Y \succeq 0 \Longleftrightarrow \left[\begin{array}{cc} I & Y \\ Y^T & X \end{array} \right] \succeq 0$$

We may define a new variable

$$Z \in S^{k+n}, Z = \left[\begin{array}{cc} I_k & Y \\ Y^T & X \end{array} \right]$$

which gives the following result.

Lemma 5.1. The quadratic constraint

$$\|y_i - y_j\|^2 = d_{ij}^2, \quad (i,j) \in E$$

has a semi-definite relaxation:

$$\begin{cases} Z_{1:k,1:k} = I \\ (0; e_i - e_j)(0; e_i - e_j)^T \bullet Z = d_{ij}^2, \quad (i,j) \in E \\ Z = \left[\begin{array}{cc} I_k & Y \\ Y^T & X \end{array} \right] \succeq 0. \end{cases}$$

where $\bullet$ denotes the Hadamard inner product, *i.e.* $A \bullet B := \sum_{i,j=1}^n A_{ij} B_{ij}$.

Note that the constraint with equalities of d_{ij}^2 can be replaced by inequalities such as $\le d_{ij}^2(1+\epsilon)$ (or $\ge d_{ij}^2(1-\epsilon)$). This is a system of linear matrix (in)-equalities with positive semidefinite variable Z. Therefore, the problem becomes a typical semidefinite programming.

Given such a SD relaxation, we can easily generalize classical MDS to the scenarios in the introduction. For example, consider the generalized MDS with anchors which is often called *sensor network localization* problem in literature [**BLT$^+$06**]. Given anchors a_k ($k = 1, \ldots, s$) with known coordinates, find x_i such that

- $\|x_i - x_j\|^2 = d_{ij}^2$ where $(i,j) \in E_x$ and x_i are unknown locations
- $\|a_k - x_j\|^2 = \widehat{d_{kj}}^2$ where $(k,j) \in E_a$ and a_k are known locations

We can exploit the following SD relaxation:

- $(0; e_i - e_j)(0; e_i - e_j)^T \bullet Z = d_{ij}$ for $(i,j) \in E_x$,
- $(a_i; e_j)(a_i; e_j)^T \bullet Z = \widehat{d_{ij}}$ for $(i,j) \in E_a$,

both of which are linear with respect to Z.

Recall that every SDP problem has a dual problem (SDD). The SDD associated with the primal problem above is

$$\min \quad I \bullet V + \sum_{i,j \in E_x} w_{ij} d_{ij} + \sum_{i,j \in E_a} \widehat{w}_{ij} \widehat{d_{ij}} \tag{61}$$

s.t.

$$S = \left(\begin{array}{cc} V & 0 \\ 0 & 0 \end{array} \right) + \sum_{i,j \in E_x} w_{ij} A_{ij} + \sum_{i,j \in E_a} \widehat{w}_{ij} \widehat{A_{ij}} \succeq 0$$

where

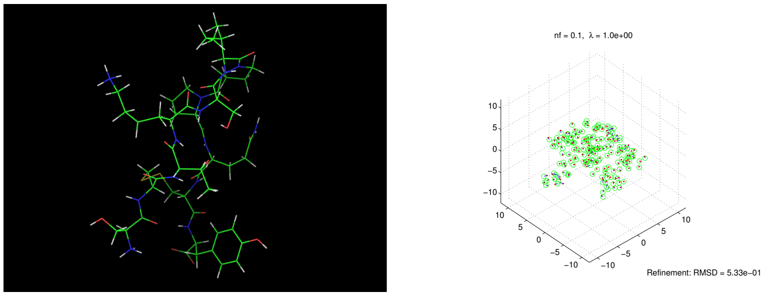
$$A_{ij} = (0; e_i - e_j)(0; e_i - e_j)^T$$
$$\widehat{A_{ij}} = (a_i; e_j)(a_i; e_j)^T.$$

The variables w_{ij} is the stress matrix on edge between unknown points i and j and $\widehat{w}_{ij}$ is the stress matrix on edge between anchor i and unknown point j. Note that the dual is always feasible, as $V = 0$, $y_{ij} = 0$ for all $(i, j) \in E_x$ and $w_{ij} = 0$ for all $(i, j) \in E_a$ is a feasible solution.

There are many matlab toolboxes for SDP, *e.g.* CVX, SEDUMI, and recent toolboxes SNLSDP (http://www.math.nus.edu.sg/~mattohkc/SNLSDP.html) and DISCO (http://www.math.nus.edu.sg/~mattohkc/disco.html) by Toh *et. al.*, adapted to MDS with uncertainty.

A crucial theoretical question is to ask, when $X = Y^T Y$ holds such that SDP embedding Y gives the same answer as the classical MDS? Before looking for answers to this question, we first present an application example of SDP embedding.

5.2. Protein 3D Structure Reconstruction. Here we show an example of using SDP to find 3-D coordinates of a protein molecule based on noisy pairwise distances for atoms in ϵ-neighbors. We use matlab package SNLSDP by Kim-Chuan Toh, Pratik Biswas, and Yinyu Ye, downladable at http://www.math.nus.edu.sg/~mattohkc/SNLSDP.html.



(a)

(b)

Figure 4. (a) 3D Protein structure of PDB-1GM2, edges are chemical bonds between atoms. (b) Recovery of 3D coordinates from SNLSDP with 5$\mathring{A}$-neighbor graph and multiplicative noise at 0.1 level. Red point: estimated position of unknown atom. Green circle: actual position of unknown atom. Blue line: deviation from estimation to the actual position.

After installation, Figure 4 shows the results of the following codes.

```
 >> startup
>> testSNLsolver

number of anchors = 0
number of sensors = 166
box scale = 20.00
radius = 5.00
multiplicative noise, noise factor = 1.00e-01
```

```
-------------------------------------------------------
estimate sensor positions by SDP
-------------------------------------------------------
num of constraints = 2552,
Please wait:
solving SDP by the SDPT3 software package
sdpobj = -3.341e+03, time = 34.2s
RMSD = 7.19e-01
-------------------------------------------------------
refine positions by steepest descent
-------------------------------------------------------
objstart = 4.2408e+02, objend = 2.7245e+02
number of iterations = 689, time = 0.9s
RMSD = 5.33e-01
-------------------------------------------------------
(noise factor)^2 = -20.0dB,
mean square error (MSE) in estimated positions = -5.0dB
-------------------------------------------------------
```

6. Exact Reconstruction and Universal Rigidity

Now we are going to answer the fundamental question, when the SDP relaxation exactly reconstruct the coordinates up to a rigid transformation. We will provide two theories, one from the optimality rank properties of SDP, and the other from a geometric criterion, *universal rigidity*.

Recall that for a standard SDP with $X, C \in \mathbb{R}^{n\times n}$

$$\begin{aligned} \min \quad & C \bullet X = \sum_{i,j} c_{ij} X_{ij} \\ s.t. \quad & A_i \bullet X = b_i, \quad \text{for } i = 1, \cdots, m \\ & X \succeq 0 \end{aligned} \tag{62}$$

whose SDD is

$$\begin{aligned} \max \quad & -b^T y \\ s.t. \quad & S = C - \sum_{i=1}^m A_i y_i \succeq 0. \end{aligned} \tag{63}$$

Such SDP has the following rank properties [Ali95]:

A. maximal rank solutions X^* or S^* exist;
B. minimal rank solutions X^* or S^* exist;
C. if complementary condition $X^*S^* = 0$ holds, then $\mathrm{rank}(X^*)+\mathrm{rank}(S^*) \leq n$ with equality holds iff strictly complementary condition holds, whence $\mathrm{rank}(S^*) \geq n-k \Rightarrow \mathrm{rank}(X^*) \leq k$.

Strong duality of SDP tells us that an interior point feasible solution in primal or dual problem will ensure the complementary condition and the zero duality gap. Now we assume that $d_{ij} = \|x_i - x_j\|$ precisely for some unknown $x_i \in \mathbb{R}^k$. Then the primal problem is feasible with $Z = (I_d; Y)^T(I_d; Y)$. Therefore the complementary condition holds and the duality gap is zero. In this case, assume that Z^* is a primal feasible solution of SDP embedding and S^* is an optimal dual solution, then

1. $\text{rank}(Z^*) + \text{rank}(S^*) \leq k + n$ and $\text{rank}(Z^*) \geq k$, whence $\text{rank}(S^*) \leq n$;
2. $\text{rank}(Z^*) = k \Longleftrightarrow X = Y^T Y$.

It follows that if an optimal dual S^* has rank n, then every primal solution Z^* has rank k, which ensures $X = Y^T Y$. Therefore it suffices to find a maximal rank dual solution S^* whose rank is n.

Above we have optimality rank condition from SDP. Now we introduce a geometric criterion based on universal rigidity.

Definition (Universal Rigidity (UR) or Unique Localization (UL))**.** $\exists ! y_i \in \mathbb{R}^k \hookrightarrow \mathbb{R}^l$ where $l \geq k$ s.t. $d_{ij}^2 = \|y_i - y_j\|^2, \widehat{d_{ij}}^2 = \|a_k - y_j\|^2$.

It simply says that there is no nontrivial extension of $y_i \in \mathbb{R}^k$ in $\mathbb{R}^l$ satisfying $d_{ij}^2 = \|y_i - y_j\|^2$ and $\widehat{d_{ij}}^2 = \|(a_k; 0) - y_j\|^2$. The following is a short history about universal rigidity.

[Schoenberg 1938] G is complete $\Longrightarrow$ UR

[So-Ye 2007] G is incomplete $\Longrightarrow$ UR $\Longleftrightarrow$ SDP has maximal rank solution $\text{rank}(Z^*) = k$.

Theorem 6.1. [SY07] The following statements are equivalent.

1. The graph is universally rigid or has a unique localization in $\mathbb{R}^k$.
2. The max-rank feasible solution of the SDP relaxation has rank k;
3. The solution matrix has $X = Y^T Y$ or $\text{trace}(X - Y^T Y) = 0$.

Moreover, the localization of a UR instance can be computed approximately in a time polynomial in n, k, and the accuracy $\log(1/\epsilon)$.

In fact, the max-rank solution of SDP embedding is unique. There are many open problems in characterizing UR conditions, see Ye's survey at ICCM'2010.

In practice, we often meet problems with noisy measurements $\alpha d_{ij}^2 \geq \tilde{d}_{ij}^2 \leq \beta d_{ij}^2$. If we relax the constraint $\|y_i - y_j\|^2 = d_{ij}^2$ or equivalently $A_i \bullet X = b_i$ to inequalities, however we can achieve arbitrary small rank solution. To see this, assume that

$$A_i X = b_i \quad \mapsto \quad \alpha b_i \leq A_i X \leq \beta b_i \quad i = 1, \ldots, m, \text{where} \quad \beta \geq 1, \alpha \in (0,1)$$

then So, Ye, and Zhang (2008) [SYZ08] show the following result.

Theorem 6.2. For every $d \geq 1$, there is a SDP solution $\widehat{X} \succeq 0$ with rank $\text{rank}(\widehat{X}) \leq d$, if the following holds,

$$\beta = \begin{cases} 1 + \dfrac{18 \ln 2m}{d} & 1 \leq d \leq 18 \ln 2m \\ 1 + \dfrac{\sqrt{18 \ln 2m}}{d} & d \geq 18 \ln 2m \end{cases}$$

$$\alpha = \begin{cases} \dfrac{1}{e(2m)^{2/d}} & 1 \leq d \leq 4 \ln 2m \\ \max\left\{ \dfrac{1}{e(2m)^{2/d}}, 1 - \sqrt{\dfrac{4 \ln 2m}{d}} \right\} & d \geq 4 \ln 2m \end{cases}$$

Note that α, β are independent to n.

7. Maximal Variance Unfolding

Here we give a special case of SDP embedding, Maximal Variance Unfolding (MVU) [WS06]. In this case we choose graph $G = (V, E)$ as k-nearest neighbor graph. As a contrast to the SDP embedding above, we did not pursue a semi-definite relaxation $X \succeq Y^T Y$, but instead define it as a positive semi-definite kernel $K = Y^T Y$ and maximize the trace of K.

Consider a set of points x_i $(i = 1, \ldots, n)$ whose pairwise distance d_{ij} is known if x_j lies in k-nearest neighbors of x_i. In other words, consider a k-nearest neighbor graph $G = (V, E)$ with $V = \{x_i : i = 1, \ldots, n\}$ and $(i, j) \in E$ if j is a member of k-nearest neighbors of i.

Our purpose is to find coordinates $y_i \in \mathbb{R}^k$ for $i = 1, 2, \ldots, n$ s.t.

$$d_{ij}^2 = \|y_i - y_j\|^2$$

wherever $(i, j) \in E$ and $\sum_i y_i = 0$.

Set $K_{ij} = \langle y_i, y_j \rangle$. Then K is symmetric and positive semidefinite, which satisfies

$$K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2.$$

There are possibly many solutions for such K, and we look for a particular one with maximal trace which characterizes the maximal variance.

$$\begin{aligned} \max \quad & \mathrm{trace}(K) = \sum_{i=1}^{n} \lambda_i(K) \\ s.t. \quad & K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2, \\ & \sum_j K_{ij} = 0, \\ & K \succeq 0 \end{aligned} \tag{64}$$

Again it is a SDP. The final embedding is obtained by using eigenvector decomposition of $K = Y^T Y$.

However we note here that maximization of trace is not a provably good approach to "unfold" a manifold. Sometimes, there are better ways than MVU, *e.g.* if original data lie on a plane then maximization of the diagonal distance between two neighboring triangles will unfold and force it to be a plane. This is a special case of the general $k+1$-lateration graphs [SY07]. From here we see that there are other linear objective functions better than trace for the purpose of "unfolding" a manifold.

CHAPTER 5

Nonlinear Dimensionality Reduction

1. Introduction

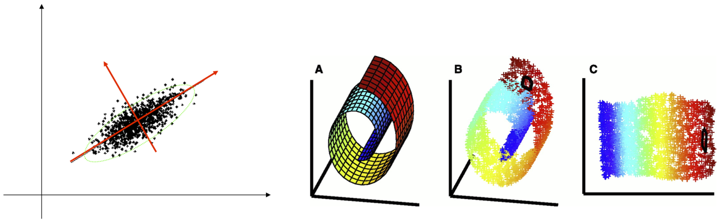
In the past month we talked about two topics: one is the sample mean and sample covariance matrix (PCA) in high dimensional spaces. We have learned that when dimension p is large and sample size n is relatively small, in contrast to the traditional statistics where p is fixed and $n \to \infty$, both 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.

All the work introduced in this chapter can be regarded as generalized PCA/MDS on nearest neighbor graphs, which has roots in manifold learning concept. Two pieces of milestone works, ISOMAP [TdSL00] and Locally Linear Embedding (LLE) [RL00], are firstly published in *science* 2000, which opens a new field called nonlinear dimensionality reduction, or manifold learning in high dimensional data analysis. Here is the development of manifold learning method:

$$\text{PCA} \longrightarrow \text{LLE} \longrightarrow \left\{ \begin{array}{c} \text{Laplacian Eigen Map} \\ \text{Diffusion Map} \\ \text{Hessian LLE} \\ \text{Local Tangent Space Alignment} \end{array} \right.$$

$$\text{MDS} \longrightarrow \text{ISOMAP}$$

To understand the motivation of such a novel methodology, let's take a brief review on PCA/MDS. Given a set of data $x_i \in \mathbb{R}^p$ $(i = 1, \ldots, n)$ or merely pairwise distances $d(x_i, x_j)$, PCA/MDS essentially looks for an affine space which best capture the variation of data distribution, see Figure 1(a). However, this scheme will not work in the scenario that data are actually distributed on a highly nonlinear curved surface, i.e. *manifolds*, see the example of Swiss Roll in Figure 1(b). Can we extend PCA/MDS in certain sense to capture intrinsic coordinate systems which charts the manifold?



(a) (b)

FIGURE 1. (a) Find an affine space to approximate data variation in PCA/MDS. (b) Swiss Roll data distributed on a nonlinear 2-D submanifold in Euclidean space $\mathbb{R}^3$. Our purpose is to capture an intrinsic coordinate system describing the submanifold.

ISOMAP and LLE, as extensions from MDS and local PCA, respectively, leads to a series of attempts to address this problem.

All the current techniques in manifold learning, as extensions of PCA and MDS, are often called as *Spectral Kernel Embedding*. The common theme of these techniques can be described in Figure 2. The basic problem is: given a set of data points $\{x_1, x_2, ..., x_n \in \mathbb{R}^p\}$, how to find out $y_1, y_2, ..., y_n \in \mathbb{R}^d$, where $d \ll p$, such that some geometric structures (local or global) among data points are best preserved.

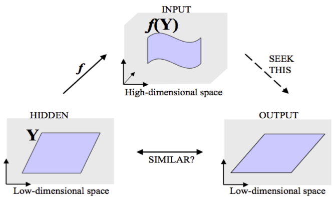


FIGURE 2. The generative model for manifold learning. Y is the hidden parameter space (like rotation angle of faces below), f is a measure process which maps Y into a sub-manifold in a high dimensional ambient space, $X = f(Y) \subset \mathbb{R}^p$. All of our purpose is to recover this hidden parameter space Y given samples $\{x_i \in \mathbb{R}^p : i = 1, \ldots, n\}$.

All the manifold learning techniques can be summarized in the following meta-algorithm, which explains precisely the name of *spectral kernel embedding*. All the methods can be called certain *eigenmaps* associated with some positive semi-definite kernels.

1. Construct a data graph $G = (V, E)$, where $V = \{x_i : i = 1, ..., n\}$.

e.g.1. ε-neighborhood, $i \sim j \Leftrightarrow d(x_i, x_j) \leqslant \varepsilon$, which leads to an undirected graph;

e.g.2. k-nearest neighbor, $(i,j) \in E \Leftrightarrow j \in \mathcal{N}_k(i)$, which leads to a directed graph.

2. Construct a positive semi-definite matrix K (kernel).

3. Eigen-decomposition $K = U\Lambda U^T$, then $Y_d = U_d \Lambda_d^{\frac{1}{2}}$, where choose d eigenvectors (top or bottom) U_d.

Example 3 (PCA)**.** G is complete, $K = \hat{\Sigma}_n$ is a covariance matrix.

Example 4 (MDS)**.** G is complete, $K = -\frac{1}{2}HDH^T$, where $D_{ij} = d^2(x_i, x_j)$.

Example 5 (ISOMAP)**.** G is incomplete.

$$D_{ij} = \begin{cases} d(x_i, x_j) & \text{if } (i,j) \in E, \\ \hat{d}_g(x_i, x_j) & \text{if } (i,j) \notin E. \end{cases}$$

where $\hat{d}_g$ is a graph shorted path. Then

$$K = -\frac{1}{2}HDH^T.$$

Note that K is positive semi-definite if and only if D is a squared distance matrix.

Example 6 (LLE)**.** G is incomplete. $K = (I - W)^T(I - W)$, where

$$W_{ij}^{n\times n} = \begin{cases} w_{ij} & j \in \mathcal{N}(i), \\ 0 & \text{other's}. \end{cases}$$

and w_{ij} solves the following optimization problem

$$\min_{\sum_j w_{ij}=1} \|X_i - \sum_{j\in\mathcal{N}(i)} w_{ij}\bar{X}_j\|^2, \quad \bar{X}_j = X_j - X_i.$$

After obtaining W, compute the global embedding d-by-n embedding matrix $Y = [Y_1, \ldots, Y_n]$,

$$\min_Y \sum_{i=1}^n \|Y_i - \sum_{j=1}^n W_{ij}Y_j\|^2 = \text{trace}((I - W)Y^TY(I - W)^T).$$

This is equivalent to find smallest eigenvectors of $K = (I - W)^T(I - W)$.

2. ISOMAP

ISOMAP is an extension of MDS, where pairwise euclidean distances between data points are replaced by geodesic distances, computed by *graph shortest path distances.*

(1) Construct a neighborhood graph $G = (V, E, d_{ij})$ such that
$V = \{x_i : i = 1, \ldots, n\}$
$E = \{(i,j) : \text{ if } j \text{ is a neighbor of } i, \text{ i.e. } j \in \mathcal{N}_i\}$, e.g. k-nearest neighbors, ϵ-neighbors
$d_{ij} = d(x_i, x_j)$, e.g. Euclidean distance when $x_i \in \mathbb{R}^p$

(2) Compute graph shortest path distances
$d_{ij} = \min_{P=(x_i,\ldots,x_j)}(\|x_i - x_{t_1}\| + \ldots + \|x_{t_{k-1}} - x_j\|)$, is the length of a graph shortest path connecting i and j
Dijkstra's algorithm ($O(kn^2\log n)$) and Floyd's Algorithm ($O(n^3)$)

(3) classical MDS with $D = (d_{ij}^2)$

construct a symmetric (positive semi-definite if D is a squared distance) $B = -0.5HDH^T$ where $H = I - \mathbf{1}\mathbf{1}^T/n$ (or $H = I - \mathbf{1}a^T$ for any $a^T\mathbf{1} = 1$).

Find eigenvector decomposition of $B = U\Lambda U^T$ and choose top d eigenvectors as embedding coordinates in $\mathbb{R}^d$, i.e. $Y_d = [y_1, \ldots, y_d] = [U_1, \ldots, U_d]\Lambda_d^{1/2} \in \mathbb{R}^{n\times d}$

Algorithm 2: ISOMAP Algorithm

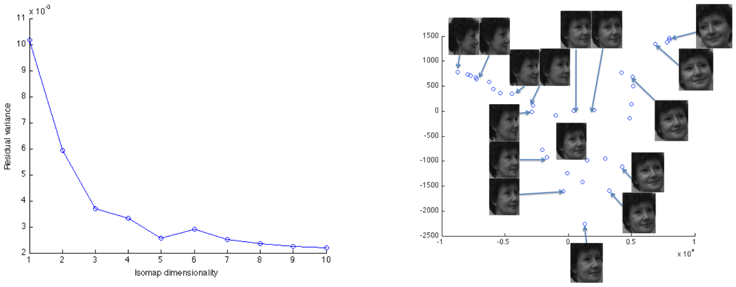
```
Input: A weighted undirected graph $G = (V, E, d)$ such that
$V = \{x_i : i = 1, \ldots, n\}$
$E = \{(i, j) :$ if $j$ is a neighbor of $i$, i.e. $j \in \mathcal{N}_i\}$, e.g. $k$-nearest neighbors, $\epsilon$-neighbors
$d_{ij} = d(x_i, x_j)$, e.g. Euclidean distance when $x_i \in \mathbb{R}^p$
Output: Euclidean $k$-dimensional coordinates $Y = [y_i] \in \mathbb{R}^{k\times n}$ of data.
Step 1: Compute graph shortest path distances
    $$d_{ij} = \min_{P=(x_i,\ldots,x_j)} (\|x_i - x_{t_1}\| + \ldots + \|x_{t_{k-1}} - x_j\|),$$
    which is the length of a graph shortest path connecting $i$ and $j$;
Step 2: Compute $K = -\frac{1}{2} H \cdot D \cdot H^T$ $(D := [d_{ij}^2])$, where H is the Householder centering matrix;
Step 3: Compute Eigenvalue decomposition $K = U\Lambda U^T$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;
Step 4: Choose top $k$ nonzero eigenvalues and corresponding eigenvectors, $\widetilde{X}_k = U_k \Lambda_k{}^{\frac{1}{2}}$ where
    $$U_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n,$$
    $$\Lambda_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$
    with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$.
```

The basic feature of ISOMAP can be described as: we find a low dimensional embedding of data such that points nearby are mapped nearby and points far away are mapped far away. In other words, we have global control on the data distance and the method is thus a *global* method. The major shortcoming of ISOMAP lies in its computational complexity, characterized by a full matrix eigenvector decomposition.

2.1. ISOMAP Example. Now we give an example of ISOMAP with matlab codes.

```
 % load 33-face data
load ../data/face.mat Y
X = reshape(Y,[size(Y,1)*size(Y,2) size(Y,3)]);
p = size(X,1);
n = size(X,2);
D = pdist(X');
DD = squareform(D);

% ISOMAP embedding with 5-nearest neighbors
[Y_iso,R_iso,E_iso]=isomapII(DD,'k',5);
```



(a) (b)

Figure 3. (a) Residual Variance plot for ISOMAP. (b) 2-D ISOMAP embedding, where the first coordinate follows the order of rotation angles of the face.

```
% Scatter plot of top 2-D embeddings
y=Y_iso.coords{2};
scatter(y(1,:),y(2,:))
```

2.2. Convergence of ISOMAP. Under dense-sample and regularity conditions on manifolds, ISOMAP is proved to show convergence to preserve geodesic distances on manifolds. The key is to approximate geodesic distance on manifold by a sequence of short Euclidean distance hops.

Consider arbitrary two points on manifold $x, y \in M$. Define

$$d_M(x,y) = \inf_{\gamma}\{\text{length}(\gamma)\}$$

$$d_G(x,y) = \min_{P}(\|x_0 - x_1\| + \ldots + \|x_{t-1} - x_t\|)$$

$$d_S(x,y) = \min_{P}(d_M(x_0,x_1) + \ldots + d_M(x_{t-1},x_t))$$

where γ varies over the set of smooth arcs connecting x to y in M and P varies over all paths along the edges of G starting at $x_0 = x$ and ending at $x_t = y$. We are going to show $d_M \approx d_G$ with the bridge d_S.

It is easy to see the following upper bounds by d_S:

$$d_M(x,y) \le d_S(x,y) \tag{65}$$

$$d_G(x,y) \le d_S(x,y) \tag{66}$$

where the first upper bound is due to triangle inequality for the metric d_M and the second upper bound is due to that Euclidean distances $\|x_i - x_{i+1}\|$ are smaller than arc-length $d_M(x_i, x_{i+1})$.

To see other directions, one has to impose additional conditions on sample density and regularity of manifolds.

Lemma 2.1 (Sufficient Sampling)**.** Let $G = (V,E)$ where $V = \{x_i : i = 1,\ldots,n\} \subseteq M$ is a ϵ-net of manifold M, *i.e.*for every $x \in M$ there exists $x_i \in V$ such that

$d_M(x, x_i) < \epsilon$, and $\{i, j\} \in E$ if $d_M(x_i, x_j) \le \alpha\epsilon$ ($\alpha \ge 4$). Then for any pair $x, y \in V$,

$$d_S(x, y) \le \max(\alpha - 1, \frac{\alpha}{\alpha - 2}) d_M(x, y).$$

Proof. Let γ be a shortest path connecting x and y on M whose length is l. If $l \le (\alpha - 2)\epsilon$, then there is an edge connecting x and y whence $d_S(x, y) = d_M(x, y)$. Otherwise split γ into pieces such that $l = l_0 + tl_1$ where $l_1 = (\alpha - 2)\epsilon$ and $\epsilon \le l_0 < (\alpha - 2)\epsilon$. This divides arc γ into a sequence of points $\gamma_0 = x$, $\gamma_1, \ldots,$ $\gamma_{t+1} = y$ such that $d_M(x, \gamma_1) = l_0$ and $d_M(\gamma_i, \gamma_{i+1}) = l_1$ ($i \ge 1$). There exists a sequence of $x_0 = x, x_1, \ldots, x_{t+1} = y$ such that $d_M(x_i, \gamma_i) \le \epsilon$ and

$$\begin{aligned} d_M(x_i, x_{i+1}) &\le d_M(x_i, \gamma_i) + d_M(\gamma_i, \gamma_{i+1}) + d_M(\gamma_{i+1}, x_{i+1}) \\ &\le \epsilon + l_1 + \epsilon \\ &= \alpha\epsilon \\ &= l_1\alpha/(\alpha - 2) \end{aligned}$$

whence $(x_i, x_{i+1}) \in E$. Similarly $d_M(x, x_1) \le d_M(x, \gamma_1) + d_M(\gamma_1, x_1) \le (\alpha - 1)\epsilon \le l_0(\alpha - 1)$.

$$\begin{aligned} d_S(x, y) &\le \sum_{i=0}^{t-1} d_M(x_i, x_{i+1}) \\ &\le l \max\left(\frac{\alpha}{\alpha - 2}, \alpha - 1\right) \end{aligned}$$

Setting $\alpha = 4$ gives rise to $d_S(x, y) \le 3 d_M(x, y)$. □

The other lower bound $d_S(x, y) \le c d_G(x, y)$ requires that for every two points x_i and x_j, Euclidean distance $\|x_i - x_j\| \le c d_M(x_i, x_j)$. This imposes a regularity on manifold M, whose curvature has to be bounded. We omit this part here and leave the interested readers to the reference by Bernstein, de Silva, Langford, and Tenenbaum 2000, as a supporting information to the ISOMAP paper.

3. Locally Linear Embedding (LLE)

In applications points nearby should be mapped nearby, while points far away should impose no constraint. This is because in applications when points are close enough, they are similar, while points are far, there is no faithful information to measure how far they are. This motivates another type of algorithm, locally linear embedding. This is a *local* method as it involves local PCA and sparse eigenvector decomposition.

(1) Construct a neighborhood graph $G = (V, E, W)$ such that

$V = \{x_i : i = 1, \ldots, n\}$

$E = \{(i, j) :$ if j is a neighbor of i, i.e. $j \in \mathcal{N}_i\}$, e.g. k-nearest neighbors, ϵ-neighbors

$W_{ij} = d(x_i, x_j)$ in Euclidean distance

(2) Local fitting:

Pick up a point x_i and its neighbors $\mathbb{N}_i$

Compute the local fitting weights

$$\min_{\sum_{j \in \mathbb{N}_i} w_{ij} = 1} \|x_i - \sum_{j \in \mathbb{N}_i} w_{ij}(x_j - x_i)\|^2.$$

This can be done by Lagrange multiplier method, *i.e.* solving

$$\min_{w_{ij}} \frac{1}{2}\|x_i - \sum_{j\in\mathbb{N}_i} w_{ij}(x_j - x_i)\|^2 + \lambda(1 - \sum_{j\in\mathbb{N}_i} w_{ij}).$$

Let $w_i = [w_{ij_1}, \ldots w_{ij_k}]^T \in \mathbb{R}^k$, $\bar{X}_i = [x_{j_1} - x_i, \ldots, x_{j_k} - x_i]$, and the local Gram (covariance) matrix $C^{(i)}_{jk} = \langle x_j - x_i, x_k - x_i\rangle$, whence the weights are

$$w_i = C_i^{\dagger}(\bar{X}_i^T x_i + \lambda \mathbf{1}),$$

where the Lagrange multiplier equals to

$$\lambda = \frac{1}{\mathbf{1}^T C_i^{\dagger} \mathbf{1}} \left(1 - \mathbf{1}^T C_i^{\dagger} \bar{X}_i^T x_i\right),$$

and $C_i^{\dagger}$ is a Moore-Penrose (pseudo) inverse of C_i. Note that C_i is often ill-conditioned and to find its Moore-Penrose inverse one can use regularization method $(C_i + \mu I)^{-1}$ for some $\mu > 0$.

(3) Global alignment

Define a n-by-n weight matrix W:

$$W_{ij} = \begin{cases} w_{ij}, & j \in \mathcal{N}_i \\ 0, & otherwise \end{cases}$$

Compute the global embedding d-by-n embedding matrix Y,

$$\min_Y \sum_i \|y_i - \sum_{j=1}^n W_{ij} y_j\|^2 = \mathrm{trace}(Y(I-W)^T(I-W)Y^T)$$

In other words, construct a positive semi-definite matrix $B = (I - W)^T(I-W)$ and find $d+1$ smallest eigenvectors of B, $v_0, v_1, \ldots, v_d$ associated smallest eigenvalues $\lambda_0, \ldots, \lambda_d$. Drop the smallest eigenvector which is the constant vector explaining the degree of freedom as translation and set $Y = [v_1/\sqrt{(\lambda_1)}, \ldots, v_d/\sqrt{\lambda_d}]^T$.

