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

Spectral Method

"Can you hear the shape of a drum?"
by Hermann Weyl
Mark Kac

⇒ Can you hear the shape of data?

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

(4) Connecting geometry and topology: Hodge Theory

Spectral clustering and graph Laplacian

Statistical Ranking and graph Helmholtzian

Game Theory

Courseweb

http://math.stanford.edu/~lyuany/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 $X = [x_1, \dots, x_n]^{p \times n}$

centered data $X_c = XH$, $H = I - \frac{1}{n} 1 1^T$, $1 = [1, \dots, 1]^T \in \mathbb{R}^n$

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

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

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

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

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

$\hat{\beta} = [\hat{\beta}_1, \dots, \hat{\beta}_n] = \hat{U}_k^T X_c \cong \hat{S}_k \hat{V}_k^T$

each column gives new coordinates

$\Leftrightarrow$ Eigenvalue Decomposition of Covariance Mat.

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

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

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

$\Leftrightarrow$ Eigenvalue Decomp. of Kernel Mat.

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

"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 ?

$n \gg 1$ $\quad$ $p \gg 1$

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

down sample $n' \Rightarrow$ good approximation of $\hat{\Sigma}_n$ & $\hat{U}_k, \hat{S}_k$

$\hat{V}_k$ restricted on subsample

high dimensionality $p \gg 1$, $\hat{\Sigma}_n \in \mathbb{R}^{p \times p}$ 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 $A_{ij} \sim \mathcal{N}(0, 1)$.

$X_c^{p \times n} \longmapsto (RX_c)^{d \times n}$, $d \ll p$.

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

- $A_{ij} = \begin{cases} 1 & p = \frac{1}{2} \\ -1 & (1-p) = \frac{1}{2} \end{cases}$
- $A_{ij} = \begin{cases} 1 & p = \frac{1}{6} \\ 0 & 1-2p = \frac{2}{3} \\ -1 & p = \frac{1}{6} \end{cases}$ sparse with many zeros!

Example (Human Genome Diversity Project / HGDP)

http://www.cephb.fr/en/hgdp_panel.php

$n = 1064$ persons $\quad p = 644,258$ 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.

$$\tilde{X}^{d \times n} = R X_c = R X H \cong \hat{U}_k \hat{S}_k \hat{V}_k^T$$

$d \times k \quad k \times k \quad k \times n$

$d = p = 600k$ $\quad d = 5k$ $\quad d = 100k$.

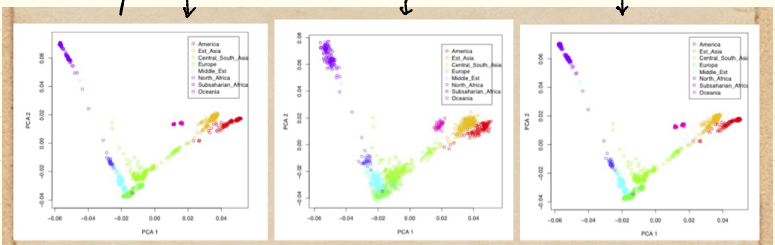


FIGURE 1. (Left) Projection of 1043 individuals on the top 2 MDS principal components. (Middle) MDS computed from 5,000 random projections. (Right) MDS computed from 100,000 random projections. Pictures are due to Qing Wang.

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 $f: X_i \mapsto Y_i \in \mathbb{R}^d$, $d = O(c(\varepsilon)\log n)$

s.t. $$1-\varepsilon \le \frac{\|Y_i - Y_j\|}{\|x_i - x_j\|} \le 1+\varepsilon.$$ with probability $\ge 1 - n^{-\alpha}$, $\alpha > 0$

$n^{-\alpha} \to 0$ as $n \to \infty$

Uniform ε-Isometry!

relative metric-distortion is uniformly bounded by ε!

f is a random projection!

1980s Johnson-Lindenstrauss Lipschitz Extension

2003 Sanjoy Dasgupta, Anupam Gupta

Dimitris Achlioptas.

