

# Topological and Geometric Data Reduction & Visualization

---

CSIC 5011

1027 LSK  
HKUST

YUAN YAO



# Lecture 1. Introduction & Syllabus

Sep. 1, 2017

This course plans to cover the following topics. Extensive experimental projects are conducted.

## I. Geometric Data Analysis

(1) Dual Geometry of Principal Component Analysis (PCA) and

Multidimensional Scaling (MDS):

(2) Robust PCA: PCA with outliers

(3) Sparse PCA: PCA with variable selection

(4) Manifold Learning:

Locally Linear Embedding (LLE) from PCA

Laplacian LLE, Diffusion map, LTSA etc.

ISOMAP from MDS

Graph Realization as manifold local MDS

(5) Supervised PCA:

Ridge Regression and PCA

Slice Inverse Regression (SIR)

Classification and Linear Discriminant Analysis (LDA)

(6) Further representation learning

tSNE

Steerable PCA

Dictionary Learning and matrix factorization

Deep learning

## II. Topological data analysis

(1) clustering

k-center

k-means

hierarchical linkage

(2) Morse theory and Topological data analysis

Reeb graph and mapper

Persistent homology and discrete Morse theory

\*Critical nodes and graphs

(3) \*Euler Calculus and signal processing

Spectral Method

"Can you hear the shape of a drum?"

by Hermann Weyl

Mark Kac

⇒ Can you hear the shape of data?

- (4) Connecting geometry and topology: Hodge Theory
- Spectral clustering and graph Laplacian
- Statistical Ranking and graph Helmholtzian
- Game Theory

Courseweb

<http://math.stanford.edu/~ynany/course/2017.fall>

Time & Venue

Mon. 3 - 4:20pm

Fri. 10:30 - 11:50 am

1027 LSK

No Final Exam!

Yes: weekly homeworks (?), no grading (no TA)

monthly mini-projects

Final Project

Peer Reviews !! (poster workshop/ github reports

                   w. doodle vote).

take a look at previous course

# Lecture 4. High dimensionality & Random Projection

Sep. 11, 2017

Recall

data  $X = [x_1, \dots, x_n]^{\text{pxn}}$

centered data  $X_c = XH$ ,  $H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^T$ ,  $\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$

k-SVD  $X_c \cong \hat{U}_k \hat{S}_k \hat{V}_k^T$  as best rank-k approximation

$\hat{U}_k \in \mathbb{R}^{p \times k}$  of  $X_c$

$\hat{V}_k \in \mathbb{R}^{n \times k}$  orthogonal column mat.

$\hat{S}_k = \text{diag}(\hat{\sigma}_1, \dots, \hat{\sigma}_k)$ ,  $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \dots \geq \hat{\sigma}_k > 0$

"k-PCA" is given by  $(\hat{U}_k, \hat{S}_k)$  with projection

$$\hat{B} = [\hat{b}_1, \dots, \hat{b}_n] = \hat{U}_k^T X_c \approx \underbrace{\hat{S}_k}_{\text{rank } k} \underbrace{\hat{V}_k^T}_{\text{rank } n}$$

each column gives new coordinates

$\Leftrightarrow$  Eigenvalue Decomposition of Covariance Mat.

$$\hat{\Sigma}_n = \frac{1}{n} X_c X_c^T \cong \hat{U}_k \hat{\Lambda}_k \hat{U}_k^T, \Lambda_k = \hat{S}_k^2$$

"k-MDS" is given by  $(\hat{S}_k, \hat{V}_k)$  with data representation

$$\underbrace{\hat{S}_k \hat{V}_k^T}_{\text{rank } k} \in \mathbb{R}^{k \times n}$$

$\Leftrightarrow$  Eigenvalue Decomp. of Kernel Mat.

$$\hat{K} = \frac{1}{n} X_c X_c^T \cong \hat{V}_k \hat{\Lambda}_k \hat{V}_k^T$$

"kernel-PCA/MDS"

$K \geq 0$  is positive semidefinite

$$B = HKH^T \cong \hat{V}_k \hat{\Lambda}_k \hat{V}_k^T$$

Problem: What about big data & high dimensionality?

big data  $n \gg 1$ ,  $\sum_n = \frac{1}{n} \sum_i (x_i - \hat{\mu})(x_i - \hat{\mu})^T$

down sample  $n' \rightarrow$  good approximation of  $\sum_n$  &  $\hat{V}_k$  restricted on subsample

high dimensionality  $p \gg 1$ ,  $\sum_n \in \mathbb{R}^{p \times p}$  too big to compute

$$K = X_C^T X_C ?$$

easy to approximate?

## Random Projection!

e.g.,  $R = \frac{1}{\sqrt{d}} A^{d \times p}$  where  $A_{ij} \sim \mathcal{N}(0, 1)$ .

$X_C^{p \times n} \mapsto (R X_C)^{d \times n}$ ,  $d \ll p$ .

$K_R = X_C^T R^T R X_C$  a good approximation of  $K$ !

$$\bullet A_{ij} = \begin{cases} 1 & p = \frac{1}{2} \\ -1 & p \neq \frac{1}{2} \end{cases}$$

$$\bullet A_{ij} = \begin{cases} 1 & p = \frac{1}{6} \\ 0 & 1-p = \frac{2}{3} \\ -1 & p = \frac{1}{6} \end{cases}$$

sparse with many zeros!

## Example (Human Genome Diversity Project / HGDP)

<http://www.cephb.fr/en/hgdp-panel.php>

$n = 1064$  persons       $p = 644,258$  SNPs

$X^{p \times n} : x_{ij} = 0 : "AA"; 1 : "AC"; 2 : "CC"; 9 : "Missing"$

Removing 21 persons with missing values.

$$X^{644,258 \times 1043}$$

$R^{d \times p}$  randomly select  $d$  rows/SNPs of  $X$ .

$$\tilde{X}^{d \times n} = RX_C = RXH \cong \hat{U}_k \hat{S}_k \hat{V}_k^T$$

$$d = p = 600K$$

$$d = 5K$$

$$d = 100K$$

$\hat{U}_k \hat{S}_k \hat{V}_k^T$

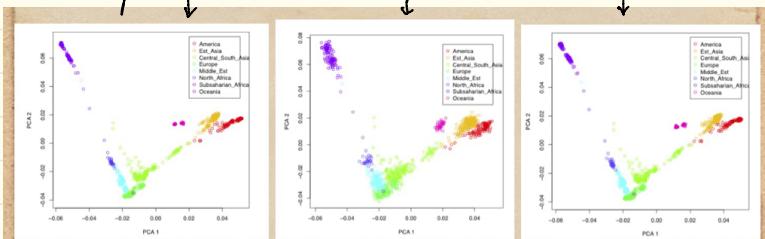


FIGURE 1. (Left) Projection of 1043 individuals on the top 2 MDS principal components. (Middle) MDS computed from 5,000 random projections. (Right) MDS computed from 100,000 random projections. Pictures are due to Qing Wang.