The benefits of LLE are:

- Neighbor graph: k-nearest neighbors is of $O(kn)$
- W is sparse: $kn/n^2 = k/n$ non-zeroes
- $B = (I-W)^T(I-W)$ is guaranteed to be positive semi-definite

However, unlike ISOMAP, it is not clear if LLE constructed above converges under certain conditions. This has to be left to some variations of basic LLE above, Hessian LLE and LTSA to finish the convergence conditions.

TABLE 1. Comparisons between ISOMAP and LLE.

ISOMAP	LLE
MDS on geodesic distance matrix	local PCA + eigen-decomposition
global approach	local approach
no for nonconvex manifolds with holes	ok with nonconvex manifolds with holes
landmark (Nystrom)	Hessian
Extensions: conformal	Extensions: Laplacian
isometric, etc.	LTSA etc.

4. Laplacian LLE (Eigenmap)

Consider the graph Laplacian with heat kernels [BN01, BN03]. Define a weight matrix $W = (w_{ij}) \in \mathbb{R}^{n\times n}$ by

$$w_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|^2}{t}} & j \in \mathcal{N}(i), \\ 0 & \text{otherwise.} \end{cases}$$

Let $D = \mathrm{diag}(\sum_{j\in\mathcal{N}_i} w_{ij})$ be the diagonal matrix with weighted degree as diagonal elements.

Define the *unnormalized graph Laplacian* by

$$L = D - W,$$

and the *normalized graph Laplacian* by

$$\mathfrak{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}.$$

Note that eigenvectors of $\mathfrak{L}$ are also *generalized eigenvectors* of L up to a scaling matrix. This can be seen in the following reasoning.

$$\mathfrak{L}\phi = \lambda\phi$$

$$\Leftrightarrow D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}\phi = \lambda\phi$$

$$\Leftrightarrow Lv = (D - W)v = \lambda Dv, \quad v = D^{-\frac{1}{2}}\phi$$

Generalized eigenvectors v of L are also right eigenvectors of row Markov matrix $P = D^{-1}W$. ($\because$ $Pv = \lambda v \Leftrightarrow D^{-1}Wv = \lambda v \Leftrightarrow (I - D^{-1}W)v = (1-\lambda)v$ $\therefore$ $(D - W)v = (1-\lambda)Dv$).

Depending on the meaning of eigenvectors above, we can always choose bottom $d + 1$ eigenvectors, and dropped the smallest eigenvector (the constant vector associated with eigenvalue 0) and use the remaining d vectors to construct a d dimensional embedding of data.

4.1. Convergence of Laplacian Eigenmap.

Why choose Laplacian? Consider a linear chain graph,

$$(df)(i) = f_{i+1} - f_i = [(z-1)f](i)$$

$$d^2 f = (z-1)^2 f = (z^2 - 2z + 1)f \to f_{i+1} - 2f_i + f_{i-1}$$

On graphs, $d^2 f = (D - W)f = Lf$

$$f^T L f = \sum_{i\geq j} w_{ij}(f_i - f_j)^2 \geq 0 \sim \int \|\nabla_M f\|^2 = \int (\mathrm{trace}(f^T \mathcal{H} f))^2$$

where $\mathcal{H} = [\partial^2/\partial_i\partial_j] \in \mathbb{R}^{d\times d}$ is the Hessian matrix.

Some rigorous results about convergence of Laplacian eigenmaps are given in [BN08]. Assume that $\mathcal{M}$ is a compact manifold with $\mathrm{vol}(\mathcal{M}) = 1$. Let the Laplacian-Beltrami operator

$$\begin{aligned} \Delta_{\mathcal{M}} \quad &: C(\mathcal{M}) &&\to L^2(\mathcal{M}) \\ &\quad f &&\mapsto -\div(\nabla f) \end{aligned}$$

Consider the following operator

$$\begin{aligned} \hat{L}_{t,n} \quad &: C(\mathcal{M}) \quad \to C(\mathcal{M}) \\ f \qquad &\mapsto \frac{1}{t(4\pi t)^{k/2}} \left(\sum_i e^{-\frac{\|y-x_i\|}{4t}} f(y) - \sum_i e^{\frac{\|y-x_i\|^2}{4t}} f(x_i) \right) \end{aligned}$$

where $(\hat{L}_{t,n} f)(y)$ is a function on $\mathcal{M}$, and

$$\begin{aligned} L_t \quad &: L^2(\mathcal{M}) \quad \to L^2(\mathcal{M}) \\ f \qquad &\mapsto \frac{1}{t(4\pi t)^{k/2}} \left(\int_{\mathcal{M}} e^{-\frac{\|y-x\|}{4t}} f(y) dx - \int_{\mathcal{M}} e^{\frac{\|y-x\|^2}{4t}} f(x) dx \right). \end{aligned}$$

Then [**BN08**] shows that when those operators have no repeated eigenvalues, the spectrum of $\hat{L}_{t,n}$ converges to L_t as $n \to \infty$ (variance), where the latter converges to that of $\Delta_{\mathcal{M}}$ with a suitable choice of $t \to \infty$ (bias). The following gives a summary.

Theorem 4.1 (Belkin-Niyogi)**.** Assume that all the eigenvalues in consideration are of multiplicity one. For small enough t, let $\hat{\lambda}^t_{n,i}$ be the i-th eigenvalue of $\hat{L}_{t,n}$ and $\hat{v}^t_{n,i}$ be the corresponding eigenfunction. Let λ_i and v_i be the corresponding eigenvalue and eigenfunction of $\Delta_{\mathcal{M}}$. Then there exists a sequence $t_n \to 0$ such that

$$\lim_{n\to\infty} \hat{\lambda}^{t_n}_{n,i} = \lambda_i$$

$$\lim_{n\to\infty} \|\hat{v}^{t_n}_{n,i} - v_i\| = 0$$

where the limits are taken in probability.

From above one can see that Laplacian LLE minimizes trace of Hessian. Is that what you desire? Why not the original Hessian?

5. Hessian LLE

Laplacian Eigenmap looks for coordinate curves

$$\min \int \|\nabla_{\mathcal{M}} f\|^2, \quad \|f\| = 1$$

while Hessian Eigenmap looks for

$$\min \int \|\mathcal{H} f\|^2, \quad \|f\| = 1$$

Donoho and Grimes (2003) [**DG03b**] replaces the graph Laplacian, or the trace of Hessian matrix, by the whole Hessian. This is because the kernel of Hessian,

$$\left\{ f(y_1, \ldots, y_d) : \frac{\partial^2 f}{\partial y_i \partial y_j} = 0 \right\}$$

must be constant function or linear functions in y_i ($i = 1, \ldots, d$). Therefore this kernel space is a linear subspace of dimension $d+1$. Minimizing Hessian will exactly leads to a basis with constant function and d independent coordinate functions.

1. G is incomplete, often k-nearest neighbor graph.
2. Local SVD on neighborhood of x_i, for $x_{i_j} \in \mathcal{N}(x_i)$,

$$\tilde{X}^{(i)} = [x_{i_1} - \mu_i, ..., x_{i_k} - \mu_i]^{p\times k} = \tilde{U}^{(i)} \tilde{\Sigma} (\tilde{V}^{(i)})^T,$$

where $\mu_i = \sum_{j=1}^{k} x_{i_j} = \frac{1}{k} X_i \mathbf{1}$, $\tilde{U}^{(i)} = [\tilde{U}_1^{(i)}, ..., \tilde{U}_k^{(i)}]$ is an approximate tangent space at x_i.

3. Hessian estimation, assumed d-dimension: define

$$M = [1, \tilde{V_1}, ..., \tilde{V_k}, \tilde{V_1}\tilde{V_2}, ..., \tilde{V}_{d-1}\tilde{V_d}] \in \mathbb{R}^{k\times(1+d+\binom{d}{2})}$$

where $\tilde{V_i}\tilde{V_j} = [\tilde{V}_{ik}\tilde{V}_{jk}]^T \in \mathbb{R}^k$ denotes the elementwise product (Hadamard product) between vector $\tilde{V_i}$ and $\tilde{V_j}$.

Now we perform a Gram-Schmidt Orthogonalization procedure on M, get

$$\tilde{M} = [1, \hat{v}_1, ..., \hat{v}_k, \hat{w}_1, \hat{w}_2, ..., \hat{w}_{\binom{d}{2}-1}] \in \mathbb{R}^{k\times(1+d+\binom{d}{2})}$$

Define Hessian by

$$[H^{(i)}]^T = [last \quad \binom{d}{2} \quad columns \quad of \quad \tilde{M}]_{k\times\binom{d}{2}}$$

as the first $d+1$ columns of $\tilde{M}$ consists an orthonormal basis for the kernel of Hessian.

Define a selection matrix $S^{(i)} \in \mathbb{R}^{n\times k}$ which selects those data in $\mathcal{N}(x_i)$, *i.e.*

$$[x_1, .., x_n]S^{(i)} = [x_{i_1}, ..., x_{i_k}]$$

Then the kernel matrix is defined to be

$$K = \sum_{i=1}^{n} S^{(i)} H^{(i)T} H^{(i)} S^{(i)T} \in \mathbb{R}^{n\times n}$$

Find smallest $d+1$ eigenvectors of K and drop the smallest eigenvector, the remaining d eigenvectors will give rise to a d dimensional embedding of data points.

5.1. Convergence of Hessian LLE. There are two assumptions for the convergence of ISOMAP:

- Isometry: the geodesic distance between two points on manifolds equals to the Euclidean distances between intrinsic parameters.
- Convexity: the parameter space is a convex subset in $\mathbb{R}^d$.

Therefore, if the manifold contains a hole, ISOMAP will not faithfully recover the intrinsic coordinates. Hessian LLE above is provable to find local orthogonal coordinates for manifold reconstruction, even in nonconvex case. Figure [**?**] gives an example.

Donoho and Grimes [DG03b] relaxes the conditions above into the following ones.

- Local Isometry: in a small enough neighborhood of each point, geodesic distances between two points on manifolds are identical to Euclidean distances between parameter points.
- Connecteness: the parameter space is an open connected subset in $\mathbb{R}^d$.

Based on the relaxed conditions above, they prove the following result.

Theorem 5.1. Supper $\mathcal{M} = \psi(\Theta)$ where Θ is an open connected subset of $\mathbb{R}^d$, and ψ is a locally isometric embedding of Θ into $\mathbb{R}^n$. Then the Hessian $\mathcal{H}(f)$ has a $d+1$ dimensional nullspace, consisting of the constant function and d-dimensional space of functions spanned by the original isometric coordinates.

Algorithm 3: Hessian LLE Algorithm

```
Input: A weighted undirected graph $G = (V, E, d)$ such that
$V = \{x_i \in \mathbb{R}^p : i = 1, \ldots, n\}$
$E = \{(i, j) :$ if $j$ is a neighbor of $i$, i.e. $j \in \mathcal{N}_i\}$, e.g. $k$-nearest neighbors
Output: Euclidean $k$-dimensional coordinates $Y = [y_i] \in \mathbb{R}^{k\times n}$ of data.
Step 1: Compute local PCA on neighborhood of $x_i$, for,
    $$\tilde{X}^{(i)} = [x_{i_1} - \mu_i, ..., x_{i_k} - \mu_i]^{p\times k} = \tilde{U}^{(i)}\tilde{\Sigma}(\tilde{V}^{(i)})^T, \quad x_{i_j} \in \mathcal{N}(x_i),$$
    where $\mu_i = \sum_{j=1}^{k} x_{i_j} = \frac{1}{k}X_i\mathbf{1}$, $\tilde{U}^{(i)} = [\tilde{U}_1^{(i)}, ..., \tilde{U}_k^{(i)}]$ is an approximate tangent
    space at $x_i$;
Step 2: Hessian estimation, assumed $d$-dimension: define
    $$M = [1, \tilde{V}_1, ..., \tilde{V}_k, \tilde{V}_1\tilde{V}_2, ..., \tilde{V}_{d-1}\tilde{V}_d] \in \mathbb{R}^{k\times(1+d+\binom{d}{2})}$$
    where $\tilde{V}_i\tilde{V}_j = [\tilde{V}_{ik}\tilde{V}_{jk}]^T \in \mathbb{R}^k$ denotes the elementwise product (Hadamard
    product) between vector $\tilde{V}_i$ and $\tilde{V}_j$. Now we perform a Gram-Schmidt
    Orthogonalization procedure on $M$, get
    $$\tilde{M} = [1, \hat{v}_1, ..., \hat{v}_k, \hat{w}_1, \hat{w}_2, ..., \hat{w}_{\binom{d}{2}-1}] \in \mathbb{R}^{k\times(1+d+\binom{d}{2})}$$
    Define Hessian by
    $$[H^{(i)}]^T = [last \quad \binom{d}{2} \quad columns \quad of \quad \tilde{M}]_{k\times\binom{d}{2}}$$
    as the first $d + 1$ columns of $\tilde{M}$ consists an orthonormal basis for the kernel of
    Hessian.
Step 3: Define
    $$K = \sum_{i=1}^{n} S^{(i)}H^{(i)T}H^{(i)}S^{(i)T} \in \mathbb{R}^{n\times n}, \quad [x_1, .., x_n]S^{(i)} = [x_{i_1}, ..., x_{i_k}],$$
    find smallest $d + 1$ eigenvectors of $K$ and drop the smallest eigenvector, the
    remaining $d$ eigenvectors will give rise to a $d$-embedding.
```

Under this theorem, the original isometric coordinates can be recovered, up to a rigid motion, by identifying a suitable basis for the null space of $\mathcal{H}(f)$.

6. Local Tangent Space Alignment (LTSA)

A shortcoming of Hessian LLE is the nonlinear construction of Hessian which requires Hadamard products between tangent vectors. This is prone to noise. Zhenyue Zhang and Hongyuan Zha (2002) [ZZ02] suggest the following procedure which does not involve nonlinear Hessian but still leave an orthogonal basis for tangent space as bottom eigenvectors. In contrast to Hessian LLE's minimization of projections on pairwise products between tangent vectors, LTSA minimizes the projection on the normal space.

LTSA looks for the following coordinates,

$$\min_{Y} \sum_{i\sim j} \|y_i - U_i U_j^T y_j\|^2$$

where U_i is a local PCA basis for tangent space at point $x_i \in \mathbb{R}^p$.

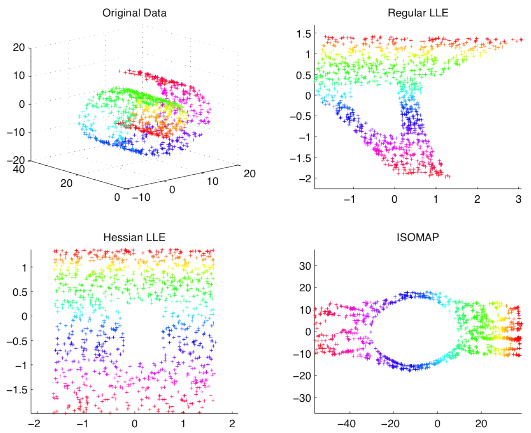


FIGURE 4. Comparisons of Hessian LLE on Swiss roll against ISOMAP and LLE. Hessian better recovers the intrinsic coordinates as the rectangular hole is the least distorted.

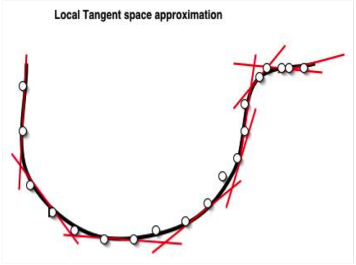


FIGURE 5. Local tangent space approximation.

Note that Connection Laplacian looks for:

$$\min_{Y} \sum_{i\sim j} \|y_i - O_{ij} y_j\|^2, \quad O_{ij} = \arg\min_{O} \|U_i - O_{ij} U_j\|^2$$

where U_i is a local PCA basis for tangent space at point $x_i \in \mathbb{R}^p$.

1. G is incomplete, taken to be k-nn graph here.
2. Local SVD on neighborhood of x_i, $x_{i_j} \in \mathcal{N}(x_i)$,

$$\tilde{X}^{(i)} = [x_{i_1} - \mu_i, ..., x_{i_k} - \mu_i]^{p\times k} = \tilde{U}^{(i)} \tilde{\Sigma} (\tilde{V}^{(i)})^T,$$

where $\mu_i = \sum_{j=1}^{k} x_{i_j} = \frac{1}{k} X_i \mathbf{1}$, $\tilde{U}^{(i)} = [\tilde{U}_1^{(i)}, ..., \tilde{U}_k^{(i)}]$ is an approximate tangent space at x_i. Define

$$G_i = [1/\sqrt{k}, \tilde{V_1}^{(i)}, ..., \tilde{V_d}^{(i)}]^{k\times(d+1)}$$

3. Alignment (kernel) matrix

$$K^{n\times n} = \Phi = \sum_{i=1}^{n} S_i W_i W_i^T S_i^T$$

where weight matrix

$$W_i^{k\times k} = I - G_i G_i^T$$

selection matrix $S_i^{n\times k} : [x_{i_1}, ..., x_{i_k}] = [x_1, ..., x_n] S_i^{n\times k}$

Similarly as above, choose bottom $d+1$ eigenvectors, and dropped smallest which gives embedding matrix $Y^{(n\times d)}$.

As the Hessian LLE, LTSA may recover the global coordinates under certain conditions where [ZZ09] presents some analysis on this.

Algorithm 4: LTSA Algorithm

```
Input: A weighted undirected graph $G = (V, E, d)$ such that
$V = \{x_i \in \mathbb{R}^p : i = 1, \ldots, n\}$
$E = \{(i, j) :$ if $j$ is a neighbor of $i$, i.e. $j \in \mathcal{N}_i\}$, e.g. $k$-nearest neighbors
Output: Euclidean $k$-dimensional coordinates $Y = [y_i] \in \mathbb{R}^{k\times n}$ of data.
Step 1: Compute local PCA on neighborhood of $x_i$, $x_{i_j} \in \mathcal{N}(x_i)$,
    $$\tilde{X}^{(i)} = [x_{i_1} - \mu_i, ..., x_{i_k} - \mu_i]^{p\times k} = \tilde{U}^{(i)} \tilde{\Sigma} (\tilde{V}^{(i)})^T,$$
    where $\mu_i = \sum_{j=1}^k x_{i_j} = \frac{1}{k} X_i \mathbf{1}$, $\tilde{U}^{(i)} = [\tilde{U}_1^{(i)}, ..., \tilde{U}_k^{(i)}]$ is an approximate tangent
    space at $x_i$. Define
    $$G_i = [1/\sqrt{k}, \tilde{V_1}^{(i)}, ..., \tilde{V_d}^{(i)}]^{k\times(d+1)};$$
Step 2: Alignment (kernel) matrix
    $$K^{n\times n} = \sum_{i=1}^n S_i W_i W_i^T S_i^T, \quad W_i^{k\times k} = I - G_i G_i^T,$$
    where selection matrix $S_i^{n\times k} : [x_{i_1}, ..., x_{i_k}] = [x_1, ..., x_n] S_i^{n\times k}$;
Step 3: Find smallest $d+1$ eigenvectors of $K$ and drop the smallest eigenvector,
    the remaining $d$ eigenvectors will give rise to a $d$-embedding.
```

7. Diffusion Map

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

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

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

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

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

and

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

Then

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

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

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 [CLL+05]

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

7.1. General Diffusion Maps and Convergence. In [CLL+05] a general class of diffusion maps are defined which involves a normalized weight matrix,

$$W_{ij}^{\alpha,t} = \frac{W_{ij}}{p_i^\alpha \cdot p_j^\alpha}, \quad p_i := \sum_k \exp\left(-\frac{d(x_i, x_k)^2}{t}\right) \tag{67}$$

where $\alpha = 0$ recovers the definition above. With this family, one can define $D_\alpha = \mathrm{diag}(\sum_j W_{ij}^{\alpha,t})$ and the row Markov matrix

$$P_{\alpha,t,n} = D_\alpha^{-1} W^\alpha, \tag{68}$$

whose right eigenvectors Φ^α lead to a family of diffusion maps parameterized by α.

Such a definition suggests the following integral operators as diffusion operators. Assume that $q(x)$ is a density on $\mathcal{M}$.

- Let $k_t(x, y) = h(\|x - y\|^2/t)$ where h is a radial basis function, *e.g.* $h(z) = \exp(-z)$.
- Define
$$q_t(x) = \int_{\mathcal{M}} k_t(x, y) q(y) dy$$
and form the new kernel
$$k_t^{(\alpha)}(x, y) = \frac{k_t(x, y)}{q_t^\alpha(x) q_t^\alpha(y)}.$$
- Let
$$d_t^{(\alpha)}(x) = \int_{\mathcal{M}} k_t^{(\alpha)}(x, y) q(y) dy$$
and define the transition kernel of a Markov chain by
$$p_{t,\alpha}(x, y) = \frac{k_t^{(\alpha)}(x, y)}{d_t^{(\alpha)}(x)}.$$
Then the Markov chain can be defined as the operator
$$P_{t,\alpha} f(x) = \int_{\mathcal{M}} p_{t,\alpha}(x, y) f(y) q(y) dy.$$
- Define the infinitesimal generator of the Markov chain
$$L_{t,\alpha} = \frac{I - P_{t,\alpha}}{t}.$$

For this, Lafon et al.[CL06] shows the following pointwise convergence results.

Theorem 7.1. Let $\mathcal{M} \in \mathbb{R}^p$ be a compact smooth submanifold, $q(x)$ be a probability density on $\mathcal{M}$, and $\Delta_{\mathcal{M}}$ be the Laplacian-Beltrami operator on $\mathcal{M}$.

$$\lim_{t \to 0} L_{t,\alpha} = \frac{\Delta_{\mathcal{M}}(f q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta_{\mathcal{M}}(q^{1-\alpha}))}{q^{1-\alpha}}. \tag{69}$$

This suggests that

- for $\alpha = 1$, it converges to the Laplacian-Beltrami operator $\lim_{t\to 0} L_{t,1} = \Delta_{\mathcal{M}}$;
- for $\alpha = 1/2$, it converges to a Schrödinger operator whose conjugation leads to a forward Fokker-Planck equation;
- for $\alpha = 0$, it is the normalized graph Laplacian.

A central question in diffusion maps is:

Why we choose right eigenvectors ϕ_i in diffusion map?

To answer this we will introduce the concept of *lumpability* in finite Markov chains on graphs.

8. Connection Laplacian and Vector Diffusion Maps

`to be finished...`

9. Comparisons

According to the comparative studies by Todd Wittman, LTSA has the best overall performance in current manifold learning techniques. Try yourself his code, `mani.m`, and enjoy your new discoveries!

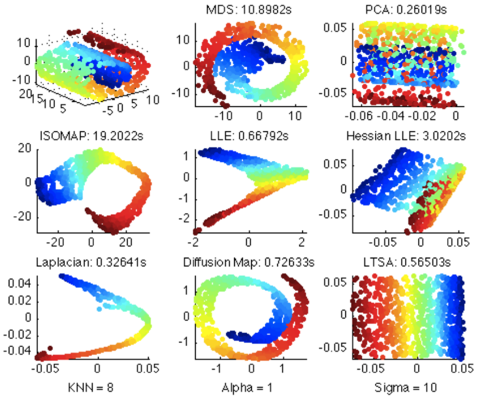


Figure 6. Comparisons of Manifold Learning Techniques on Swiss Roll

CHAPTER 6

Random Walk on Graphs

We have talked about Diffusion Map as a model of Random walk or Markov Chain on data graph. Among other methods of Manifold Learning, the distinct feature of Diffusion Map lies in that it combines both geometry and stochastic process. In the next few sections, we will talk about general theory of random walks or finite Markov chains on graphs which are related to data analysis. From this one can learn the origin of many ideas in diffusion maps.

Random Walk on Graphs.

- Perron-Frobenius Vector and Google's PageRank: this is about Perron-Frobenius theory for nonnegative 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 constant right eigenvector, and Multiple spectral clustering ("MNcut" 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. Introduction to Perron-Frobenius Theory and PageRank

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 $\Leftrightarrow x^T A x > 0 \quad \forall x \neq 0$
$A \succeq 0 \Leftrightarrow$ A is semi-positive-definite $\Leftrightarrow x^T A x \geq 0 \quad \forall x \neq 0$

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) Collatz-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) Collatz-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 (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 (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. Irreducible matrices imply an unique primary eigenvector, but not unique primary eigenvalue.

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) Collatz-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 (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 $\pi \geq 0, \quad 1^T \pi = 1$; 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 [BP98], 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. For more details about PageRank, readers may refer to Langville and Meyer's book [LM06].

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

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

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 [**Kle99**], 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 (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 (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 (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

1.1. Proof of Perron Theorem for Positive Matrices. A complete proof can be found in Meyer's book [Mey00], Chapter 8. Our proof below is based on optimization view, which is related to the Collatz-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 1.4. $Az = \lambda z, |\lambda| = \lambda^*, z \neq 0 \quad \Rightarrow \quad A|z| = \lambda^*|z|. \quad \lambda^* = \max_i |\lambda_i(A)|$

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). □

1.2. 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.

$$\lambda^* = \max_{x>0} \min_i \frac{[Ax^{[m-1]}]_i}{x_i^{m-1}} = \min_{x>0} \max_i \frac{[Ax^{[m-1]}]_i}{x_i^{m-1}}.$$

2. Introduction to Fiedler Theory and Cheeger Inequality

In this class, we introduced the random walk on graphs. The last lecture shows Perron-Frobenius theory to the analysis of primary eigenvectors which is the stationary distribution. In this lecture we will study the second eigenvector. To analyze the properties of the graph, we construct two matrices: one is (unnormalized) graph Laplacian and the other is normalized graph Laplacian. In the first part, we introduce Fiedler Theory for the unnormalized graph Laplacian, which shows the second eigenvector can be used to bipartite the graph into two connected components. In the second part, we study the eigenvalues and eigenvectors of normalized Laplacian matrix to show its relations with random walks or Markov chains on graphs. In the third part, we will introduce the Cheeger Inequality for second eigenvector of normalized Laplacian, which leads to an approximate algorithm for Normalized graph cut (NCut) problem, an NP-hard problem itself.

2.1. Unnormalized Graph Laplacian and Fiedler Theory. Let $G = (V, E)$ be an undirected, unweighted simple[1] graph. Although the edges here are unweighted, the theory below still holds when weight is added. We can get a similar conclusion with the weighted adjacency matrix. However the extension to directed graphs will lead to different pictures.

We use $i \sim j$ to denote that node $i \in V$ is a neighbor of node $j \in V$.

Definition (Adjacency Matrix)**.**

$$A_{ij} = \left\{ \begin{array}{cc} 1 & i \sim j \\ 0 & otherwise \end{array} \right. .$$

Remark. We can use the weight of edge $i \sim j$ to define A_{ij} if the graph is weighted. That indicates $A_{ij} \in \mathbb{R}^+$. We can also extend A_{ij} to $\mathbb{R}$ which involves both positive and negative weights, like correlation graphs. But the theory below can not be applied to such weights being positive and negative.

The degree of node i is defined as follows.

$$d_i = \sum_{j=1}^{n} A_{ij}.$$

[1] Simple graph means for every pair of nodes there are at most one edge associated with it; and there is no self loop on each node.

Define a diagonal matrix $D = \mathrm{diag}(d_i)$. Now let's come to the definition of Laplacian Matrix L.

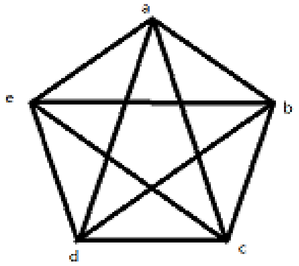
Definition (Graph Laplacian).

$$L_{ij} = \begin{cases} d_i & i = j, \\ -1 & i \sim j \\ 0 & otherwise \end{cases}$$

This matrix is often called *unnormalized graph Laplacian* in literature, to distinguish it from the normalized graph Laplacian below. In fact, $L = D - A$.

Example. $V = \{1, 2, 3, 4\}$, $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$. This is a linear chain with four nodes.

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$



Example. A complete graph of n nodes, K_n. $V = \{1, 2, 3...n\}$, every two points are connected, as the figure above with $n = 5$.

$$L = \begin{pmatrix} n-1 & -1 & -1 & ... & -1 \\ -1 & n-1 & -1 & ... & -1 \\ -1 & ... & -1 & n-1 & -1 \\ -1 & ... & -1 & -1 & n-1 \end{pmatrix}.$$

From the definition, we can see that L is symmetric, so all its eigenvalues will be real and there is an orthonormal eigenvector system. Moreover L is positive semi-definite (p.s.d.). This is due to the fact that

$$\begin{aligned} v^T L v &= \sum_i \sum_{j:j\sim i} v_i(v_i - v_j) = \sum_i \left(d_i v_i^2 - \sum_{j:j\sim i} v_i v_j \right) \\ &= \sum_{i\sim j} (v_i - v_j)^2 \geq 0, \qquad \forall v \in \mathbb{R}^n. \end{aligned}$$

In fact, L admits the decomposition $L = BB^T$ where $B \in \mathbb{R}^{|V|\times|E|}$ is called *incidence matrix* (or *boundary map* in algebraic topology) here, for any $1 \leq j < k \leq n$,

$$B(i, \{j,k\}) = \begin{cases} 1, & i = j, \\ -1, & i = k, \\ 0, & \text{otherwise} \end{cases}$$

These two statements imply the eigenvalues of L can't be negative. That is to say $\lambda(L) \geq 0$.

Theorem 2.1 (Fiedler theory)**.** Let L has n eigenvectors

$$Lv_i = \lambda_i v_i, \quad v_i \neq 0, \quad i = 0, \ldots, n-1$$

where $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. For the second smallest eigenvector v_1, define

$$\begin{aligned} N_- &= \{i : v_1(i) < 0\}, \\ N_+ &= \{i : v_1(i) > 0\}, \\ N_0 &= V - N_- - N_+. \end{aligned}$$

We have the following results.

(1) $\#\{i, \lambda_i = 0\} = \#\{$*connected components of* $G\}$;

(2) If G is connected, then both N_- and N_+ are connected. $N_- \cup N_0$ and $N_+ \cup N_0$ might be disconnected if $N_0 \neq \emptyset$.

This theorem tells us that the second smallest eigenvalue can be used to tell us if the graph is connected, *i.e.*G is connected iff $\lambda_1 \neq 0$, *i.e.*

$$\lambda_1 = 0 \Leftrightarrow \textit{there are at least two connected components.}$$
$$\lambda_1 > 0 \Leftrightarrow \textit{the graph is connected.}$$

Moreover, the second smallest eigenvector can be used to bipartite the graph into two connected components by taking N_- and N_+ when N_0 is empty. For this reason, we often call the second smallest eigenvalue λ_1 as the *algebraic connectivity.* More materials can be found in Jim Demmel's Lecture notes on Fiedler Theory at UC Berkeley: why we use unnormalized Laplacian eigenvectors for spectral partition (http://www.cs.berkeley.edu/~demmel/cs267/lecture20/lecture20.html).

We can calculate eigenvalues by using Rayleigh Quotient. This gives a sketch proof of the first part of the theory.

