MDS and PCA

Lecture 1.

Yuan Yao

Peking Univesity

Geometric Embedding

- A Fundamental Problem in Data Representation
- Unstructured data -> Euclidean Space
- a.k.a. 'feature' learning (e.g. deep learning)
- speech, text, image, video...

Multidimensional Scaling

- Given pairwise distances between data points, can we find a *system* of Euclidean coordinates for those points whose pairwise distances meet given constraints?

		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

Inverse problem: D->X?

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 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.

From inner product to squared distance

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

Centered the data

Define the mean and the centered data

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

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

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

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

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*.

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

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,

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

Inner product matrix!

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

Note that often we define the covariance matrix

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

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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$;

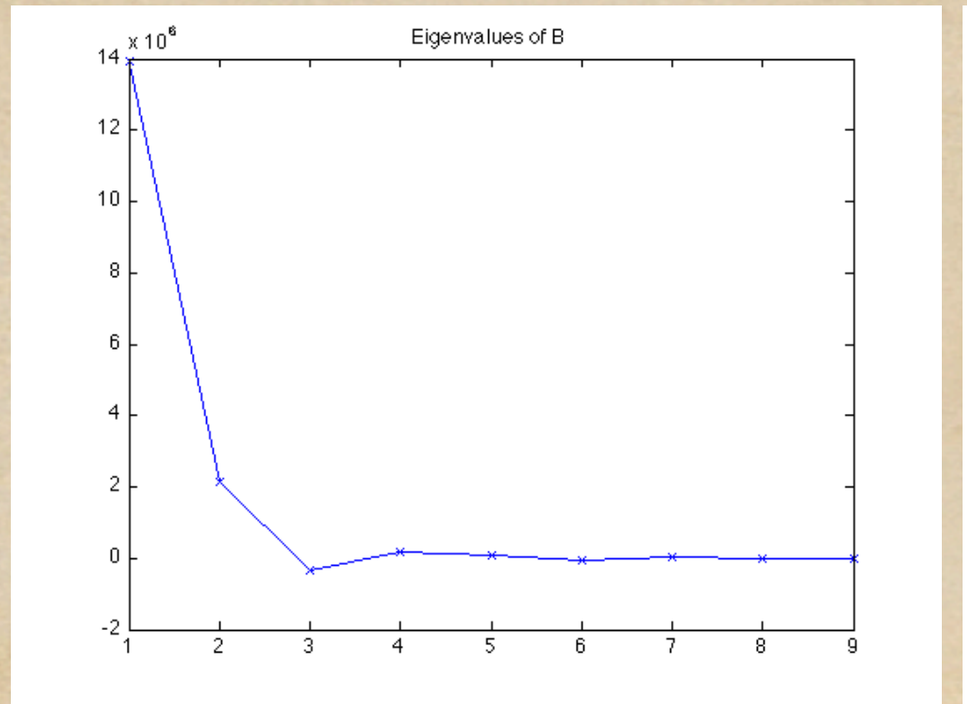
3 Choose top k nonzero eigenvalues and corresponding eigenvectors, $\widetilde{X}_k = U_k \Lambda_k{}^{\frac{1}{2}}$ where

$$U_k = [u_1, \ldots, u_k], \quad u_k \in \mathbb{R}^n,$$
$$\Lambda_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

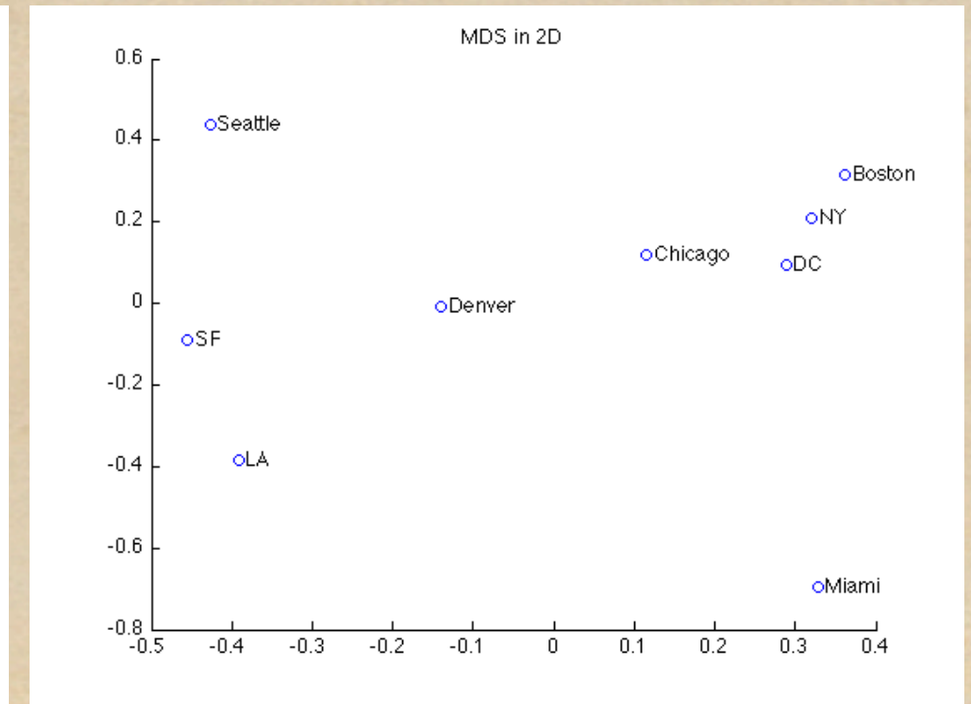
with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$.

