A Mathematical Introduction to Data Science

Lecture 2: Generalized MDS/PCA

Yuan Yao

Peking University

2015.5.6.

National Central University

Main Content

- Lecture 1: Sample mean and Covariance (PCA/MDS): Fisher's Principle of Maximum Likelihood Estimate, yet things might go wrong --
  - Stein's phenomenon and shrinkage
  - Random matrix theory and failure of PCA
- Lecture 2: Beyond MDS/PCA
  - Random projections and compressed sensing
  - Semidefinite Programming (SDP) extensions of MDS/PCA:
    - MDS with uncertainty (Sensor Network Localization)
    - Robust PCA
    - Sparse PCA
- Lecture 3: Geometric and Topological structures of data
  - Manifold learning: ISOMAP/LLE etc.
  - From graphs to simplicial complexes: persistent homology, Hodge theory, etc.

Random Projections and Compressed Sensing

A Blessing of Dimensionality: Concentration of Measure

Reflection of MDS

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

- MDS looks for Euclidean embedding of data whose *total* or *average* metric distortion are minimized.
- MDS embedding basis is *adaptive* to the data, namely as a function of data via eigen-decomposition.

- Can we find an universal embedding whose metric distortions are uniformly small for all pairs

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

Johnson-Lindenstrauss Lemma

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

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

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

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

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

Remarks

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

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

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

where the following random matrices R can cater our needs.

- $R = [r_1, \cdots, r_k] \quad r_i \in S^{d-1} \quad r_i = (a_1^i, \cdots, a_d^i)/\parallel a^i \parallel \quad a_k^i \sim N(0,1)$
- $R = A/\sqrt{k} \quad A_{ij} \sim N(0,1)$
- $R = A/\sqrt{k} \quad A_{ij} = \begin{cases} 1 & p = 1/2 \\ -1 & p = 1/2 \end{cases}$
- $R = A/\sqrt{k/3} \quad A_{ij} = \begin{cases} 1 & p = 1/6 \\ 0 & p = 2/3 \\ -1 & p = 1/6 \end{cases}$

Remarks

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

Example: Human Genome Diversity Project

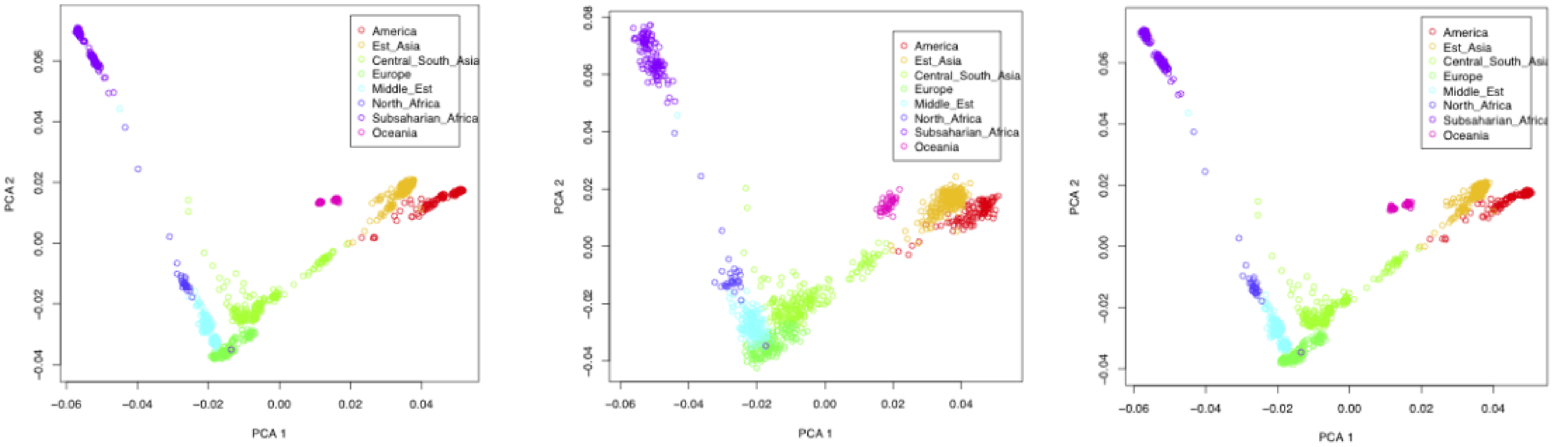
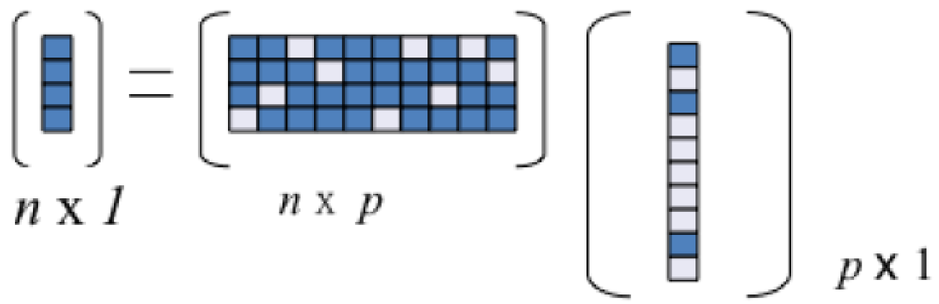