Proof of Part I. Let (λ, v) be a pair of eigenvalue-eigenvector, *i.e.*$Lv = \lambda v$. Since $L1 = 0$, so the constant vector $1 \in \mathbb{R}^n$ is always the eigenvector associated with $\lambda_0 = 0$. In general,

$$\lambda = \frac{v^T L v}{v^T v} = \frac{\sum\limits_{i\sim j} (v_i - v_j)^2}{\sum\limits_i {v_i}^2}.$$

Note that

$$0 = \lambda_1 \Leftrightarrow v_i = v_j \ (\textit{j is path connected with i}).$$

Therefore v is a piecewise constant function on connected components of G. If G has k components, then there are k independent piecewise constant vectors in the span of characteristic functions on those components, which can be used as eigenvectors of L. In this way, we proved the first part of the theory. □

2.2. Normalized graph Laplacian and Cheeger's Inequality.

Definition (Normalized Graph Laplacian)**.**

$$\mathcal{L}_{ij} = \begin{cases} 1 & i = j, \\ -\frac{1}{\sqrt{d_i d_j}} & i \sim j, \\ 0 & otherwise. \end{cases}$$

In fact $\mathcal{L} = D^{-1/2}(D - A)D^{-1/2} = D^{-1/2}LD^{-1/2} = I - D^{-1/2}(D - A)D^{-1/2}$. From this one can see the relations between eigenvectors of normalized $\mathcal{L}$ and unnormalized L. For eigenvectors $\mathcal{L}v = \lambda v$, we have

$$(I - D^{-1/2}LD^{-1/2})v = \lambda v \Leftrightarrow Lu = \lambda Du, \quad u = D^{-1/2}v,$$

whence eigenvectors of $\mathcal{L}$, v after rescaling by $D^{-1/2}v$, become generalized eigenvectors of L.

We can also use the Rayleigh Quotient to calculate the eigenvalues of $\mathcal{L}$.

$$\begin{aligned} \frac{v^T\mathcal{L}v}{v^Tv} &= \frac{v^TD^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}v}{v^v} \\ &= \frac{u^TLu}{u^TDu} \\ &= \frac{\sum\limits_{i\sim j}(u_i - u_j)^2}{\sum\limits_j {u_j}^2 d_j}. \end{aligned}$$

Similarly we get the relations between eigenvalue and the connected components of the graph.

$$\#\{\lambda_i(\mathcal{L}) = 0\} = \#\{connected\ components\ of\ G\}.$$

Next we show that eigenvectors of $\mathcal{L}$ are related to random walks on graphs. This will show you why we choose this matrix to analysis the graph.

We can construct a random walk on G whose transition matrix is defined by

$$P_{ij} \sim \frac{A_{ij}}{\sum\limits_j A_{ij}} = \frac{1}{d_i}.$$

By easy calculation, we see the result below.

$$P = D^{-1}A = D^{-1/2}(I - \mathcal{L})D^{1/2}.$$

Hence P is similar to $I - \mathcal{L}$. So their eigenvalues satisfy $\lambda_i(P) = 1 - \lambda_i(\mathcal{L})$. Consider the right eigenvector ϕ and left eigenvector ψ of P.

$$u^TP = \lambda u,$$

$$Pv = \lambda v.$$

Due to the similarity between P and $\mathcal{L}$,

$$u^TP = \lambda u^T \Leftrightarrow u^TD^{-1/2}(I - \mathcal{L})D^{1/2} = \lambda u^T.$$

Let $\bar{u} = D^{-1/2}u$, we will get:

$$\bar{u}^T(I - \mathcal{L}) = \lambda\bar{u}^T$$

$$\Leftrightarrow \mathcal{L}\bar{u} = (1 - \lambda)\bar{u}.$$

You can see $\bar{u}$ is the eigenvector of $\mathcal{L}$, and we can get left eigenvectors of P from $\bar{u}$ by multiply it with $D^{1/2}$ on the left side. Similarly for the right eigenvectors $v = D^{-1/2}\bar{u}$.

If we choose $u_0 = \pi_i \sim \frac{d_i}{\sum d_i}$, then:

$$\begin{aligned}\bar{u}_0(i) &\sim \sqrt{d_i},\\ \bar{u}_k^T \bar{u}_l &= \delta_{kl},\\ u_k^T D v_l &= \delta_{kl},\\ \pi_i P_{ij} = \pi_j P_{ji} \sim A_{ij} &= A_{ji},\end{aligned}$$

where the last identity says the Markov chain is time-reversible.

All the conclusions above show that the normalized graph Laplacian $\mathcal{L}$ keeps some connectivity measure of unnormalized graph Laplacian L. Furthermore, $\mathcal{L}$ is more related with random walks on graph, through which eigenvectors of P are easy to check and calculate. That's why we choose this matrix to analysis the graph.

Now we are ready to introduce the Cheeger's inequality with normalized graph Laplacian.

Let G be a graph, $G = (V, E)$ and S is a subset of V whose complement is $\bar{S} = V - S$. We define $Vol(S)$, $CUT(S)$ and $NCUT(S)$ as below.

$$Vol(S) = \sum_{i \in S} d_i.$$

$$CUT(S) = \sum_{i \in S, j \in \bar{S}} A_{ij}.$$

$$NCUT(S) = \frac{CUT(S)}{\min(Vol(S), Vol(\bar{S}))}.$$

$NCUT(S)$ is called normalized-cut. We define the Cheeger constant

$$h_G = \min_S NCUT(S).$$

Finding minimal normalized graph cut is NP-hard. It is often defined that

$$\text{Cheeger ratio (expander): } h_S := \frac{CUT(S)}{Vol(S)}$$

and

$$\text{Cheeger constant: } h_G := \min_S \max\{h_S, h_{\bar{S}}\}.$$

Cheeger Inequality says the second smallest eigenvalue provides both upper and lower bounds on the minimal normalized graph cut. Its proof gives us a constructive polynomial algorithm to achieve such bounds.

Theorem 2.2 (Cheeger Inequality)**.** For every undirected graph G,

$$\frac{h_G^2}{2} \le \lambda_1(\mathcal{L}) \le 2h_G.$$

Proof. (1) Upper bound:

Assume the following function f realizes the optimal normalized graph cut,

$$f(i) = \begin{cases} \frac{1}{Vol(S)} & i \in S,\\ \frac{-1}{Vol(\bar{S})} & i \in \bar{S}, \end{cases}$$

By using the Rayleigh Quotient, we get

$$\begin{aligned}\lambda_1 = &\inf_{g\perp D^{1/2}e} \frac{g^T\mathcal{L}g}{g^Tg}\\ \leq&\frac{\sum_{i\sim j}(f_i-f_j)^2}{\sum f_i^2 d_i}\\ =&\frac{(\frac{1}{Vol(S)}+\frac{1}{Vol(\bar{S})})^2 CUT(S)}{Vol(S)\frac{1}{Vol(S)^2}+Vol(\bar{S})\frac{1}{Vol(\bar{S})^2}}\\ =&(\frac{1}{Vol(S)}+\frac{1}{Vol(\bar{S})})CUT(S)\\ \leq&\frac{2CUT(S)}{\min(Vol(S),Vol(\bar{S}))}=:2h_G.\end{aligned}$$

which gives the upper bound.

(2) Lower bound: the proof of lower bound actually gives a constructive algorithm to compute an approximate optimal cut as follows.

Let v be the second eigenvector, i.e. $\mathcal{L}v=\lambda_1 v$, and $f=D^{-1/2}v$. Then we reorder node set V such that $f_1\leq f_2\leq ...\leq f_n$). Denote $V_-=\{i;v_i<0\}, V_+=\{i;v_i\geq v_r\}$. Without Loss of generality, we can assume

$$\sum_{i\in V_-} d_v \geq \sum_{i\in V_+} d_v$$

Define new functions f^+ to be the magnitudes of f on V_+.

$$f_i^+=\begin{cases} f_i & i\in V_+,\\ 0 & otherwise,\end{cases}$$

Now consider a series of particular subsets of V,

$$S_i=\{v_1,v_2,...v_i\},$$

and define

$$\widetilde{Vol}(S)=\min(Vol(S),Vol(\bar{S})).$$

$$\alpha_G=\min_i NCUT(S_i).$$

Clearly finding the optimal value α just requires comparison over $n-1$ NCUT values.

Below we shall show that

$$\frac{h_G^2}{2}\leq\frac{\alpha_G^2}{2}\leq\lambda_1.$$

First, we have $Lf=\lambda_1 Df$, so we must have

$$\sum_{j:j\sim i} f_i(f_i-f_j)=\lambda_1 d_i f_i^2. \quad (70)$$

From this we will get the following results,

$$\begin{aligned}
\lambda_1 &= \frac{\sum_{i\in V_+} f_i \sum_{j:j\sim i}(f_i - f_j)}{\sum_{i\in V_+} d_i f_i^2}, \\
&= \frac{\sum_{i\sim j\ i,j\in V_+}(f_i - f_j)^2 + \sum_{i\in V_+} f_i \sum_{j\sim i\ j\in V_-}(f_i - f_j)}{\sum_{i\in V_+} d_i f_i^2}, (f_i - f_j)^2 = f_i(f_i - f_j) + f_j(f_j - f_i) \\
&> \frac{\sum_{i\sim j\ i,j\in V_+}(f_i - f_j)^2 + \sum_{i\in V_+} f_i \sum_{j\sim i\ j\in V_-}(f_i)}{\sum_{i\in V_+} d_i f_i^2}, \\
&= \frac{\sum_{i\sim j}(f_i^+ - f_j^+)^2}{\sum_{i\in V} d_i {f_i^+}^2}, \\
&= \frac{(\sum_{i\sim j}(f_i^+ - f_j^+)^2)(\sum_{i\sim j}(f_i^+ + f_j^+)^2)}{(\sum_{i\in V} {f_i^+}^2 d_i)(\sum_{i\sim j}(f_i^+ + f_j^+)^2)} \\
&\geq \frac{(\sum_{i\sim j} {f_i^+}^2 - {f_j^+}^2)^2}{(\sum_{i\in V} {f_i^+}^2 d_i)(\sum_{i\sim j}(f_i^+ + f_j^+)^2)}, \quad \text{Cauchy-Schwartz Inequality} \\
&\geq \frac{(\sum_{i\sim j} {f_i^+}^2 - {f_j^+}^2)^2}{2(\sum_{i\in V} {f_i^+}^2 d_i)^2},
\end{aligned}$$

where the second last step is due to the Cauchy-Schwartz inequality $|\langle x, y\rangle|^2 \leq \langle x, x\rangle \cdot \langle y, y\rangle$, and the last step is due to $\sum_{i\sim j\in V}(f_i^+ + f_j^+)^2 = \sum_{i\sim j\in V}({f_i^+}^2 + {f_j^+}^2 + 2f_i^+ f_j^+) \leq 2\sum_{i\sim j\in V}({f_i^+}^2 + {f_j^+}^2) \leq 2\sum_{i\in V} {f_i^+}^2 d_i$. Continued from the last inequality,

$$\begin{aligned}
\lambda_1 &\geq \frac{(\sum_{i\sim j} {f_i^+}^2 - {f_j^+}^2)^2}{2(\sum_{i\in V} {f_i^+}^2 d_i)^2}, \\
&\geq \frac{(\sum_{i\in V}({f_i^+}^2 - {f_{i-1}^+}^2)CUT(S_{i-1}))^2}{2(\sum_{i\in V} {f_i^+}^2 d_i)^2}, \quad \text{since } f_1 \leq f_2 \leq \ldots \leq f_n \\
&\geq \frac{(\sum_{i\in V}({f_i^+}^2 - {f_{i-1}^+}^2)\alpha_G \widetilde{Vol}(S_{i-1}))^2}{2(\sum_{i\in V} {f_i^+}^2 d_i)^2} \\
&= \frac{\alpha_G^2}{2} \cdot \frac{(\sum_{i\in V} {f_i^+}^2(\widetilde{Vol}(S_{i-1}) - \widetilde{Vol}(S_i)))^2}{(\sum_{i\in V} {f_i^+}^2 d_i)^2}, \\
&= \frac{\alpha_G^2}{2} \frac{(\sum_{i\in V} {f_i^+}^2 d_i)^2}{(\sum_{i\in V} {f_i^+}^2 d_i)^2} = \frac{\alpha_G^2}{2}.
\end{aligned}$$

where the last inequality is due to the assumption $Vol(V_-) \geq Vol(V_+)$, whence $\widetilde{Vol}(S_i) = Vol(\bar{S}_i)$ for $i \in V_+$.

This completes the proof. □

Fan Chung gives a short proof of the lower bound in Simons Institute workshop, 2014.

Short Proof. The proof is based on the fact that

$$h_G = \inf_{f \neq 0} \sup_{c \in \mathbb{R}} \frac{\sum_{x \sim y} |f(x) - f(y)|}{\sum_x |f(x) - c| d_x}$$

where the supreme over c is reached at $c^* = median(f(x) : x \in V)$.

$$\begin{aligned}
\lambda_1 &= R(f) = \sup_c \frac{\sum_{x \sim y} (f(x) - f(y))^2}{\sum_x (f(x) - c)^2 d_x}, \\
&\geq \frac{\sum_{x \sim y} (g(x) - g(y))^2}{\sum_x g(x)^2 d_x}, \quad g(x) = f(x) - c \\
&= \frac{(\sum_{x \sim y} (g(x) - g(y))^2)(\sum_{x \sim y} (g(x) + g(y))^2)}{(\sum_{x \in V} g^2(x) d_x)((\sum_{x \sim y} (g(x) + g(y))^2)} \\
&\geq \frac{(\sum_{x \sim y} |g^2(x) - g^2(y)|)^2}{(\sum_{x \in V} g^2(x) d_x)((\sum_{x \sim y} (g(x) + g(y))^2)}, \quad \text{Cauchy-Schwartz Inequality} \\
&\geq \frac{(\sum_{x \sim y} |g^2(x) - g^2(y)|)^2}{2(\sum_{x \in V} g^2(x) d_x)^2}, \quad (g(x) + g(y))^2 \leq 2(g^2(x) + g^2(y)) \\
&\geq \frac{h_G^2}{2}.
\end{aligned}$$

□

3. *Laplacians and the Cheeger inequality for directed graphs

The following section is mainly contained in [Chu05], which described the following results:

(1) Define Laplacians on directed graphs.
(2) Define Cheeger constants on directed graphs.
(3) Give an example of the singularity of Cheeger constant on directed graph.
(4) Use the eigenvalue of Lapacian and the Cheeger constant to estimate the convergence rate of random walk on a directed graph.

Another good reference is [LZ10].

3.1. Definition of Laplacians on directed graphs.

On a finite and strong connected directed graph $G = (V, E)$ (A directed graph is strong connected if there is a path between any pair of vertices), a weight is a function

$$w : \quad E \quad \rightarrow \quad \mathbb{R}_{\geq 0}$$

The in-degree and out-degree of a vertex are defined as

$$\begin{array}{cccc}
d^{in} : & V & \rightarrow & \mathbb{R}_{\geq 0} \\
& d_i^{in} & = & \sum_{j \in V} w_{ji} \\
d^{out} : & V & \rightarrow & \mathbb{R}_{\geq 0} \\
& d_i^{out} & = & \sum_{j \in V} w_{ij}
\end{array}$$

Note that d_i^{in} may be different from d_i^{out}.

A random walk on the weighted G is a Markov chain with transition probability

$$P_{ij} = \frac{w_{ij}}{d_i^{out}}.$$

Since G is strong connected, P is irreducible, and consequently there is a unique stationary distribution ϕ. (And the distribution of the Markov chain will converge to it if and only if P is aperiodic.)

Example (undirected graph)**.**

$$\phi(x) = \frac{d_x}{\sum_y d_y}.$$

Example (Eulerian graph)**.** If $d_x^{in} = d_x^{out}$ for every vertex x, then $\phi(x) = \frac{d_x^{out}}{\sum_y d_y^{out}}$.

This is because d_x^{out} is an unchanged measure with

$$\sum_x d_x^{out} P_{xy} = \sum_x w_{xy} = d_y^{in} = d_y^{out}.$$

Example (exponentially small stationary dist.)**.** G is a directed graph with $n+1$ vertices formed by the union of a directed circle $v_0 \to v_1 \to \cdots \to v_n$ and edges $v_i \to v_0$ for $i = 1, 2, \cdots, n$. The weight on any edge is 1. Checking from v_n to v_0 with the prerequisite of stationary distribution that the inward probability flow equals to the outward probability flow, we can see that

$$\phi(v_0) = 2^n \phi(v_n), i.e. \phi(v_n) = 2^{-n}\phi(v_0).$$

This exponentially small stationary distribution cannot occur in undirected graph cases for then

$$\phi(i) = \frac{d_i}{\sum_j d_j} \geq \frac{1}{n(n-1)}.$$

However, the stationary dist. can be no smaller than exponential, because we have

Theorem 3.1. If G is a strong connected directed graph with $w \equiv 1$, and $d_x^{out} \leq k, \forall x$, then $\max\{\phi(x) : x \in V\} \leq k^D \min\{\phi(y) : y \in V\}$, where D is the diameter of G.

It can be easily proved using induction on the path connecting x and y.

Now we give a definition on those balanced weights.

Definition (circulation)**.**

$$F : \quad E \quad \to \quad \mathbb{R}_{\geq 0}$$

If F satisfies

$$\sum_{u, u \to v} F(u, v) = \sum_{w, v \to w} F(v, w), \forall v,$$

then F is called a circulation.

Note. A circulation is a flow with no source or sink.

Example. For a directed graph, $F_\phi(u, v) = \phi(u)P(u, v)$ is a circulation, for

$$\sum_{u, u \to v} F_\phi(u, v) = \phi(v) = \sum_{w, v \to w} F_\phi(v, w).$$

Definition (Rayleigh quotient)**.** For a directed graph G with transition probability matrix P and stationary distribution ϕ, the Rayleigh quotient for any $f : V \to \mathbb{C}$ is defined as

$$R(f) = \frac{\sum_{u \to v} \mid f(u) - f(v) \mid^2 \phi(u)P(u, v)}{\sum_v \mid f(v) \mid^2 \phi(v)}.$$

Note. Compare with the undirected graph condition where

$$R(f) = \frac{\sum_{u \sim v} \mid f(u) - f(v) \mid^2 w_{uv}}{\sum_v \mid f(v) \mid^2 d(v)}.$$

If we look on every undirected edge (u, v) as two directed edges $u \to v, v \to u$, then we get a Eulerian directed graph. So $\phi(u) \sim d_u^{out}$ and $d_u^{out} P(u, v) = w_{uv}$, as a result $R(f)(directed) = 2R(f)(undirected)$. The factor 2 is the result of looking on every edge as two edges.

The next step is to extend the definition of Laplacian to directed graphs. First we give a review on Lapalcian on undirected graphs. On an undirected graph, adjacent matrix is

$$A_{ij} = \begin{cases} 1, & i \sim j; \\ 0, & i \not\sim j. \end{cases}$$

$$D = \operatorname{diag}(d(i)),$$

$$\mathcal{L} = D^{-1/2}(D - A)D^{-1/2}.$$

On a directed graph, however, there are two degrees on a vertex which are generally inequivalent. Notice that on an undirected graph, stationary distribution $\phi(i) \sim d(i)$, so $D = c\Phi$, where c is a constant and $\Phi = \operatorname{diag}(\phi(i))$.

$$\begin{aligned} \mathcal{L} &= I - D^{-1/2} A D^{-1/2} \\ &= I - D^{1/2} P D^{-1/2} \\ &= I - c^{1/2}\Phi^{1/2} P c^{-1/2}\Phi^{-1/2} \\ &= I - \Phi^{1/2} P \Phi^{-1/2} \end{aligned}$$

Extending and symmetrizing it, we define Laplacian on a directed graph

Definition (Laplacian)**.**

$$\mathcal{L} = I - \frac{1}{2}(\Phi^{1/2} P \Phi^{-1/2} + \Phi^{-1/2} P^* \Phi^{1/2}).$$

Suppose the eigenvalues of $\mathcal{L}$ are $0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$. Like the undirected case, we can calculate λ_1 with the Rayleigh quotient.

Theorem 3.2.

$$\lambda_1 = \inf_{\sum f(x)\phi(x) = 0} \frac{R(f)}{2}.$$

Before proving that, we need

Lemma 3.3.

$$R(f) = 2\frac{g\mathcal{L}g^*}{\parallel g \parallel^2}, \text{ where } g = f\Phi^{1/2}.$$

Proof.

$$
\begin{aligned}
R(f) &= \frac{\sum_{u\to v} \mid f(u)-f(v) \mid^2 \phi(u)P(u,v)}{\sum_v \mid f(v) \mid^2 \phi(v)} \\
&= \frac{\sum_{u\to v} \mid f(u) \mid^2 \phi(u)P(u,v) + \sum_v \mid f(v) \mid^2 \phi(v) - \sum_{u\to v}(\overline{f(u)}f(v) + f(u)\overline{f(v)})\phi(u)P(u,v)}{f\Phi f^*} \\
&= \frac{\sum_u \mid f(u) \mid^2 \phi(u) + \sum_v \mid f(v) \mid^2 \phi(v) - (f^*\Phi P f + f\Phi P f^*}{f\Phi f^*} \\
&= 2 - \frac{f(P^*\Phi + \Phi P)f^*}{f\Phi f^*} \\
&= 2 - \frac{(g\Phi^{-1/2})(P^*\Phi + \Phi P)(\Phi^{-1/2}g^*)}{(g\Phi^{-1/2})\Phi(\Phi^{-1/2}g^*)} \\
&= 2 - \frac{g(\Phi^{-1/2}P^*\Phi^{1/2} + \Phi^{1/2}P\Phi^{-1/2})g^*}{gg^*} \\
&= 2 \cdot \frac{g\mathcal{L}g^*}{\parallel g \parallel^2}
\end{aligned}
$$

□

Proof of Theorem 3.2. With Lemma 3.3 and $\mathcal{L}(\phi(x)^{1/2})_{n\times 1} = 0$, we have

$$
\begin{aligned}
\lambda_1 &= \inf_{\sum g(x)\phi(x)^{1/2}=0} \frac{R(f)}{2} \\
&= \inf_{\sum f(x)\phi(x)=0} \frac{R(f)}{2}.
\end{aligned}
$$

□

Note.

$$
\begin{aligned}
\lambda_1 &= \inf_{f,\sum f(x)\phi(x)=0} \frac{R(f)}{2} \\
&= \inf_{f,\sum f(x)\phi(x)=0} \frac{\sum_{u\to v} \mid f(u)-f(v) \mid^2 \phi(u)P(u,v)}{2\sum_v \mid f(v) \mid^2 \phi(v)} \\
&= \inf_{f,\sum f(x)\phi(x)=0} \sup_c \frac{\sum_{u\to v} \mid f(u)-f(v) \mid^2 \phi(u)P(u,v)}{2\sum_v \mid f(v)-c \mid^2 \phi(v)}
\end{aligned}
$$

Theorem 3.4. Suppose the eigenvalues of P are $\rho_0, \cdots, \rho_{n-1}$ with $\rho_0 = 1$, then

$$\lambda_1 \leq \min_{i\neq 0}(1 - Re\rho_i).$$

3.2. Definition of Cheeger constants on directed graphs.

We have a circulation $F_\phi(u,v) = \phi(u)P(u,v)$. Define

$$F(\partial S) = \sum_{u\in S, v\notin S} F(u,v), F(v) = \sum_{u, u\to v} F(u,v) = \sum_{w, v\to w} F(v,w), F(S) = \sum_{v\in S} F(v),$$

then $F(\partial S) = F(\partial \bar{S})$.

Definition (Cheeger constant)**.** The Cheeger constant of a graph G is defined as

$$h(G) = \inf_{S\subset V} \frac{F(\partial S)}{\min\big(F(S), F(\bar{S})\big)}$$

Note. Compare with the undirected graph condition where

$$h_G = \inf_{S\subset V} \frac{\mid \partial S \mid}{\min\big(\mid S \mid, \mid \bar{S} \mid\big)}.$$

Similarly, we have

$$\begin{aligned}
h_G(undirected) &= \inf_{S\subset V} \frac{\mid \partial S \mid}{\min\big(\mid S \mid, \mid \bar{S} \mid\big)} \\
&= \inf_{S\subset V} \frac{\sum_{u\in S, v\in \bar{S}} w_{uv}}{\min\big(\sum_{u\in S} d(u), \sum_{u\in \bar{S}} d(u)\big)} \\
h_G(directed) &= \inf_{S\subset V} \frac{\sum_{u\in S, v\in \bar{S}} \phi(u)P(u,v)}{\min\big(\sum_{u\in S} \phi(u), \sum_{u\in \bar{S}} \phi(u)\big)} \\
&= \inf_{S\subset V} \frac{F(\partial S)}{\min\big(F(S), F(\bar{S})\big)}.
\end{aligned}$$

Theorem 3.5. For every directed graph G,

$$\frac{h^2(G)}{2} \leq \lambda_1 \leq 2h(G).$$

The proof is similar to the undirected case using Rayleigh quotient and Theorem 3.2.

3.3. An example of the singularity of Cheeger constant on a directed graph. We have already given an example of a directed graph with $n+1$ vertices and stationary distribution ϕ satisfying $\phi(v_n) = 2^{-n}\phi(v_0)$. Now we make a copy of this graph and denote the new $n+1$ vertices $u_0, \ldots, u_n$. Joining the two graphs together by two edges $v_n \to u_n$ and $u_n \to v_n$, we get a bigger directed graph. Let $S = (v_0, \cdots, v_n)$, we have $h(G) \sim 2^{-n}$. In comparison, $h(G) \geq \frac{2}{n(n-1)}$ for undirected graph.

3.4. Estimate the convergence rate of random walks on directed graphs. Define the distance of P after s steps and ϕ as

$$\Delta(s) = \max_{y\in V} \left(\sum_{x\in V} \frac{(P^s(y,x) - \phi(x))^2}{\phi(x)} \right)^{1/2}.$$

Modify the random walk into a lazy random walk $\tilde{P} = \frac{I+P}{2}$, so that it is aperiodic.

Theorem 3.6.

$$\Delta(t)^2 \leq C(1 - \frac{\lambda_1}{2})^t.$$

3.5. Random Walks on Digraphs, The Generalized Digraph Laplacian, and The Degree of Asymmetry. In this paper the following have been discussed:

(1) Define an asymmetric Laplacian $\tilde{\mathcal{L}}$ on directed graph;
(2) Use $\tilde{\mathcal{L}}$ to estimate the hitting time and commute time of the corresponding Markov chain;
(3) Introduce a metric to measure the asymmetry of $\tilde{\mathcal{L}}$ and use this measure to give a tighter bound on the Markov chain mixing rate and a bound on the Cheeger constant.

Let P be the transition matrix of Markov chain, and $\pi = (\pi_1, \dots, \pi_n)^T$(column vector) denote its stationary distribution (which is unique if the Markov chain is irreducible, or if the directed graph is strongly connected). Let $\Pi = \mathrm{diag}\{\pi_1, \dots, \pi_n\}$, then we define the normalized Laplacian $\tilde{\mathcal{L}}$ on directed graph:

$$\tilde{\mathcal{L}} = I - \Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}} \tag{71}$$

3.5.1. *Hitting time, commute time and fundamental matrix.* We establish the relations between $\tilde{\mathcal{L}}$ and the hitting time and commute time of random walk on directed graph through the fundamental matrix $Z = [z_{ij}]$, which is defined as:

$$z_{ij} = \sum_{t=0}^{\infty} (p_{ij}^t - \pi_j),\ 1 \le i, j \le n \tag{72}$$

or alternatively as an infinite sum of matrix series:

$$Z = \sum_{t=0}^{\infty} (P^t - \mathbf{1}\pi^T) \tag{73}$$

With the fundamental matrix, the hitting time and commute time can be expressed as follows:

$$H_{ij} = \frac{z_{jj} - z_{ij}}{\pi_j} \tag{74}$$

$$C_{ij} = H_{ij} + H_{ji} = \frac{z_{jj} - z_{ij}}{\pi_j} + \frac{z_{ii} - z_{ji}}{\pi_i} \tag{75}$$

Using (73), we can write the fundamental matrix Z in a more explicit form. Notice that

$$(P - \mathbf{1}\pi^T)(P - \mathbf{1}\pi^T) = P^2 - \mathbf{1}\pi^T P - P\mathbf{1}\pi^T + \mathbf{1}\pi^T\mathbf{1}\pi^T = P^2 - \mathbf{1}\pi^T \tag{76}$$

We use the fact that $\mathbf{1}$ and π are the right and left eigenvector of the transition matrix P with eigenvalue 1, and that $\pi^T\mathbf{1} = 1$ since π is a distribution. Then

$$Z + \mathbf{1}\pi^T = \sum_{t=0}^{\infty} (P - \mathbf{1}\pi^T)^t = (I - P + \mathbf{1}\pi^T)^{-1} \tag{77}$$

3.5.2. *Green's function and Laplacian for directed graph.* If we treat the directed graph Laplacian $\tilde{\mathcal{L}}$ as an asymmetric operator on a directed graph G, then we can define the Green's Function $\tilde{\mathcal{G}}$ (without boundary condition) for directed graph. The entries of G satisfy the conditions:

$$(\tilde{\mathcal{G}}\tilde{\mathcal{L}})_{ij} = \delta_{ij} - \sqrt{\pi_i \pi_j} \tag{78}$$

or in the matrix form

$$\tilde{\mathcal{G}}\tilde{\mathcal{L}} = I - \pi^{\frac{1}{2}} \pi^{\frac{1}{2}^T} \tag{79}$$

The central theorem in the second paper associate the Green's Function $\tilde{\mathcal{G}}$, the fundamental matrix Z and the normalize directed graph Laplacian $\tilde{\mathcal{L}}$:

Theorem 3.7. Let $\tilde{\mathcal{Z}} = \Pi^{\frac{1}{2}} Z \Pi^{-\frac{1}{2}}$ and $\tilde{\mathcal{L}}^\dagger$ denote the Moore-Penrose pseudo-inverse $\tilde{\mathcal{L}}$, then

$$\tilde{\mathcal{G}} = \tilde{\mathcal{Z}} = \tilde{\mathcal{L}}^\dagger \tag{80}$$

3.6. measure of asymmetric and its relation to Cheeger constant and mixing rate. To measure the asymmetry in directed graph, we write the $\tilde{\mathcal{L}}$ into the sum of a symmetric part and a skew-symmetric part:

$$\tilde{\mathcal{L}} = \frac{1}{2}(\tilde{\mathcal{L}} + \tilde{\mathcal{L}}^T) + \frac{1}{2}(\tilde{\mathcal{L}} - \tilde{\mathcal{L}}^T) \tag{81}$$

$\frac{1}{2}(\tilde{\mathcal{L}} + \tilde{\mathcal{L}}^T) = \mathcal{L}$ is the symmetrized Laplacian introduced in the first paper. Let $\Delta = \frac{1}{2}(\tilde{\mathcal{L}} - \tilde{\mathcal{L}}^T)$, the Δ captures the difference between $\tilde{\mathcal{L}}$ and its transpose. Let σ_i, λ_i and $\delta_i(1 \le i \le n)$ denotes the i-th singular value of $\mathcal{L}$, $\mathcal{L}$, Δ in ascending order ($\sigma_1 = \lambda_1 = \delta_1 = 0$). Then the relation $\tilde{\mathcal{L}} = \mathcal{L} + \Delta$ implies

$$\lambda_i \le \sigma_i \le \lambda_i + \delta_n \tag{82}$$

Therefore $\delta_n = \|\Delta\|_2$ is used to measure the degree of asymmetry in the directed graph.

The following two theorems are application of this measure.

Theorem 3.8. The second singular of $\tilde{\mathcal{L}}$ has bounds :

$$\frac{h(G)^2}{2} \le \sigma_2 \le (1 + \frac{\delta_n}{\lambda_2}) \cdot 2h(G) \tag{83}$$

where $h(G)$ is the Cheeger constant of graph G

Theorem 3.9. For a aperiodic Markov chain P,

$$\delta_n^2 \le \max\{\frac{\|\tilde{P}f\|^2}{\|f\|^2} : f \perp \pi^{\frac{1}{2}}\} \le (1 - \lambda_2)^2 + 2\delta_n\lambda_n + \delta_n^2 \tag{84}$$

where $\tilde{P} = \Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}}$

4. Lumpability of Markov Chain

Let P be the transition matrix of a Markov chain on graph $G = (V, E)$ with $V = \{1, 2, \cdots, n\}$, i.e. $P_{ij} = \Pr\{x_t = j : x_{t-1} = i\}$. Assume that V admits a partition Ω:

$$V = \cup_{i=1}^k \Omega_i, \quad \Omega_i \cap \Omega_j = \emptyset, \ i \ne j.$$
$$\Omega = \{\Omega_s : s = 1, \cdots, k\}.$$

Observe a sequence$\{x_0, x_1, \cdots, x_t\}$ sampled from the Markov chain with initial distribution π_0.