Computer Science, data compression

nearest neighbor search

Thm 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 $x_i \in \mathbb{R}^D$ $(i = 1, \dots, n)$, there exists a map $f: \mathbb{R}^D \to \mathbb{R}^k$ s.t. $\forall x_i, x_j$

$$1-\varepsilon \le \frac{\|f(x_i) - f(x_j)\|^2}{\|x_i - x_j\|^2} \le 1+\varepsilon \qquad (*)$$

with prob $\ge 1 - n^{-\alpha}$.

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

(random projections)

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

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

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

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

$\Longleftrightarrow$ (dist.) random vector on S^{d-1} restricted onto top k-coord.

$(\underbrace{a_1^i \cdots a_k^i}_{k} \cdots a_d^i)$ $\quad \|a^i\|_2 = 1$

$(a_1^i, \cdots a_k^i, 0)$

$\chi_i \sim \mathcal{N}(0,1)$, $i = 1, \ldots, d$.

$X = (\chi_1 \cdots \chi_d)$ $\quad Y = \frac{X}{\|X\|} \in S^{d-1}$ uniformly distributed.

$Z = (\chi_1, \cdots, \chi_k, \cdots 0) \in \mathbb{R}^d$ $\quad$ k-subspace

$L := \|Z\|^2$ $\quad \mathbb{E}L = \frac{k}{d} \; (= \mu) \cdot \mathbb{E}\|X\|^2 = \mu \cdot d$. $\quad \mathbb{E}(\chi_1^2 + \cdots \chi_d^2) = d$

Lemma Concentration Inequality $\quad k < d$.

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

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

(b) $\beta > 1$ upper bound

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

$\mathbb{E}L = \mu$. exponentially $\sim \mu$ $\downarrow$

Proof of J-L Lemma

$d<k$. trivial

$d>k$.

k-subspace $\quad x_i \in \mathbb{R}^d \to \in \mathbb{R}^k \quad i=1,\dots,n.$

$L = \|y_i - y_j\|^2 \qquad \mu = \mathbb{E}L = \frac{k}{d}\|x_i - x_j\|^2 \qquad y_i = (x_i^1 \cdots x_i^k)$

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

β

$\ln(1-\varepsilon) \le -\varepsilon - \varepsilon^2/2 \qquad \varepsilon \in [0,1)$

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

$$\le \exp\left(-(2+\alpha)\ln n\right) \qquad k \ge 4(1+\alpha/2)(\varepsilon^2/2)^{-1}\ln n$$

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

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

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

$\ln(1+\varepsilon) \le \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3}, \quad \varepsilon \ge 0$

$$\le \exp\left(-(2+\alpha)\ln n\right) = n^{-(2+\alpha)} \to 0$$

$$k \ge 4(1+\alpha/2)\left(\varepsilon^2/2 - \varepsilon^3/3\right)^{-1}\ln n$$

Given pair (i,j).

$$\text{Prob}\left\{1-\varepsilon \le \frac{\|y_i - y_j\|^2}{\|x_i - x_j\|^2} \le (1+\varepsilon)\right\} \le 1 - \frac{2}{n^{2+\alpha}}$$

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

$$\text{Prob}\left\{\forall (i,j),\ 1-\varepsilon \le \frac{\|y_i - y_j\|^2}{\|x_i - x_j\|^2} \le 1+\varepsilon\right\} \le 1 - \binom{n}{2}\frac{2}{n^{2+\alpha}} \le 1 - n^{-\alpha}$$

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

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

3/29/2016 Introduction to Sparse Recovery: OMP. BP. LASSO. ℓ_1 LSS.

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

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

Given $b = Ax^*$ $\in \mathbb{R}^n$ $\quad b = Ax + \varepsilon$

Find x^* ? $\quad x^*$?

↑ noise-free ↑ noisy.