FIGURE 1. (Left) Projection of 1043 individuals on the top 2 MDS principal components. (Middle) MDS computed from 5,000 random projections. (Right) MDS computed from 100,000 random projections. Pictures are due to Qing Wang.

Application: Compressed Sensing

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

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



$$\begin{array}{lll} (P_0) & \min & \|x\|_0 \\ & s.t. & \Phi x = b. \end{array}$$

$$\begin{array}{lll} (P_1) & \min & \|x\|_1 := \sum |x_i| \\ & s.t. & \Phi x = b. \end{array}$$

- P0 is NP-hard, while P1 is tractable as linear programming
- Under what conditions, P0 or P1 have solutions exactly meets the true signal?

Restricted Isometry Property

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

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

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

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

Exact Recovery Theory (e.g. Candes'08)

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

(1) If $\delta_{2k} < 1$, then problem P_0 has a unique solution x^*;

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

From Johnson-Lindenstrauss to RIP

- The key step to establish Johnson-Lindenstrauss Lemma is the following

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

- Then [Baraniuk-Davenport-DeVore-Wakin'2008]

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

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

the following RIP holds

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

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

Summary

- Johnson-Lindenstrauss Lemma shows that random projections provide
  - universal basis (non-adaptive) with
  - almost isometry uniformly for all data pairs (rather than average distortion)
- Johnson-Lindenstrauss Lemma implies Restricted Isometry Property in compressed sensing

SDP extensions of MDS/PCA

Incomplete MDS with uncertainty, Robust PCA, Sparse PCA

SDP is like LP with Linear Matrix Inequalities

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

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

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

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

LD (Dual Linear Programming):

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

SDD (Dual Semi-definite Programming):

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

- So SDP is a generalization of LP
- Tractable in theory, but in practice not easy to deal with large matrices (e.g. 100-by-100 for CVX)

Extension I: MDS with missing values (Graph Realization Problem)

Given a graph $G = (V, E)$ and sets of non–negative weights, say $\{d_{ij} : (i, j) \in E\}$ on edges, the goal is to compute a realization of G in the Euclidean space $\mathbf{R}^d$ for a given low dimension d. That is,

- to place the vertexes of G in $\mathbf{R}^d$ such that
- the Euclidean distance between a pair of adjacent vertexes (i, j) equals to (or bounded by) the prescribed weight $d_{ij} \in E$.

Classical MDS (complete graph with squared distance d_{ij}):
Young/Householder 1938, Shoenberg 1938

MDS with Uncertainty Quadratic equality and inequality system

Given graph $G = (V, E)$ and $d_{ij} \in E$, find $\mathbf{x}_j \in \mathbf{R}^d$ such that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (\leq) = (\geq) \quad d_{ij}^2, \ \forall \ (i,j) \in E, \ i < j.$$