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.*

$$\mathrm{Prob}_{\pi_0}\{x_t \in \Omega_{k_t} : x_{t-1} \in \Omega_{k_{t-1}}, \cdots, x_0 \in \Omega_{k_0}\} = \mathrm{Prob}\{x_t \in \Omega_{k_t} : x_{t-1} \in \Omega_{k_{t-1}}\}. \tag{85}$$

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. The lumpability condition above can be rewritten as

$$\mathrm{Prob}_{\pi_0}\{y_t = k_t : y_{t-1} = k_{t-1}, \cdots, y_0 = k_0\} = \mathrm{Prob}\{y_t = k_t : y_{t-1} = k_{t-1}\}. \tag{86}$$

Theorem 4.1. **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 $\hat{P}_{i\Omega_t} = \sum_{j\in\Omega_t} P_{ij}$.

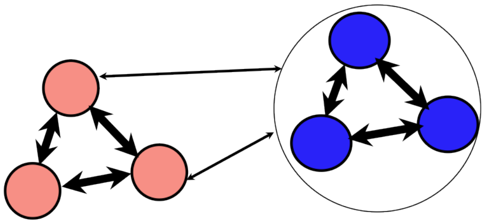


FIGURE 1. Lumpability condition $\hat{P}_{i\Omega_t} = \hat{P}_{j\Omega_t}$

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 span$\{\chi_{\Omega_s} : s = 1, \cdots, k\}$, χ is the characteristic function.

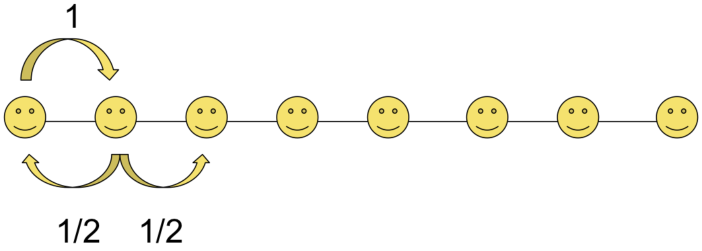


FIGURE 2. A linear chain of $2n$ nodes with a random walk.

Example. 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.

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.

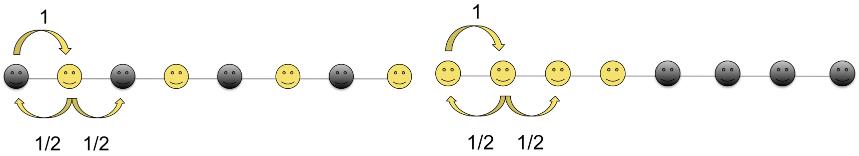


FIGURE 3. Left: two lumpable states; Right: optimal-bipartition of Ncut.

Example. 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* ↦ *middle schools* ↦ *high schools* ↦ *college/university* ↦ *work unit*, etc.

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

$$VUPV = PV, \tag{87}$$

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

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

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

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

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.** "⇒" To see the necessity, P is lumpable w.r.t. partition Ω, then it is necessary that

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

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

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

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

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

depends only on $\Omega_s, \Omega_t \in \Omega$. Probability above can be written as $\text{Prob}_{\pi_{t-1}}(y_t = t)$ 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 $\text{Prob}_i(y_t = t) = \hat{p}_{i\Omega_t} \equiv \hat{p}_{st}$ for all $i \in \Omega_s$, then $\text{Prob}_{\pi_{t-1}}(y_t = t) = \sum_{i\in\Omega_s} \pi_{t-1}\hat{p}_{i\Omega_t} = \hat{p}_{st}$ 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 $\text{span}\{\chi_{\Omega_i}, i = 1, \cdots, k\}$. 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 $\text{span}\{\mathcal{X}_{\Omega_i}, i = 1, \cdots, k\}$. 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 $VUPV = PV$. □

5. Applications of Lumpability: MNcut and Optimal Reduction of Complex Networks

If the random walk on a graph P has *top k* nearly piece-wise constant right eigenvectors, then the Markov chain P is approximately lumpable. Some spectral clustering algorithms are proposed in such settings.

5.1. MNcut.

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 $Y^{n\times k} = [\phi_1, \cdots, \phi_k] \rightarrow$ diffusion map when $\lambda_i \approx 1$.
3) k-means (or other suitable clustering methods) on Y to k-clusters.

5.2. Optimal Reduction and Complex Network.

5.2.1. *Random Walk on Graph.* Let $G = G(S, E)$ denotes an undirected graph. Here S has the meaning of "states". $|S| = n \gg 1$. Let $A = e(x, y)$ denotes its adjacency matrix, that is,

$$e(x,y) = \begin{cases} 1 & x \sim y \\ 0 & otherwise \end{cases}$$

Here $x \sim y$ means $(x, y) \in E$. Here, weights on different edges are the same 1. They may be different in some cases.

Now we define a random walk on G . Let

$$p(x,y) = \frac{e(x,y)}{d(x)} \quad \text{where} \quad d(x) = \sum_{y \in S} e(x,y)$$

We can check that $P = p(x, y)$ is a stochastic matrix and (S, P) is a Markov chain. If G is connected, this Markov chain is irreducible and if G is not a tree, the chain is even primitive. We assume G is connected from now on. If it is not, we can focus on each of its connected component.So the Markov chain has unique invariant distributionμ by irreducibility:

$$\mu(x) = \frac{d(x)}{\sum\limits_{z \in S} d(z)} \quad \forall x \in S$$

A Markov chain defined as above is reversible. That is, detailed balance condition is satisfied:

$$\mu(x)p(x,y) = \mu(y)p(y,x) \quad \forall x, y \in S$$

Define an inner product on space$\mathcal{L}^2_\mu$:

$$< f, g >_\mu = \sum_{x \in S} \sum_{y \in S} f(x)g(x)\mu(x) \quad f, g \in \mathcal{L}^2_\mu$$

$\mathcal{L}^2_\mu$ is a Hilbert space with this inner product. If we define an operator T on it:

$$Tf(x) = \sum_{y \in S} p(x,y)f(y) = \mathbb{E}_{[y|x]} f(y)$$

We can check that T is a self adjoint operator on $\mathcal{L}^2_\mu$:

$$\begin{aligned} < Tf(x), g(x) >_\mu &= \sum_{x \in S} Tf(x)g(x)\mu(x) \\ &= \sum_{x \in S} \sum_{y \in S} p(x,y)f(y)g(x)\mu(x) \quad \text{with detailed balance condition} \\ &= \sum_{y \in S} \sum_{x \in S} p(y,x)f(y)g(x)\mu(y) \\ &= \sum_{y \in S} f(y)Tg(y)\mu(y) \\ &= < f(x), Tg(x) >_\mu \end{aligned}$$

That means T is self-adjoint. So there is a set of orthonormal basis $\{\phi_j(x)\}_{j=0}^{n-1}$ and a set of eigenvalue $\{\lambda_j\}_{j=0}^{n-1} \subset [-1, 1], 1 = \lambda_0 > \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{n-1}, s.t.\text{Prob}\phi_j = \lambda_j \phi_j, j = 0, 1, \ldots n-1,$and $< \phi_i, \phi_j >_\mu = \delta_{ij}, \forall i, j = 0, 1, \ldots n-1$.So $\phi_j(x)$ is right eigenvectors. The corresponding left eigenvectors are denoted by $\{\psi_j(x)\}_{j=0}^{n-1}$. One can obtain that $\psi_j(x) = \phi_j(x)\mu(x)$. In fact,because $T\phi_j = \lambda_j\phi_j$,

$$\begin{aligned}\mu(x)\sum_{y\in S} p(x,y)\phi_j(y) = \lambda_j\phi_j(x)\mu(x) \quad \text{with detailed balance condition}\\ \sum_{y\in S} p(y,x)\mu(y)\phi_j(y) = \lambda_j\phi_j(x)\mu(x) \quad \text{that is}\\ \psi_j\text{Prob}(x) = \sum_{y\in S} p(y,x)\phi(y) = \lambda_j(x)\psi(x)\end{aligned}$$

Generally, T has spectral decomposition

$$p(x,y) = \sum_{i=0}^{n-1}\lambda_i\psi_i(x)\phi(y) = \sum_{i=0}^{n-1} p(x,y)\phi_i(x)\phi_i(y)\mu(x)$$

Since P is a stochastic matrix, we have $\lambda_0 = 1$,the corresponding right eigenvector is $\phi_0(x) \equiv 1$,and left eigenvector is the invariant distribution $\psi_0(x) = \mu(x)$

5.2.2. *Optimal Reduction.* This section is by [**ELVE08**]. Suppose the number of states n is very large. The scale of Markov chain is so big that we want a smaller chain to present its behavior. That is, we want to decompose the state space S: Let $S = \bigcup_{i=1}^{N} S_i, s.t. N \ll n, S_i \bigcap S_j = \emptyset, \forall i \neq j$, and define a transition probability $\hat{P}$ on it. We want the Markov chain $(\{S_i\}, \hat{P})$ has similar property as chain (S, P).

We call $\{S_i\}$ coarse space. The first difficult we're facing is whether $(\{S_i\}, \hat{P})$ really Markovian. We want

$$\Pr(X_{i_{t+1}} \in S_{i_{t+1}} | x_{i_t} \in S_{i_t}, \ldots X_0 \in S_{i_0}) = \Pr(X_{i_{t+1}} \in S_{i_{t+1}} | x_{i_t} \in S_{i_t})$$

and this probability is independent of initial distribution. This property is so-called lumpability, which you can refer Lecture 9. Unfortunately, lumpability is a strick constraint that it seldom holds.

So we must modify our strategy of reduction. One choice is to do a optimization with some norm on $\mathcal{L}^2_\mu$. First, Let us introduce Hilbert-Schmidt norm on $\mathcal{L}^2_\mu$. Suppose F is an operator on $\mathcal{L}^2_\mu$, and $Ff(x) = \sum_{y\in S} K(x,y)f(y)\mu(y)$. Here K is called a kernel function. If K is symmetric, F is self adjoint. In fact,

$$\begin{aligned} < Ff(x), g(x) >_\mu &= \sum_{x\in S}\sum_{y\in S} K(x,y)f(y)\mu(y)g(x)\mu(x)\\ &= \sum_{y\in S}\sum_{x\in S} K(y,x)f(y)\mu(y)g(x)\mu(x)\\ &= < f(x), Fg(x) >_\mu\end{aligned}$$

So F guarantee a spectral decomposition. Let $\{\lambda_j\}_{j=0}^{n-1}$ denote its eigenvalue and $\{\phi_j(x)\}_{j=0}^{n-1}$denote its eigenvector, then $k(x,y)$ can be represented as $K(x,y) = \sum_{j=0}^{n-1}\lambda_j\phi_j(x)\phi_j(y)$. Hilbert-Schmidt norm of F is defined as follow:

$$\|F\|^2_{HS} = tr(F^*F) = tr(F^2) = \sum_{i=0}^{n-1}\lambda_i^2$$

One can check that $\|F\|_{HS}^2 = \sum\limits_{x,y\in S} K^2(x,y)\mu(x)\mu(y)$. In fact,

$$\begin{aligned} RHS &= \sum_{x,y\in S}\left(\sum_{j=0}^{n-1}\lambda_j\phi_j(x)\phi_j(y)\right)^2\mu(x)\mu(y) \\ &= \sum_{j=0}^{n-1}\sum_{k=0}^{n-1}\lambda_j\lambda_k\sum_{x,y\in S}\phi_j(x)\phi_k(x)\phi_j(y)\phi_k(y)\mu(x)\mu(y) \\ &= \sum_{j=0}^{n-1}\lambda_j^2 \end{aligned}$$

the last equal sign dues do the orthogonality of eigenvectors. It is clear that if $\mathcal{L}_\mu^2 = \mathcal{L}^2$, Hilbert-Schmidt norm is just Frobenius norm.

Now we can write our T as

$$Tf(x) = \sum_{y\in S} p(x,y)f(y) = \sum_{y\in S}\frac{p(x,y)}{\mu(y)}f(y)\mu(y)$$

and take $K(x,y) = \frac{p(x,y)}{\mu(y)}$. By detailed balance condition, K is symmetric. So

$$\|T\|_{HS}^2 = \sum_{x,y\in S}\frac{p^2(x,y)}{\mu^2(y)}\mu(x)\mu(y) = \sum_{x,y\in S}\frac{\mu(x)}{\mu(y)}p^2(x,y)$$

We'll rename $\|P\|_{HS}$ to $\|P\|_\mu$ in the following paragraphs.

Now go back to our reduction problem. Suppose we have a coarse space $\{S_i\}_{i=1}^N$, and a transition probability $\hat{P}(k,l), k,l = 1,2,\dots N$ on it. If we want to compare $(\{S_i\},\hat{P})$ with (S,P), we must "lift" the coarse process to fine space. One nature consideration is as follow: if $x\in S_k, y\in S_l$, first, we transit from x to S_l follow the rule $\hat{P}(k,l)$, and in S_l, we transit to y "randomly". To make "randomly" rigorously, one may choose the lifted transition probably as follow:

$$\tilde{P}(x,y) = \sum_{k,l=1}^N 1_{S_k}(x)\hat{P}(k,l)1_{S_l}(y)\frac{1}{|S_l|}$$

One can check that this $\tilde{P}$ is a stochastic matrix, but it is not reversible. One more convenient choice is transit "randomly" by invariant distribution:

$$\tilde{P}(x,y) = \sum_{k,l=1}^N 1_{S_k}(x)\hat{P}(k,l)1_{S_l}(y)\frac{\mu(y)}{\hat{\mu}(S_l)}$$

where

$$\hat{\mu}(S_l) = \sum_{z\in S_l}\mu(z)$$

Then you can check this matrix is not only a stochastic matrix, but detailed balance condition also hold provides $\hat{P}$ on $\{S_i\}$ is reversible.

Now let us do some summary. Given a decomposition of state space $S = \bigcup_{i=1}^N S_i$, and a transition probability $\hat{P}$ on coarse space, we may obtain a lifted transition probability $\tilde{P}$ on fine space. Now we can compare $(\{S_i\}, \hat{P})$ and (S, P) in a clear way: $\|P - \tilde{P}\|_\mu$. So our optimization problem can be defined clearly:

$$E = \min_{S_1 \dots S_N} \min_{\hat{P}} \|P - \hat{P}\|_\mu^2$$

That is, given a partition of S, find the optimal $\hat{P}$ to minimize $\|P - \hat{P}\|_\mu^2$, and find the optimal partition to minimize E.

5.2.3. *Community structure of complex network.* Given a partition $S = \bigcup_{k=1}^{N} S_k$, the solution of optimization problem

$$\min_{\hat{p}} \|p - \hat{p}\|_\mu^2$$

is

$$\hat{p}_{kl}^* = \frac{1}{\hat{\mu}(S_k)} \sum_{x \in S_k, y \in S_l} \mu(x) p(x, y)$$

It is easy to show that $\{\hat{p}_{kl}^*\}$ form a transition probability matrix with detailed balance condition:

$$\begin{aligned}
\hat{p}_{kl}^* &\geq 0 \\
\sum_l \hat{p}_{kl}^* &= \frac{1}{\hat{\mu}(S_k)} \sum_{x \in S_k} \mu(x) \sum_l \sum_{y \in S_l} p(x, y) \\
&= \frac{1}{\hat{\mu}(S_k)} \sum_{x \in S_k} \mu(x) = 1 \\
\hat{\mu}(S_k) \hat{p}_{kl}^* &= \sum_{x \in S_k, y \in S_l} \mu(x) p(x, y) \\
&= \sum_{x \in S_k, y \in S_l} \mu(y) p(y, x) \\
&= \hat{\mu}(S_l) \hat{p}_{lk}^*
\end{aligned}$$

The last equality implies that $\hat{\mu}$ is the invariant distribution of the reduced Markov chain. Thus we find the optimal transition probability in the coarse space. $\hat{p}^*$ has the following property

$$\|p - p^*\|_\mu^2 = \|p\|_\mu^2 - \|\hat{p}^*\|_{\hat{\mu}}^2$$

However, the partition of the original graph is not given in advance, so we need to minimize E^* with respect to all possible partitions. This is a combinatorial optimization problem, which is extremely difficult to find the exact solution. An effective approach to obtain an approximate solution, which inherits ideas of K-means clustering, is proposed as following: First we rewrite E^* as

$$\begin{aligned}
E^* &= \sum_{x, y \in S} \frac{\mu(x)}{\mu(y)} |p(x, y) - \sum_{k,l=1}^{N} 1_{S_k}(x) \frac{\hat{p}_{kl}^*}{\hat{\mu}(S_k)} 1_{S_l}(y) \mu(y)|^2 \\
&= \sum_{k,l=1}^{N} \sum_{x \in S_k, y \in S_l} \mu(x) \mu(y) \left| \frac{p(x, y)}{\mu(y)} - \frac{\hat{p}_{kl}^*}{\hat{\mu}(S_k)} \right|^2 \\
&\triangleq \sum_{k=1}^{N} \sum_{x \in S_k} E^*(x, S_k)
\end{aligned}$$

where

$$E^*(x, S_k) = \sum_{l=1}^{N} \sum_{y \in S_l} \mu(x)\mu(y) \left| \frac{p(x,y)}{\mu(y)} - \frac{\hat{p}^*_{kl}}{\hat{\mu}(S_k)} \right|^2$$

Based on above expression, a variation of K-means is designed:

E step: Fix partition $\overset{N}{\underset{k=1}{\cup}} S_k$, compute $\hat{p}^*$.

M step: Put x in $S_k^{(n+1)}$ such that

$$E^*(x, S_k) = \min_j E^*(x, S_j)$$

5.2.4. *Extensions: Fuzzy Partition.* This part is in [**LLE09**, **LL11**]. It is unnecessary to require that each vertex belong to a definite class. We introduce $\rho_k(x)$ as the probability of a vertex x belonging to class k, and we lift the Markov chain in coarse space to fine space using the following transition probability

$$\tilde{p}(x,y) = \sum_{k,l=1}^{N} \rho_k(x) \hat{p}_{kl} \rho_l(y) \frac{\mu(y)}{\hat{\mu}_l}$$

Now we solve

$$\min_{\hat{p}} \|p - \tilde{p}\|_\mu^2$$

to obtain a optimal reduction.

5.2.5. *Model selection.* Note the number of partition N should also not be given in advance. But in strategies similar to K-means, the value of minimal E^* is monotone decreasing with N. This means larger N is always preferred.

A possible approach is to introduce another quantity which is monotone increasing with N. We take K-means clustering for example. In K-means clustering, only compactness is reflected. If another quantity indicates separation of centers of each cluster, we can minimize the ratio of compactness and separation to find an optimal N.

6. Mean First Passage Time

Consider a Markov chain P on graph $G = (V, E)$. In this section we study the *mean first passage time* between vertices, which exploits the unnormalized graph Laplacian and will be useful for commute time map against diffusion map.

Definition.

(1) *First passage time (or hitting time)*: $\tau_{ij} := \inf(t \geq 0 | x_t = j, x_0 = i)$;
(2) *Mean First Passage Time*: $T_{ij} = \mathbb{E}_i \tau_{ij}$;
(3) $\tau_{ij}^+ := \inf(t > 0 | x_t = j, x_0 = i)$, where τ_{ii}^+ is also called *first return time*;
(4) $T_{ij}^+ = \mathbb{E}_i \tau_{ij}^+$, where T_{ii}^+ is also called *mean first return time.*

Here $\mathbb{E}_i$ denotes the conditional expectation with fixed initial condition $x_0 = i$.

Theorem 6.1. Assume that P is irreducible. Let $L = D - W$ be the unnormalized graph Laplacian with Moore-Penrose inverse $L^\dagger$, where $D = \mathrm{diag}(d_i)$ with $d_i = \sum_{j:j\ \ i} W_{ij}$ being the degree of node i. Then

(1) Mean First Passage Time is given by

$$T_{ii} = 0,$$
$$T_{ij} = \sum_k L^\dagger_{ik} d_k - L^\dagger_{ij} vol(G) + L^\dagger_{jj} vol(G) - \sum_k L^\dagger_{jk} d_k, \quad i \neq j.$$

(2) Mean First Return Time is given by

$$T^+_{ii} = \frac{1}{\pi_i}, \quad T^+_{ij} = T_{ij}.$$

PROOF. Since P is irreducible, then the stationary distribution is unique, denoted by π. By definition, we have

$$T^+_{ij} = P_{ij} \cdot 1 + \sum_{k \neq j} P_{ik}(T^+_{kj} + 1) \tag{88}$$

Let $E = 1 \cdot 1^T$ where $1 \in \mathbb{R}^n$ is a vector with all elements one, $T^+_d = \mathrm{diag}(T^+_{ii})$. Then 127 becomes

$$T^+ = E + P(T^+ - T^+_d). \tag{89}$$

For the unique stationary distribution π, $\pi^T P = P$, whence we have

$$\begin{array}{rcl} \pi^T T^+ & = & \pi^T 1 \cdot 1^T + \pi^T P(T^+ - T^+_d) \\ \pi^T T^+ & = & 1^T + \pi^T T^+ - \pi^T T^+_d \\ 1 & = & T^+_d \pi \\ T^+_{ii} & = & \dfrac{1}{\pi_i} \end{array}$$

Before proceeding to solve equation (127), we first show its solution is unique.

Lemma 6.2. *P is irreducible $\Rightarrow$ T^+ and T are both unique.*

PROOF. Assume S is also a solution of equation (128), then

$$(I - P)S = E - P\mathrm{diag}(1/\pi_i) = (I - P)T^+$$

$$\Leftrightarrow ((I - P)(T^+ - S) = 0.$$

Therefore for irreducible P, S and T^+ must satisfy

$$\left\{ \begin{array}{rcl} \mathrm{diag}(T^+ - S) & = & 0 \\ T^+ - S & = & 1u^T, \quad \forall u \end{array} \right.$$

which implies $T^+ = S$. T's uniqueness follows from $T = T^+ - T^+_d$. □

Now we continue with the proof of the main theorem. Since $T = T^+ - T^+_d$, then (127) becomes

$$\begin{array}{rcl} T & = & E + PT - T^+_d \\ (I - P)T & = & E - T^+_d \\ (I - D^{-1}W)T & = & F \\ (D - W)T & = & DF \\ LT & = & DF \end{array}$$

where $F = E - T^+_d$ and $L = D - W$ is the (unnormalized) graph Laplacian. Since L is symmetric and irreducible, we have $L = \sum_{k=1}^n \mu_k \nu_k \nu_k^T$, where $0 = \mu_1 < \mu_2 \leq \cdots \leq \mu_n, \nu_1 = 1/||1||, \nu_k^T \nu_l = \delta_{kl}$. Let $L^\dagger = \sum_{k=2}^n \frac{1}{\mu_k} \nu_k \nu_k^T$, $L^\dagger$ is called the *pseudo-inverse* (or *Moore-Penrose inverse*) of L. We can test and verify $L^\dagger$ satisfies the following four conditions

$$\begin{cases} L^\dagger L L^\dagger & = & L^\dagger \\ L L^\dagger L & = & L \\ (L L^\dagger)^T & = & L L^\dagger \\ (L^\dagger L)^T & = & L^\dagger L \end{cases}$$

From $LT = D(E - T_d^+)$, multiplying both sides by $L^\dagger$ leads to

$$T = L^\dagger D E - L^\dagger D T_d^+ + 1 \cdot u^T,$$

as $1 \cdot u^T \in \ker(L)$, whence

$$\begin{aligned} T_{ij} &= \sum_{k=1}^n L_{ik}^\dagger d_k - L_{ij}^\dagger d_j \cdot \frac{1}{\pi_j} + u_j \\ u_i &= -\sum_{k=1}^n L_{ik}^\dagger d_k + L_{ii}^\dagger vol(G), \quad j = i \\ T_{ij} &= \sum_k L_{ik}^\dagger d_k - L_{ij}^\dagger vol(G) + L_{jj}^\dagger vol(G) - \sum_k L_{jk}^\dagger d_k \end{aligned}$$

Note that $vol(G) = \sum_i d_i$ and $\pi_i = d_i / vol(G)$ for all i. □

As $L^\dagger$ is a positive definite matrix, this leads to the following corollary.

Corollary 6.3.

$$T_{ij} + T_{ji} = vol(G)(L_{ii}^\dagger + L_{jj}^\dagger - 2L_{ij}^\dagger). \tag{90}$$

Therefore the average commute time between i and j leads to an Euclidean distance metric

$$d_c(x_i, x_j) := \sqrt{T_{ij} + T_{ji}}$$

often called *commute time distance*.

7. Transition Path Theory

The transition path theory was originally introduced in the context of continuous-time Markov process on continuous state space [**EVE06**] and discrete state space [**MSVE09**], see [**EVE10**] for a review. Another description of discrete transition path theory for molecular dynamics can be also found in [**NSVE$^+$09**]. The following material is adapted to the setting of discrete time Markov chain with transition probability matrix P [**?**]. We assume reversibility in the following presentation, which can be extended to non-reversible Markov chains.

Assume that an irreducible Markov Chain on graph $G = (V, E)$ admits the following decomposition $P = D^{-1}W = \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$. Here $V_l = V_0 \cup V_1$ denotes the labeled vertices with source set V_0 (e.g. reaction state in chemistry) and sink set V_1 (e.g. product state in chemistry), and V_u is the unlabeled vertex set (intermediate states). That is,

- $V_0 = \{i \in V_l : f_i = f(x_i) = 0\}$
- $V_1 = \{i \in V_l : f_i = f(x_i) = 1\}$
- $V = V_0 \cup V_1 \cup V_u$ where $V_l = V_0 \cup V_1$

Given two sets V_0 and V_1 in the state space V, the transition path theory tells how these transitions between the two sets happen (mechanism, rates, etc.). If we view V_0 as a reactant state and V_1 as a product state, then one transition from V_0 to V_1 is a reaction event. The reactve trajectories are those part of the equilibrium trajectory that the system is going from V_0 to V_1.

Let the hitting time of V_l be

$$\tau_i^k = \inf\{t \geq 0 : x(0) = i, x(t) \in V_k\}, \quad k = 0, 1.$$

The central object in transition path theory is the committor function. Its value at $i \in V_u$ gives the probability that a trajectory starting from i will hit the set V_1 first than V_0, *i.e.*, the success rate of the transition at i.

Proposition 7.1. For $\forall i \in V_u$, define the *committor function*

$$q_i := Prob(\tau_i^1 < \tau_i^0) = Prob(\text{trajectory starting from } x_i \text{ hit } V_1 \text{ before } V_0)$$

which satisfies the following Laplacian equation with Dirichlet boundary conditions

$$(Lq)(i) = [(I - P)q](i) = 0, \quad i \in V_u$$
$$q_{i \in V_0} = 0, q_{i \in V_1} = 1.$$

The solution is

$$q_u = (D_u - W_{uu})^{-1} W_{ul} q_l.$$

PROOF. By definition,

$$q_i = Prob(\tau_i^1 < \tau_i^0) = \begin{cases} 1 & x_i \in V_1 \\ 0 & x_i \in V_0 \\ \sum_{j \in V} P_{ij} q_j & i \in V_u \end{cases}$$

This is because $\forall i \in V_u$,

$$\begin{aligned} q_i &= Pr(\tau_{iV_1} < \tau_{iV_0}) \\ &= \sum_j P_{ij} q_j \\ &= \sum_{j \in V_1} P_{ij} q_j + \sum_{j \in V_0} P_{ij} q_j + \sum_{j \in V_u} P_{ij} q_j \\ &= \sum_{j \in V_1} P_{ij} + \sum_{j \in V_u} P_{ij} q_j \end{aligned}$$

$$\therefore \quad q_u = P_{ul} q_l + P_{uu} q_u = D_u^{-1} W_{ul} q_l + D_u^{-1} W_{uu} q_u$$

multiply D_u to both side and reorganize

$$(D_u - W_{uu}) q_u = W_{ul} q_l$$

If $D_u - W_{uu}$ is reversible, we get

$$q_u = (D_u - W_{uu})^{-1} W_{ul} q_l.$$

□

The committor function provides natural decomposition of the graph. If $q(x)$ is less than 0.5, i is more likely to reach V_0 first than V_1; so that $\{i \mid q(x) < 0.5\}$ gives the set of points that are more attached to set V_0.

Once the committor function is given, the statistical properties of the reaction trajectories between V_0 and V_1 can be quantified. We state several propositions characterizing transition mechanism from V_0 to V_1. The proof of them is an easy adaptation of [**EVE06, MSVE09**] and will be omitted.

Proposition 7.2 (Probability distribution of reactive trajectories)**.** The probability distribution of reactive trajectories

$$\pi_R(x) = \mathbb{P}(X_n = x, n \in R) \tag{91}$$

is given by

$$\pi_R(x) = \pi(x)q(x)(1 - q(x)). \tag{92}$$

The distribution π_R gives the equilibrium probability that a reactive trajectory visits x. It provides information about the proportion of time the reactive trajectories spend in state x along the way from V_0 to V_1.

Proposition 7.3 (Reactive current from V_0 to V_1)**.** The reactive current from A to B, defined by

$$J(xy) = \mathbb{P}(X_n = x, X_{n+1} = y, \{n, n+1\} \subset R), \tag{93}$$

is given by

$$J(xy) = \begin{cases} \pi(x)(1 - q(x))P_{xy}q(y), & x \neq y; \\ 0, & \text{otherwise.} \end{cases} \tag{94}$$

The reactive current $J(xy)$ gives the average rate the reactive trajectories jump from state x to y. From the reactive current, we may define the effective reactive current on an edge and transition current through a node which characterizes the importance of an edge and a node in the transition from A to B, respectively.

Definition. The *effective current* of an edge xy is defined as

$$J^+(xy) = \max(J(xy) - J(yx), 0). \tag{95}$$

The *transition current* through a node $x \in V$ is defined as

$$T(x) = \left\{ \begin{array}{ll} \sum_{y \in V} J^+(xy), & x \in A \\ \sum_{y \in V} J^+(yx), & x \in B \\ \sum_{y \in V} J^+(xy) = \sum_{y \in V} J^+(yx), & x \notin A \cup B \end{array} \right. \tag{96}$$

In applications one often examines partial transition current through a node connecting two communities $V^- = \{x : q(x) < 0.5\}$ and $V^+ = \{x : q(x) \geq 0.5\}$, *e.g.* $\sum_{y \in V^+} J^+(xy)$ for $x \in V^-$, which shows relative importance of the node in bridging communities.

The reaction rate ν, defined as the number of transitions from V_0 to V_1 happened in a unit time interval, can be obtained from adding up the probability current flowing out of the reactant state. This is stated by the next proposition.

Proposition 7.4 (Reaction rate)**.** The reaction rate is given by

$$\nu = \sum_{x \in A, y \in V} J(xy) = \sum_{x \in V, y \in B} J(xy). \tag{97}$$

Finally, the committor functions also give information about the time proportion that an equilibrium trajectory comes from A (the trajectory hits A last rather than B).