		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)

P.S.D. as inner product

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 $(A \circ B)_{i,j} := A_{i,j} \times B_{i,j}$.

C.N.D.

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 $v^T A v \leq 0$

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: $H_\alpha = \mathbf{I} - \mathbf{1} \cdot \alpha^T$.

Young-Household-Schoenberg Theorem

Theorem 2.2 (Classical MDS)**.** Let $D^{n\times n}$ a real symmetric matrix. $C = D - \frac{1}{2}d\cdot \mathbf{1}^T - \frac{1}{2}\mathbf{1}\cdot d^T$, $d = \mathrm{diag}(D)$. 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. $C_{i,j} = \sum_{m=1}^{k}(y_{i,m} - y_{j,m})^2$)

Schoenberg Transform

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

$$C \text{ is a square distance matrix} \iff \Phi \text{ is Shoenberg Transform.}$$

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]).

Positive Definite functions

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.

Reproducing Kernel Hilbert Spaces

- k(x,y)=k(y,x) p.d.
- define functions $k_x(\cdot) = k(x, \cdot)$
- take the span
- define inner product $\langle k_x, k_y \rangle = k(x, y)$
- take the closer => RKHS

Universality of RKHS

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**].

Dimensionality Reduction

Find low dimensional embedding

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

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 $Y_i \in \mathbb{R}^k$:

$$\min \|Y^T Y - B\|_F^2$$

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 $\widetilde{X} = USV^T$.

B is Gram matrix or kernel matrix

PCA

- PCA is given by the top k eigenvector of covariance matrix

$$\widehat{\Sigma}_n = \frac{1}{n-1}\widetilde{X} \cdot \widetilde{X}^T$$

Both MDS and PCA are given by SVD of centered data matrix.

Geometry of PCA

Let $X = [X_1|X_2|\cdots|X_n] \in \mathbb{R}^{p\times n}$.

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

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 repre-

Best k-affine space approximation of data

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

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

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

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

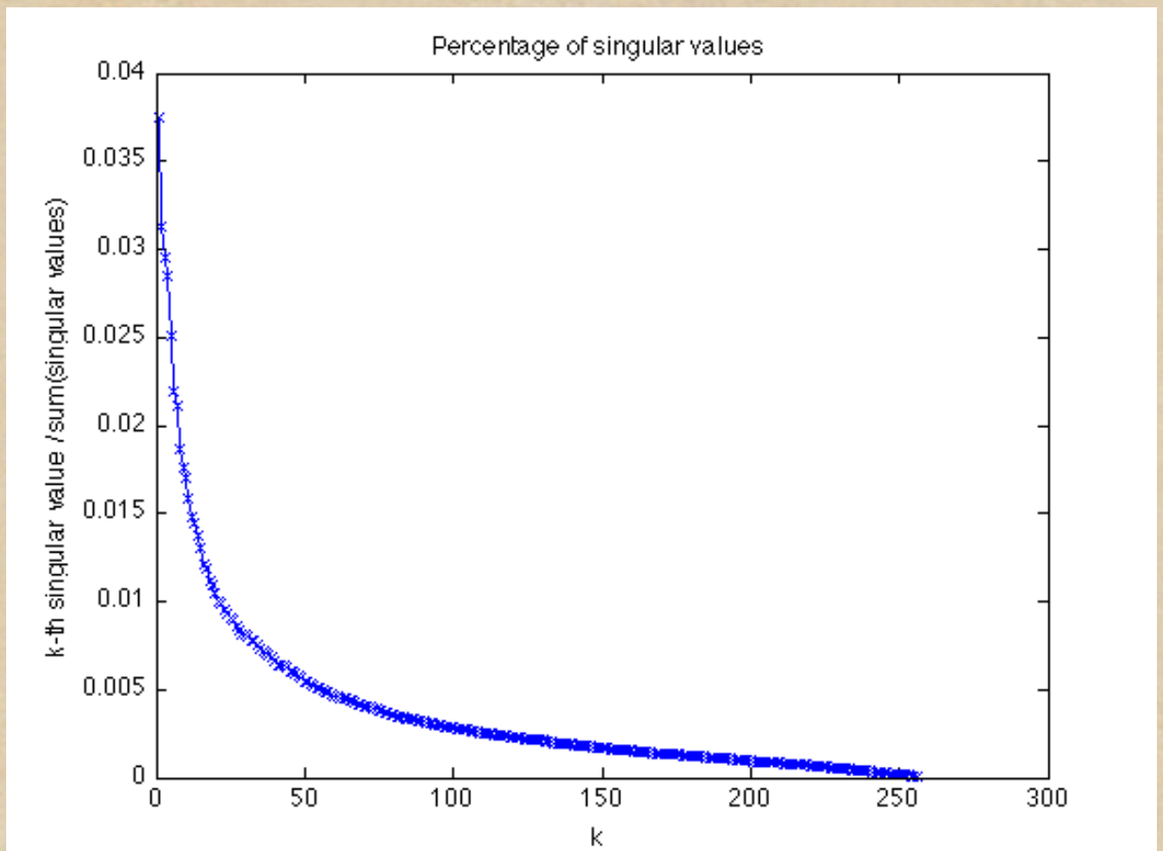
Above we use cyclic property of trace and idempotent property of projection.

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

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

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

(a)


Percentage of singular values
k-th singular value /sum(singular values)
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
50
100
150
200
250
300
k

(b)


≈
- 2.52
- 0.64
+ 2.02

(c)

MDS and PCA=SVD

(SVD) of $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times p}$ in the following sense,

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

- top k left singular vectors give MDS (Kernel spectrum)
- top k right singular vectors give PCA (Covariance spectrum)