In all cases,  $(\hat{S}_{k,d}, \hat{V}_{k,d})$  are good results!

Here: PCA coordinates:  $\hat{S}_{k,d} \hat{V}_{k,d}^T \in \mathbb{R}^{K \times n}$ ,  $k=2$ .

Why does it work?

## Johnson - Lindenstrauss Lemma

Idea:  $x_i \in \mathbb{R}^P$ ,  $d_{ij} = \|x_i - x_j\|$ ,  $i=1, \dots, n$

Look for a transform  $f: X_i \mapsto Y_i \in \mathbb{R}^d$ ,  $d = O(c \alpha \log n)$

s.t.

$$1 - \varepsilon \leq \frac{\|Y_i - Y_j\|}{\|x_i - x_j\|} \leq 1 + \varepsilon \text{ with probability } \geq 1 - n^{-\alpha}, \alpha > 0$$

Uniform  $\varepsilon$ -Isometry!

relative metric-distortion is uniformly bounded by  $\varepsilon$ !

$f$  is a random projection!

1980s Johnson - Lindenstrauss Lipschitz Extension

2003 Sanjoy Dasgupta, Anupam Gupta  
Dimitris Achlioptas

Computer Science, data compression  
nearest neighbor search

Then Given  $\varepsilon \in (0, 1)$ ,  $n, \alpha > 0$ .

Let

$$k = c(\alpha, \varepsilon) \log n = (4+2\alpha) \left( \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right)^{-1} \log n$$

Then for any  $n$  points  $x_i \in \mathbb{R}^P$  ( $i=1, \dots, n$ ), there exists a map  $f: \mathbb{R}^P \rightarrow \mathbb{R}^k$  s.t.  $\|f(x_i) - f(x_j)\|^2 \leq \frac{\|x_i - x_j\|^2}{1 - \varepsilon}$

$$\text{with prob } \geq 1 - n^{-\alpha} \quad (*)$$

- (x) holds with probability at least  $1 - n^{-\alpha}$ ,
- f can be found in randomized polynomial time

(random projections)

e.g.  $f(x) = Rx$ ,  $R = [r_1, \dots, r_k]^T \in \mathbb{R}^{k \times D}$ ,  $x \in \mathbb{R}^D$

$r_i \in S^{D-1}$  sphere of  $D-1$  dim.

e.g.  $r_i = \frac{(a_1^i, \dots, a_D^i)}{\|a^i\|}$ ,  $a_j^i \sim N(0, 1)$

$\mathbb{R}^d$  uniformly projected to  $k$ -subspace

$\xleftarrow{\text{dist}}$  random vector on  $S^{d-1}$  restricted onto top  $k$ -coord.

$$\begin{aligned} & (\underbrace{a_1^i \dots a_d^i}_k) \quad \|a^i\|_2 = 1 \\ & (\underbrace{a_1^i \dots a_k^i}_k, 0) \end{aligned}$$

$X_i \sim \mathcal{N}(0, 1)$ ,  $i=1, \dots, d$

$X = (x_1 \dots x_d)$   $Y = \frac{X}{\|X\|_2} \in S^{d-1}$  uniformly distributed

$Z = (x_1, \dots, x_k, 0) \in \mathbb{R}^d$   $k$ -subspace  $E(x_1^2 + \dots + x_d^2) = d$

$$L := \|Z\|^2 \quad E[L] = \frac{k}{d} \cdot (\mu) \cdot E[\|X\|^2] = \mu \cdot d$$

Lemma Concentration Inequality  $k < d$ .

(a)  $\beta < 1$  lower bound.

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

(b)  $\beta > 1$  upper bound

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

$E L = \mu$ . exponentially  $\sim \mu$   $\downarrow$

# Proof of JL Lemma

Copyright by e板书

$d < k$ , trivial

$d > k$ .

$k$ -subspace  $\mathbb{X}_i \in \mathbb{R}^d \rightarrow \mathbb{Y}_i \in \mathbb{R}^k$ .  $i=1, \dots, n$ .

$$L = \|y_i - \hat{y}_i\|^2 \quad \mu = \mathbb{E}L = \frac{k}{d} \|x_i - \bar{x}_i\|^2 \quad \hat{y}_i = (x_i^T \cdots x_k^T)^T$$

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

$$\stackrel{\beta}{\uparrow} \quad \ln(1-\varepsilon) \leq -\varepsilon - \varepsilon^2/2 \quad \varepsilon \in [0,1]$$

$$\leq \exp\left[-\frac{k}{2}\left(\varepsilon - \left(\varepsilon + \frac{\varepsilon^2}{2}\right)\right)\right] = \exp\left(-\frac{k\varepsilon^2}{4}\right)$$

$$\leq \exp(-k(\alpha) \ln n) \quad \underline{k \geq 4(1+\alpha/2)(\varepsilon^2/2)^{1/\alpha}}$$

$$= n^{-(2+\alpha)} \rightarrow 0$$

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

$$\leq \exp\left[\frac{k}{2}(\varepsilon^2/2 + \varepsilon^3/3)\right],$$

$$\ln(1+\varepsilon) \leq \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}, \varepsilon \geq$$

$$\leq \exp(-(2+\alpha) \ln n) = n^{-(2+\alpha)} \rightarrow 0$$

$$\underline{k \geq 4(1+\alpha/2)(\varepsilon^2/2 + \varepsilon^3/3)^{-1} \ln n}$$

Given pair  $(i, j)$ .

$$\text{Prob}\left\{1-\varepsilon \leq \frac{\|y_i - \hat{y}_i\|^2}{\|x_i - \bar{x}_j\|^2} \leq 1+\varepsilon\right\} \leq 1 - \frac{2}{n^{2+\alpha}}$$

$\forall (i, j)$  from  $\{i=1, \dots, n; j=1, \dots, n\}$  of  $\binom{n}{2}$

$$\text{Prob}\left\{\forall (i, j), 1-\varepsilon \leq \frac{\|y_i - \hat{y}_i\|^2}{\|x_i - \bar{x}_j\|^2} \leq 1+\varepsilon\right\} \leq 1 - \binom{n}{2} \frac{2}{n^{2+\alpha}} \leq 1 - n^{-\alpha}$$

$$\binom{n}{2} = n(n-1)/2$$

# Lecture 5. Random Projections & Compressed Sensing 9/15/2017

3/29/2016 Introduction to Sparse Recovery: OMP, BP, LASSO, L1SS.