Proposition 7.5. The proportion of time that the trajectory comes from A (resp. from B) is given by

$$\rho^A = \sum_{x\in V} \pi(x) q(x), \quad \rho^B = \sum_{x\in V} \pi(x)(1 - q(x)). \tag{98}$$

CHAPTER 7

Diffusion Map

Finding meaningful low-dimensional structures hidden in high-dimensional observations is an fundamental task in high-dimensional statistics. The classical techniques for dimensionality reduction, principal component analysis (PCA) and multidimensional scaling (MDS), guaranteed to discover the true structure of data lying on or near a linear subspace of the high-dimensional input space. PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space. Classical MDS finds an embedding that preserves the interpoint distances, equivalent to PCA when those distances are Euclidean [**TdL00**]. However, these linear techniques cannot adequately handle complex nonlinear data. Recently more emphasis is put on detecting non-linear features in the data. For example, ISOMAP [**TdL00**] *etc.* extends MDS by incorporating the geodesic distances imposed by a weighted graph. It defines the geodesic distance to be the sum of edge weights along the shortest path between two nodes. The top n eigenvectors of the geodesic distance matrix are used to represent the coordinates in the new n-dimensional Euclidean space. Nevertheless, as mentioned in [**EST09**], in practice robust estimation of geodesic distance on a manifold is an awkward problem that require rather restrictive assumptions on the sampling. Moreover, since the MDS step in the ISOMAP algorithm intends to preserve the geodesic distance between points, it provides a correct embedding if submanifold is isometric to a convex open set of the subspace. If the submanifold is not convex, then there exist a pair of points that can not be joined by a straight line contained in the submanifold. Therefore,their geodesic distance can not be equal to the Euclidean distance. Diffusion maps [**CLL$^+$05**] leverages the relationship between heat diffusion and a random walk (Markov Chain); an analogy is drawn between the diffusion operator on a manifold and a Markov transition matrix operating on functions defined on a weighted graph whose nodes were sampled from the manifold. A diffusion map, which maps coordinates between data and diffusion space, aims to re-organize data according to a new metric. In this class, we will discuss this very metric-diffusion distance and it's related properties.

1. Diffusion map and Diffusion Distance

Viewing the data points x_1,x_2,...,x_n as the nodes of a weighted undirected graph $G = (V, E_W)(W = (W_{ij}))$, where the weight W_{ij} is a measure of the similarity between x_i and x_j. There are many ways to define W_{ij}, such as:

(1) **Heat kernel**. If x_i and x_j are connected, put:

$$W_{ij}^{\varepsilon} = e^{\frac{-\|x_i - x_j\|^2}{\varepsilon}} \tag{99}$$

with some positive parameter $\varepsilon \in \mathbb{R}_0^+$.

(2) **Cosine Similarity**

$$W_{ij} = \cos(\angle(x_i, x_j)) = \frac{x_i}{\|x_i\|} \cdot \frac{x_j}{\|x_j\|} \tag{100}$$

(3) **Kullback-Leibler divergence**. Assume x_i and x_j are two nonvanishing probability distribution, *i.e.* $\sum_k x_i^k = 1$ and $x_i^k > 0$. Define *Kullback-Leibler divergence*

$$D^{(KL)}(x_i||x_j) = \sum_k x_i^{(k)} \log \frac{x_i^{(k)}}{x_j^{(k)}}$$

and its symmetrization $\bar{D} = D^{(KL)}(x_i||x_j) + D^{KL}(x_j||x_i)$, which measure a kind of 'distance' between distributions; *Jensen-Shannon divergence* as the symmetrization of KL-divergence between one distribution and their average,

$$D^{(JS)}(x_i, x_j) = D^{(KL)}(x_i||(x_i + x_j)/2) + D^{(KL)}(x_j||(x_i + x_j)/2)$$

A similarity kernel can be

$$W_{ij} = -D^{(KL)}(x_i||x_j) \tag{101}$$

or

$$W_{ij} = -D^{(JS)}(x_i, x_j) \tag{102}$$

The similarity functions are widely used in various applications. Sometimes the matrix W is positive semi-definite (psd), that for any vector $x \in \mathbb{R}^n$,

$$x^T W x \geq 0. \tag{103}$$

PSD kernels includes heat kernels, cosine similarity kernels, and JS-divergence kernels. But in many other cases (e.g. KL-divergence kernels), similarity kernels are not necessarily PSD. For a PSD kernel, it can be understood as a generalized covariance function; otherwise, diffusions as random walks on similarity graphs will be helpful to disclose their structures.

Define $A := D^{-1}W$, where $D = \mathrm{diag}(\sum_{j=1}^{n} W_{ij}) \triangleq \mathrm{diag}(d_1, d_2, \cdots, d_n)$ for symmetric $W_{ij} = W_{ji} \geq 0$. So

$$\sum_{j=1}^{n} A_{ij} = 1 \ \forall i \in \{1, 2, \cdot \cdot \cdot, n\} \ (A_{ij} \geq 0) \tag{104}$$

whence A is a row Markov matrix of the following discrete time Markov chain $\{X_t\}_{t \in N}$ satisfying

$$P(X_{t+1} = x_j \mid X_t = x_i) = A_{ij}. \tag{105}$$

1.1. Spectral Properties of A. We may reach a spectral decomposition of A with the aid of the following symmetric matrix S which is similar to A. Let

$$S := D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{106}$$

which is symmetric and has an eigenvalue decomposition

$$S = V \Lambda V^T, \qquad \text{where } VV^T = I_n, \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \tag{107}$$

So

$$A = D^{-1} W = D^{-1}(D^{\frac{1}{2}} S D^{\frac{1}{2}}) = D^{-\frac{1}{2}} S D^{\frac{1}{2}}$$

which is similar to S, whence sharing the same eigenvalues as S. Moreover

$$A = D^{-\frac{1}{2}} V \Lambda V^T D^{\frac{1}{2}} = \Phi \Lambda \Psi^T \tag{108}$$

where $\Phi = D^{-\frac{1}{2}} V$ and $\Psi = D^{\frac{1}{2}} V$ give right and left eigenvectors of A respectively, $A\Phi = \Phi\Lambda$ and $\Psi^T A = \Lambda \Psi^T$, and satisfy $\Psi^T \Phi = I_n$.

The Markov matrix A satisfies the following properties by Perron-Frobenius Theory.

Proposition 1.1. (1) A has eigenvalues $\lambda(A) \subset [-1, 1]$.

(2) A is irreducible, if and only if $\forall (i,j)\ \exists t\ s.t.\ (A^t)_{ij} > 0 \Leftrightarrow$ Graph $G = (V, E)$ is connected
(3) A is irreducible $\Rightarrow \lambda_{\max} = 1$
(4) A is primitive, if and only if $\exists t > 0\ s.t.\ \forall (i,j)\ (A^t)_{ij} > 0 \Leftrightarrow$ Graph $G = (V, E)$ is path-t connected, i.e. any pair of nodes are connected by a path of length no more than t
(5) A is irreducible and $\forall i$, $A_{ii} > 0 \Rightarrow$ A is primitive
(6) A is primitive $\Rightarrow -1 \notin \lambda(A)$
(7) W_{ij} is induced from the heat kernel, or any positive definite function $\Rightarrow \lambda(A) \geq 0$

Proof. (1) 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:

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

(7) Let $S = D^{-1/2} W D^{-1/2}$. As W is positive semi-definite, so S has eigenvalues $\lambda(S) \geq 0$. Note that $A = D^{-1/2} S D^{1/2}$, i.e. similar to S, whence A shares the same eigenvalues with S. □

Sort the eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq -1$. Denote $\Phi = [\phi_1, \ldots, \phi_n]$ and $\Psi = [\psi_1, \ldots, \psi_n]$. So the primary (first) right and left eigenvectors are

$$\phi_1 = \mathbf{1},$$

$$\psi_1 = \pi$$

as the stationary distribution of the Markov chain, respectively.

1.2. Diffusion Map and Distance. Diffusion map of a point x is defined as the weighted Euclidean embedding via right eigenvectors of Markov matrix A. From the interpretation of the matrix A as a Markov transition probability matrix

$$A_{ij} = Pr\{s(t+1) = x_j | s(t) = x_i\} \tag{109}$$

it follows that

$$A_{ij}^t = Pr\{s(t+1) = x_j | s(0) = x_i\} \tag{110}$$

We refer to the i'th row of the matrix A^t, denoted $A_{i,*}^t$, as the transition probability of a t-step random walk that starts at x_i. We can express A^t using the decomposition of A. Indeed, from

$$A = \Phi \Lambda \Psi^T \tag{111}$$

with $\Psi^T \Phi = I$, we get

$$A^t = \Phi \Lambda^t \Psi^T. \tag{112}$$

Written in a component-wise way, this is equivalent to

$$A_{ij}^t = \sum_{k=1}^n \lambda_k^t \phi_k(i) \psi_k(j). \tag{113}$$

Therefore Φ and Ψ are right and left eigenvectors of A^t, respectively.

Let the diffusion map $\Phi_t : V \mapsto \mathbb{R}^n$ at scale t be

$$\Phi_t(x_i) := \begin{pmatrix} \lambda_1^t \phi_1(i) \\ \lambda_2^t \phi_2(i) \\ \vdots \\ \lambda_n^t \phi_n(i) \end{pmatrix} \tag{114}$$

The mapping of points onto the diffusion map space spanned the right eigenvectors of the row Markov matrix has a well defined probabilistic meaning in terms of the random walks. Lumpable Markov chains with Piece-wise constant right eigenvectors thus help us understand the behavior of diffusion maps and distances in such cases.

The diffusion distance is defined to be the Euclidean distances between embedded points,

$$d_t(x_i, x_j) := \|\Phi_t(x_i) - \Phi_t(x_j)\|_{\mathbb{R}^n} = \left(\sum_{k=1}^n \lambda_k^{2t} (\phi_k(i) - \phi_k(j))^2 \right)^{1/2}. \tag{115}$$

The main intuition to define diffusion distance is to describe "perceptual distances" of points in the same and different clusters. For example Figure 1 shows that points within the same cluster have small diffusion distances while in different clusters have large diffusion distances. This is because the metastability phenomenon of random walk on graphs where each cluster represents a metastable state. The main properties of diffusion distances are as follows.

- Diffusion distances reflect average path length connecting points via random walks.
- Small t represents local random walk, where diffusion distances reflect local geometric structure.
- Large t represents global random walk, where diffusion distances reflect large scale cluster or connected components.

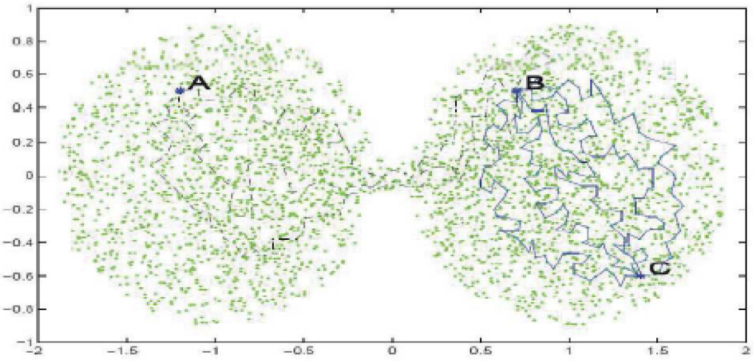


FIGURE 1. Diffusion Distances $d_t(A,B) >> d_t(B,C)$ while graph shortest path $d_{geod}(A,B) \sim d_{geod}(B,C)$.

1.3. 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. Let

$$k_\varepsilon(x,y) = \exp\left(-\frac{\|x-y\|^2}{\varepsilon}\right),$$

$W_{ij}^\varepsilon = k_\varepsilon(x_i, x_j)$ and $A_\varepsilon = D^{-1}W^\varepsilon$ where $D = \text{diag}(\sum_j W_{ij}^\varepsilon)$. Define a graph Laplacian (recall that $L = D^{-1}A - I$)

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

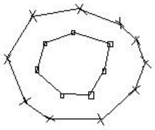


FIGURE 2. Two circles

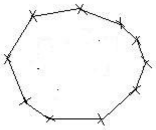


FIGURE 3. EX2 single circle

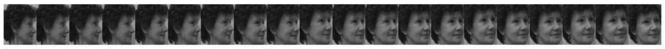


FIGURE 4. Order the face

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

Where $V(s)$ is the Gibbs free energy 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

$$V(s) = const \Rightarrow \phi^{''}(s) = 2\lambda\phi(s) \Rightarrow \phi_k(s) = \cos(k\pi s), 2\lambda_k = -k^2\pi^2 \tag{116}$$

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 $\phi_1(s)$.

1.4. Properties of Diffusion Distance.

Lemma 1.2. The diffusion distance is equal to a ℓ^2 distance between the probability clouds $A^t_{i,*}$ and $A^t_{j,*}$ with weights $1/d_l$,i.e.,

$$d_t(x_i, x_j) = \|A^t_{i,*} - A^t_{j,*}\|_{\ell^2(\mathbb{R}^n,1/d)} \tag{117}$$

[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 $(0, 1)$.

PROOF.

$$\begin{aligned}
\|A^t_{i,*} - A^t_{j,*}\|^2_{\ell^2(\mathbb{R}^n,1/d)} &= \sum_{l=1}^n (A^t_{il} - A^t_{jl})^2 \frac{1}{d_l} \\
&= \sum_{l=1}^n [\sum_{k=1}^n \lambda_k^t \phi_k(i)\psi_k(l) - \lambda_k^t \phi_k(j)\psi_k(l)]^2 \frac{1}{d_l} \\
&= \sum_{l=1}^n \sum_{k,k'}^n \lambda_k^t (\phi_k(i) - \phi_k(j))\psi_k(l)\lambda_{k'}^t(\phi_{k'}(i) - \phi_{k'}(j))\psi_{k'}(l)\frac{1}{d_l} \\
&= \sum_{k,k'}^n \lambda_k^t \lambda_{k'}^t (\phi_k(i) - \phi_k(j))(\phi_{k'}(i) - \phi_{k'}(j)) \sum_{l=1}^n \frac{\psi_k(l)\psi_{k'}(l)}{d_l} \\
&= \sum_{k,k'}^n \lambda_k^t \lambda_{k'}^t (\phi_k(i) - \phi_k(j))(\phi_{k'}(i) - \phi_{k'}(j))\delta_{kk'} \\
&= \sum_{k=1}^n \lambda_k^{2t}(\phi_k(i) - \phi_k(j))^2 \\
&= d_t^2(x_i, x_j)
\end{aligned}$$

□

In practice we usually do not use the mapping Φ_t but rather the truncate diffusion map Φ_t^δ that makes use of fewer than n coordinates. Specifically, Φ_t^δ uses only the eigenvectors for which the eigenvalues satisfy $|\lambda_k|^t > \delta$. When t is enough large, we can use the truncated diffusion distance:

$$d_t^\delta(x_i, x_j) = \|\Phi_t^\delta(x_i) - \Phi_t^\delta(x_j)\| = [\sum_{k:|\lambda_k|^t > \delta} \lambda_k^{2t}(\phi_k(i) - \phi_k(j))^2]^{\frac{1}{2}} \tag{118}$$

as an approximation of the weighted ℓ^2 distance of the probability clouds. We now derive a simple error bound for this approximation.

Lemma 1.3 (Truncated Diffusion Distance)**.** The truncated diffusion distance satisfies the following upper and lower bounds.

$$d_t^2(x_i, x_j) - \frac{2\delta^2}{d_{min}}(1 - \delta_{ij}) \leq [d_t^\delta(x_i, x_j)]^2 \leq d_t^2(x_i, x_j),$$

where $d_{min} = \min_{1\leq i\leq n} d_i$ with $d_i = \sum_j W_{ij}$.

PROOF. Since, $\Phi = D^{-\frac{1}{2}}V$, where V is an orthonormal matrix ($VV^T = V^TV = I$), it follows that

$$\Phi\Phi^T = D^{-\frac{1}{2}}VV^TD^{-\frac{1}{2}} = D^{-1} \tag{119}$$

Therefore,

$$\sum_{k=1}^n \phi_k(i)\phi_k(j) = (\Phi\Phi^T)_{ij} = \frac{\delta_{ij}}{d_i} \tag{120}$$

and

$$\sum_{k=1}^n (\phi_k(i) - \phi_k(j))^2 = \frac{1}{d_i} + \frac{1}{d_j} - \frac{2\delta_{ij}}{d_i} \tag{121}$$

clearly,

$$\sum_{k=1}^{n}(\phi_k(i)-\phi_k(j))^2 \leq \frac{2}{d_{min}}(1-\delta_{ij}),\ forall\ i,j=1,2,\cdots,n \tag{122}$$

As a result,

$$\begin{aligned}
[d_t^\delta(x_i,x_j)]^2 &= d_t^2(x_i,x_j) - \sum_{k:|\lambda_k|^t<\delta}\lambda_k^{2t}(\phi_k(i)-\phi_k(j))^2 \\
&\geq d_t^2(x_i,x_j) - \delta^2\sum_{k:|\lambda_k|^t<\delta}(\phi_k(i)-\phi_k(j))^2 \\
&\geq d_t^2(x_i,x_j) - \delta^2\sum_{k=1}^{n}(\phi_k(i)-\phi_k(j))^2 \\
&\geq d_t^2(x_i,x_j) - \frac{2\delta^2}{d_{min}}(1-\delta_{ij})
\end{aligned}$$

on the other hand, it is clear that

$$[d_t^\delta(x_i,x_j)]^2 \leq d_t^2(x_i,x_j) \tag{123}$$

We conclude that

$$d_t^2(x_i,x_j) - \frac{2\delta^2}{d_{min}}(1-\delta_{ij}) \leq [d_t^\delta(x_i,x_j)]^2 \leq d_t^2(x_i,x_j) \tag{124}$$

□

Therefore, for small δ the truncated diffusion distance provides a very good approximation to the diffusion distance. Due to the fast decay of the eigenvalues, the number of coordinates used for the truncated diffusion map is usually much smaller than n, especially when t is large.

1.5. Is the diffusion distance really a distance? A distance function $d: X\times X\to\mathbb{R}$ must satisfy the following properties:

(1) Symmetry: $d(x,y)=d(y,x)$
(2) Non-negativity: $d(x,y)\geq 0$
(3) Identity of indiscernibles: $d(x,y)=0\Leftrightarrow x=y$
(4) Triangle inequality: $d(x,z)+d(z,y)\geq d(x,y)$

Since the diffusion map is an embedding into the Euclidean space $\mathbb{R}^n$, the diffusion distance inherits all the metric properties of $\mathbb{R}^n$ such as symmetry, non-negativity and the triangle inequality. The only condition that is not immediately implied is $d_t(x,y)=0\Leftrightarrow x=y$. Clearly, $x_i=x_j$ implies that $d_t(x_i,x_j)=0$. But is it true that $d_t(x_i,x_j)=0$ implies $x_i=x_j$? Suppose $d_t(x_i,x_j)=0$, Then,

$$0=d_t^2(x_i,x_j)=\sum_{k=1}^{n}\lambda_k^{2t}(\phi_k(i)-\phi_k(j))^2 \tag{125}$$

It follows that $\phi_k(i)=\phi_k(j)$ for all k with $\lambda_k\neq 0$. But there is still the possibility that $\phi_k(i)\neq\phi_k(j)$ for k with $\lambda_k=0$. We claim that this can happen only whenever i and j have the exact same neighbors and proportional weights, that is:

Proposition 1.4. The situation $d_t(x_i, x_j) = 0$ with $x_i \neq x_j$ occurs if and only if node i and j have the exact same neighbors and proportional weights

$$W_{ik} = \alpha W_{jk}, \alpha > 0, for\ all\ k \in V.$$

Proof. (Necessity) If $d_t(x_i, x_j) = 0$, then $\sum_{k=1}^{n} \lambda_k^{2t}(\phi_k(i) - \phi_k(j))^2 = 0$ and $\phi_k(i) = \phi_k(j)$ for k with $\lambda_k \neq 0$ This implies that $d_{t'}(x_i, x_j) = 0$ for all t', because

$$d_{t'}(x_i, x_j) = \sum_{k=1}^{n} \lambda_k^{2t'}(\phi_k(i) - \phi_k(j)^2 = 0. \tag{126}$$

In particular, for $t' = 1$, we get $d_1(x_i, x_j) = 0$. But

$$d_1(x_i, x_j) = \|A_{i,*} - A_{j,*}\|_{\ell^2(\mathbb{R}^n, 1/d)},$$

and since $\|\cdot\|_{\ell^2(\mathbb{R}^n, 1/d)}$ is a norm, we must have $A_{i,*} = A_{j,*}$, which implies for each $k \in V$,

$$\frac{W_{ik}}{d_i} = \frac{W_{jk}}{d_j}, \quad \forall k \in V$$

whence $W_{ik} = \alpha W_{jk}$ where $\alpha = d_i/d_j$, as desired.

(Sufficiency) If $A_{i,*} = A_{j,*}$, then $0 = \sum_{k=1}^{n} (A_{i,k} - A_{j,k})^2/d_k = d_1^2(x_i, x_j) == \sum_{k=1}^{n} \lambda_k^2(\phi_k(i) - \phi_k(j))^2$ and therefore $\phi_k(i) = \phi_k(j)$ for k with $\lambda_k \neq 0$, from which it follows that $d_t(x_i, x_j) = 0$ for all t. □

Example 7. In a graph with three nodes $V = \{1, 2, 3\}$ and two edges, say $E = \{(1, 2), (2, 3)\}$, the diffusion distance between nodes 1 and 3 is 0. Here the transition matrix is

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}.$$

2. Commute Time Map and Distance

Diffusion distance depends on time scale parameter t which is hard to select in applications. In this section we introduce another closely related distance, namely *commute time distance*, derived from *mean first passage time* between points. For such distances we do not need to choose the time scale t.

Definition.

(1) *First passage time (or hitting time)*: $\tau_{ij} := \inf(t \geq 0 | x_t = j, x_0 = i)$;
(2) *Mean First Passage Time*: $T_{ij} = \mathbb{E}_i \tau_{ij}$;
(3) $\tau_{ij}^+ := \inf(t > 0 | x_t = j, x_0 = i)$, where τ_{ii}^+ is also called *first return time*;
(4) $T_{ij}^+ = \mathbb{E}_i \tau_{ij}^+$, where T_{ii}^+ is also called *mean first return time.*

Here $\mathbb{E}_i$ denotes the conditional expectation with fixed initial condition $x_0 = i$.

All the below will show that the (average) commute time between x_i and x_j, *i.e.*$T_{ij} + T_{ji}$, in fact leads to an Euclidean distance metric which can be used for embedding.

Theorem 2.1. $d_c(x_i, x_j) := \sqrt{T_{ij} + T_{ji}}$ is an Euclidean distance metric, called *commute time distance.*

Proof. For simplicity, we will assume that P is irreducible such that the stationary distribution is unique. We will give a constructive proof that $T_{ij}+T_{ji}$ is a squared distance of some Euclidean coordinates for x_i and x_j.

By definition, we have

$$T^{+}_{ij} = P_{ij}\cdot 1 + \sum_{k\neq j} P_{ik}(T^{+}_{kj}+1) \tag{127}$$

Let $E = 1\cdot 1^T$ where $1\in\mathbb{R}^n$ is a vector with all elements one, $T^{+}_{d} = \text{diag}(T^{+}_{ii})$. Then 127 becomes

$$T^{+} = E + P(T^{+} - T^{+}_{d}). \tag{128}$$

For the unique stationary distribution π, $\pi^T P = P$, whence we have

$$\begin{aligned} \pi^T T^{+} &= \pi^T 1\cdot 1^T + \pi^T P(T^{+}-T^{+}_{d}) \\ \pi^T T^{+} &= 1^T + \pi^T T^{+} - \pi^T T^{+}_{d} \\ 1 &= T^{+}_{d}\pi \\ T^{+}_{ii} &= \frac{1}{\pi_i} \end{aligned}$$

Before proceeding to solve equation (127), we first show its solution is unique.

Lemma 2.2. *P is irreducible $\Rightarrow$ T^{+} and T are both unique.*

Proof. Assume S is also a solution of equation (128), then

$$(I-P)S = E - P\text{diag}(1/\pi_i) = (I-P)T^{+}$$

$$\Leftrightarrow ((I-P)(T^{+}-S) = 0.$$

Therefore for irreducible P, S and T^{+} must satisfy

$$\begin{cases} \text{diag}(T^{+}-S) &= 0 \\ T^{+}-S &= 1u^T, \quad \forall u \end{cases}$$

which implies $T^{+}=S$. T's uniqueness follows from $T = T^{+} - T^{+}_{d}$. □

Now we continue with the proof of the main theorem. Since $T = T^{+} - T^{+}_{d}$, then (127) becomes

$$\begin{aligned} T &= E + PT - T^{+}_{d} \\ (I-P)T &= E - T^{+}_{d} \\ (I - D^{-1}W)T &= F \\ (D-W)T &= DF \\ LT &= DF \end{aligned}$$

where $F = E - T^{+}_{d}$ and $L = D - W$ is the (unnormalized) graph Laplacian. Since L is symmetric and irreducible, we have $L = \sum_{k=1}^{n}\mu_k\nu_k\nu_k^T$, where $0=\mu_1<\mu_2\leq\cdots\leq\mu_n, \nu_1 = 1/||1||, \nu_k^T\nu_l = \delta_{kl}$. Let $L^{+} = \sum_{k=2}^{n}\frac{1}{\mu_k}\nu_k\nu_k^T$, L^{+} is called the *pseudo-inverse* (or *Moore-Penrose inverse*) of L. We can test and verify L^{+} satisfies the following four conditions

$$\begin{cases} L^{+}LL^{+} &= L^{+} \\ LL^{+}L &= L \\ (LL^{+})^T &= LL^{+} \\ (L^{+}L)^T &= L^{+}L \end{cases}$$

From $LT = D(E - T_d^+)$, multiplying both sides by L^+ leads to

$$T = L^+ DE - L^+ DT_d^+ + 1 \cdot u^T,$$

as $1 \cdot u^T \in \ker(L)$, whence

$$\begin{aligned} T_{ij} &= \sum_{k=1}^{n} L_{ik}^+ d_k - L_{ij}^+ d_j \cdot \frac{1}{\pi_j} + u_j \\ u_i &= -\sum_{k=1}^{n} L_{ik}^+ d_k + L_{ii}^+ vol(G), \quad j = i \\ T_{ij} &= \sum_{k} L_{ik}^+ d_k - L_{ij}^+ vol(G) + L_{jj}^+ vol(G) - \sum_{k} L_{jk}^+ d_k \end{aligned}$$

Note that $vol(G) = \sum_i d_i$ and $\pi_i = d_i / vol(G)$ for all i.

Then

$$T_{ij} + T_{ji} = vol(G)(L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+). \tag{129}$$

To see it is a squared Euclidean distance, we need the following lemma.

Lemma 2.3. If K is a symmetric and positive semidefinite matrix, then

$$K(x,x)+K(y,y)-2K(x,y) = d^2(\Phi(x), \Phi(y)) = \langle \Phi(x), \Phi(x) \rangle + \langle \Phi(y), \Phi(y) \rangle - 2\langle \Phi(x), \Phi(y) \rangle$$

where $\Phi = (\phi_i : i = 1, \ldots, n)$ are orthonormal eigenvectors with eigenvalues $\mu_i \geq 0$, such that $K(x,y) = \sum_i \mu_i \phi_i(x) \phi_i(y)$.

Clearly L^+ is a positive semidefinite matrix and we define the *commute time map* by its eigenvectors,

$$\Psi(x_i) = \left(\frac{1}{\sqrt{\mu_2}} \nu_2(i), \cdots, \frac{1}{\sqrt{\mu_n}} \nu_n(i) \right)^T \in \mathbb{R}^{n-1}.$$

then $L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+ = ||\Psi(x_i) - \Psi(x_j))||_{l^2}^2$, and we call $d_r(x_i, x_j) = \sqrt{L_{ii}^+ + L_{jj}^+ - 2L_{ij}^+}$ the *resistance distance*.

So we have $d_c(x_i, x_j) = \sqrt{T_{ij} + T_{ji}} = \sqrt{vol(G)} d_r(x_i, x_j)$. □

Table 1. Comparisons between diffusion map and commute time map. Here $x \sim y$ means that x and y are in the same cluster and $x \nsim y$ for different clusters.

Diffusion Map	Commute Time Map
P's right eigenvectors	L^+'s eigenvectors
scale parameters: t and ε	scale: ε
$\exists t$, s.t. $x \sim y$, $d_t(x,y) \to 0$ and $x \nsim y$, $d_t(x,y) \to \infty$	$x \sim y$, $d_c(x,y)$ small and $x \nsim y$, $d_c(x,y)$ large?

2.1. Comparisons between diffusion map and commute time map.

However, recently Radl, von Luxburg, and Hein give a *negative* answer for the last desired property of $d_c(x,y)$ in geometric random graphs. Their result is as follows. Let $\mathcal{X} \subseteq \mathbb{R}^p$ be a compact set and let $k : \mathcal{X} \times \mathcal{X} \to (0, +\infty)$ be a symmetric and continuous function. Suppose that $(x_i)_{i \in \mathbb{N}}$ is a sequence of data points drawn i.i.d. from $\mathcal{X}$ according to a density function $p > 0$ on $\mathcal{X}$. Define $W_{ij} = k(x_i, x_j)$, $P = D^{-1}W$, and $L = D - W$. Then Radl et al. shows

$$\lim_{n\to\infty} nd_r(x_i, x_j) = \frac{1}{d(x_i)} + \frac{1}{d(x_j)}$$

where $d(x) = \int_X k(x, y)dp(y)$ is a smoothed density at x, $d_r(x_i, x_j) = \frac{d_c(x_i,x_j)}{\sqrt{vol(G)}}$ is the resistance distance. This result shows that in this setting commute time distance has no information about cluster information about point cloud data, instead it simply reflects density information around the two points.

3. Diffusion Map: Convergence Theory

Diffusion distance depends on both the geometry and density of the dataset. The key concepts in the analysis of these methods, that incorporates the density and geometry of a dataset. This section we will prove the convergence of diffusion map with heat kernels to its geometric limit, the eigenfunctions of Laplacian-Beltrami operators.

This is left by previous lecture. W is positive definite if using Gaussian Kernel.

One can check that, when

$$Q(x) = \int_{\mathbb{R}} e^{-ix\xi} d\mu(\xi),$$

for some positive finite Borel measure $d\mu$ on $\mathbb{R}$, then the (symmetric/Hermitian) integral kernel

$$k(x, y) = Q(x - y)$$

is positive definite, that is, for any function $\phi(x)$ on $\mathbb{R}$,

$$\int\int \bar{\phi}(x)\phi(y)k(x, y) \geq 0.$$

Proof omitted. The reverse is also true, which is Bochner theorem. High dimensional case is similar.

Take 1-dimensional as an example. Since the Gaussian distribution $e^{-\xi^2/2}d\xi$ is a positive finite Borel measure, and the Fourier transform of Gaussian kernel is itself, we know that $k(x, y) = e^{-|x-y|^2/2}$ is a positive definite integral kernel. The matrix W as an discretized version of $k(x, y)$ keeps the positive-definiteness (make this rigorous? Hint: take $\phi(x)$ as a linear combination of n delta functions).

3.1. Main Result. In this lecture, we will study the bias and variance decomposition for sample graph Laplacians and their asymptotic convergence to Laplacian-Beltrami operators on manifolds.

Let $\mathcal{M}$ be a smooth manifold without boundary in $\mathbb{R}^p$ (*e.g.* a d-dimensional sphere). Randomly draw a set of n data points, $x_1, ..., x_n \in M \subset \mathbb{R}^p$, according to distribution $p(x)$ in an independent and identically distributed (i.i.d.) way. We can extract an $n \times n$ weight matrix W_{ij} as follows:

$$W_{ij} = k(x_i, x_j)$$

where $k(x, y)$ is a symmetric $k(x, y) = k(y, x)$ and positivity-preserving kernel $k(x, y) \geq 0$. As an example, it can be the *heat kernel* (or Gaussian kernel),

$$k_\epsilon(x_i,x_j) = \exp\left(-\frac{||x_i-x_j||^2}{2\epsilon}\right),$$

where $||\,.\,||^2$is the Euclidean distance in space R^p and ϵ is the bandwidth of the kernel. W_{ij} stands for similarity function between x_i and x_j. A diagonal matrix D is defined with diagonal elements are the row sums of W:

$$D_{ii} = \sum_{j=1}^n W_{ij}.$$

Let's consider a family of re-weighted similarity matrix, with superscript (α),

$$W^{(\alpha)} = D^{-\alpha} W D^{-\alpha}$$

and

$$D_{ii}^{(\alpha)} = \sum_{j=1}^n W_{ij}^{(\alpha)}.$$

Denote $A^{(\alpha)} = (D^{(\alpha)})^{-1}W$, and we can verify that $\sum_{j=1}^n A_{ij}^{(\alpha)} = 1$, *i.e.*a row Markov matrix. Now define $L^{(\alpha)} = A^{(\alpha)} - I = (D^{(\alpha)})^{-1}W^{(\alpha)} - I$; and

$$L_{\epsilon,\alpha} = \frac{1}{\epsilon}(A_\epsilon^{(\alpha)} - I)$$

when $k_\epsilon(x,y)$ is used in constructing W. In general, $L^{(\alpha)}$ and $L_{\epsilon,\alpha}$ are both called *graph Laplacians*. In particular $L^{(0)}$ is the unnormalized graph Laplacian in literature.