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

P_0: s.t. $Ax = b$

NP-hard. computational burden

Algorithm 1 (Orthogonal-Matching-Pursuit, Mallat-Zhang '1993)

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

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

$j_{t+1} = \arg\max_{1\le j\le p} |\langle A_j, r_t\rangle|$ % maximal correlation with residue

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

$x_{t+1} = \arg\min_{x_{S_{t+1}}} \|b - Ax\|$ % $A_{S_{t+1}}$ restricts to S_{t+1} column index

$= (A_{S_{t+1}}^* A_{S_{t+1}})^{-1} A_{S_{t+1}}^* b$.

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

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-Saunders '99)

P_1: $\min \|x\|_1$ s.t. $Ax = b$

P_1^ε (BPDN): $\min \|x\|_1$ s.t. $\|Ax - b\|_2 \le \lambda$

Tibshirani '96

Algorithm. (LASSO) $\min \frac{1}{2n}\|Ax-b\|_{\ell_2}^2 + \lambda\|x\|_1$

Algorithm. (Dantzig Selector, Candes-Tao '2007)

$$\min \|x\|_1$$
$$\text{s.t. } \|A^T(b-Ax)\|_\infty \le \lambda$$

Algorithm: ~~Inverse~~-Scale-Space method ?
Linearized Bregman Iteration

How does it work?

Important conditions.

1) $A_S^* A_S \ge \gamma I$, $\gamma > 0$. non-singular $A_S^T A_S$. $\Rightarrow$ uniqueness of P_0.

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

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

$$\max_{j \in S^c} \|A_j^* A_S (A_S^* A_S)^{-1}\|$$

"Irrepresentable": regress A_{S^c} by A_S. regression representation coefficients has maximum magnitude

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

Thm (Tropp '04) If $M<1$, then BP. . . . $M>1$, $\|\eta\| \le \|\varepsilon\|_2$ stop. $|x_i^*| \ge \frac{2\|\varepsilon\|_2}{1-(k-1)\mu}$

Cai-Wang '11 extends OMP to noisy case.
Yu-Zhao '06 solves LASSO case

Other conditions

2') (Incoherence, Donoho-Huo '01)

$$\mu = \max_{i \ne j} |\langle A_i, A_j \rangle|$$

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

Recht

3' (Restricted-Isometry-Property, Candes-Tao '05)

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

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

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^*. "recovers"

*: Johnson-Lindestrauss Lemma $\Rightarrow$ RIP with high Probability

Proofs (Tropp '04 OMP)

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

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

$$\rho(r_t) = \frac{\|\langle r_t, A_{S^c}\rangle\|_\infty}{\|\langle r_t, A_S\rangle\|_\infty} < 1$$

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

$$r_t = b - Ax_t \in \mathrm{im}(A_S)$$

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

$\Rightarrow r_t = P_t r_t$

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

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

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

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

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

$S=\text{supp}(x^*)$ $\{\hat{S} \setminus S\} \neq \phi$.

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

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

$$= \| \because A_S(\hat{x}_S + \hat{x}_{\hat{S}\setminus S}) \|_1$$

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

$$\overset{M<1}{<} \|\hat{x}_S\|_1 + \|\hat{x}_{\hat{S}\setminus S}\|_1 \quad \because \hat{S}\setminus S \neq \phi$$

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

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

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

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

$\because A_S^* A_S = I_k + \Delta$ $\max_j |\Delta_{ij}| \leq \mu$ $\text{diag}(\Delta)=0$

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

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

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

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

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

Note

How LASSO works?

$$\min_x \frac{1}{2}\|b-Ax\|_2^2 + \lambda\|x\|_1$$

条件 $\begin{cases} M \le 1-\eta. \\ \min_{i\in S}|x_i^*| \ge \dfrac{b_\infty}{\eta\gamma}(1+\eta)\sqrt{k}. \end{cases}$ $\hat{\lambda} = \dfrac{b_\infty}{\eta}$.