$$x^* \in \mathbb{R}^p, A^{n \times p}, n < p, \|A_j\|_2 = 1$$

$$S = \text{supp}(x^*) |S| = k < n < p$$

$$\begin{array}{ll} \text{Given } b = Ax \in \mathbb{R}^n & b = Ax + \varepsilon \\ \text{Find } x^* ? & x^* ? \\ \uparrow \text{noise-free} & \uparrow \text{noisy.} \end{array}$$

$$\text{ideal: } \min \|x\|_0 = \#\{i : x_i \neq 0\}$$

$$\text{P}_0: \text{s.t. } Ax = b$$

NP-hard. computational burden

Algorithm 1 (Orthogonal-Matching-Pursuit, Mallat-Zhang '93)

$$r_0 = b, x_0 = 0, S_0 = \emptyset$$

for  $t \in \mathbb{Z}_+$

$$j_{t+1} = \underset{1 \leq j \leq p}{\arg \max} |\langle A_j, r_t \rangle| \quad \% \text{ maximal correlation with residue}$$

$$S_{t+1} = S_t \cup \{j_{t+1}\}$$

$$\begin{aligned} x_{t+1} &= \underset{\mathcal{X}_{S_{t+1}}}{\arg \min} \|b - Ax\| \quad \% A_{S_{t+1}} \text{ restricts } \text{column index} \\ &= (A_{S_{t+1}}^* A_{S_{t+1}})^{-1} A_{S_{t+1}}^* b. \end{aligned}$$

$$r_{t+1} = b - Ax_{t+1}$$

end for stop if  $\|r_{t+1}\|$  is small enough. (0' in noise-free case)

Q1: Can it find  $x^*$  at  $\|r_{t+1}\| = 0$  in noise-free case? Tropp '04

$\|r_{t+1}\|$  small in noisy-case? Cai-Wang '11

Algorithm 2 (Basis Pursuit, Chen-Donoho-Sauders '99)

$$\begin{array}{ll} P_1: & \min \|x\|_1 \\ \text{s.t. } & Ax = b \end{array}$$

$$\begin{array}{ll} P_1^\varepsilon: & \min \|x\|_1 \\ (\text{BP DN}) & \text{s.t. } \|Ax - b\|_2 \leq \lambda \end{array}$$

Tibshirani '96

Algorithm 3.

$$(\text{LASSO}) \quad \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{x}\|_1$$

Algorithm 4. (Dantzig Selector, Candes-Tao '07)

$$\min \|\mathbf{x}\|_1$$

$$\text{s.t. } \|\mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x})\|_\infty \leq \lambda$$

Algorithm 5. L1-Biased-Scale-Space method  
Linearized Bregman Iteration

How does it work?

Important conditions.

1)  $\mathbf{A}^* \mathbf{A}_S \geq \gamma \mathbf{I}$ ,  $\gamma > 0$ . non-singular  $\mathbf{A}^* \mathbf{A}_S \Rightarrow$  uniqueness of  $\mathbf{x}_0$ .

2) (Exact Recovery Condition - Tropp '04; Irrepresentable Condition Yu-Zhai '06)

$$M = \|\mathbf{A}_S^* \mathbf{A}_S (\mathbf{A}_S^* \mathbf{A}_S)^{-1}\|_\infty < 1$$

"irrepresentable":  
regress  $\mathbf{A}_S^* \mathbf{A}_S$  by  $\mathbf{A}_S$ .  
representation coefficients

Thm (Tropp '04) If  $M < 1$ , then OMP recovers  $\mathbf{x}^*$  after  $k$  steps.

Thm (Tropp '04) If  $M \leq 1$ , then BP. ..

$$M \geq 1, \|\mathbf{r}_k\| \leq \|\mathbf{e}\|_2 \text{ stop.}$$

Cai-Wang '11 extends OMP to noisy case.  $|\mathbf{x}_i^*| \geq \frac{2\|\mathbf{e}\|_2}{1-(k-1)\mu}$

Yu-Zhai '06 solves LASSO case

Other conditions

2) (Incoherence, Donoho-Tao '01)

$$\mu = \max_{i \neq j} |\langle \mathbf{A}_i, \mathbf{A}_j \rangle|$$

LEM (Tropp '04)  $\mu < \frac{1}{2k-1} \Rightarrow M \leq \frac{k\mu}{1-(k-1)\mu} < 1$

Right

3) (Restricted-Isometry-Property, Candes-Tao '06)

For all  $k$ -sparse  $x \in \mathbb{R}^P$ ,  $\exists \delta_K \in (0, 1)$

$$(1 - \delta_K) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_K) \|x\|_2^2$$

Thm (Candes '08)

1)  $\delta_{2k} < 1 \Rightarrow P_0$  has a unique sol'n  $x^*$

2)  $\delta_{2k} < \sqrt{2} - 1 \Rightarrow P_1$  has a unique sol'n  $x^*$ . "reverses"

\* Johnson-Lindenstrauss Lemma  $\Rightarrow$  RIP with high Probability

Proofs (Tropp '04 OMP)

(I)  $M < 1 \Rightarrow$  OMP recovers  $x^*$ ?

key: for  $t \leq k$ , every step, it selects from  $S$  rather than  $S^c$ .  
i.e., examines

$$p(r_t) = \frac{\|k r_t, A_S x^*\|_\infty}{\|k r_t, A_{S^c} x^*\|_\infty} < 1$$

In noise-free case  $b = Ax^* \in \text{im}(A_S)$ ,  $r_t \in \text{im}(A_S)$

$$r_t = b - Ax^* \in \text{im}(A_S)$$

$P_t = A_S (A_S^* A_S)^{-1} A_S^*$  is the projection on  $\text{im}(A_S)$

$$\Rightarrow r_t = P_t r_t$$

$$\Rightarrow p(r_t) = \frac{\|(P_t r_t)^* A_S x^*\|_\infty}{\|r_t^* A_S x^*\|_\infty} = \frac{\|r_t^* P_t A_S (A_S^* A_S)^{-1} A_S^* x^*\|_\infty}{\|r_t^* A_S x^*\|_\infty}$$

$$= \frac{\|A_{S^c}^* A_S (A_S^* A_S)^{-1} \tilde{r}_t^*\|_\infty}{\|\tilde{r}_t^*\|_\infty}$$

$$= \|A_{S^c}^* A_S (A_S^* A_S)^{-1}\|_\infty = M$$

(II)  $M < 1 \Rightarrow$  BP recovers  $\hat{x}^*$ ?