The target is to show that graph Laplacian $L_{\epsilon,\alpha}$ converges to continuous differential operators acting on smooth functions on $\mathcal{M}$ the manifold. The convergence can be roughly understood as: we say a sequence of n-by-n matrix $L^{(n)}$ as $n\to\infty$ converges to a limiting operator $\mathcal{L}$, if for $\mathcal{L}$'s eigenfunction $f(x)$ (a smooth function on $\mathcal{M}$) with eigenvalue λ, that is

$$\mathcal{L}f = \lambda f,$$

the length-n vector $f^{(n)} = (f(x_i)), (i = 1, \cdots, n)$ is approximately an eigenvector of $L^{(n)}$ with eigenvalue λ, that is

$$L^{(n)} f^{(n)} = \lambda f^{(n)} + o(1),$$

where $o(1)$ goes to zero as $n\to\infty$.

Specifically, (the convergence is in the sense of multiplying a positive constant)

(I) $L_{\epsilon,0} = \frac{1}{\epsilon}(A_\epsilon - I) \to \frac{1}{2}(\Delta_{\mathcal{M}} + 2\frac{\nabla p}{p}\cdot\nabla)$ as $\epsilon\to 0$ and $n\to\infty$. $\Delta_{\mathcal{M}}$ is the Laplace-Beltrami operator of manifold M. At a point on M which is d-dimensional, in local (orthogonal) geodesic coordinate $s_1,\cdots,s_d$, the Laplace-Beltrami operator has the same form as the laplace in calculus

$$\Delta_{\mathcal{M}} f = \sum_{i=1}^d \frac{\partial^2}{\partial s_i^2} f;$$

∇ denotes the gradient of a function on M, and $\cdot$ denotes the inner product on tangent spaces of $\mathcal{M}$. Note that $p = e^{-V}$, so $\frac{\nabla p}{p} = -\nabla V$.

(Ignore this part if you don't know stochastic process) Suppose we have the following diffusion process

$$dX_t = -\nabla V(X_t)dt + \sigma dW_t^{(M)},$$

where $W_t^{(M)}$ is the Brownian motion on M, and σ is the volatility, say a positive constant, then the backward Kolmogorov operator/Fokker-Plank operator/infinitesimal generator of the process is

$$\frac{\sigma^2}{2}\Delta_{\mathcal{M}} - \nabla V \cdot \nabla,$$

so we say in (I) the limiting operator is the Fokker-Plank operator. Notice that in Lafon '06 paper they differ the case of $\alpha = 0$ and $\alpha = 1/2$, and argue that only in the later case the limiting operator is the Fokker-Plank. However the difference between $\alpha = 0$ and $\alpha = 1/2$ is a $1/2$ factor in front of $-\nabla V$, and that can be unified by changing the volatility σ to another number. (Actually, according to Thm 2. on Page 15 of Lafon'06, one can check that $\sigma^2 = \frac{1}{1-\alpha}$.) So here we say for $\alpha = 0$ the limiting operator is also Fokker-Plank. (not talked in class, open to discussion...)

(II) $L_{\epsilon,1} = \frac{1}{\epsilon}(A_\epsilon^{(1)} - I) \to \frac{1}{2}\Delta_{\mathcal{M}}$ as $\epsilon \to 0$ and $n \to \infty$. Notice that this case is of important application value: whatever the density $p(x)$ is, the Laplacian-Beltrami operator of $\mathcal{M}$ is approximated, so the geometry of the manifold can be understood.

A special case is that samples x_i are uniformly distributed on $\mathcal{M}$, whence $\nabla p = 0$. Then (I) and (II) are the same up to multiplying a positive constant, due to that D's diagonal entries are almost the same number and the re-weight does not do anything.

Convergence results like these can be found in Coifman and Lafon [**CL06**], *Diffusion maps, Applied and Computational Harmonic Analysis.*

We also refer [**Sin06**] *From graph to manifold Laplacian: The convergence rate, Applied and Computational Harmonic Analysis* for a complete analysis of the variance error, while the analysis of bias is very brief in this paper.

3.2. Proof. For a smooth function $f(x)$ on $\mathcal{M}$, let $f = (f_i) \in \mathbb{R}^n$ as a vector defined by $f_i = f(x_i)$. At a given fixed point x_i, we have the formula:

$$\begin{aligned}(Lf)^i &= \frac{1}{\epsilon}\left(\frac{\sum_{j=1}^n W_{ij}f_j}{\sum_{j=1}^n W_{ij}} - f_i\right) = \frac{1}{\epsilon}\left(\frac{\frac{1}{n}\sum_{j=1}^n W_{ij}f_j}{\frac{1}{n}\sum_{j=1}^n W_{ij}} - f_i\right)\\ &= \frac{1}{\epsilon}\left(\frac{\frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j).f(x_j)}{\frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j)} - f(x_i) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right)\end{aligned}$$

where in the last step the diagonal terms $j = i$ are excluded from the sums resulting in an $O(n^{-1}\epsilon^{-\frac{d}{2}})$ error. Later we will see that compared to the variance error, this term is negligible.

We rewrite the Laplacian above as

$$(Lf)^i = \frac{1}{\epsilon}\left(\frac{F(x_i)}{G(x_i)} - f(x_i) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right) \tag{130}$$

where

$$F(x_i) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i, x_j)f(x_j), \quad G(x_i) = \frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i,x_j).$$

depends only on the other $n-1$ data points than x_i. In what follows we treat x_i as a fixed chosen point and write as x.

Bias-Variance Decomposition. The points $x_j, j \neq i$ are independent identically distributed (i.i.d), therefore every term in the summation of $F(x)$ $(G(x))$ are i.i.d., and by the Law of Large Numbers (LLN) one should expect $F(x) \approx \mathbb{E}_{x_1}[k(x,x_1)f(x_1)] = \int_{\mathcal{M}} k(x,y)f(y)p(y)dy$ (and $G(x) \approx \mathbb{E}k(x,x_1) = \int_{\mathcal{M}} k(x,y)p(y)dy$). Recall that given a random variable x, and a sample estimator $\hat{\theta}$ (*e.g.* sample mean), the bias-variance decomposition is given by

$$\mathbb{E}\|x-\hat{\theta}\|^2 = \mathbb{E}\|x-\mathbb{E}x\|^2 + \mathbb{E}\|\mathbb{E}x-\hat{\theta}\|^2.$$

If we use the same strategy here (though not exactly the same, since $\mathbb{E}[\frac{F}{G}] \neq \frac{\mathbb{E}[F]}{\mathbb{E}[G]}$!), we can decompose Eqn. (130) as

$$\begin{aligned}(Lf)^i &= \frac{1}{\epsilon}\left(\frac{\mathbb{E}[F]}{\mathbb{E}[G]} - f(x_i) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right) + \frac{1}{\epsilon}\left(\frac{F(x_i)}{G(x_i)} - \frac{\mathbb{E}[F]}{\mathbb{E}[G]}\right)\\ &= bias + variance.\end{aligned}$$

In the below we shall show that for case (I) the estimates are

(131)
$$bias = \frac{1}{\epsilon}\left(\frac{\mathbb{E}[F]}{\mathbb{E}[G]} - f(x) + f(x_i)O(\frac{1}{n\epsilon^{\frac{d}{2}}})\right) = \frac{m_2}{2}(\Delta_{\mathcal{M}}f + 2\nabla f\cdot\frac{\nabla p}{p}) + O(\epsilon) + O\left(n^{-1}\epsilon^{-\frac{d}{2}}\right).$$

(132)
$$variance = \frac{1}{\epsilon}\left(\frac{F(x_i)}{G(x_i)} - \frac{\mathbb{E}[F]}{\mathbb{E}[G]}\right) = O(n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}),$$

whence

$$bias + variance = O(\epsilon, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}) = C_1\epsilon + C_2 n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}.$$

As the bias is a monotone increasing function of ϵ while the variance is decreasing w.r.t. ϵ, the optimal choice of ϵ is to balance the two terms by taking derivative of the right hand side equal to zero (or equivalently setting $\epsilon \sim n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}$) whose solution gives the optimal rates

$$\epsilon^* \sim n^{-1/(2+d/2)}.$$

[CL06] gives the bias and [HAvL05] contains the variance parts, which are further improved by [Sin06] in both bias and variance.

3.3. The Bias Term.

Now focus on $\mathbb{E}[F]$

$$\mathbb{E}[F] = \mathbb{E}\left[\frac{1}{n}\sum_{j\neq i} k_\epsilon(x_i, x_j)f(x_j)\right] = \frac{n-1}{n}\int_{\mathcal{M}} k_\epsilon(x,y)f(y)p(y)dy$$

$\frac{n-1}{n}$ is close to 1 and is treated as 1.

(1) the case of one-dimensional and flat (which means the manifold $\mathcal{M}$ is just a real line, $i.e.\mathcal{M} = \mathbb{R}$)

Let $\tilde{f}(y) = f(y)p(y)$, and $k_\epsilon(x,y) = \frac{1}{\sqrt{\epsilon}}e^{-\frac{(x-y)^2}{2\epsilon}}$, by change of variable

$$y = x + \sqrt{\epsilon}z,$$

we have

$$\Box = \int_{\mathbb{R}} \tilde{f}(x+\sqrt{\epsilon}z)e^{-\frac{\epsilon^2}{2}}dz = m_0\tilde{f}(x) + \frac{1}{2}m_2 f''(x)\epsilon + O(\epsilon^2)$$

where $m_0 = \int_{\mathbb{R}} e^{-\frac{\epsilon^2}{2}}dz$, and $m_2 = \int_{\mathbb{R}} z^2 e^{-\frac{\epsilon^2}{2}}dz$.

(2) 1 Dimensional & Not flat:

Divide the integral into 2 parts:

$$\int_m k_\epsilon(x,y)\tilde{f}(y)p(y)dy = \int_{||x-y||>c\sqrt{\epsilon}} \cdot + \int_{||x-y||<c\sqrt{\epsilon}} \cdot$$

First part $= \circ$

$$|\circ| \leq ||\tilde{f}||_\infty \frac{1}{\epsilon^{\frac{a}{2}}} e^{-\frac{\epsilon^2}{2\epsilon}},$$

due to $||x-y||^2 > c\sqrt{\epsilon}$

$$c \sim ln(\frac{1}{\epsilon}).$$

so this item is tiny and can be ignored.

Locally, that is $u \sim \sqrt{\epsilon}$, we have the curve in a plane and has the following parametrized equation

$$(x(u), y(u)) = (u, au^2 + qu^3 + \cdots),$$

then the chord length

$$\frac{1}{\epsilon}||x-y||^2 = \frac{1}{\epsilon}[u^2 + (au^2 + qu^3 + ...)^2] = \frac{1}{\epsilon}[u^2 + a^2u^4 + q_5(u) + \cdots],$$

where we mark $a^2u^4 + 2aqu^5 + ... = q_5(u)$. Next, change variable $\frac{u}{\sqrt{\epsilon}} = z$, then with $h(\xi) = e^{-\frac{\xi}{2}}$

$$h(\frac{||x-y||}{\epsilon})^2 = h(z^2) + h'(z^2)(\epsilon^2 az^4 + \epsilon^{\frac{3}{2}}q_5 + O(\epsilon^2)),$$

also

$$\tilde{f}(s) = \tilde{f}(x) + \frac{d\tilde{f}}{ds}(x)s + \frac{1}{2}\frac{d^2\tilde{f}}{ds^2}(x)s^2 + \cdots$$

and

$$s = \int_0^u \sqrt{1 + (2au + 3quu^2 + ...)^2 du} + \cdots$$

and

$$\frac{ds}{du} = 1 + 2a^2u^2 + q_2(u) + O(\epsilon^2), \quad s = u + \frac{2}{3}a^2u^3 + O(\epsilon^2).$$

Now come back to the integral

$$
\begin{aligned}
&\int_{|x-y|<c\sqrt{\epsilon}} \frac{1}{\sqrt{\epsilon}} h(\frac{x-y}{\epsilon})\tilde{f}(s)ds \\
\approx &\int_{-\infty}^{+\infty} [h(z^2)+h'(z^2)(\epsilon^2 a z^4+\epsilon^{\frac{3}{2}}q_5]\cdot[\tilde{f}(x)+\frac{d\tilde{f}}{ds}(x)(\sqrt{\epsilon}z+\frac{2}{3}a^2z^2\epsilon^{\frac{3}{2}}) \\
&+\frac{1}{2}\frac{d^2\tilde{f}}{ds^2}(x)\epsilon z^2]\cdot[1+2a^2+\epsilon^3 y_3(z)]dz \\
=&m_0\tilde{f}(x)+\epsilon\frac{m_2}{2}(\frac{d^2\tilde{f}}{ds^2}(x)+a^2\tilde{f}(x))+O(\epsilon^2),
\end{aligned}
$$

where the $O(\epsilon^2)$ tails are omitted in middle steps, and $m_0=\int h(z^2)dz$,$m_2=\int z^2h(z^2)dz$, are positive constants. In what follows we normalize both of them by m_0, so only m_2 appears as coefficient in the $O(\epsilon)$ term. Also the fact that $h(\xi)=e^{-\frac{\xi}{2}}$, and so $h'(\xi)=-\frac{1}{2}h(\xi)$, is used.

(3) For high dimension, $\mathcal{M}$ is of dimension d,

$$k_\epsilon(x,y)=\frac{1}{\epsilon^{\frac{d}{2}}}e^{-\frac{|x-y|^2}{2\epsilon}},$$

the corresponding result is (Lemma 8 in Appendix B of Lafon '06 paper)

$$\int_{\mathcal{M}} k_\epsilon(x,y)\tilde{f}(y)dy=\tilde{f}(x)+\epsilon\frac{m_2}{2}(\Delta_{\mathcal{M}}\tilde{f}+E(x)\tilde{f}(x))+O(\epsilon^2), \tag{133}$$

where

$$E(x)=\sum_{i=1}^{d}a_i(x)^2-\sum_{i_1\neq i_2}a_{i_1}(x)a_{i_2}(x),$$

and $a_i(x)$ are the curvatures along coordinates s_i $(i=1,\cdots,d)$ at point x.

Now we study the limiting operator and the bias error:

$$
\begin{aligned}
\frac{\mathbb{E}F}{\mathbb{E}G} &= \frac{\int k_\epsilon(x,y)f(y)p(y)dy}{\int k_\epsilon(x,y)p(y)dy} \approx \frac{f+\epsilon\frac{m_2}{2}(f''+2f'\frac{p'}{p}+f\frac{p^2}{p}+Ef)+O(\epsilon^2)}{1+\epsilon\frac{m_2}{2}(\frac{p''}{p}+E)+O(\epsilon^2)} \\
&= f(x)+\epsilon\frac{m_2}{2}(f''+2f'\frac{p'}{p})+o(\epsilon^2), \qquad (134)
\end{aligned}
$$

and as a result, for generally d-dim case,

$$\frac{1}{\epsilon}\left(\frac{\mathbb{E}F}{\mathbb{E}G}-f(x)\right)=\frac{m_2}{2}(\Delta_{\mathcal{M}}f+2\nabla f\cdot\frac{\nabla p}{p})+O(\epsilon).$$

Using the same method and use Eqn. (133), one can show that for case (II) where $\alpha=1$, the limiting operator is exactly the Laplace-Beltrami operator and the bias error is again $O(\epsilon)$ (homework).

About $\mathcal{M}$ with boundary: firstly the limiting differential operator bears Newmann/no-flux boundary condition. Secondly, the convergence at a belt of width $\sqrt{\epsilon}$ near $\partial\mathcal{M}$ is slower than the inner part of $\mathcal{M}$, see more in Lafon'06 paper.

3.4. Variance Term. Our purpose is to derive the large deviation bound for[2]

$$Prob\left(\frac{F}{G}-\frac{\mathbb{E}[F]}{\mathbb{E}[G]}\geq\alpha\right) \tag{135}$$

where $F=F(x_i)=\frac{1}{n}\sum_{j\neq i}k_\epsilon(x_i,x_j)f(x_j)$ and $G=G(x_i)=\frac{1}{n}\sum_{j\neq i}k_\epsilon(x,x_j)$. With $x_1,x_2,...,x_n$ as i.i.d random variables, F and G are sample means (up to a scaling constant). Define a new random variable

$$Y=\mathbb{E}[G]F-\mathbb{E}[F]G-\alpha\mathbb{E}[G](G-\mathbb{E}[G])$$

which is of mean zero and Eqn. (135) can be rewritten as

$$Prob(Y\geq\alpha\mathbb{E}[G]^2).$$

For simplicity by *Markov (Chebyshev) inequality*[3] ,

$$Prob(Y\geq\alpha\mathbb{E}[G]^2)\leq\frac{\mathbb{E}[Y^2]}{\alpha^2\mathbb{E}[G]^4}$$

and setting the right hand side to be $\delta\in(0,1)$, then with probability at least $1-\delta$ the following holds

$$\alpha\leq\frac{\sqrt{\mathbb{E}[Y^2]}}{\mathbb{E}[G]^2\sqrt{\delta}}\sim O\left(\frac{\sqrt{\mathbb{E}[Y^2]}}{\mathbb{E}[G]^2}\right).$$

It remains to bound

$$\begin{aligned}\mathbb{E}[Y^2]&=(\mathbb{E}G)^2\mathbb{E}(F^2)-2(\mathbb{E}G)(\mathbb{E}F)\mathbb{E}(FG)+(\mathbb{E}F)^2\mathbb{E}(G^2)+...\\&+2\alpha(\mathbb{E}G)[(\mathbb{E}F)\mathbb{E}(G^2)-(\mathbb{E}G)\mathbb{E}(FG)]+\alpha^2(\mathbb{E}G)^2(\mathbb{E}(G^2)-(\mathbb{E}G)^2).\end{aligned}$$

So it suffices to give $\mathbb{E}(F)$, $\mathbb{E}(G)$, $\mathbb{E}(FG)$, $\mathbb{E}(F^2)$, and $\mathbb{E}(G^2)$. The former two are given in bias and for the variance parts in latter three, let's take one simple example with $\mathbb{E}(G^2)$.

Recall that $x_1,x_2,...,x_n$ are distributed i.i.d according to density $p(x)$, and

$$G(x)=\frac{1}{n}\sum_{j\neq i}k_\epsilon(x,x_j),$$

so

$$Var(G)=\frac{1}{n^2}(n-1)\left[\int_{\mathcal{M}}k_\epsilon(x,y))^2p(y)dy-(\mathbb{E}k_\epsilon(x,y))^2\right].$$

Look at the simplest case of 1-dimension flat $\mathcal{M}$ for an illustrative example:

$$\int\limits_{\mathcal{M}}(k_\epsilon(x,y))^2p(y)dy=\int\limits_{\mathbb{R}}\frac{1}{\sqrt{\epsilon}}h^2(z^2)(p(x)+p'(x)(\sqrt{\epsilon}z+O(\epsilon)))dz,$$

let $M_2=\int\limits_{\mathbb{R}}h^2(z^2)dz$

$$\int\limits_{\mathcal{M}}(k_\epsilon(x,y))^2p(y)dy=p(x)\cdot\frac{1}{\sqrt{\epsilon}}M_2+O(\sqrt{\epsilon}).$$

Recall that $\mathbb{E}k_\epsilon(x,y)=O(1)$, we finally have

$$Var(G)\sim\frac{1}{n}\left[\frac{p(x)M_2}{\sqrt{\epsilon}}+O(1)\right]\sim\frac{1}{n\sqrt{\epsilon}}.$$

[2] The opposite direction is omitted here.

[3] It means that $Prob(X>\alpha)\leq\mathbb{E}(X^2)/\alpha^2$. A Chernoff bound with exponential tail can be found in Singer'06.

Generally, for d-dimensional case, $Var(G) \sim n^{-1}\epsilon^{-\frac{d}{2}}$. Similarly one can derive estimates on $Var(F)$.

Ignoring the joint effect of $\mathbb{E}(FG)$, one can somehow get a rough estimate based on $F/G = [\mathbb{E}(F) + O(\sqrt{\mathbb{E}(F^2)})]/[\mathbb{E}(G) + O(\sqrt{\mathbb{E}(G^2)})]$ where we applied the Markov inequality on both the numerator and denominator. Combining those estimates together, we have the following,

$$\begin{aligned} \frac{F}{G} &= \frac{fp + \epsilon\frac{m_2}{2}(\Delta(fp) + \mathbb{E}[fp]) + O(\epsilon^2, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}})}{p + \epsilon\frac{m_2}{2}(\Delta p + \mathbb{E}[p]) + O(\epsilon^2, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}})} \\ &= f + \epsilon\frac{m_2}{2}(\Delta p + \mathbb{E}[p]) + O(\epsilon^2, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}}), \end{aligned}$$

here $O(B_1, B_2)$ denotes the dominating one of the two bounds B_1 and B_2 in the asymptotic limit. As a result, the error (bias + variance) of $L_{\epsilon,\alpha}$ (dividing another ϵ) is of the order

$$O(\epsilon, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-1}). \tag{136}$$

In [Sin06] paper, the last term in the last line is improved to

$$O(\epsilon, n^{-\frac{1}{2}}\epsilon^{-\frac{d}{4}-\frac{1}{2}}), \tag{137}$$

where the improvement is by carefully analyzing the large deviation bound of $\frac{F}{G}$ around $\frac{\mathbb{E}F}{\mathbb{E}G}$ shown above, making use of the fact that F and G are correlated. Technical details are not discussed here.

In conclusion, we need to choose ϵ to balance bias error and variance error to be both small. For example, by setting the two bounds in Eqn. (137) to be of the same order we have

$$\epsilon \sim n^{-1/2}\epsilon^{-1/2-d/4},$$

that is

$$\epsilon \sim n^{-1/(3+d/2)},$$

so the total error is $O(n^{-1/(3+d/2)})$.

4. *Vector Diffusion Map

In this class, we introduce the topic of vector Laplacian on graphs and vector diffusion map.

The ideas for vector Laplacian on graphs and vector diffusion mapping are a natural extension from graph Laplacian operator and diffusion mapping on graphs. The reason why diffusion mapping is important is that previous dimension reduction techniques, such as the PCA and MDS, ignore the intrinsic structure of the manifold. By contrast, diffusion mapping derived from graph Laplacian is the optimal embedding that preserves locality in a certain way. Moreover, diffusion mapping gives rise to a kind of metic called diffusion distance. Manifold learning problems involving vector bundle on graphs provide the demand for vector diffusion mapping. And since vector diffusion mapping is an extension from diffusion mapping, their properties and convergence behavior are similar.

The application of vector diffusion mapping is not restricted to manifold learning however. Due to its usage of optimal registration transformation, it is also a valuable tool for problems in computer vision and computer graphics, for example, optimal matching of 3D shapes.

The organization of this lecture notes is as follows: We first review graph Laplacian and diffusion mapping on graphs as the basis for vector diffusion mapping. We then introduce three examples of vector bundles on graphs. After that, we come to vector diffusion mapping. Finally, we introduce some conclusions about the convergence of vector diffusion mapping.

4.1. graph Laplacian and diffusion mapping.

4.2. graph Laplacian. The goal of graph Laplacian is to discover the intrinsic manifold structure given a set of data points in space. There are three steps of constructing the graph Laplacian operator:

- construct the graph using either the $\epsilon-$*neighborhood way* (for any data point, connect it with all the points in its $\epsilon-$neighborhood) or the *k-nearest neighbor way* (connect it with its k-nearest neighbors);
- construct the the weight matrix. Here we can use the simple-minded *binary weight* (0 or 1), or use the *heat kernel weight.* For undirected graph, the weight matrix is symmetric;
- denote $\mathcal{D}$ as the diagonal matrix with $\mathcal{D}(i,i) = deg(i),\ deg(i) := \sum_j w_{ij}$. The graph Laplacian operator is:
$$L = \mathcal{D} - W$$

The graph Laplacian has the following properties:

- $\forall f : V \to \mathbb{R},\ f^T L f = \sum_{(i,j)\in E} w_{ij}(f_i - f_j)^2 \geq 0$
- G is connected $\Leftrightarrow f^T L f > 0, \forall f^T \vec{1}$, where $\vec{1} = (1, \cdots, 1)^T$
- G has k-connected components $\Leftrightarrow$ dim(ker(L))=k
(this property is compatible with the previous one, since $L\vec{1} = 0$)
- Kirchhofff's Matrix Tree theorem:
Consider a connected graph G and the binary weight matrix: $w_{ij} = \begin{cases} 1, & (i,j) \in E \\ 0, & otherwise \end{cases}$, denote the eigenvalues of L as $0 = \lambda_1 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, then #{T: T is a spanning tree of G}$=\frac{1}{n}\lambda_2 \cdots \lambda_n$
- Fieldler Theory, which will be introduced in later chapters.

We can have a further understanding of Graph Laplacian using the language of exterior calculus on graph.

We give the following denotations:

$V = \{1, 2, \cdots, |V|\}$. $\vec{E}$ is the oriented edge set that for $(i,j) \in E$ and $i < j$, $\langle i,j \rangle$ is the positive orientation, and $\langle j,i \rangle$ is the negative orientation.

$\delta_0 :\ \mathbb{R}^V \to \mathbb{R}^{\vec{E}}$ is a coboundary map, such that

$$\delta_0 \circ f(i,j) = \begin{cases} f_i - f_j, & \langle i,j \rangle \in \vec{E} \\ 0, & otherwise \end{cases}$$

It is easy to see that $\delta_0 \circ f(i,j) = -\delta_0 \circ f(j,i)$

The inner product of operators on $\mathbb{R}^{\vec{E}}$ is defined as:

$$\langle u, v \rangle = \sum_{i,j} w_{ij} u_{ij} v_{ij}$$

$$u^* := u\ diag(w_{ij})$$

where $diag(w_{ij}) \in \mathbb{R}^{\frac{n(n-1)}{2} \times \frac{n(n-1)}{2}}$ is the diagonal matrix that has w_{ij} on the diagonal position corresponding to $\langle i, j \rangle$.

$$u^* v = \langle u, v \rangle$$

Then,

$$L = D - W = \delta_0^T diag(w_{ij}) \delta_0 = \delta_0^* \delta_0$$

We first look at the graph Laplacian operator. We solve the generalized eigenvalue problem:

$$Lf = \lambda \mathcal{D} f$$

denote the generalized eigenvalues as:

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

and the corresponding generalized eigenvectors:

$$f_1, \cdots, f_n$$

we have already obtained the m-dimensional Laplacian eigenmap:

$$\mathbf{x_i} \to (\mathbf{f_1(i)}, \cdots, \mathbf{f_m(i)})$$

We now explains that this is the optimal embedding that preserves locality in the sense that connected points stays as close as possible. Specifically speaking, for the one-dimensional embedding, the problem is:

$$min \sum_{i,j} (y_i - y_j)^2 w_{ij} = 2min_{\mathbf{y}} \mathbf{y}^T L \mathbf{y}$$

$$y^T L y = y^T \mathcal{D}^{-\frac{1}{2}} (I - \mathcal{D}^{-\frac{1}{2}} W \mathcal{D}^{-\frac{1}{2}}) \mathcal{D}^{-\frac{1}{2}} y$$

Since $I - \mathcal{D}^{-\frac{1}{2}} W \mathcal{D}^{-\frac{1}{2}}$ is symmetric, the object is minimized when $\mathcal{D}^{-\frac{1}{2}}) \mathcal{D}^{-\frac{1}{2}} y$ is the eigenvector for the second smallest eigenvalue(the first smallest eigenvalue is 0) of $I - \mathcal{D}^{-\frac{1}{2}} W \mathcal{D}^{-\frac{1}{2}}$, which is the same with λ_2, the second smallest generalized eigenvalue of L.
Similarly, the m-dimensional optimal embedding is given by $Y = (\mathbf{f_1}, \cdots, \mathbf{f_m})$.

In diffusion map, the weights are used to define a discrete random walk. The transition probability in a single step from i to j is:

$$a_{ij} = \frac{w_{ij}}{deg(i)}$$

Then the transition matrix $A = \mathcal{D}^{-1} W$.

$$A = \mathcal{D}^{-\frac{1}{2}} (\mathcal{D}^{-\frac{1}{2}} W \mathcal{D}^{-\frac{1}{2}}) \mathcal{D}^{\frac{1}{2}}$$

Therefore, A is similar to a symmetric matrix, and has n real eigenvalues $\mu_1, \cdots, \mu_n$ and the corresponding eigenvectors $\phi_1, \cdots, \phi_n$.

$$A \phi_i = \mu_i \phi_i$$

A^t is the transition matrix after t steps. Thus, we have:

$$A^t \phi_i = \mu_i^t \phi_i$$

Define Λ as the diagonal matrix with $\Lambda(i,i) = \mu_i$, $\Phi = [\phi_1, \cdots, \phi_n]$. The diffusion map is given by:

$$\Phi_t := \Phi \Lambda^t = [\mu_1^t \phi_1, \cdots, \mu_n^t \phi_n]$$

4.3. the embedding given by diffusion map. $\Phi_t(i)$ denotes the ith row of Φ_t.

$$\langle \Phi_t(i), \Phi(j) \rangle = \sum_{k=1}^{n} \frac{A^t(i,k)}{\sqrt{deg(k)}} \frac{A^t(j,k)}{\sqrt{deg(k)}}$$

we can thus define a distance called *diffusion distance*

$$d^2_{DM,t}(i,j) := \langle \Phi_t(i), \Phi(i) \rangle + \langle \Phi_t(j), \Phi(j) \rangle - 2\langle \Phi_t(i), \Phi(j) \rangle = \sum_{k=1}^{n} \frac{(A^t(i,k) - A^t(j,k))^2}{deg(k)}$$

4.4. Examples of vector bundles on graph.

(1) Wind velocity field on globe:

To simplify the problem, we consider the two dimensional mesh on the globe(the latitude and the longitude). Each node on the mesh has a vector $\vec{f}$ which is the wind velocity at that place.

(2) Local linear regression:

The goal of local linear regression is to give an approximation of the regression function at an arbitrary point in the variable space.

Given the data $(y_i, \vec{x_i})_{i=1}^n$ and an arbitrary point $\vec{x}$,$\vec{x}, \vec{x_1}, \cdots, \vec{x_n} \in \mathbb{R}^p$, we want to find $\vec{\beta} := (\beta_0, \beta_1, \cdots, \beta_p)^T$ that minimize $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots \beta_p x_{ip})^2 K_n(\vec{x_i}, \vec{x})$. Here $K_n(\vec{x_i}, \vec{x})$ is a kernel function that defines the weight for $\vec{x_i}$ at the point $\vec{x}$. For example, we can use the Nadaraya-Watson kernel $K_n(\vec{x_i}, \vec{x}) = e^{\frac{||x - x_i||^2}{n}}$.

For a graph G=(V,E), each point $\vec{x} \in V$ has a corresponding vector $\vec{\beta}(\vec{x})$. We therefore get a vector bundle on the graph G(V,E).

Here $\vec{\beta}$ is kind of a gradient. In fact, if y and $\vec{x}$ has the relationship $y = f(\vec{x})$, then $\beta = (f(\vec{x}), \nabla f(\vec{x}))^T$.