Pf

KKT: $\lambda\rho(x) = A^T(b-Ax)$, $\rho(x) \in \partial\|x\|_1$

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

$$\Rightarrow \begin{cases} \hat{\lambda}\,\mathrm{sign}(x_S^*) = A_S^T(b-A_S\hat{x}) & (1) \\ \|A_{S^c}^T(b-A_S\hat{x})\|_\infty \le \hat{\lambda}. & (2) \end{cases}$$

$b = A_S x_S^* + \varepsilon$

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

代入(2) $\Rightarrow$ $\|A_{S^c}^TA_Sx^* + A_{S^c}^T\varepsilon - A_{S^c}^TA_Sx^* - A_{S^c}^TA_S(A_S^TA_S)^{-1}A_S^T\varepsilon + \lambda A_{S^c}^TA_S(A_S^TA_S)^{-1}\mathrm{sign}(x_S^*)\|_\infty \le \hat{\lambda}$

$$\|A_{S^c}^T(I-P_S)\varepsilon + \lambda A_{S^c}^TA_S(A_S^TA_S)^{-1}\mathrm{sign}(x_S^*)\|_\infty \le \hat{\lambda}$$

$M \le 1-\eta \Rightarrow \|A_{S^c}^TA_S(A_S^TA_S)^{-1}\mathrm{sign}(x_S^*)\|_\infty \le 1-\eta$.

It suffices $\|A_{S^c}^T(I-P_S)\varepsilon\|_\infty < \hat{\lambda}\eta$

usually $\|A_{S^c}^T\varepsilon\|_\infty \le b_\infty$ w.h.p. @ Gaussian noise

$b_\infty = c\sigma\sqrt{\log p}$

so $\hat{\lambda} = \dfrac{b_\infty}{\eta}$ (2) is satisfied.

from (1) $\mathrm{sign}(x_S^*) = \mathrm{sign}(\hat{x}_S)$

$$\Rightarrow = \mathrm{sign}\left(x^* - \lambda(A_S^TA_S)^{-1}\mathrm{sign}(x_S^*) + (A_S^TA_S)^{-1}A_S^T\varepsilon\right)$$

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

but $\|\hat{\lambda}(A_S^TA_S)^{-1}\mathrm{sign}(x_S^*)\|_\infty \le \hat{\lambda}\sqrt{k}$ $\|(A_S^TA_S)^{-1}A_S^T\varepsilon\|_\infty \le \dfrac{\hat{\lambda}\eta\sqrt{k}}{\gamma}$.

i.e. $\min|x_i^*| \ge \dfrac{\hat{\lambda}}{\gamma}(1+\eta)\sqrt{k}$

$$\mathcal{H}_C = \{ \psi \in \mathbb{R}^{n \times n \times n} : \exists Y \in M_C \text{ with } \psi = \delta_1[Y] \}$$

where

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

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

Xia, Jiacheng 2/20/2017.

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

MLE → { sample mean $\hat{\mu}_n^{MLE} = \frac{1}{n}\sum_{i=1}^n X_i$

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

~~unbiased, yet large variance~~, good/bad?

Use risk to evaluate, e.g. mean

$R(\hat{\mu}_n) := \mathbb{E}_{(X_i)_1^n} \|\hat{\mu}_n - \mu\|_n^2$ ⟺ prediction error $Y \sim N(\mu, \Sigma)$.

$$\mathbb{E}\|\hat{\mu}_n - Y\| = \mathbb{E}_{(X_i)_1^n} \mathbb{E}_{Y|(X_i)_1^n} \|\hat{\mu}_n - Y\|^2$$

($\hat{\mu}_n$: random, iid.)

where $\mathbb{E}_{Y|X_i} \|\hat{\mu}_n - (\mu + \varepsilon)\|_2^2 = \mathbb{E}_{\varepsilon|X_i} \|\hat{\mu}_n - \mu\|^2 - 2\mathbb{E}_{\varepsilon|X_i} \langle \hat{\mu}_n - \mu, \varepsilon \rangle + \mathbb{E}_{\varepsilon|X_i} \|\varepsilon\|^2$

($\mathbb{E}_{\varepsilon|X_i}\varepsilon = 0$; $\mathbb{E}_{\varepsilon|X_i}\|\varepsilon\|^2$: const)

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

So optimization of Risk ⟺ optimization of prediction error.