Assume  $\hat{x} \neq x^*$ . solves  $\min_{\text{s.t. } Ax=b} \|x\|_1$

$$\hat{S} = \text{supp}(\hat{x}) \quad \{\hat{S} \setminus S\} \neq \emptyset.$$

$$\|\hat{x}\|_1 = \left\| \underbrace{A_S \hat{x}}_{\hat{x}_S} + \underbrace{(A_S^* A_S)^{-1} A_S^* (A_S \hat{x} - x^*)}_{\hat{x}_{S^c}} \right\|_1 = \|(A_S^* A_S)^{-1} A_S^* b\|_1,$$

$$= \|(A_S^* A_S)^{-1} A_S^* A_S \hat{x}\|_1,$$

$$= \left\| A_S (\hat{x}_S + \hat{x}_{S^c}) \right\|_1,$$

$$\leq \|\hat{x}_S\|_1 + \|(A_S^* A_S)^{-1} A_S^* A_S \hat{x}_{S^c}\|_1,$$

$$\stackrel{M<1}{\leftarrow} \|\hat{x}_S\|_1 + \|\hat{x}_{S^c}\|_1 \stackrel{M<1}{\rightarrow} 0$$

$$= \|\hat{x}\|_1$$

so.  $\hat{x}$  is impossible to be a minimizer.

$$\text{Note } M < \frac{1}{2k-1} \Rightarrow M \leq \frac{k\mu}{1-(k-1)\mu}.$$

$$\underline{\text{Pf}} \quad M = \|(A_S^* A_S)^{-1} A_S^* A_S\|_\infty = \underbrace{\|(A_S^* A_S)^{-1}\|_\infty}_{?} \underbrace{\|A_S^* A_S\|_\infty}_{\text{sym}}$$

$$\text{Q) } A_S^* A_S = I_k + \Delta \quad \max_j \|\Delta\|_{j,j} \leq \mu \quad \text{diag}(\Delta) = 0.$$

$$(A_S^* A_S)^{-1} = (I_k + \Delta)^{-1} = \sum_{j=0}^{\infty} (-\Delta)^j$$

$$\|(A_S^* A_S)^{-1}\|_\infty = \left\| \sum_{j=0}^{\infty} (-\Delta)^j \right\|_\infty \leq \sum_{j=0}^{\infty} \|\Delta\|_\infty^j = \frac{1}{1-\|\Delta\|_\infty}$$

$$\therefore M \leq \frac{k\mu}{1-(k-1)\mu}.$$

$$\|A^T \varepsilon\|_\infty \leq b_\infty$$

Note

How LASSO works?

$$\min_x \|b - Ax\|_2^2 + \lambda \|x\|_1$$

$$\left\{ \begin{array}{l} M \leq 1 - \eta \\ \hat{\lambda} = \frac{b_\infty}{\eta} \end{array} \right. \quad \left\{ \begin{array}{l} \min_{x \in \mathbb{R}^k} \|x\|_1 \geq \frac{b_\infty}{(1+\eta) \sqrt{k}} \end{array} \right.$$

Pf

$$\Leftrightarrow \lambda p(x) = A^T(b - Ax), \quad p(x) \in \partial \|x\|_1$$

sign consistency at  $(\hat{\lambda}, \hat{x})$

$$\Rightarrow \left\{ \begin{array}{l} \lambda \text{sign}(x_S^*) = A_S^T(b - A_S \hat{x}) \end{array} \right. \quad (1)$$

$$\left\| A_S^T(b - A_S \hat{x}) \right\|_\infty \leq \hat{\lambda}. \quad (2)$$

$$b = A_S x_S^* + \varepsilon$$

$$(1) \Rightarrow \hat{x} = \lambda (A_S^T A_S)^{-1} \text{sign}(x_S^*) + (A_S^T A_S)^{-1} A_S^T \varepsilon$$

$$\begin{aligned} (1) \lambda (2) \Rightarrow & \left\| A_S^T A_S \hat{x}^* + A_S^T \varepsilon - A_S^T A_S \hat{x}^* - A_S^T (A_S^T A_S)^{-1} A_S^T \varepsilon \right. \\ & \left. + \lambda A_S^T A_S (A_S^T A_S)^{-1} \text{sign}(x_S^*) \right\|_\infty \leq \lambda \end{aligned}$$

$$\left\| A_S^T (I - P_S) \varepsilon + \lambda A_S^T A_S (A_S^T A_S)^{-1} \text{sign}(x_S^*) \right\|_\infty \leq \hat{\lambda}$$

$$M \leq 1 - \eta \Rightarrow \left\| A_S^T A_S (A_S^T A_S)^{-1} \text{sign}(x_S^*) \right\|_\infty \leq 1 - \eta.$$

$$\text{It suffices: } \left\| A_S^T (I - P_S) \varepsilon \right\|_\infty < \hat{\lambda} \eta$$

Usually -  $\|A_S^T \varepsilon\|_\infty \leq b_\infty$  w.h.p. @ Gaussian noise

$$\text{so } \hat{\lambda} = \frac{b_\infty}{\eta} \quad (2) \text{ is satisfied.} \quad b_\infty = C \sigma \sqrt{k} \log p$$

$$\text{from (1) } \text{sign}(x_S^*) = \text{sign}(x_S^*)$$

$$\Rightarrow \text{sign}(x^* - \lambda (A_S^T A_S)^{-1} \text{sign}(x_S^*) + (A_S^T A_S)^{-1} A_S^T \varepsilon)$$

$$\text{requires } \min_{i \in S} |x_i^*| > \lambda (A_S^T A_S)^{-1} \text{sign}(x_S^*) + (A_S^T A_S)^{-1} A_S^T \varepsilon \|_\infty$$

$$\text{but. } \|\lambda (A_S^T A_S)^{-1} \text{sign}(x_S^*)\|_\infty \leq \frac{\lambda}{\sqrt{k}} \quad \|(A_S^T A_S)^{-1} A_S^T \varepsilon\|_\infty \leq \frac{\lambda \sqrt{k} \cdot \varepsilon}{\sqrt{k}}$$

$$\text{只要 } \min_{i \in S} |x_i^*| \geq \frac{\lambda}{\sqrt{(1+\eta) k}}$$

$$\mathcal{H}_C = \{ \psi \in \mathbb{R}^{n \times n \times n} : \exists Y \in M_C \text{ with } \psi = \delta_0(Y) \}$$

where

$$M_C = \{ Y \in \mathbb{R}^{n \times n} : Y_{il} = - \sum_{j \neq l} Y_{ij} \text{ and } \exists l \text{ with } Y^{(l)} = M_G^{(l)} \}$$

$$M_G^{(l)} = \{ X \in \mathbb{R}^{(n-1) \times (n-1)} : \exists s : V^{(l)} \rightarrow \mathbb{R} \text{ with } X_{ik} = \delta_0(s) \}$$

Xia, Jiacheng 2/20/2017

Summary:  $x_i \sim N(\mu, \Sigma)$

MLE  $\rightarrow$  Sample mean  $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i$

Sample covariance matrix  $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T$

unbiased, yet large variance, good/bad?