(3) Social networks:

If we see users as vertices and the relationship bonds that connected users as edges, then a social network naturally gives rise to a graph G=(V,E). Each user has an attribute profile containing all kinds of personal information, and a certain kind of information can be described by a vector $\vec{f}$ recording different aspects. Again, we get a vector bundle on graph.

4.5. optimal registration transformation. Like in graph eigenmap, we expect the embedding $\vec{f}$ to be preserve locality to a certain extent, which means that we expect the embedding of connected points to be sufficiently close. In the graph Laplacian case, we use $\sum_{i \sim j} w_{ij} ||\vec{f_i} - \vec{f_j}||^2$. However, for vector bundle on graphs, subtraction of vectors at different points may not be done directly due to the curvature of the manifold. What makes sense should be the difference of vectors compared with the tangent spaces at the certain points. Therefore, we borrow the idea of parallel transport from differential geometry. Denote O_{ij} as the parallel transport operator from the tangent space at x_j to the tangent space at x_i. We want to find out the embedding that minimizes

$$\sum_{i \sim j} w_{ij} ||\vec{f_i} - O_{ij} \vec{f_j}||^2$$

we will later define the vector diffusion mapping, and using the similar argument as in diffusion mapping, it is easy to see that vector diffusion mapping gives the optimal embedding that preserves locality in this sense.

we now discuss how we get the approximation of parallel transport operator given the data set.
The approximation of the tangent space at a certain point x_i is given by local PCA. Choose ϵ_i to be sufficiently small, and denote $x_{i_1}, \cdots, x_{i_{N_i}}$ as the data points in the ϵ_i-neighborhood of x_i. Define

$$X_i := [x_{i_1} - x_i, \cdots, x_{i_{N_i}} - x_i]$$

Denote D_i as the diagonal matrix with

$$D_i(j,j) = \sqrt{K(\frac{||x_{i_j} - x_i||}{\epsilon_i})},\ j = 1, \cdots, N_i$$

$$B_i := X_i D_i$$

Perform SVD on B_i:

$$B_i = U_i \Sigma_i V_i^T$$

We use the first d columns of U_i (which are the left eigenvectors of the d largest eigenvalues of B_i) to form an approximation of the tangent space at x_i. That is,

$$O_i = [u_{i_1}, \cdots, u_{i_d}]$$

Then O_i is a numerical approximation to an orthonormal basis of the tangent space at x_i.

For connected points x_i and x_j, since they are sufficiently close to each other, their tangent space should be close. Therefore, $O_i O_{ij}$ and O_j should also be close. We there use the closest orthogonal matrix to $O_i^T O_j$ as the approximation of the parallel transport operator from x_j to x_i:

$$\rho_{ij} := argmin_{O orthogonol} ||O - O_i^T O_j||_{HS}$$

where $||A||_{HS}^2 = Tr(AA^T)$ is the Hilbert-Schimidt norm.

4.6. Vector Laplacian. Given the weight matrix $W = (w_{ij})$, we denote

$$D := \begin{pmatrix} deg(1)I_p & & \\ & \ddots & \\ & & deg(n)I_p \end{pmatrix} \in \mathbb{R}^{np \times np}$$

where $deg(i) = \sum_j w_{ij}$ as in graph Laplacian.
Define S as the block matrix with

$$S_{ij} = \begin{cases} w_{ij}\rho_{ij}, & i \sim j \\ 0, & otherwise \end{cases}$$

The vector Laplacian is then defined as $\mathcal{L} = D - S$

Like Graph Laplacian, we introduce an orientation on E and a coboundary map $\delta_0 : (\mathbb{R}^d)^V \to (\mathbb{R}^d)^{\vec{E}}$

$$\delta_0 \circ f(i,j) = \begin{cases} \vec{f_i} - \rho_{ij}\vec{f_j}, & \langle i,j\rangle \in \vec{E} \\ 0, & otherwise \end{cases}, \ where\ f = (\vec{f_1}, \cdots, \vec{f_n})^T$$

Inner product on $(\mathbb{R}^d)^{\vec{E}}$ is defined as

$$\langle u, v\rangle = \sum_{i,j} w_{ij} u_{ij}^T v_{ij}$$

$$u^* := u\ diag(w_{ij}),\ u^* v = \langle u, v\rangle$$

If we let ρ_{ij} be orthogonal, $\forall i, j,\ s.t. \langle i,j\rangle \in \vec{E}$, then, $L = D - W = \delta_0^T diag(w_{ij})\delta_0 = \delta_0^* \delta_0$.

Analogous properties with Graph Laplacian:

- G has k connected components $\Leftrightarrow dim\ ker(\mathcal{L}) = kp$
- generalized Matrix tree theorem.

4.7. Vector diffusion mapping.

$$\mathcal{L} = D - S = D(I - D^{-1}S)$$

$$D^{-1}S = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

Denote

$$\tilde{S} := D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$$

$\tilde{S}$ has nd real eigenvalues $\lambda_1, \cdots, \lambda_{nd}$ and the corresponding eigenvectors $v_1, \cdots, v_{nd}$. Thinking of these vectors of length nd in blocks of d, we denote $v_k(i)$ as the ith block of v_k.

The spectral decompositions of $\tilde{S}(i,j)$ and $\tilde{S}^{2t}(i,j)$ are given by:

$$\tilde{S}(i,j) = \sum_{k=1}^{nd} \lambda_k v_k(i) v_k(j)^T$$

$$\therefore \tilde{S}^{2t}(i,j) = \sum_{k=1}^{nd} \lambda_k^{2t} v_k(i) v_k(j)^T$$

We use $||\tilde{S}^{2t}(i,j)||_{HS}^2$ to measure the affinity between i and j. Thus,

$$\begin{aligned} ||\tilde{S}^{2t}(i,j)||_{HS}^2 &= Tr(\tilde{S}^{2t}(i,j)\tilde{S}^{2t}(i,j)^T) \\ &= \textstyle\sum_{k,l=1}^{nd} (\lambda_k\lambda_l)^{2t} Tr(v_k(i)v_k(j)^T v_l(j) v_l(i)^T) \\ &= \textstyle\sum_{k,l=1}^{nd} (\lambda_k\lambda_l)^{2t} Tr(v_k(j)^T v_l(j) v_l(i)^T v_k(i)) \\ &= \textstyle\sum_{k,l=1}^{nd} (\lambda_k\lambda_l)^{2t} \langle v_k(j), v_l(j)\rangle \langle v_k(i), v_l(i)\rangle \end{aligned}$$

The vector diffusion mapping is defined as:

$$V_t : i \to ((\lambda_k\lambda_l)^t \langle v_k(i), v_l(i)\rangle)_{k,l=1}^{nd}$$

Like graph Laplacian, $||\tilde{S}^{2t}(i,j)||_{HS}^2$ is actually an inner product:

$$||\tilde{S}^{2t}(i,j)||_{HS}^2 = \langle V_t(i), V_t(j)\rangle$$

This gives rise to a distance called vector diffusion distance:

$$d_{VDM,t}^2 = \langle V_t(i), V_t(i)\rangle + \langle V_t(j), V_t(j)\rangle - 2\langle V_t(i), V_t(j)\rangle$$

4.8. Normalized Vector Diffusion Mappings.

An important kind of normalized VDM is obtained as follows:
Take $0 \leq \alpha \leq 1$,

$$W_\alpha := \mathcal{D}^{-\alpha} W \mathcal{D}^{-\alpha}$$
$$S_\alpha := D^{-\alpha} S D^{-\alpha}$$
$$deg_\alpha(i) := \sum_{j=1}^{n} W_\alpha(i,j)$$

We define $\mathcal{D}_\alpha \in \mathbb{R}^{n \times n}$ as the diagonal matrix with

$$\mathcal{D}_\alpha(i,i) = deg_\alpha(i)$$

and $D_\alpha \in \mathbb{R}^{nd \times nd}$ as the block diagonal matrix with

$$D_\alpha(i,i) = deg_\alpha(i) I_d$$

We can then get the vector diffusion mapping $V_{\alpha,t}$ using S_α and D_α instead of S and D.

4.9. Convergence of VDM.

We first introduce some concepts.

Suppose $\mathcal{M}$ is a smooth manifold, and $T_\mathcal{M}$ is a *tensor bundle* on $\mathcal{M}$. When the rank of $T_\mathcal{M}$ is 0, it is the set of functions on $\mathcal{M}$. When the rank of $T_\mathcal{M}$ is 1, it is the set of vector fields on $\mathcal{M}$.

The*connection Laplacian operator* is:

$$\nabla^2_{X,Y} T = -(\nabla_X \nabla_Y T - \nabla_{\nabla_X Y} T)$$

where $\nabla_X Y$ is the covariant derivative of Y over X.
Intuitively, we can see the first item of the connection Laplacian operator as the sum of the change of T over X and over Y, and the second item as the overlapped part of the change of T over X and over Y. The remainder can be seen as an operator that differentiates the vector fields in the direction of two orthogonal vector fields.

Now we introduce some results about convergence.
The normalized graph Laplacian converges to the Laplace-Beltrami operator:

$$(\mathcal{D}^{-1} W - I) f \rightarrow c \Delta f$$

for sufficiently smooth f and some constant c.

For VDM, $D_\alpha^{-1} S_\alpha - I$ converges to the connection Laplacian operator [SW12] plus some potential terms. When $\alpha = 1$, $D_1^{-1} S_1 - I$ converges to exactly the connection Laplacian operator:

$$(D_1^{-1} S_1 - I) X \rightarrow c \nabla^2 X$$

CHAPTER 8

Semi-supervised Learning

1. Introduction

Problem: $x_1, x_2, ..., x_l \in V_l$ are labled data, that is data with the value $f(x_i), f \in V \to \mathbb{R}$ observed. $x_{l+1}, x_{l+2}, ..., x_{l+u} \in V_u$ are unlabled. Our concern is how to fully exploiting the information (like geometric structure in disbution) provided in the labeled and unlabeled data to find the unobserved labels.

This kind of problem may occur in many situations, like ZIP Code recognition. We may only have a part of digits labeled and our task is to label the unlabeled ones.

2. Harmonic Extension of Functions on Graph

Suppose the whole graph is $G = (V, E, W)$, where $V = V_l \cup V_u$ and weight matrix is partitioned into blocks $W = \begin{pmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{pmatrix}$. As before, we define $D = \mathrm{diag}(d_1, d_2, ..., d_n) = \mathrm{diag}(D_l, D_u)$, $d_i = \sum_{j=1}^n W_{ij}$, $L = D - W$ The goal is to find $f_u = (f_{l+1}, ..., f_{l+u})^T$ such that

$$\begin{aligned} \min \quad & f^T L f \\ s.t. \quad & f(V_l) = f_l \end{aligned}$$

where $f = \begin{pmatrix} f_l \\ f_u \end{pmatrix}$. Note that

$$f^T L f = (f_l^T, f_u^T) L \begin{pmatrix} f_l \\ f_u \end{pmatrix} = f_u^T L_{uu} f_u + f_l^T L_{ll} f_l + 2 f_u^T L_{ul} f_l$$

So we have:

$$\frac{\partial f^T L f}{\partial f_u} = 0 \Rightarrow 2L_{uu} f_u + 2L_{lu} f_u = 0 \Rightarrow f_u = -L_{uu}^{-1} L_{ul} f_l = (D_u - W_{uu})^{-1} W_{ul} f_l$$

3. Explanation from Gaussian Markov Random Field

If we consider $f : V \to \mathbb{R}$ are Gaussian random variables on graph nodes whose inverse covariance matrix (precision matrix) is given by unnormalized graph Laplacian L (sparse but singular), i.e. $f \sim \mathcal{N}(0, \Sigma)$ where $\Sigma^{-1} = L$ (interpreted as a pseudo inverse). Then the conditional expectation of f_u given f_l is:

$$f_u = \Sigma_{ul} \Sigma_{ll}^{-1} f_l$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{ll} & \Sigma_{lu} \\ \Sigma_{ul} & \Sigma_{uu} \end{bmatrix}$$

Block matrix inversion formula tells us that when A and D are invertible,

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \cdot \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = I \Rightarrow \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} S_D^{-1} & -A^{-1}BS_A^{-1} \\ -D^{-1}CS_D^{-1} & S_A^{-1} \end{bmatrix}$$

$$\begin{bmatrix} X & Y \\ Z & W \end{bmatrix} \cdot \begin{bmatrix} A & B \\ C & D \end{bmatrix} = I \Rightarrow \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \begin{bmatrix} S_D^{-1} & -S_D^{-1}BD^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{bmatrix}$$

where $S_A = D - CA^{-1}B$ and $S_D = A - BD^{-1}C$ are called Schur complements of A and D, respectively. The matrix expressions for inverse are equivalent when the matrix is invertible.

The graph Laplacian

$$L = \begin{bmatrix} D_l - W_{ll} & W_{lu} \\ W_{ul} & D_u - W_{uu} \end{bmatrix}$$

is not invertible. $D_l - W_{ll}$ and $D_u - W_{uu}$ are both strictly diagonally dominant, i.e. $D_l(i,i) > \sum_j |W_{ll}(i,j)|$, whence they are invertible by Gershgorin Circle Theorem. However their Schur complements $S_{D_u - W_{uu}}$ and $S_{D_l - W_{ll}}$ are still not invertible and the block matrix inversion formula above can not be applied directly. To avoid this issue, we define a regularized version of graph Laplacian

$$L_\lambda = L + \lambda I, \quad \lambda > 0$$

and study its inverse $\Sigma_\lambda = L_\lambda^{-1}$.

By the block matrix inversion formula, we can set Σ as its right inverse above,

$$\Sigma_\lambda = \begin{bmatrix} S^{-1}_{\lambda + D_u - W_{uu}} & -(\lambda + D_l - W_{ll})^{-1} W_{lu} S^{-1}_{\lambda + D_l - W_{ll}} \\ -(\lambda + D_u - W_{uu})^{-1} W_{ul} S^{-1}_{\lambda + D_u - W_{uu}} & S^{-1}_{\lambda + D_l - W_{ll}} \end{bmatrix}$$

Therefore,

$$f_{u,\lambda} = \Sigma_{ul,\lambda} \Sigma^{-1}_{ll,\lambda} f_l = (\lambda + D_u - W_{uu})^{-1} W_{ul} f_l,$$

whose limit however exits $\lim_{\lambda \to 0} f_{u,\lambda} = (D_u - W_{uu})^{-1} W_{ul} f_l = f_u$. This implies that f_u can be regarded as the conditional mean given f_l.

4. Explanation from Transition Path Theory

We can also view the problem as a random walk on graph. Constructing a graph model with transition matrix $P = D^{-1}W = \begin{pmatrix} P_{ll} & P_{lu} \\ P_{ul} & P_{uu} \end{pmatrix}$. Assume that the labeled data are binary (classification). That is, for $x_i \in V_l$, $f(x_i) = 0 \, or \, 1$. Denote

- $V_0 = \{i \in V_l : f_i = f(x_i) = 0\}$
- $V_1 = \{i \in V_l : f_i = f(x_i) = 1\}$
- $V = V_0 \cup V_1 \cup V_u$ where $V_l = V_0 \cup V_1$

With this random walk on graph P, f_u can be interpreted as hitting time or first passage time of V_1.

Proposition 4.1. Define hitting time

$$\tau_i^k = \inf\{t \geq 0 : x(0) = i, x(t) \in V_k\}, \quad k = 0, 1$$

Then for $\forall i \in V_u$,

$$f_i = Prob(\tau_i^1 < \tau_i^0)$$

i.e.

$$f_i = Prob(\text{trajectory starting from } x_i \text{ hit } V_1 \text{ before } V_0)$$

Note that the probability above also called committor function in Transition Path Theory of Markov Chains.

Proof. Define the committor function,

$$q_i^+ = Prob(\tau_i^1 < \tau_i^0) = \begin{cases} 1 & x_i \in V_1 \\ 0 & x_i \in V_0 \\ \sum_{j \in V} P_{ij} q_j^+ & i \in V_u \end{cases}$$

This is because $\forall i \in V_u$,

$$\begin{aligned} q_i^+ &= Pr(\tau_{iV_1} < \tau_{iV_0}) \\ &= \sum_j P_{ij} q_j^+ \\ &= \sum_{j \in V_1} P_{ij} q_j^+ + \sum_{j \in V_0} P_{ij} q_j^+ + \sum_{j \in V_u} P_{ij} q_j^+ \\ &= \sum_{j \in V_1} P_{ij} + \sum_{j \in V_u} P_{ij} q_j^+ \end{aligned}$$

$$\therefore \quad q_u^+ = P_{ul} f_l + P_{uu} q_u^+ = D_u^{-1} W_{ul} f_l + D_u^{-1} W_{uu} q_u^+$$

multiply D_u to both side and reorganize:

$$(D_u - W_{uu}) q_u^+ = W_{ul} f_l$$

If $D_u - W_{uu}$ is reversible, we get:

$$q_u^+ = (D_u - W_{uu})^{-1} W_{ul} f_l = f_u$$

i.e. f_u is the committor function on V_u. □

The result coincides with we obtained through the view of gaussian markov random field.

5. Well-posedness

One natural problem is: if we only have a fixed amount of labeled data, can we recover labels of an infinite amount of unobserved data? This is called well-posedness. [Nadler-Srebro 2009] gives the following result:

- If $x_i \in \mathbb{R}^1$, the problem is well-posed.
- If $x_i \in \mathbb{R}^d (d \geq 3)$, the problem is ill-posed in which case $D_u - W_{uu}$ becomes singular and f becomes a bump function (f_u is almost always zeros or ones except on some singular points).

Here we can give a brief explanation:

$$f^T L f \sim \int \|\nabla f\|_2$$

If we have $V_l = \{0, 1\}$, $f(x_0) = 0$, $f(x_1) = 1$ and let $f_\epsilon(x) = \begin{cases} \frac{\|x - x_0\|_2^2}{\epsilon^2} & \|x - x_0\|_2 < \epsilon \\ 1 & otherwise \end{cases}$.

From multivariable calculus,

$$\int \|\nabla f\|_2 = c\epsilon^{d-2}.$$

Since $d \geq 3$, so $\epsilon \to 0 \Rightarrow \int \|\nabla f\|_2 \to 0$. So $f_\epsilon(x)$ $(\epsilon \to 0)$ converges to a bump function which is one almost everywhere except x_0 whose value is 0. No generalization ability is learned for such bump functions.

This means in high dimensional case, to obtain a smooth generalization, we have to add constraints more than the norm of the first order derivatives. We also have a theorem to illustrate what kind of constraint is enough for a good generalization:

Theorem 5.1 (Sobolev embedding Theorem)**.** $f \in \mathbf{W}^{s,p}(\mathbb{R}^d) \iff f$ has s'th order weak derivative $f^{(s)} \in \mathbf{L}_p$,

$$s > \frac{d}{2} \Rightarrow \mathbf{W}^{s,2} \hookrightarrow \mathbf{C}(\mathbb{R}^d).$$

So in $\mathbb{R}^d$, to obtain a continuous function, one needs smoothness regularization $\int \|\nabla^s f\|$ with degree $s > d/2$. To implement this in discrete Laplacian setting, one may consider iterative Laplacian L^s which might converge to high order smoothness regularization.

CHAPTER 9

Beyond graphs: high dimensional topological/geometric analysis

1. From Graph to Simplicial Complex

Definition (Simplicial Complex)**.** An abstract simplicial complex is a collection Σ of subsets of V which is closed under inclusion (or deletion), i.e. $\tau \in \Sigma$ and $\sigma \subseteq \tau$, then $\sigma \in \Sigma$.

We have the following examples:

- Chess-board Complex
- Point cloud data:
  Nerve complex
  Cech, Rips, Witness complex
  Mayer-Vietoris Blowup
- Term-document cooccurance complex
- Clique complex in pairwise comparison graphs
- Strategic complex in flow games

Example (Chess-board Complex)**.** Let V be the positions on a Chess board. Σ collects position subsets of V where one can place queens (rooks) without capturing each other. It is easy to check the closedness under deletion: if $\sigma \in \Sigma$ is a set of "safe" positions, then any subset $\tau \subseteq \sigma$ is also a set of "safe" positions

Example (Nerve Complex)**.** Define a cover of X, $X = \cup_\alpha U_\alpha$. $V = \{U_\alpha\}$ and define $\Sigma = \{U_I : \cap_{\alpha \in I} U_I \neq \emptyset\}$.

- Closedness under deletion
- Can be applied to any topological space X
- In a metric space (X, d), if $U_\alpha = B_\epsilon(t_\alpha) := \{x \in X : d(x - t_\alpha) \leq \epsilon\}$, we have Cech complex C_ϵ.
- Nerve Theorem: if every U_I is contractible, then X has the same homotopy type as Σ.
- Cech complex is hard to compute, even in Euclidean space
- One can easily compute an upper bound for Cech complex

  Construct a Cech subcomplex of 1-dimension, i.e. a graph with edges connecting point pairs whose distance is no more than ϵ.

  Find the clique complex, i.e. maximal complex whose 1-skeleton is the graph above, where every k-clique is regarded as a $k - 1$ simplex

Example (Vietoris-Rips Complex)**.** Let $V = \{x_\alpha \in X\}$. Define $VR_\epsilon = \{U_I \subseteq V : d(x_\alpha, x_\beta) \leq \epsilon, \alpha, \beta \in I\}$.

- Rips is easier to compute than Cech

even so, Rips is exponential to dimension generally

- However Vietoris-Rips CAN NOT preserve the homotopy type as Cech
- But there is still a hope to find a lower bound on homology –

Theorem 1.1 ("Sandwich").

$$VR_\epsilon \subseteq C_\epsilon \subseteq VR_{2\epsilon}$$

- If a homology group "persists" through $R_\epsilon \to R_{2\epsilon}$, then it must exists in C_ϵ; but not the vice versa.
- All above gives rise to a filtration of simplicial complex

$$\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \dots$$

- Functoriality of inclusion: there are homomorphisms between homology groups

$$0 \to H_1 \to H_2 \to \dots$$

- A persistent homology is the image of H_i in H_j with $j > i$.

Example (Strong Witness Complex)**.** Let $V = \{t_\alpha \in X\}$. Define $W^s_\epsilon = \{U_I \subseteq V : \exists x \in X, \forall \alpha \in I, d(x, t_\alpha) \leq d(x, V) + \epsilon\}$.

Example (Week Witness Complex)**.** Let $V = \{t_\alpha \in X\}$. Define $W^w_\epsilon = \{U_I \subseteq V : \exists x \in X, \forall \alpha \in I, d(x, t_\alpha) \leq d(x, V_{-I}) + \epsilon\}$.

- V can be a set of landmarks, much smaller than X
- Monotonicity: $W^*_\epsilon \subseteq W^*_{\epsilon'}$ if $\epsilon \leq \epsilon'$
- But not easy to control homotopy types between W^* and X

Example (Term-Document Occurrence complex, Li & Kwong 2009)**.** Left is a term-document co-occurrence matrix; Right is a simplicial complex representation of terms. Connectivity analysis captures more information than Latent Semantic Index.

	c_1	c_2	c_3	c_4	c_5
r_1	1	0	0	0	0
r_2	1	1	1	0	0
r_3	0	0	1	1	0
r_4	0	0	1	1	0
r_5	0	0	0	0	1
r_6	0	0	0	0	1

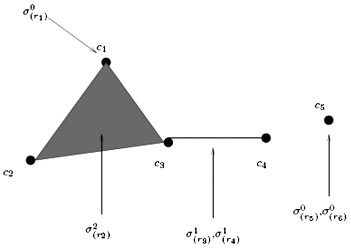

$\sigma^0_{(r_1)}$
c_1
c_2
c_3
c_4
c_5
$\sigma^2_{(r_2)}$
$\sigma^1_{(r_3)}, \sigma^1_{(r_4)}$
$\sigma^0_{(r_5)}, \sigma^0_{(r_6)}$

Figure 1. Term-Document Occurrence complex

Example (Flag Complex of Paired Comparison Graph, Jiang-Lim-Yao-Ye 2011[JLYY11])**.** Let V be a set of alternatives to be compared and undirected pair $(i, j) \in E$ if the pair is comparable. A flag complex χ_G consists all cliques as simplices or faces (e.g. 3-cliques as 2-faces and $k+1$-cliques as k-faces), also called clique complex of G.

Example (Strategic Simplicial Complex for Flow Games, Candogan-Menache-Ozdaglar–Parrilo 2011 [CMOP11])**.** Strategic simplicial complex is the clique complex of pairwise comparison graph $G = (V, E)$ of strategic profiles, where V consists of all strategy profiles of players and a pair of strategy $(x, x') \in E$ is comparable if only one player changes strategy from x to x'. Every finite game can be decomposed as the direct sum of potential games and zero-sum games (harmonic games).

	O	F
O	3, 2	0, 0
F	0, 0	2, 3

(a) Battle of the sexes

$$\begin{tikzcd}
(O,O) & (O,F) \arrow[l, "2"'] \arrow[d, "2"] \\
(F,O) \arrow[u, "3"] \arrow[r, "3"] & (F,F)
\end{tikzcd}$$

Figure 2. Illustration of Game Strategic Complex: Battle of Sex

2. Persistent Homology and Discrete Morse Theory

Recall that

Theorem 2.1 ("Sandwich")**.**

$$VR_\epsilon \subseteq C_\epsilon \subseteq VR_{2\epsilon}$$

- If a homology group "persists" through $R_\epsilon \to R_{2\epsilon}$, then it must exists in C_ϵ; but not the vice versa.
- All above gives rise to a filtration of simplicial complex
$$\emptyset = \Sigma_0 \subseteq \Sigma_1 \subseteq \Sigma_2 \subseteq \ldots$$
- Functoriality of inclusion: there are homomorphisms between homology groups
$$0 \to H_1 \to H_2 \to \ldots$$
- A persistent homology is the image of H_i in H_j with $j > i$.

Persistent Homology is firstly proposed by Edelsbrunner-Letscher-Zomorodian, with an algebraic formulation by Zomorodian-Carlsson. The algorithm is equivalent to Robin Forman's discrete Morse theory.

`to be continued...`

3. Exterior Calculus on Complex and Combinatorial Hodge Theory

We are going to study functions on simplicial complex, $l^2(V^d)$.

A basis of "forms":

- $l^2(V)$: e_i $(i \in V)$, so $f \in l^2(V)$ has a representation $f = \sum_{i \in V} f_i e_i$, e.g. global ranking score on V.
- $l^2(V^2)$: $e_{ij} = -e_{ji}$, $f = \sum_{(i,j)} f_{ij} e_{ij}$ for $f \in l^2(V^2)$, e.g. paired comparison scores on V^2.
- $l^2(V^3)$: $e_{ijk} = e_{jki} = e_{kij} = -e_{jik} = -e_{kji} = -e_{ikj}$, $f = \sum_{ijk} f_{ijk} e_{ijk}$
- $l^2(V^{d+1})$: $e_{i_0,\ldots,i_d}$ is an alternating d-form
$$e_{i_0,\ldots,i_d} = \mathrm{sign}(\sigma) e_{\sigma(i_0),\ldots,\sigma(i_d)},$$
where $\sigma \in \mathfrak{S}_d$ is a permutation on $\{0, \ldots, d\}$.

Vector spaces of functions $l^2(V^{d+1})$ represented on such basis with an inner product defined, are called d-forms (cochains).

Example. In the crowdsourcing ranking of world universities,
http://www.allourideas.org/worldcollege/,
V consists of world universities, E are university pairs in comparison, $l^2(V)$ consists of ranking scores of universities, $l^2(V^2)$ is made up of paired comparison data.

Discrete differential operators: k-dimensional coboundary maps $\delta_k : L^2(V^k) \to L^2(V^{k+1})$ are defined as the alternating difference operator

$$(\delta_k u)(i_0, \ldots, i_{k+1}) = \sum_{j=0}^{k+1} (-1)^{j+1} u(i_0, \ldots, i_{j-1}, i_{j+1}, \ldots, i_{k+1})$$

- δ_k plays the role of differentiation
- $\delta_{k+1} \circ \delta_k = 0$

So we have chain map

$$L^2(V) \xrightarrow{\delta_0} L^2(V^2) \xrightarrow{\delta_1} L^2(V^3) \to \ldots L^2(V^k) \xrightarrow{\delta_{k-1}} L^2(V^{k+1}) \xrightarrow{\delta_k} \ldots$$

with $\delta_k \circ \delta_{k-1} = 0$.

Example (Gradient, Curl, and Divergence)**.** We can define discrete gradient and curl, as well as their adjoints

- $(\delta_0 v)(i,j) = v_j - v_i =: (\mathrm{grad}\, v)(i,j)$
- $(\delta_1 w)(i,j,k) = (\pm)(w_{ij} + w_{jk} + w_{ki}) =: (\mathrm{curl}\, w)(i,j,k)$, which measures the total flow-sum along the loop $i \to j \to k \to i$ and $(\delta_1 w)(i,j,k) = 0$ implies the paired comparison data is path-independent, which defines the triangular transitivity subspace
- for each alternative $i \in V$, the combinatorial divergence

  $$(\mathrm{div}\, w)(i) := -(\delta_0^T w)(i) := \sum w_{i*}$$

  which measures the inflow-outflow sum at i and $(\delta_0^T w)(i) = 0$ implies alternative i is preference-neutral in all pairwise comparisons as a cyclic ranking passing through alternatives.

Definition (Combinatorial Hodge Laplacian)**.** Define the k-dimensional combinatorial Laplacian, $\Delta_k : L^2(V^{k+1}) \to L^2(C^{k+1})$ by

$$\Delta_k = \delta_{k-1}\delta_{k-1}^T + \delta_k^T \delta_k, \qquad k > 0$$

- $k = 0$, $\Delta_0 = \delta_0^T \delta_0$ is the well-known graph Laplacian
- $k = 1$,

  $$\Delta_1 = \mathrm{curl} \circ \mathrm{curl}^* - \mathrm{div} \circ \mathrm{grad}$$

- Important Properties:
  Δ_k positive semi-definite
  $\ker(\Delta_k) = \ker(\delta_{k-1}^T) \cap \ker(\delta_k)$: k-Harmonics, dimension equals to k-th Betti number
  Hodge Decomposition Theorem

Theorem 3.1 (Hodge Decomposition)**.** The space of k-forms (cochains) $C^k(\mathcal{K}(G), \mathbb{R})$, admits an orthogonal decomposition into three

$$C^k(\mathcal{K}(G), \mathbb{R}) = \mathrm{im}(\delta_{k-1}) \oplus H_k \oplus \mathrm{im}(\delta_k^T)$$

where

$$H_k = \ker(\delta_{k-1}) \cap \ker(\delta_k^T) = \ker(\Delta_k).$$

- $\dim(H_k) = \beta_k$.

A simple understanding is possible via Dirac operator:

$$D = \delta + \delta^* : \oplus_k L^2(V^k) \to \oplus_k L^2(V^k)$$

Hence $D = D^*$ is self-adjoint. Combine the chain map

$$L^2(V) \xrightarrow{\delta_0} L^2(V^2) \xrightarrow{\delta_1} L^2(V^3) \to \ldots L^2(V^k) \xrightarrow{\delta_{k-1}} L^2(V^{k+1}) \xrightarrow{\delta_k} \ldots$$

into a big operator: Dirac operator.

Abstract Hodge Laplacian:

$$\Delta = D^2 = \delta\delta^* + \delta^*\delta,$$

since $\delta^2 = 0$.

By the Fundamental Theorem of Linear Algebra (Closed Range Theorem in Banach Space),

$$\oplus_k L^2(V^k) = \mathrm{im}(D) \oplus \ker(D)$$

where

$$\mathrm{im}(D) = \mathrm{im}(\delta) \oplus \mathrm{im}(\delta^*)$$

and $\ker(D) = \ker(\Delta)$ is the space of harmonic forms.