We have learned, given $Y \sim N_p(\mu, I_p)$, $n=1$

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

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

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

(bias)

example Ridge regression $\min_\mu \frac{1}{2}\|\mu - Y\|^2 + \lambda\|\mu\| \Rightarrow \hat{\mu}_\lambda = \frac{1}{1+\lambda} Y$, $= c_i$

$$R(\hat{\mu}_\lambda) = \mathbb{E}\frac{p}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\sum_i \mu_i^2 = \frac{A\lambda^2 + p}{(1+\lambda)^2} \qquad A = \sum_i \mu_i^2$$

$$0 = R' = \frac{(A\lambda^2+p)'(1+\lambda)^2 - 2(A\lambda^2+p)(\lambda+1)}{(1+\lambda)^4} = \frac{(1+\lambda)(2A\lambda(1+\lambda) - 2A\lambda^2 - 2p)}{(1+\lambda)^4}$$

$$2A\lambda(1+\lambda)^2 - 2(A\lambda^2+p)(1+\lambda)$$

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

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

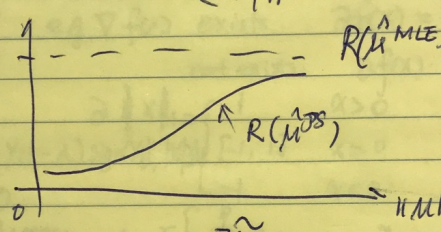
$$\inf_\lambda R(\hat\mu_\lambda) = \frac{A\cdot(\frac{p}{A})^2 + p}{(1+\frac{p}{A})^2} = \frac{p(\frac{p}{A}+1)}{(1+\frac{p}{A})^2} = \frac{p}{p/A+1} < p$$

esp. $A = \sum_i A_i^2 \ll p$ s.t. highly concentrated on top-k components

$A \sim O(k)$ $\quad p/A \gg 1$

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

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



$\hat\mu$ is <u>inadmissible</u> if $\forall\mu$, $R(\tilde\mu) = \mathbb{E}\|\tilde\mu - \mu\|^2 \le R(\hat\mu)$

& $\exists\mu_0$. s.t. $R(\tilde\mu) < R(\mu_0)$

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

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

Admissible estimaters? Yes.

<u>Thm</u> [Lemma 2.8. Johnstone (GE)] $Y \sim N(\mu, I_p)$

$\hat\mu_C = CY$ admissible iff.

(1) C is symmetric, ~~and~~

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

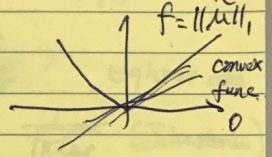
(3) $\rho_i(C) = 1$ for at most two i.

4) LASSO / Soft-thresholding

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

1st derivative $\Rightarrow 0 \in (\hat{\mu}-Y) + \lambda\, \partial\|\hat{\mu}\|_1 \ni 0$

$\partial\|\hat{\mu}\|_1$: subgradient.



convex $f(x)$. $\partial f(x) = \{u \in \mathbb{R}^p : f(x') \geq f(x) + \langle u, (x'-x)\rangle\}$

eg. $\nabla f(x)$ exists. $\partial f(x) = \nabla f(x)$

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

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

$$\Rightarrow \hat{\mu}_\lambda^{LASSO}(i) = \begin{cases} (|Y_i|-\lambda)\,\mathrm{sign}(Y_i), & |Y_i|>\lambda \\ 0 & \end{cases}$$

$\mathbb{E}(\hat{\mu}) \neq \mu_i$ biased

$$R(\hat{\mu}_\lambda^{LASSO}) = p - \mathbb{E}\left[k - \sum_{i=1}^{p} \mathbb{I}(|Y_i|\leq\lambda) + \sum_{i=1}^{p} Y_i^2 \wedge\lambda\right]$$

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

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

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

$\ll p$

5) L_0 / Hard-thresholding

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

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

$\min_{\theta}(Y_i-\theta)^2 + \lambda^2\|\theta\|_0 = \min(Y_i^2, \lambda^2)$

By inspection.