Use risk to evaluate, e.g. mean

$R(\hat{\mu}_n) := \mathbb{E}_{(x_i, y)} \| \hat{\mu}_n - \mu \|^2 \Leftrightarrow$  prediction error  $y \sim N(\mu, \Sigma)$

$$\mathbb{E} \| \hat{\mu}_n - y \|^2 = \mathbb{E}_{(x_i, y)} \mathbb{E}_{y|x_i} \| \hat{\mu}_n - y \|^2$$

↑ random, i.i.d.

$$\text{where } \mathbb{E}_{y|x_i} \| \hat{\mu}_n - (\mu + \varepsilon) \|^2 = \mathbb{E}_{y|x_i} \| \hat{\mu}_n - \mu \|^2$$

-  $2 \mathbb{E}_{\varepsilon|x_i} \langle \hat{\mu}_n - \mu, \varepsilon \rangle + \mathbb{E}_{\varepsilon|x_i} \|\varepsilon\|^2$

$$\mathbb{E} \| \hat{\mu} - y \|^2 = R(\hat{\mu}_n) + \text{const.}$$

So optimization of Risk  $\Leftrightarrow$  optimization of prediction error.  
We have learned, given  $Y \sim N(\mu, I_p)$ ,  $n=1$

1) MLE  $R(\hat{\mu}_{MLE}) = p$ . unbiased.

2) Linear  $\hat{\mu}_C = CY$ ,  $C = [\text{diag}(c_i)]_{q \times p}$  biased

$$R(\hat{\mu}_C) = \sum_{i=1}^p c_i^2 + \sum_{i=1}^p (1-c_i)^2 \hat{\mu}_i^2 \neq R(\hat{\mu}_{MLE})$$

bias

examp Ridge regression  $\frac{1}{n} \| \hat{\mu} - Y \|^2 + \lambda \| \hat{\mu} \|^2 \Rightarrow \hat{\mu}_\lambda = \frac{1}{1+\lambda} Y$

$$R(\hat{\mu}_\lambda) = \frac{p}{(1+\lambda)^2} + \sum_{i=1}^p \hat{\mu}_i^2 = \frac{A\lambda^2 + p}{(1+\lambda)^2} \quad A = \sum_i \hat{\mu}_i^2$$

$$R' = \frac{(A\lambda^2 + p)'(1+\lambda)^2 - 2(A\lambda^2 + p)(\lambda + 1)}{(1+\lambda)^4} = (1+\lambda)(2A\lambda(1+\lambda) - 2A\lambda^2 - 2p)$$

$$\lambda^* = p/A$$

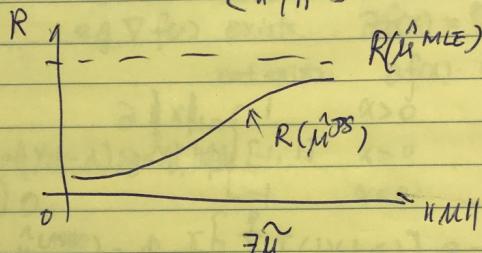
$$2A\lambda + 2A\lambda - 2A\lambda - 2p$$

$$\inf_{\hat{\mu}} R(\hat{\mu}_N) = \frac{A \cdot \left(\frac{p}{A}\right)^2 + p}{(1 + p/A)^2} = \frac{p \left(\frac{p}{A} + 1\right)}{(1 + \frac{p}{A})^2} = \frac{p}{p/A + 1} < p$$

$\Rightarrow A = \sum_i A_i \ll p$  s.t. highly concentrated on top- $k$  components  
 $A \sim O(k)$   $p/A \gg 1$

3) JS-estimate.  $\hat{\mu}^{JS} = \left(1 - \frac{p(p-2)}{\|\hat{\mu}^{MLE}\|}\right) \hat{\mu}^{MLE} = \left(1 - \frac{p-2}{\|\gamma\|^2}\right) \gamma$

$$R(\hat{\mu}^{JS}) = p - \mathbb{E} \left[ \frac{(p-2)^2}{\|\gamma\|^2} \right] < p = R(\hat{\mu}^{MLE}) \quad \text{biased}$$



$\hat{\mu}$  is inadmissible if  $A\hat{\mu}$ ,  $R(\hat{\mu}) = \mathbb{E} \|\hat{\mu} - \mu\|^2 \leq R(\hat{\mu})$   
&  $\exists \mu_0$  s.t.  $R(\hat{\mu}) < R(\mu_0)$

So  $\hat{\mu}^{MLE}$  is inadmissible

$\hat{\mu}^{JS}$  is also inadmissible.  $\hat{\mu}^{JS+} = \left(1 - \frac{(p-2)}{\|\gamma\|^2}\right) \gamma$  is better

Admissible estimators? Yes.

Thm [Lemma 2.8. Johnstone (GE)]  $\gamma \sim N(\mu, I_p)$

$\hat{\mu}_C = C\gamma$  admissible iff.

(1)  $C$  is symmetric, and

(2) Eigenvalues of  $C$ :  $0 \leq \rho_i(C) \leq 1$

(3)  $\rho_i(C) = 1$  for at most two  $i$ .

