Mathematics for Data Sciences	**November 8, 2011**

Lecture 11. PCA vs. MDS: Schoenberg Theory

Instructor: Yuan Yao, Peking University *Scribe: Deng, Yanzhen; Li, Changcheng; Ren, Jie*

Introduction

In this lecture, we shall introduce Multi-Dimensional Scaling (MDS) which is equivalent to PCA when pairwise Euclidean distances are known among data points.

First, recall PCA: given a set of points $x_1, x_2, ..., x_n \in \mathbb{R}^p$, let

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

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

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

Define

$$\widetilde{X} = X - \frac{1}{n}X\cdot\mathbf{1}\cdot\mathbf{1}^T \text{ or } \widetilde{x}_i = x_i - \widehat{u}_n$$

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

Then PCA is given by eigen-decomposition of $\widetilde{X}\cdot\widetilde{X}^T{}_{p\times p}$. Top k eigenvectors give rise to a k-dimensional embedding of data.

For PCA, given (centralized) Euclidean coordinate $\widetilde{X}$, we can get the inner product (or pairwise distances between points) $\widetilde{X}\cdot\widetilde{X}^T$ which is a $p\times p$ matrix.

For MDS, an inverse problem is raised: given inner product, or equivalently pairwise distances between points, can we find a system of Euclidean coordinates for data points? Such a metric embedding problem has a long history tracing back to 1930s and leads to many important modern fields in data analysis.

1 Classical MDS

In this section we study classical MDS, or metric Multidimensional scaling problem.

The distance between point x_i and x_j is

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

General ideas of classic (metric) MDS is:

1. transform squared distance matrix D 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

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

Originally, we have

$$\widetilde{x_i} = x_i - \hat{\mu}_n = x_i - \frac{1}{n} \cdot X \cdot \mathbf{1}.$$

or

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

Thus,

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

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,

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

That is,

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

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

2 MDS algorithm

Based on the theory above, we could conclude a algorithm of MDS as following,

Given the squared distance matrix $D^{n \times n}$, which is symmetric matrix,

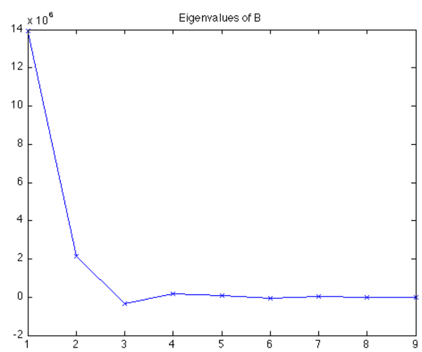
(1) Compute $B = -\frac{1}{2} H \cdot D \cdot H^T$, where H is a centering matrix.

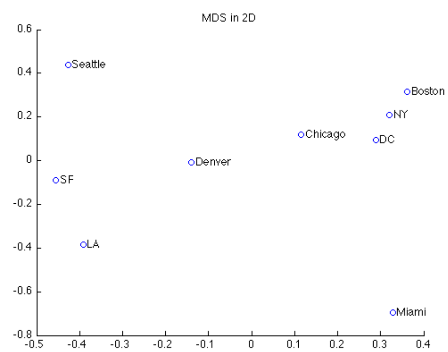
(2) Do eigen-decomposition of B, such that $B = U \Lambda U^T$. Choose the top-k eigenvectors, define $\widetilde{X} = {\Lambda_k}^{\frac{1}{2}} U_k^T$ where $U_k = [u_1, \ldots, u_k]$ ($u_k \in \mathbb{R}^n$) and $\Lambda_k = \text{diag}(\lambda_1, \ldots, \lambda_k)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$. This is the k-dimensional Euclidean coordinations for the n points in lower dimension.

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.

		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 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.

3 Theory of MDS (Young/Househölder/Schoenberg'1938)

Definition (Positive Semi-definite)**.** Suppose $A^{n \times n}$ is a real symetric marix,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 symetric marix, $A \succeq 0$, $B \succeq 0$. Then we have:

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)**.** Suppose $A^{n\times n}$ is a real symetric marix,then:
A is c.n.d.(conditionally negative definite) $\iff \forall v \in \mathbb{R}^n, \mathbf{1}v^T = \sum_{i=1}^n v_i = 0$, we have $v^T A v \le 0$

Lemma 3.1 (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.

$\Leftarrow$ $\forall x \in \mathbb{R}^n$

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

Since $\mathbf{1}^T \cdot H_\alpha^T x = \mathbf{1}^T \cdot (\mathbf{I} - \alpha \cdot \mathbf{1}^T)x = (1 - \mathbf{1}^T \cdot \alpha)\mathbf{1}^T \cdot x = 0$ ($\mathbf{1}^T \cdot \alpha = 1$ for signed probability measure) and C is c.n.d., we have:

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

So B_α is p.s.d.

$\Rightarrow$ $\forall x \in \mathbb{R}^n$ satisfies $\mathbf{1}^T \cdot x = 0$, we have:

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

Thus,

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

So, C is c.n.d.

This completes the proof. □

Theorem 3.2 (Classical MDS)**.** Let $D^{n\times n}$ a real symmetric matrix. $C = D - \frac{1}{2} d \cdot \mathbf{1}^T - \frac{1}{2}\mathbf{1} \cdot d^T$, $d = \text{diag}(D)$. 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. Accoding to Lemma 3.1 and the first part of Theorem 3.2: C c.n.d. $\iff$ B p.s.d $\iff$ D c.n.d.

4. Accoding to Lemma 3.1 and the second part of Theorem 3.2:
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 matrix.

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

- $\phi_1(d) = \dfrac{1 - \exp(-ad)}{a}$ 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 $g_3(\lambda) = \lambda \exp(-a\lambda)$ (see more in Bavoud 2010).

Note that Schoeberg transform satisfies $\phi'(d) = \int_0^\infty exp(-\lambda d) g(\lambda) d\lambda$ and $\phi''(d) = -\int_0^\infty \exp(-\lambda d)\lambda g(\lambda) d\lambda$, etc. In other words, ϕ is related to the so called *completely monotonic functions* $(-1)^n f^{(n)}(x) \geq 0$.

Theorem 3.3 (Schoenberg Transform)**.** Given D a square distance matrix, $C_{i,j} = \Phi(D_{i,j})$. Then:
C is a square distance matrix $\iff$ Φ is Shoenberg Transform.