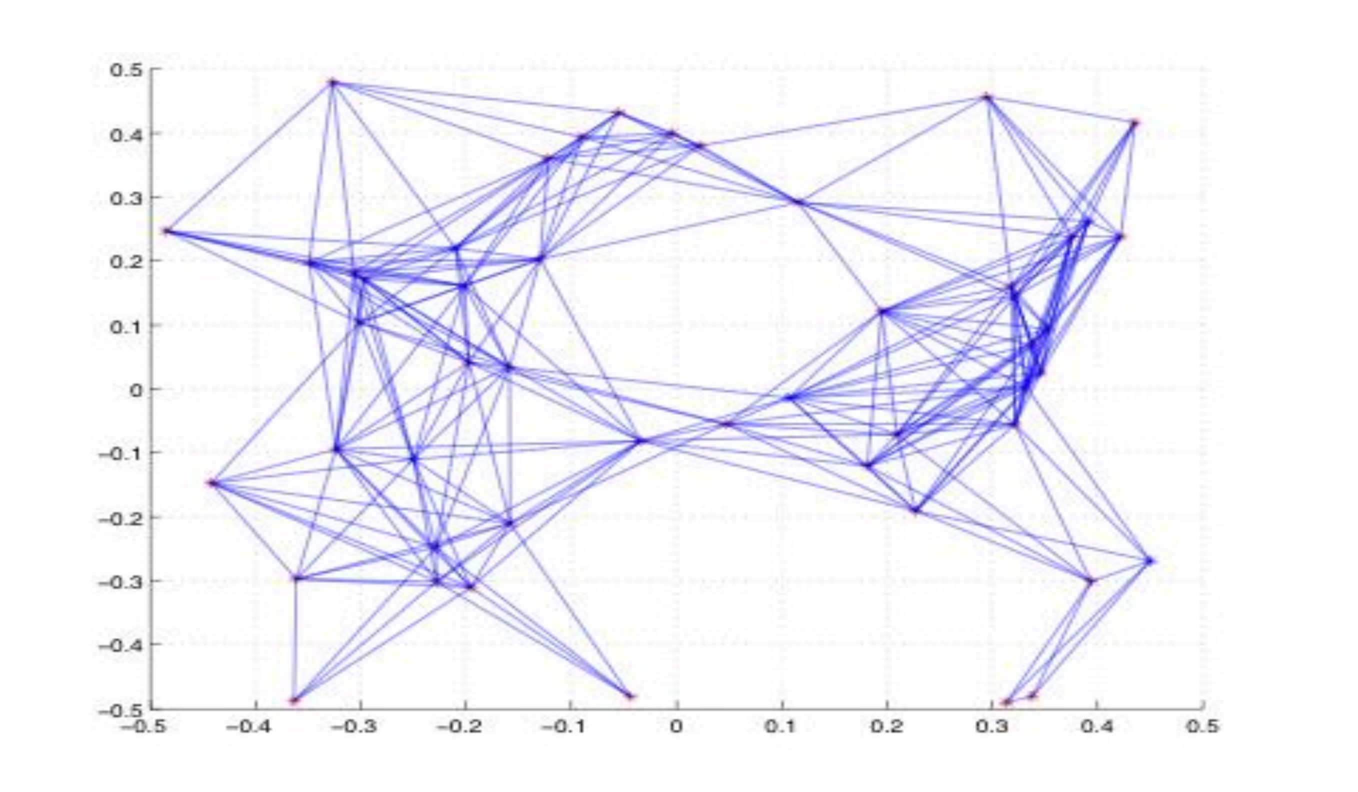
MDS in Sensor Network Localization: anchor points

Or given $\mathbf{a}_k \in \mathbf{R}^d$, $d_{ij} \in N_x$, and $\hat{d}_{kj} \in N_a$, find $\mathbf{x}_i \in \mathbf{R}^d$ such that