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

maybe unbiased.

not continuous

variance?

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

Concentration Ineq. $Y_i \sim N(0, \sigma^2)$ i.i.d.

$$Pr\left(\max_{1\le i\le p} |Y_i| \ge \varepsilon\right) \le 2p\, e^{-\varepsilon^2/2\sigma^2}$$

holds for subgaussian

$\sigma=1$. $\varepsilon = \sqrt{2\log p}$ $\quad Pr\left(\max_{1\le i\le p} |Y_i| \ge \sqrt{2\log p}\right) \le \frac{2}{\sqrt{p}}$? tighter one $\frac{1}{\sqrt{\pi \log p}}$ (Johnstone 8.3b)

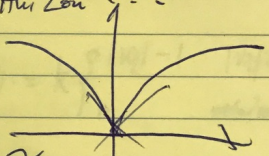
$|\bar{x} = \sqrt{2\log p}$ $\quad |\mu_i| \ne 0, |\mu_i| \ge$

Sparsity: $k = \#\{i : |Y_i = \mu_i + \varepsilon_i| \ge \sqrt{2\log p}\} << p$.

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

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

$$\hat{\mu}_\lambda = \arg\min \frac{1}{2}\|Y-\tilde{\mu}\|^2 + \lambda p(|\tilde{\mu}|)$$



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

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

NP-hard for global optimizer

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

7') Differential Inclusion (Osher-Ruan-Xiong-Yao-Yin '16)

LASSO. KKT.

$$\Rightarrow (\tilde{\mu} - y) + \lambda \partial \|\tilde{\mu}\|_1 \ni 0 \quad , \quad p(\tilde{\mu}) \in \partial \|\tilde{\mu}\|_1$$

$\lambda = t$

$$\Rightarrow \frac{p(\tilde{\mu})}{t} = -(\tilde{\mu} - y)$$

biased

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

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

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

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

① ~~Under nearly~~ Simple Algorithm

$$\frac{\partial \phi(\tilde{\mu})}{dt} + \frac{1}{k}\frac{d\tilde{\mu}}{dt} = -\nabla \ell(\tilde{\mu})$$ ← loss.

Euler Discretization → $w^t = \phi^t + \frac{1}{k}\tilde{\mu}^t$

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

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

* Model Selection Problem

$\mu = X\beta$ (X: feature/covariate matrix) $\quad X^{n \times p}$ $\quad n < p$

underdetermined in general

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

$\begin{cases} y = X\beta \Rightarrow \beta ? & \text{compressed sensing, talk later} \\ y = X\beta \Rightarrow \hat{\beta} \text{ s.t.} \begin{cases} \text{supp}(\hat{\beta}) = \text{supp}(\beta) \\ \frac{1}{n}\|\hat{\beta} - \beta\|^2 \text{ small?} \end{cases} & \text{Model Selection} \\ & \text{LASSO: Consistency} \\ & L_2\text{-consistency} \end{cases}$

LASSO $\arg\min_{\beta} \frac{1}{n}\|y - X\tilde{\beta}\|^2 + \lambda\|\tilde{\beta}\|_1 = \hat{\beta}^{\text{LASSO}}_{\lambda}$

Model-Selection Consistency: $\begin{cases} \text{Irrepresentable Condition} \\ \text{Strong Signal} \end{cases}$ $\quad S = \text{supp}(\beta_i)$