#### 4) LASSO / Soft-thresholding

$$\hat{\mu}^{\text{LASSO}} = \arg \min_{\tilde{\mu}} \frac{1}{2} \|Y - \tilde{\mu}\|^2 + \lambda \|\tilde{\mu}\|_1$$

$$\stackrel{\text{1st derivative}}{\Rightarrow} (\hat{\mu} - Y) + \lambda \text{sign}(\tilde{\mu}) \geq 0.$$

↑  
Subgradient.

convex  $f(x)$ .  $\partial f(x) \subseteq \mathbb{R}^p$ .

$$f(x) \geq f(x) + \langle u, (x' - x) \rangle$$

e.g.  $\nabla f(x)$  exists.  $\partial f(x) = \nabla f(x)$

not exists.  $\partial f(x)$  is a set.

$$\partial \|x\|_1 = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ 0 & x = 0 \end{cases}$$

$$\hat{\mu}_{\lambda}^{(i)} = \begin{cases} (y_i - \lambda) \text{sign}(x_i), & |x_i| \geq \lambda \\ 0, & |x_i| < \lambda \end{cases} \quad \mathbb{E}[\hat{\mu}] \neq \mu \text{ biased}$$

$$R(\hat{\mu}_{\lambda}^{\text{LASSO}}) = p \cdot \mathbb{E} \left[ \sum_{i=1}^p I(|x_i| \leq \lambda) \right] - \sum_{i=1}^p x_i^2 / \lambda$$

Donoho-Johnstone 95%  $\lambda = \sqrt{2 \log p}$

$$\leq 1 + (2 \log p + 1) \sum_{i=1}^p \mu_i^2 / \lambda$$

$\mu_i$  is sparse.  $k = \#\{i : \mu_i \neq 0\} \ll p$ ,  $R(\hat{\mu}_{\lambda}^{\text{LASSO}}) \sim O(k \log p)$

$\ll p$ .

#### 5) L<sub>0</sub> / Hard-thresholding

$$\hat{\mu}^{\text{Hard}} = \arg \min_{\tilde{\mu}} \|Y - \tilde{\mu}\|^2 + \lambda^2 \|\tilde{\mu}\|_0$$

$$\|\tilde{\mu}\|_0 = \#\{i : \tilde{\mu}_i \neq 0\}$$

By inspection,

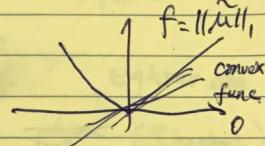
$$\hat{\mu}_{\lambda}^{(i)} = \begin{cases} y_i & |y_i| \geq \lambda \\ 0 & |y_i| < \lambda \end{cases}$$

$$\min(y_i - \theta)^2 + \lambda |\theta|_0 \\ = \min(y_i^2 / \lambda)$$

maybe unbiased.

not continuous  
variance?

Risk? not applied by Stein's Lemma. (Johnstone [GE])



Concentration Ineq.  $y_i \sim N(0, \sigma_p^2)$  i.i.d.

$$\Pr(\max_{1 \leq i \leq p} |y_i| \geq \varepsilon) \leq 2p e^{-\varepsilon^2/2\sigma_p^2}$$

holds subgaussian for

$$\sigma = 1, \varepsilon = \sqrt{2 \log p} \quad \Pr(\max_{1 \leq i \leq p} |y_i| \geq \sqrt{2 \log p}) \leq \frac{2}{\sqrt{2 \log p}} ? \text{ tighter one}$$

$$|\lambda| = \sqrt{2 \log p}$$

$$|\mu_i| \neq 0, |\mu_i| \geq$$

$\sqrt{2 \log p}$  (Johnstone 8.30)

Sparisty:  $k = \#\{i : |y_i - \mu_i + \varepsilon_i| \geq \sqrt{2 \log p}\} \ll p$ .

$$R(\hat{\mu}^{\text{hard}}) = O(k \log p)$$

6) Nonconvex Fan-Li '90 C.H. Zhang, Hui Zou

$$\hat{\mu}_i = \arg \min_{\tilde{\mu}_i} \|y - \tilde{\mu}\|^2 + \lambda p(\tilde{\mu})$$

$$\text{for large } |\tilde{\mu}_i|, \frac{\partial p(\tilde{\mu}_i)}{\partial \tilde{\mu}_i} = 0 \Rightarrow \tilde{\mu}_i = y_i \text{ unbiased}$$

for small  $|\tilde{\mu}_i|$ , close singularity at "0"  $\Rightarrow \tilde{\mu}_i = 0$ .

NP-hard. for global optimizer

Initial choice by  $\hat{\mu}^{\text{LASSO}}$   $\xrightarrow{\text{local convergence}} \underline{\text{global }} \mu$ .

7) Differential Inclusion (Osher-Ruan-Xiang-Yao-Yen '16)

LASSO-KKT.

$$\begin{aligned} & \Rightarrow ((\tilde{\mu} - y)) + \lambda \nabla \phi(\tilde{\mu}) = 0, \quad \phi(\tilde{\mu}) \in \partial \|\tilde{\mu}\|_1, \\ & \Rightarrow \frac{\phi(\tilde{\mu})}{\lambda} = -(\tilde{\mu} - y) \end{aligned}$$

$\phi(\tilde{\mu}) \in \partial \|\tilde{\mu}\|_1$

biased

$$\Rightarrow \frac{\partial \hat{f}(\tilde{\mu})}{\partial t} = -(\tilde{\mu} - y)$$

$$f(\tilde{\mu}) \leq \partial \|\tilde{\mu}\|_1$$

If.  $\tilde{\mu}_i \neq 0$ .  $f(\tilde{\mu}_i) \in \{\pm 1\}$ . constant.  $\frac{\partial f(\tilde{\mu}_i)}{\partial t} = 0 \Rightarrow \tilde{\mu}_i = y_i$  Unbiased?

$\Theta$  Under nearly Simple Algorithm

$$\frac{\partial f(\tilde{\mu})}{\partial t} + \frac{1}{k} \frac{\partial \tilde{\mu}}{\partial t} = -\nabla l(\tilde{\mu}). \quad \downarrow \text{loss.}$$

Euler Discretization  $\Rightarrow w^t = p^t + \frac{1}{k} \nabla l^t$

$$w^{t+1} = w^t - \alpha_t \nabla l(\tilde{\mu}^t)$$

$$\tilde{\mu}^t = k \cdot \text{Shrink}(w^t, 1) = k \begin{cases} |w_i| - 1 & |w_i| > 1 \\ 0 & \text{otherwise} \end{cases}$$

\* Model Selection Problem

$$y = X^\top \beta \quad X^{n \times p} \quad n < p$$

feature/covariate matrix

underdetermined in general

$\beta$  is  $k$ -sparse?  $\#\{i : \beta_i \neq 0\} = k \ll p$

$\left\{ \begin{array}{l} y = X\beta \Rightarrow \beta? \text{ compressed sensing} \\ \text{talk later} \end{array} \right.$

$y = X\beta \Rightarrow \hat{\beta} \text{ s.t. } \text{supp}(\hat{\beta}) = \text{supp}(\beta)$

Model Selection  
LASSO Consistency

$\|\hat{\beta} - \beta\|^2$  small?