4. Applications of Hodge Theory: Statistical Ranking

4.1. HodgeRank on Graphs.

Let $\wedge = \{1, ..., m\}$ be a set of participants and $V = \{1, ..., n\}$ be the set of videos to be ranked. Paired comparison data is collected as a function on $\wedge \times V \times V$, which is *skew-symmetric* for each participant α, *i.e.*, $Y_{ij}^\alpha = -Y_{ji}^\alpha$ representing the degree that α prefers i to j. The simplest setting is the binary choice, where

$$Y_{ij}^\alpha = \begin{cases} 1 & \text{if } \alpha \text{ prefers } i \text{ to } j, \\ -1 & \text{otherwise.} \end{cases}$$

In general, Y_{ij}^α can be used to represent paired comparison grades, *e.g.*, $Y_{ij}^\alpha > 0$ refers to the degree that α prefers i to j and the vice versa $Y_{ji}^\alpha = -Y_{ij}^\alpha < 0$ measures the dispreference degree [JLYY11].

In this paper we shall focus on the binary choice, which is the simplest setting and the data collected in this paper belongs to this case. However the theory can be applied to the more general case with multiple choices above.

Such paired comparison data can be represented by a directed graph, or hypergraph, with n nodes, where each directed edge between i and j refers the preference indicated by Y_{ij}^α.

A nonnegative weight function $\omega : \wedge \times V \times V \longrightarrow [0, \infty)$ is defined as,

$$\omega_{ij}^\alpha = \begin{cases} 1 & \text{if } \alpha \text{ makes a comparison for } \{i, j\}, \\ 0 & \text{otherwise.} \end{cases} \tag{138}$$

It may reflect the confidence level that a participant compares $\{i, j\}$ by taking different values, and this is however not pursued in this paper.

Our statistical rank aggregation problem is to look for some global ranking score $s : V \to R$ such that

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j,\alpha} \omega_{ij}^{\alpha} (s_i - s_j - Y_{ij}^{\alpha})^2, \tag{139}$$

which is equivalent to the following weighted least square problem

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{i,j} \omega_{ij} (s_i - s_j - \hat{Y}_{ij})^2, \tag{140}$$

where $\hat{Y}_{ij} = (\sum_\alpha \omega_{ij}^\alpha Y_{ij}^\alpha)/(\sum_\alpha \omega_{ij}^\alpha)$ and $\omega_{ij} = \sum_\alpha \omega_{ij}^\alpha$. For the principles behind such a choice, readers may refer [**JLYY11**].

A graph structure arises naturally from ranking data as follows. Let $G = (V, E)$ be a paired ranking graph whose vertex set is V, the set of videos to be ranked, and whose edge set is E, the set of video pairs which receive some comparisons, *i.e.*,

$$E = \left\{ \{i,j\} \epsilon \binom{V}{2} \mid \sum_{\alpha} \omega_{i,j}^{\alpha} > 0 \right\}. \tag{141}$$

A pairwise ranking is called *complete* if each participant α in $\wedge$ gives a total judgment of all videos in V; otherwise it is called *incomplete*. It is called *balanced* if the paired comparison graph is k-regular with equal weights $\omega_{ij} = \sum_\alpha \omega_{ij}^\alpha \equiv c$ for all $\{i,j\} \in E$; otherwise it is called *imbalanced*. A complete and balanced ranking induces a complete graph with equal weights on all edges. The existing paired comparison methods in VQA often assume complete and balanced data. However, this is an unrealistic assumption for real world data, *e.g.* randomized experiments. Moreover in crowdsourcing, raters and videos come in an unspecified way and it is hard to control the test process with precise experimental designs. Nevertheless, as to be shown below, it is efficient to utilize some random sampling design based on random graph theory where for each participant a fraction of video pairs are chosen randomly. The HodgeRank approach adopted in this paper enables us a unified scheme which can deal with incomplete and imbalanced data emerged from random sampling in paired comparisons.

The minimization problem (140) can be generalized to a family of *linear models* in paired comparison methods [**Dav88**]. To see this, we first rewrite (140) in another simpler form. Assume that for each edge as video pair $\{i,j\}$, the number of comparisons is n_{ij}, among which a_{ij} participants have a preference on i over j (a_{ji} carries the opposite meaning). So $a_{ij} + a_{ji} = n_{ij}$ if no tie occurs. Therefore, for each edge $\{i,j\} \in E$, we have a preference probability estimated from data $\hat{\pi}_{ij} = a_{ij}/n_{ij}$. With this definition, the problem (140) can be rewritten as

$$\min_{s \in \mathrm{R}^{|V|}} \sum_{\{i,j\} \in E} n_{ij} (s_i - s_j - (2\hat{\pi}_{ij} - 1))^2, \tag{142}$$

since $\hat{Y}_{ij} = (a_{ij} - a_{ji})/n_{ij} = 2\hat{\pi}_{ij} - 1$ due to Equation (138).

General *linear models*, which are firstly formulated by G. Noether [**Noe60**], assume that the true preference probability can be fully decided by a linear scaling function on V, *i.e.*,

$$\pi_{ij} = \mathrm{Prob}\{i \text{ is preferred over } j\} = F(s_i^* - s_j^*), \tag{143}$$

for some $s^* \in \mathrm{R}^{|V|}$. F can be chosen as any symmetric cumulated distributed function. When only an empirical preference probability $\hat{\pi}_{ij}$ is observed, we can map it to a skew-symmetric function by the inverse of F,

$$\hat{Y}_{ij} = F^{-1}(\hat{\pi}_{ij}), \tag{144}$$

where $\hat{Y}_{ij} = -\hat{Y}_{ji}$. However, in this case, one can only expect that

$$\hat{Y}_{ij} = s_i^* - s_j^* + \varepsilon_{ij}, \tag{145}$$

where ε_{ij} accounts for the noise. The case in (142) takes a linear F and is often called a *uniform model*. Below we summarize some well known models which have been studied extensively in [Dav88].

1. *Uniform* model:

$$\hat{Y}_{ij} = 2\hat{\pi}_{ij} - 1. \tag{146}$$

2. *Bradley-Terry* model:

$$\hat{Y}_{ij} = \log \frac{\hat{\pi}_{ij}}{1 - \hat{\pi}_{ij}}. \tag{147}$$

3. *Thurstone-Mosteller* model:

$$\hat{Y}_{ij} = F^{-1}(\hat{\pi}_{ij}). \tag{148}$$

where F is essentially the Gauss error function

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-x/[2\sigma^2(1-\rho)]^{1/2}}^{\infty} e^{-\frac{1}{2}t^2} dt. \tag{149}$$

Note that constants σ and ρ will only contribute to a rescaling of the solution of (140).

4. *Angular transform* model:

$$\hat{Y}_{ij} = \arcsin(2\hat{\pi}_{ij} - 1). \tag{150}$$

This model is created for the so called variance stabilization property: asymptotically $\hat{Y}_{ij}$ has variance only depending on number of ratings on edge $\{i, j\}$ or the weight ω_{ij}, but not on the true probability p_{ij}.

Different models will give different $\hat{Y}_{ij}$ from the same observation $\hat{\pi}_{ij}$, followed by the same weighted least square problem (140) for the solution. Therefore, a deeper analysis of problem (140) will disclose more properties about the ranking problem.

HodgeRank on graph $G = (V, E)$ provides us such a tool, which characterizes the solution and residue of (140), adaptive to topological structures of G. The following theorem adapted from [JLYY11] describes a decomposition of $\hat{Y}$, which can be visualized as edge flows on graph G with direction $i \to j$ if $\hat{Y}_{ij} > 0$ and vice versa. Before the statement of the theorem, we first define the triangle set of G as all the 3-cliques in G.

$$T = \left\{ \{i, j, k\} \epsilon \binom{V}{3} | \{i, j\}, \{j, k\}, \{k, i\} \epsilon E \right\}. \tag{151}$$

Equipped with T, graph G becomes an abstract simplicial complex, the clique complex $\chi(G) = (V, E, T)$.

Theorem 1 [Hodge Decomposition of Paired Ranking] Let $\hat{Y}_{ij}$ be a paired comparison flow on graph $G = (V, E)$, *i.e.*, $\hat{Y}_{ij} = -\hat{Y}_{ji}$ for $\{i, j\} \in E$, and $\hat{Y}_{ij} = 0$ otherwise. There is a unique decomposition of $\hat{Y}$ satisfying

$$\hat{Y} = \hat{Y}^g + \hat{Y}^h + \hat{Y}^c, \tag{152}$$

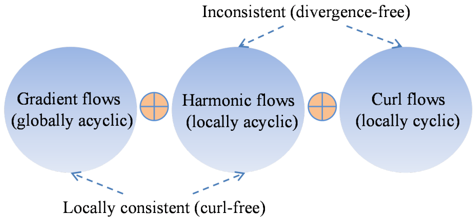

Inconsistent (divergence-free)
Gradient flows (globally acyclic)
Harmonic flows (locally acyclic)
Curl flows (locally cyclic)
Locally consistent (curl-free)

FIGURE 3. Hodge decomposition (three orthogonal components) of paired rankings [JLYY11].

where

$$\hat{Y}^g_{ij} = \hat{s}_i - \hat{s}_j, \quad \text{for some } \hat{s} \in \mathbb{R}^V, \tag{153}$$

$$\hat{Y}^h_{ij} + \hat{Y}^h_{jk} + \hat{Y}^h_{ki} = 0, \ \text{for each } \{i,j,k\} \in T, \tag{154}$$

$$\sum_{j\sim i} \omega_{ij} \hat{Y}^h_{ij} = 0, \ \text{for each } i \in V. \tag{155}$$

The decomposition above is *orthogonal* under the following inner product on $\mathbb{R}^{|E|}$, $\langle u, v\rangle_\omega = \sum_{\{i,j\}\in E} \omega_{ij} u_{ij} v_{ij}$.

The following provides some remarks on the decomposition.

1. When G is connected, $\hat{Y}^g_{ij}$ is a rank two skew-symmetric matrix and gives a linear score function $\hat{s} \in \mathbb{R}^V$ up to translations. We thus call $\hat{Y}^g$ a *gradient flow* since it is given by the difference (discrete gradient) of the score function $\hat{s}$ on graph nodes,

$$\hat{Y}^g_{ij} = (\delta_0 \hat{s})(i,j) := \hat{s}_i - \hat{s}_j, \tag{156}$$

where $\delta_0 : \mathbb{R}^V \to \mathbb{R}^E$ is a finite difference operator (matrix) on G. $\hat{s}$ can be chosen as any least square solution of (140), where we often choose the minimal norm solution,

$$\hat{s} = \Delta_0^\dagger \delta_0^* \hat{Y}, \tag{157}$$

where $\delta_0^* = \delta_0^T W$ ($W = \mathrm{diag}(\omega_{ij})$), $\Delta_0 = \delta_0^* \cdot \delta_0$ is the unnormalized graph Laplacian defined by $(\Delta_0)_{ii} = \sum_{j\sim i} \omega_{ij}$ and $(\Delta_0)_{ij} = -\omega_{ij}$, and $(\cdot)^\dagger$ is the Moore-Penrose (pseudo) inverse. On a complete and balanced graph, (157) is reduced to $\hat{s}_i = \frac{1}{n-1}\sum_{j\neq i} \hat{Y}_{ij}$, often called *Borda Count* as the earliest preference aggregation rule in social choice [JLYY11]. For expander graphs like regular graphs, graph Laplacian Δ_0 has small condition numbers and thus the global ranking is stable against noise on data.

2. $\hat{Y}^h$ satisfies two conditions (154) and (155), which are called *curl-free* and *divergence-free* conditions respectively. The former requires the triangular trace of $\hat{Y}$ to be zero, on every 3-clique in graph G; while the later requires the total sum (inflow minus outflow) to be zero on each node of G. These two conditions characterize a linear subspace which is called *harmonic flows*.

3. The residue $\hat{Y}^c$ actually satisfies (155) but not (154). In fact, it measures the amount of intrinsic (local) inconsistancy in $\hat{Y}$ characterized by the triangular

trace. We often call this component *curl flow*. In particular, the following relative curl,

$$\mathrm{curl}^r_{ijk} = \frac{|\hat{Y}_{ij} + \hat{Y}_{jk} + \hat{Y}_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} = \frac{|\hat{Y}^c_{ij} + \hat{Y}^c_{jk} + \hat{Y}^c_{ki}|}{|\hat{Y}_{ij}| + |\hat{Y}_{jk}| + |\hat{Y}_{ki}|} \in [0, 1], \tag{158}$$

can be used to characterize triangular intransitivity; $\mathrm{curl}^r_{ijk} = 1$ iff $\{i, j, k\}$ contains an intransitive triangle of $\hat{Y}$. Note that computing the percentage of $\mathrm{curl}^r_{ijk} = 1$ is equivalent to calculating the Transitivity Satisfaction Rate (TSR) in complete graphs.

Figure 3 illustrates the Hodge decomposition for paired comparison flows and Algorithm 5 shows how to compute global ranking and other components. The readers may refer to [JLYY11] for the detail of theoretical development. Below we just make a few comments on the application of HodgeRank in our setting.

Algorithm 5: Procedure of Hodge decomposition in Matlab Pseudocodes

```
Input: A paired comparison hypergraph $G$ provide by assessors.
Output: Global score $\hat{s}$, gradient flow $\hat{Y}^g$, curl flow $\hat{Y}^c$, and harmonic flow $\hat{Y}^h$.
Initialization:
$\hat{Y}$ (a numEdge-vector consisting $\hat{Y}_{ij}$ defined),
$W$ (a numEdge-vector consisting $\omega_{ij}$).
Step 1:
Compute $\delta_0$, $\delta_1$; // $\delta_0$ = gradient, $\delta_1$ = curl
$\delta_0^* = \delta_0^T * diag(W)$; // the conjugate of $\delta_0$
$\triangle_0 = \delta_0^* * \delta_0$; // Unnormalized Graph Laplacian
$div = \delta_0^* * \hat{Y}$; // divergence operator
$\hat{s} = lsqr(\triangle_0, div)$; // global score
Step 2:
Compute 1st projection on gradient flow: $\hat{Y}^g = \delta_0 * \hat{s}$;
Step 3:
$\delta_1^* = \delta_1^T * diag(1./W)$;
$\triangle_1 = \delta_1 * \delta_1^*$;
$curl = \delta_1 * \hat{Y}$;
$z = lsqr(\triangle_1, curl)$;
Compute 3rd projection on curl flow: $\hat{Y}^c = \delta_1^* * z$;
Step 4:
Compute 2nd projection on harmonic flow: $\hat{Y}^h = \hat{Y} - \hat{Y}^g - \hat{Y}^c$.
```

1. To find a global ranking $\hat{s}$ in (157), the recent developments of Spielman-Teng [ST04] and Koutis-Miller-Peng [KMP10] suggest fast (almost linear in $|E|\mathrm{Poly}(\log|V|)$) algorithms for this purpose.

2. Inconsistency of $\hat{Y}$ has two parts: global inconsistency measured by harmonic flow $\hat{Y}^h$ and local inconsistency measured by curls in $\hat{Y}^c$. Due to the orthogonal decomposition, $\|\hat{Y}^h\|^2_\omega/\|\hat{Y}\|^2_\omega$ and $\|\hat{Y}^c\|^2_\omega/\|\hat{Y}\|^2_\omega$ provide percentages of global and local inconsistencies, respectively.

3. A nontrivial harmonic component $\hat{Y}^h \neq 0$ implies the fixed tournament issue, *i.e.*, for any candidate $i \in V$, there is a paired comparison design by removing some of the edges in $G = (V, E)$ such that i is the overall winner.

4. One can control the harmonic component by controlling the topology of clique complex $\chi(G)$. In a loop-free clique complex $\chi(G)$ where $\beta_1 = 0$, harmonic component vanishes. In this case, there are no cycles which traverse all the nodes, *e.g.*, $1 \succ 2 \succ 3 \succ 4 \succ \ldots \succ n \succ 1$. All the inconsistency will be summarized in those triangular cycles, *e.g.*, $i \succ j \succ k \succ i$.

Theorem 2. The linear space of harmonic flows has the dimension equal to β_1, *i.e.*, the number of independent loops in clique complex $\chi(G)$, which is called the first order Betti number.

Fortunately, with the aid of some random sampling principles, it is not hard to obtain graphs whose β_1 are zero.

4.2. Random Graphs.

In this section, we first describe two classical random models: Erdös-Rényi random graph and random regular graph; then we investigate the relation between them.

4.2.1. *Erdös-Rényi Random Graph.* Erdös-Rényi random graph $G(n, p)$ starts from n vertices and draws its edges independently according to a fixed probability p. Such random graph model is chosen to meet the scenario that in crowdsourcing ranking raters and videos come in an unspecified way. Among various models, Erdös-Rényi random graph is the simplest one equivalent to I.I.D. sampling. Therefore, such a model is to be systematically studied in the paper.

However, to exploit Erdös-Rényi random graph in crowdsourcing experimental designs, one has to meet some conditions depending on our purpose:

1. *The resultant graph should be connected, if we hope to derive global scores for all videos in comparison*;

2. *The resultant graph should be loop-free in its clique complex, if we hope to get rid of the global inconsistency in harmonic component.*

The two conditions can be easily satisfied for large Erdös-Rényi random graph.

Theorem 3. Let $G(n, p)$ be the set of Erdös-Rényi random graphs with n nodes and edge appearance probability p. Then the following holds as $n \to \infty$,

1. [Erdös-Rényi 1959] [**ER59**] if $p \succ \log n/n$, then $G(n, p)$ is almost always connected; and if $p \prec \log n/n$ then $G(n, p)$ is almost always disconnected;

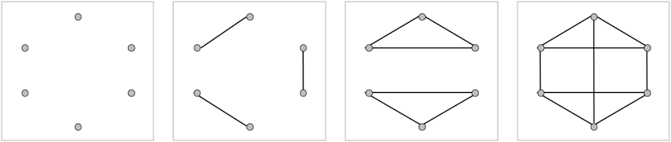
2. [Kahle 2009] [**Kah09, Kah13**] if $p = O(n^{\alpha})$, with $\alpha < -1$ or $\alpha > -1/2$, then the expected β_1 of the clique complex $\chi(G(n, p))$ is almost always equal to zero, *i.e.*, loop-free.

These theories imply that when p is large enough, Erdös-Rényi random graph will meet the two conditions above with high probability. In particular, almost linear $O(n \log n)$ edges suffice to derive a global ranking, and with $O(n^{3/2})$ edges harmonic-free condition is met.

Despite such an asymptotic theory for large random graphs, it remains a question how to ensure that a given graph instance satisfies the two conditions? Fortunately, the recent development in computational topology provides us such a tool, persistent homology, which will be illustrated in Section **??**.

5. Euler-Calculus

```
to be finished...
```



(a) 0-regular graph (b) 1-regular graph (c) 2-regular graph (d) 3-regular graph

Figure 4. Examples of k-regular graphs.

Bibliography

[AC09] R DeVore A Cohen, W Dahmen, *Compressed sensing and best k-term approximation*, J. Amer. Math. Soc **22** (2009), no. 1, 211–231.

[Ach03] Dimitris Achlioptas, *Database-friendly random projections: Johnson-lindenstrauss with binary coins*, Journal of Computer and System Sciences **66** (2003), 671687.

[Ali95] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim. **5** (1995), no. 1, 13–51.

[Aro50] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society **68** (1950), no. 3, 337–404.

[Bav11] Francois Bavaud, *On the schoenberg transformations in data analysis: Theory and illustrations*, Journal of Classification **28** (2011), no. 3, 297–314.

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin, *A simple proof of the restricted isometry property for random matrices*, Constructive Approximation **28** (2008), no. 3, 253–263.

[BLT$^+$06] P. Biswas, T.-C. Liang, K.-C. Toh, T.-C. Wang, and Y. Ye, *Semidefinite programming approaches for sensor network localization with noisy distance measurements*, IEEE Transactions on Automation Science and Engineering **3** (2006), 360–371.

[BN01] Mikhail Belkin and Partha Niyogi, *Laplacian eigenmaps and spectral techniques for embedding and clustering*, Advances in Neural Information Processing Systems (NIPS) 14, MIT Press, 2001, pp. 585–591.

[BN03] Mikhail Belkin and Partha Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computation **15** (2003), 1373–1396.

[BN08] Mikhail Belkin and Partha Niyogi, *Convergence of laplacian eigenmaps*, Tech. report, 2008.

[BP98] Sergey Brin and Larry Page, *The anatomy of a large-scale hypertextual web search engine*, Proceedings of the 7th international conference on World Wide Web (WWW) (Australia), 1998, pp. 107–117.

[BS10] Zhidong Bai and Jack W. Silverstein, *Spectral analysis of large dimensional random matrices*, Springer, 2010.

[BTA04] Alain Berlinet and Christine Thomas-Agnan, *Reproducing kernel hilbert spaces in probability and statistics*, Kluwer Academic Publishers, 2004.

[Can08] E. J. Candès, *The restricted isometry property and its implications for compressed sensing*, Comptes Rendus de l'Académie des Sciences, Paris, Série I **346** (2008), 589–592.

[CDS98] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing **20** (1998), 33–61.

[Chu05] Fan R. K. Chung, *Laplacians and the cheeger inequality for directed graphs*, Annals of Combinatorics **9** (2005), no. 1, 1–19.

[CL06] Ronald R. Coifman and Stéphane. Lafon, *Diffusion maps*, Applied and Computational Harmonic Analysis **21** (2006), 5–30.

[CLL$^+$05] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps i*, Proceedings of the National Academy of Sciences of the United States of America **102** (2005), 7426–7431.

[CLMW09] E. J. Candès, Xiaodong Li, Yi Ma, and John Wright, *Robust principal component analysis*, Journal of ACM **58** (2009), no. 1, 1–37.

[CMOP11] Ozan Candogan, Ishai Menache, Asuman Ozdaglar, and Pablo A. Parrilo, *Flows and decompositions of games: Harmonic and potential games*, Mathematics of Operations Research **36** (2011), no. 3, 474–503.

[CPW12] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky, *Latent variable graphical model selection via convex optimization (with discussion)*, Annals of Statistics (2012), to appear, http://arxiv.org/abs/1008.1290.

[CR09] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Foundation of Computational Mathematics **9** (2009), no. 6, 717772.

[CRPW12] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, Foundation of Computational Mathematics (2012), to appear, http://arxiv.org/abs/1012.0621.

[CRT06] Emmanuel. J. Candès, Justin Romberg, and Terrence Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. on Info. Theory **52** (2006), no. 2, 489–509.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A. Willsky, *Rank-sparsity incoherence for matrix decomposition*, SIAM Journal on Optimization **21** (2011), no. 2, 572596, http://arxiv.org/abs/0906.2220.

[CST03] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, 2003.

[CT05] E. J. Candès and Terrence Tao, *Decoding by linear programming*, IEEE Trans. on Info. Theory **51** (2005), 4203–4215.

[CT06] Emmanuel. J. Candès and Terrence Tao, *Near optimal signal recovery from random projections: Universal encoding strategies*, IEEE Trans. on Info. Theory **52** (2006), no. 12, 5406–5425.

[CT10] E. J. Candès and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Transaction on Information Theory **56** (2010), no. 5, 2053–2080.

[Dav88] H. David, *The methods of paired comparisons, 2nd ed.*, Griffin's Statistical Monographs and Courses, 41, Oxford University Press, New York, NY, 1988.

[DG03a] Sanjoy Dasgupta and Anupam Gupta, *An elementary proof of a theorem of johnson and lindenstrauss*, Random Structures and Algorithms **22** (2003), no. 1, 60–65.

[DG03b] David L. Donoho and Carrie Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data*, Proceedings of the National Academy of Sciences of the United States of America **100** (2003), no. 10, 5591–5596.

[dGJL07] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet, *A direct formulation for sparse pca using semidefinite programming*, SIAM Review **49** (2007), no. 3, http://arxiv.org/abs/cs/0406021.

[DH01] David L. Donoho and Xiaoming Huo, *Uncertainty principles and ideal atomic decomposition*, IEEE Transactions on Information Theory **47** (2001), no. 7, 2845–2862.

[EB01] M. Elad and A.M. Bruckstein, *On sparse representations*, International Conference on Image Processing (ICIP) (Tsaloniky, Greece), November 2001.

[ELVE08] Weinan E, Tiejun Li, and Eric Vanden-Eijnden, *Optimal partition and effective dynamics of complex networks*, Proc. Nat. Acad. Sci. **105** (2008), 7907–7912.

[ER59] P. Erdos and A. Renyi, *On random graphs i*, Publ. Math. Debrecen **6** (1959), 290–297.

[EST09] Ioannis Z. Emiris, Frank J. Sottile, and Thorsten Theobald, *Nonlinear computational geometry*, Springer, New York, 2009.

[EVE06] Weinan E and Eric Vanden-Eijnden, *Towards a theory of transition paths*, J. Stat. Phys. **123** (2006), 503–523.

[EVE10] Weinan E and Eric Vanden-Eijnden, *Transition-path theory and path-finding algorithms for the study of rare events*, Annual Review of Physical Chemistry **61** (2010), 391–420.

[Gro11] David Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Transaction on Information Theory **57** (2011), 1548, arXiv:0910.1879.

[HAvL05] M. Hein, J. Audibert, and U. von Luxburg, *From graphs to manifolds: weak and strong pointwise consistency of graph laplacians*, COLT, 2005.

[JL84] W. B. Johnson and J. Lindenstrauss, *Extensions of lipschitz maps into a hilbert space*, Contemp Math **26** (1984), 189–206.

[JLYY11] Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye, *Statistical ranking and combinatorial hodge theory*, Mathematical Programming **127** (2011), no. 1, 203–244, arXiv:0811.1067 [stat.ML].

[Joh06] I. Johnstone, *High dimensional statistical inference and random matrices*, Proc. International Congress of Mathematicians, 2006.

[JYLG12] Xiaoye Jiang, Yuan Yao, Han Liu, and Leo Guibas, *Detecting network cliques with radon basis pursuit*, The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS) (La Palma, Canary Islands), April 21-23 2012.

[Kah09] Matthew Kahle, *Topology of random clique complexes*, Discrete Mathematics **309** (2009), 1658–1671.

[Kah13] ______, *Sharp vanishing thresholds for cohomology of random flag complexes*, Annals of Mathematics (2013), arXiv:1207.0149.

[Kle99] Jon Kleinberg, *Authoritative sources in a hyperlinked environment*, Journal of the ACM **46** (1999), no. 5, 604–632.

[KMP10] Ioannis Koutis, G. Miller, and Richard Peng, *Approaching optimality for solving sdd systems*, FOCS '10 51st Annual IEEE Symposium on Foundations of Computer Science, 2010.

[KN08] S. Kritchman and B. Nadler, *Determining the number of components in a factor model from limited noisy data*, Chemometrics and Intelligent Laboratory Systems **94** (2008), 19–32.

[LL11] Jian Li and Tiejun Li, *Probabilistic framework for network partition*, Phys. A **390** (2011), 3579.

[LLE09] Tiejun Li, Jian Liu, and Weinan E, *Probabilistic framework for network partition*, Phys. Rev. E **80** (2009), 026106.

[LM06] Amy N. Langville and Carl D. Meyer, *Google's pagerank and beyond: The science of search engine rankings*, Princeton University Press, 2006.

[LZ10] Yanhua Li and Zhili Zhang, *Random walks on digraphs, the generalized digraph laplacian, and the degree of asymmetry*, Algorithms and Models for the Web-Graph, Lecture Notes in Computer Science, vol. 6516, 2010, pp. 74–85.

[Mey00] Carl D. Meyer, *Matrix analysis and applied linear algebra*, SIAM, 2000.

[MSVE09] Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden, *Transition path theory for markov jump processes*, Multiscale Model. Simul. **7** (2009), 1192.

[MY09] Nicolai Meinshausen and Bin Yu, *Lasso-type recovery of sparse representations for high-dimensional data*, Annals of Statistics **37** (2009), no. 1, 246–270.

[NBG10] R. R. Nadakuditi and F. Benaych-Georges, *The breakdown point of signal subspace estimation*, IEEE Sensor Array and Multichannel Signal Processing Workshop (2010), 177–180.

[Noe60] G. Noether, *Remarks about a paired comparison model*, Psychometrika **25** (1960), 357–367.

[NSVE$^+$09] Frank Noè, Christof Schütte, Eric Vanden−Eijnden, Lothar Reich, and Thomas R. Weikl, *Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations*, Proceedings of the National Academy of Sciences of the United States of America **106** (2009), no. 45, 19011–19016.

[RL00] Sam T. Roweis and Saul K. Lawrence, *Locally linear embedding*, Science **290** (2000), no. 5500, 2319–2323.

[Sch37] I. J. Schoenberg, *On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in hilbert space*, The Annals of Mathematics **38** (1937), no. 4, 787–793.

[Sch38a] ______, *Metric spaces and completely monotone functions*, The Annals of Mathematics **39** (1938), 811–841.

[Sch38b] ______, *Metric spaces and positive denite functions*, Transactions of the American Mathematical Society **44** (1938), 522–536.

[Sin06] Amit Singer, *From graph to manifold laplacian: The convergence rate*, Applied and Computational Harmonic Analysis **21** (2006), 128–134.

[ST04] D. Spielman and Shang-Hua Teng, *Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems*, STOC '04 Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, 2004.

[Ste56] Charles Stein, *Inadmissibility of the usual estimator for the mean of a multivariate distribution*, Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability **1** (1956), 197–206.

[SW12] Amit Singer and Hau-Tieng Wu, *Vector diffusion maps and the connection laplacian*, Comm. Pure Appl. Math. **65** (2012), no. 8, 1067–1144.

[SY07] Anthony Man-Cho So and Yinyu Ye, *Theory of semidefinite programming for sensor network localization*, Mathematical Programming, Series B **109** (2007), no. 2-3, 367–384.

[SYZ08] Anthony Man-Cho So, Yinyu Ye, and Jiawei Zhang, *A unified theorem on sdp rank reduction*, Mathematics of Operations Research **33** (2008), no. 4, 910–920.

[Tao11] Terrence Tao, *Topics in random matrix theory*, Lecture Notes in UCLA, 2011.

[TdL00] J. B. Tenenbaum, Vin deSilva, and John C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), 2319–2323.

[TdSL00] J. Tenenbaum, V. de Silva, and J. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), no. 5500, 2323–2326.

[Tib96] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. of the Royal Statistical Society, Series B **58** (1996), no. 1, 267–288.

[Tro04] Joel A. Tropp, *Greed is good: Algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory **50** (2004), no. 10, 2231–2242.

[Tsy09] Alexandre Tsybakov, *Introduction to nonparametric estimation*, Springer, 2009.

[Vap98] V. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.

[Vem04] Santosh Vempala, *The random projection method*, Am. Math. Soc., Providence, 2004.

[Wah90] Grace Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics 59, SIAM, 1990.

[WS06] Killian Q. Weinberger and Lawrence K. Saul, *Unsupervised learning of image manifolds by semidefinite programming*, International Journal of Computer Vision **70** (2006), no. 1, 77–90.

[YH41] G. Young and A. S. Householder, *A note on multidimensional psycho-physical analysis*, Psychometrika **6** (1941), 331–333.

[ZHT06] H. Zou, T. Hastie, and R. Tibshirani, *Sparse principal component analysis*, Journal of Computational and Graphical Statistics **15** (2006), no. 2, 262–286.

[ZY06] Peng Zhao and Bin Yu, *On model selection consistency of lasso*, J. Machine Learning Research **7** (2006), 2541–2567.

[ZZ02] Zhenyue Zhang and Hongyuan Zha, *Principal manifolds and nonlinear dimension reduction via local tangent space alignment*, SIAM Journal of Scientific Computing **26** (2002), 313–338.

[ZZ09] Hongyuan Zha and Zhenyue Zhang, *Spectral properties of the alignment matrices in manifold learning*, SIAM Review **51** (2009), no. 3, 545–566.