$\forall j \notin S \quad \left\|(X_S^T X_S)^{-1} X_S^T X_j\right\|_1 < 1$

$\min(\beta_i) > O\left(\sqrt{\frac{\log p}{n}}\right)$

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

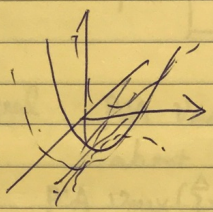
L_2 - Consistency: RE (restricted eigenvalue) condition — Bickel-Ritov-Tsybakov

Hessian matrix $H_n = \frac{X^T X}{n}$ is not p.s.d.

but restricted p.s.d. on a cone near β^*

$\Delta = \hat{\beta} - \beta^*$ $\left\{ \Delta^T H_n \Delta \geq \gamma \|\Delta\| : \|\Delta_{S^c}\|_1 \leq c\|\Delta_S\| \right\}$

IRR $\Rightarrow$ RE.



Note: Bregman ISS (LBI) nearly the same as LASSO.

Split LBI/ISS: better than LASSO / genLasso.

(NIPS 2016: Huang - Sun - Xiong - Yao).

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

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

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

fixed p, $n \to \infty$ $\hat{\Sigma}_n \Rightarrow \Sigma = \sigma_u^2 u u^T + \sigma_\varepsilon^2 I_p$

$= \frac{1}{n} \sum_{i=1}^n Y_i Y_i^T$

$\sigma_u^2 u u^T$: low rank; $\sigma_\varepsilon^2 I_p$: diagonal: sparse

If $\sigma_u^2 > \sigma_\varepsilon^2$, top-1 eval. $\sigma_u^2 + \sigma_\varepsilon^2$ eve. u.

How about $(p, n) \to +\infty$?

$\gamma = \lim_{p,n\to\infty} \frac{p}{n}$ ~~γ~~ ~~$p,n\to\infty$~~ $\sigma_\varepsilon^2 = 1$.

$$\lambda_{max}(\hat{\Sigma}_n) \to \begin{cases} (1+\sqrt{\gamma})^2 = b & \sigma_x^2 \le \sqrt{\gamma} \\ (1+\sigma_x^2)\left(1+\frac{\gamma}{\sigma_x^2}\right), & \sigma_x^2 > \sqrt{\gamma} \end{cases}$$

$$\left|\langle u, v_{max}(\hat{\Sigma}_n)\rangle\right| \to \begin{cases} 0 & \sigma_x^2 \le \sqrt{\gamma} \\ \dfrac{1-\gamma/\sigma_x^4}{1+\gamma/\sigma_x^2} & \sigma_x^2 > \sqrt{\gamma} \end{cases}$$

$\sigma_x^2 \le \sqrt{\gamma}$ 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 a biased est. of $\sigma_x^2 + \sigma_\varepsilon^2 = 1$

Marcenko Pastur Dist. of Random matrices

$$X_i \overset{i.id}{\sim} N(0, I_p) \quad i = 1, \cdots n.$$

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

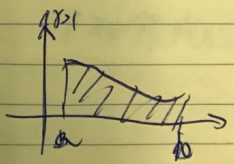
$\gamma = \frac{p}{n}$ eigvalue dist.

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

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

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

$$\int_{t=0} \left(1-\frac{1}{\gamma}\right) I(\gamma \ge 1)$$



Lecture 5. Robust PCA

Sep. 15, 2017

Lecture Graph Realization V. MDS with Uncertainty

MDS with Uncertainty : Sensor Network Localization

Recall Classical MDS

Given d_{ij} pair distance $\forall (i,j)$ Complete

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

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

$$\Leftrightarrow \min_{Y} \| YY^T - B \|^2 \qquad B = -\frac{1}{2} H D H^T$$

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

$$\Leftrightarrow \text{eigen-decomp}(B)$$

新的特点 : SNL

$G = (V, E)$

V : sensor

① $(i,j) \in E$ iff d_{ij}

incomplete

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

③ anchor point partial $x_i = a_i$

目标:

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

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

? $\min_{Y} \sum_{(i,j)\in E} \left(\|y_i - y_j\|^2 - d_{ij}^2 \right)^2$ { gradient method nonlinear opt.