$L_2$ -consistency

$$\text{LASSO} \arg_{\beta} \min_{\beta} \|y - X\beta\|^2 + \lambda \|\beta\|_1 = \hat{\beta}_{\lambda}$$

Model-Selection Consistency:  $\left\{ \begin{array}{l} \text{Irrepresentable Condition} \\ \text{Strong Signal} \end{array} \right.$

$$S = \text{Supp}(\beta_i)$$

$$\min(\beta_i) > 0 \quad \left( \frac{\log p}{n} \right)$$

$$\Rightarrow \text{sign}(\hat{\beta}_i) = \text{sign}(\beta_i) \quad \text{w.h.p.}$$

$L_2$ -Consistency: RE (restricted eigenvalue) condition Bickel-Ritov  
Hessian matrix  $H_n = \frac{X^T X}{n}$  is not p.s.d. - Tsybakov

but restricted p.s.d. on a cone near  $\beta^*$

$$\Delta = \hat{\beta} - \beta^* \quad \left\{ \Delta^T H_n \Delta \geq r \|\Delta\| : \|\Delta\|_1 \leq C \|\Delta\|_2 \right\}$$

IRR  $\Rightarrow$  RE.



Note:

Bregman IIS (LB1) nearly the same as LASSO.

Split. LB1/IIS.: better than LASSO / genLasso.  
(NIP'2016: Huang <sup>sun</sup> - Xiong - Yao).

PCA Back to  $\sum_n^{MLE}$ , PCA is a hard-thresholding on eigenspace of  $\sum_n^{MLE}$ , why?

Example  $y_i = \alpha u + \varepsilon_i, i=1, \dots, n.$

$$\|\alpha\|_2 = 1, \alpha \sim N(0, \sigma_\alpha^2), \varepsilon_i \sim N(0, \sigma_\varepsilon^2 I_p)$$

$$\begin{aligned} \text{fixed } p, n \rightarrow \infty \quad & \sum_n \Rightarrow \sum = \hat{\alpha} u u^T + \sigma_\varepsilon^2 I_p \\ & = \frac{1}{n} \sum_{i=1}^n y_i y_i^T \quad \begin{matrix} \uparrow & \uparrow \\ \text{low rank} & \text{diagonal: sparse} \end{matrix} \end{aligned}$$

$$\text{If } \sigma_u^2 > \sigma_\varepsilon^2, \text{ top-1 eval. } \sigma_u^2 \sigma_\varepsilon^{-2}$$

eval.  $\sigma_u$ .

How about  $(p, n) \rightarrow \infty$ ?

$$\delta = \lim_{n \rightarrow \infty} \frac{\rho}{n} \Rightarrow \delta \quad \text{if } n \rightarrow \infty \quad \sigma_x^2 = 1.$$

$$\lambda_{\max}(\hat{\Sigma}_n) \rightarrow \begin{cases} (1 + \sqrt{\delta})^2 = b & \sigma_x^2 \leq \sqrt{\delta} \\ (1 + \frac{\sigma_x^2}{\delta})(1 + \frac{\delta}{\sigma_x^2}) & \sigma_x^2 > \sqrt{\delta} \end{cases}$$

$$|K(u, v_{\max}(\hat{\Sigma}_n))| \rightarrow \begin{cases} 0 & \sigma_x^2 \leq \sqrt{\delta} \\ 1 - \frac{\delta}{\sigma_x^2} & \sigma_x^2 > \sqrt{\delta} \end{cases}$$

$\sigma_x^2 \leq \sqrt{\delta}$ . small signal : PCA  $v_{\max}(\hat{\Sigma}_n)$  provides no information about  $u$ .

large signal : PCA  $v_{\max}(\hat{\Sigma}_n)$  is a biased est. of  $u$ .  
in a conic neighbor of  $u$ .  
eval is unbiased est. of  $\sigma_u^2 + \sigma_v^2$

Marcenko-Pastur Dist. of Random matrices

$$X_i \stackrel{i.i.d.}{\sim} N(0, I_p) \quad i=1, \dots, n.$$

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \quad \text{Wishart random matrix}$$

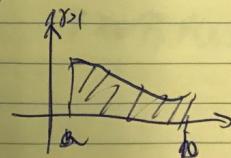
$\frac{\rho}{n}$  eigenvalue dist.

$$a = (1 - \sqrt{\delta})^2$$

$$b = (1 + \sqrt{\delta})^2$$

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

$$\sum_{t=0}^{\rho} (1 - \frac{t}{\rho}) I(\gamma \geq 1)$$



# Lecture 5 . Robust PCA

Sep. 15, 2017

# Lecture Graph Realization V. MDS with Uncertainty

## Recall Classical MDS

Given dig pair distance

$\forall (i, j)$  complete

?  $y_i \in \mathbb{R}^n$  s.t.

$$\min_{Y \in \mathbb{R}^{n \times n}} \sum_{i,j} ( \|y_i - y_j\|^2 - d_{ij}^2 )^2$$

$$\Leftrightarrow \min_Y \|YY^T - B\|^2$$

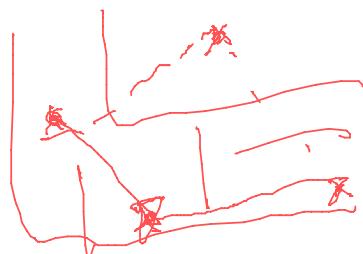
$$B = -\frac{1}{2} H D H^T$$

$$D = [d_{ij}^2]$$

$\Leftrightarrow$  eigen-decomp (B)

新特点 : SNL

$$G = (V, E)$$



V : Sensor

①  $(i, j) \in E$  iff  $d_{ij}$

incomplete

② noise:  $\tilde{d}_{ij} = d_{ij} + e_{ij}$  noise

③ anchor point  $x_i = a_i$   
partial

目标:

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

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

$$? \min_Y \sum_{(i, j) \in E} (\|y_i - y_j\|^2 - d_{ij}^2)^2 \quad \left. \begin{array}{l} \text{gradient method} \\ \text{nonlinear opt.} \end{array} \right\}$$

不能 Eigendecomposition  
[www.ebanshu.com](http://www.ebanshu.com)

练习:  $\|y_i - y_j\|^2 = d_{ij}^2$  Copyright by 板书  
 Quadratic equation  
~~SDR~~ Relax  $\Rightarrow$  "Linear"  
 $\downarrow$

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

$y_i \in \mathbb{R}^k$

$$Y = [y_1 \dots y_n]^{k \times n}$$

$$= (e_i - e_j)^T [Y^T Y]^{n \times n} (e_i - e_j)$$

$$e_i = (0 \dots 1 \dots 0)$$

$$y_i = Y \cdot e_i$$

$$y_i - y_j = Y(e_i - e_j)$$

$$= (e_i - e_j) (e_i - e_j)^T \bullet [Y^T Y]$$

$$X = Y^T Y \succeq 0$$

$$= (e_i - e_j) (e_i - e_j)^T \bullet X$$

LMI. SD  
 $X \succeq 0$

### Relaxation

$$X = Y^T Y \Rightarrow X \succeq Y^T Y \quad (X - Y^T Y \succeq 0)$$

$$\Leftrightarrow \begin{bmatrix} I_k & Y \\ Y^T & X \end{bmatrix} \succeq 0, X \succeq 0 \quad \text{Linear Matrix Inequality}$$

### Lemmas

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

SDR.