不能 Eigen

练习: $\|y_i - y_j\|^2 = d_{ij}^2$ Quadratic equality

SDP Relax $\Rightarrow$ "Linear"

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

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

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

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

$e_i = (0 \cdots 0 \underset{i\text{-th}}{\downarrow} 1 \; 0 \cdots 0)$

$y_i = Y \cdot e_i$

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

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

$X = Y^T Y \succeq 0$

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

LMI: $X \succeq 0$ SD

Relaxation

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

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

Lemma

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

SDR $\Rightarrow$ LMI

$$\begin{cases} Z = \begin{bmatrix} I & Y \\ Y^T & X \end{bmatrix} \succeq 0 \\ Z_{1:k,1:k} = I \\ (0; e_i - e_j)(0; e_i - e_j)^T \cdot Z = d_{ij}^2 \quad (i,j) \in E \end{cases}$$

Note:

① $d_{ij}^2(1-\epsilon_{ij}) \le \|y_i - y_j\|^2 \le d_{ij}^2(1+\epsilon_{ij}) \Rightarrow$ LMI. noise

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

SPP approach → MDS

MatLab Protein 3-D structure Reconstruction

CMDS Schoenberg

SDP-MDS Exact Recovery ? Yinyu Ye group.

P. Biswas.

A. So.

Recall SDP

(SPP) $\min\ C \bullet X$ $C, X \in \mathbb{R}^{n\times n}$

s.t. $A_i \bullet X = b_i$ $i = 1, \dots, m$ $b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$

$X \succeq 0$

(SDD) $\max\ -b^T y$ $y \in \mathbb{R}^m$, $S \in \mathbb{R}^{n\times n}$

s.t. $S = C - \sum_{i=1}^{m} A_i y_i \succeq 0$

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

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

primal obj: $C \bullet X$ dual obj $b^T y$

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

$\Rightarrow$ $C \bullet X - b^T y \geq 0$ $\forall X \in F_P$, $\forall (y, S) \in F_d$

duality gap

Strong Duality of SDP

(1) $F_P \neq \phi$, $F_D \neq \phi$

(2) F_P or F_D has an interior solution

$\Rightarrow$ $X^* \in F_P$, $(y^*, S^*) \in F_D$ is optimal soln

iff $C \bullet X^* = b^T y^*$ duality gap is zero

$X^* S^* = 0$. Complementary

存在内点 Soln

X^*, (y^*, S^*) $X^* S^* = 0$ opt.

"witness" primal-dual pair

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

SDP-MDS.

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

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

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

$\text{rank}(Z^*) \geq k$

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

$\Rightarrow \text{rank}(Z^*) + \text{rank}(S^*) \leq k + n$

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

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

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

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

几何：Universal Rigidity (UR)：

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

$(\underbrace{y_i}_{k}, \underbrace{0; \cdots; 0}_{l-k})$ s.t. $\|y_i - y_j\| = d_{ij}, (ij) \in E$

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

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

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

- (几何) G is UR or has a unique embedding in $\mathbb{R}^k$
- (代数1) SDP has a max-rank feasible soln $\text{rank}(Z^*) = k$ ($\text{rank}(S^*) = n$)
- $X^* = Y^T Y$ or $\text{trace}(X - Y^T Y) \Rightarrow$ eigendecomp of X^*.

UR $\Leftrightarrow$ polynomial $(n, k, \log(\frac{1}{\varepsilon}))$

Ye, ICCM '2010 ; Fields '2011

Maximum Variance Unfolding (Manifold Learning)

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

MVU

$K = Y^T Y$ 变量

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

$K_{ij} = \langle y_i, y_j \rangle$

"SDP"

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

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

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

$K \succeq 0$

$K \Rightarrow Y^T Y$

Why max tr(K) ? 不能 work.

[So-Ye '2007]

unfolding

SDP.

k-manifold

(k+1)-lateration graph

"Unfold manifold"