LMI

$$\left\{ \begin{array}{l} Z = \begin{bmatrix} I & Y \\ Y^T & X \end{bmatrix} \succeq 0 \\ Z_{1:k, 1:k} = I \end{array} \right.$$

$$(0; e_i - e_j) (0; e_i - e_j)^T \bullet Z = d_{ij}^2 \quad (i, j) \in E$$

Note:  $d_{ij}^2 \leq \|y_i - y_j\|^2 \leq d_{ij}^2 + \epsilon_{ij} \Rightarrow \text{LMI. noise}$

①

anchor point.  $y_i = a_i, \|a_i - y_j\|^2 = d_{ij}^2 \Rightarrow (a_i; e_j) (a_i; e_j)^T \bullet Z = d_{ij}^2$

SPP approach  $\rightarrow$  MDS

Matlab

Protein 3-D structure Reconstruction

CMDS

Schoenberg

SDP-MDS Exact Recovery ? Yin Yu Ye group.  
P. Biswas. A. So.

Recall SPP

$$\begin{array}{lll} (\text{SPP}) & \min C \cdot X & C, X \in \mathbb{R}^{n \times n} \\ \text{st.} & A_i \cdot X = b_i & i=1, \dots, m \\ & X \succeq 0 & b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \end{array}$$

$$\begin{array}{lll} (\text{SDD}) & \max -b^T y & y \in \mathbb{R}^m, S \in \mathbb{R}^{n \times n} \\ \text{st.} & S = C - \sum_{i=1}^m A_i b_i \succeq 0 & \end{array}$$

$$F_P = \{X \geq 0 : A_i X = b_i\}$$

$$F_D = \{(y, S) : S = C - \sum A_i y_i \succeq 0\}$$

$$\text{primal obj. } C \cdot X \quad \text{dual obj. } b^T y$$

Weak Duality  $F_P \neq \emptyset, F_D \neq \emptyset$

$$\Rightarrow \underbrace{C \cdot X - b^T y \geq 0}_{\text{duality gap}} \quad \forall X \in F_P, \forall (y, S) \in F_D$$

# Strong Duality of SDP

(1)  $F_p \neq \emptyset, F_d \neq \emptyset$

(2)  $F_p$  or  $F_d$  has an interior solution

$\Rightarrow x^* \in F_p, (y^*, s^*) \in F_d$  is optimal soln.

iff.  $Cx^* = b^T y^*$  duality gap is zero

$y^* \cdot s^* = 0$ . Complementary

存在内点 Soln

$x^*, (y^*, s^*)$        $\underbrace{x^* \cdot s^* = 0}$       opt.  
 "witness" primal-dual pair

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

SDP-MDS.

$$Z = \begin{bmatrix} I_k & X \\ Y^T & X \end{bmatrix} \geq 0$$

假设  $d_{ij} = \|x_i - x_j\| \quad x_i \in \mathbb{R}^k$

$\exists Z^* = \begin{bmatrix} I_k & Y^* \\ Y & X^* \end{bmatrix} \geq 0$  interior point soln

$\text{rank}(Z^*) \geq k$        $\text{rank}(X^*) + \text{rank}(S^*) \leq n$        $\left. \right\} \Rightarrow \text{rank}(Z^*) + \text{rank}(S^*) \leq k + n$   
 $\text{rank}(S^*) \leq n$        $\left. \right\} \Rightarrow \text{rank}(Z^*) \geq k$

$$X^* = Y^T Y \quad Y \in \mathbb{R}^{k \times n}$$

$\Rightarrow \text{rank}(Z^*) = k \quad (\geq k) \quad \text{minimal rank.}$

代数  $\text{rank}(S^*) = n \quad (\leq n) \quad \text{maximal rank.}$

UR: Universal Rigidity (UR)

there is a unique embedding  $y_i \in \mathbb{R}^k \hookrightarrow \mathbb{R}^l \quad l \geq k$

$$\left( \frac{y_i}{k}, \underbrace{\dots}_{l-k} \right) \quad \text{s.t.} \quad \|y_i - y_j\| = d_{ij}, \quad i, j \in E$$

minimal dim embedding  $k: \mathbb{R}^k$

spectrum

Schoenberg '1938 :  $G$  is complete  $\Rightarrow$  UR.

$G$  is incomplete

[So-Ye '2007]  $G$  general.

UR  $\Leftrightarrow$  SDP maximal rank soln  
 $\text{rank}(S^*) = n, \text{rank}(Z^*) = k$

Theorem Equivalent statements

• (UR)  $G$  is UR or has a unique embedding in  $\mathbb{R}^k$

• (代数) SDP has a max-rank feasible soln  $\text{rank}(Z^*) = k$   
 $(\text{rank}(S^*) = n)$

•  $X^* = Y^T Y$  or  $\text{trace}(X - Y^T Y) \Rightarrow$   
 eigen decomp of  $X^*$ .

UR is polynomial ( $n, k, \log(\frac{1}{\epsilon})$ )

Ye, ICCM '2010; Fields '2011  
[www.ebanshu.com](http://www.ebanshu.com)

# Maximum Variance Unfolding (Manifold Learning)

Copyright by 板书 2006.

$$X = Y^T Y \Rightarrow X \geq Y^T Y$$

MVU

$$\underline{K} = Y^T Y$$

度量

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

$$K_{ij} = \langle Y_i, Y_j \rangle$$

$$\max \text{trace}(K)$$

"SDP"

$$\text{s.t. } K_{ii} + K_{jj} - 2K_{ij} = d_{ij}^2$$

$$\sum_j K_{ij} = 0$$

$$K \geq 0$$

$$K \Rightarrow Y^T Y$$

Why  $\max \text{tr}(K)$ ?

不能 work.

[So-Ye 2007]



SDP.

$k$ -manifold

$(k+1)$ -lateration graph

"Unfold manifold"