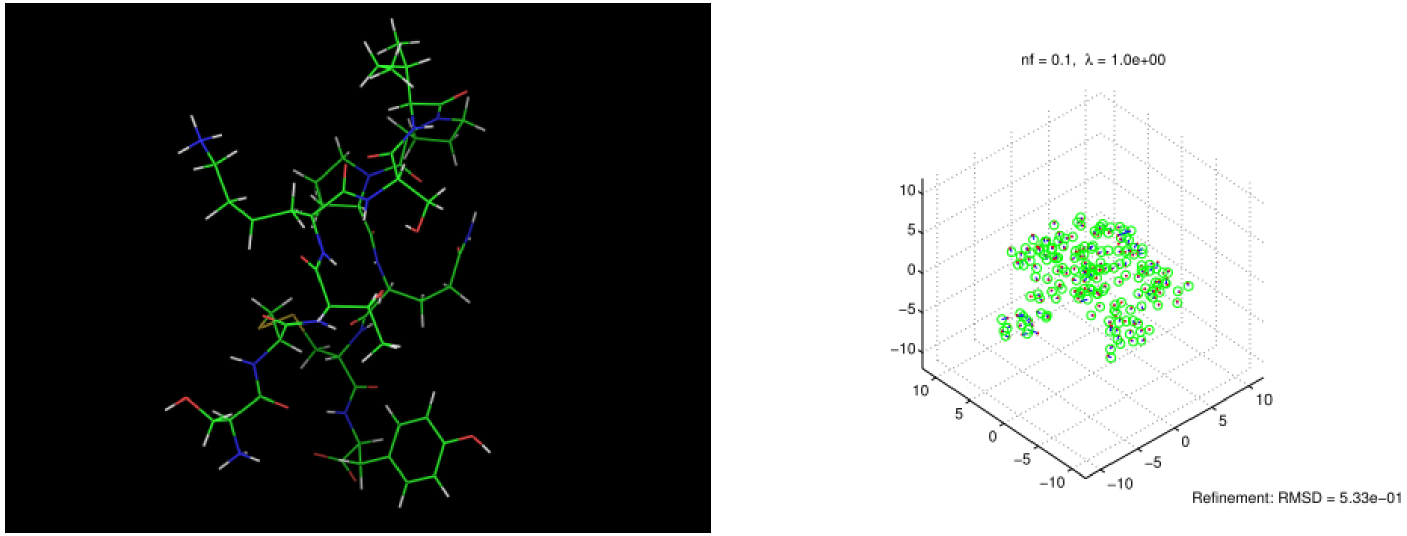
$$\begin{aligned} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (\leq) = (\geq) \quad & d_{ij}^2, \ \forall\, (i,j) \in N_x, \ i < j, \\ \|\mathbf{a}_k - \mathbf{x}_j\|^2 \quad (\leq) = (\geq) \quad & \hat{d}_{kj}^2, \ \forall\, (k,j) \in N_a; \end{aligned}$$

that is, edge (ij) (or (kj)) connects sensors i and j (or anchor k and sensor j) with the Euclidean length equal to d_{ij} (or $\hat{d}_{kj}$).

Example: 50-node 2-D Sensor Network Localiztion



Example: Protein 3D reconstruction



(a) (b)

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

Key Problems

Recall the system:

$$\begin{aligned} \|\mathbf{x}_i - \mathbf{x}_j\|^2 &= d_{ij}^2, \ \forall\, (i,j) \in N_x, \ i < j, \\ \|\mathbf{a}_k - \mathbf{x}_j\|^2 &= \hat{d}_{kj}^2, \ \forall\, (k,j) \in N_a, \end{aligned}$$

- Does the system have a localization or realization of all $\mathbf{x}_j$'s?
- Is the localization unique or the framework $(G, D, \mathbf{x})$ is rigid, and it can be certified?
- Is the system partially localizable or rigid with a certification?

Nonlinear Least Squares

$$\min_{\mathbf{x}} \quad \sum_{(i,j)\in N_x}(\|\mathbf{x}_i - \mathbf{x}_j\|^2 - d_{ij}^2)^2 + \sum_{(k,j)\in N_a}(\|\mathbf{a}_k - \mathbf{x}_j\|^2 - \hat{d}_{kj}^2)^2$$

Or

$$\min_{\mathbf{x}} \quad \sum_{(i,j)\in N_x}(\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij})^2 + \sum_{(k,j)\in N_a}(\|\mathbf{a}_k - \mathbf{x}_j\| - \hat{d}_{kj})^2.$$

A difficult global optimization problem.

Matrix Represenation I

For simplicity, let $d = 2$ and $X = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_n]$ be the $2 \times n$ matrix that needs to be determined and $\mathbf{e}_j$ be the vector of all zero except 1 at the jth position. Then

$$\mathbf{x}_i - \mathbf{x}_j = X(\mathbf{e}_i - \mathbf{e}_j) \quad \text{and} \quad \mathbf{a}_k - \mathbf{x}_j = [I \quad X](\mathbf{a}_k; -\mathbf{e}_j)$$

so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = (\mathbf{e}_i - \mathbf{e}_j)^T X^T X(\mathbf{e}_i - \mathbf{e}_j)$$

$$\|\mathbf{a}_k - \mathbf{x}_j\|^2 = (\mathbf{a}_k; -\mathbf{e}_j)^T [I \quad X]^T [I \quad X](\mathbf{a}_k; -\mathbf{e}_j) =$$

$$(\mathbf{a}_k; -\mathbf{e}_j)^T \begin{pmatrix} I & X \\ X^T & X^T X \end{pmatrix} (\mathbf{a}_k; -\mathbf{e}_j).$$

Matrix Represenation II

Or, equivalently,

$$
\begin{aligned}
(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T \bullet Y &= d_{ij}^2, \ \forall\ i,j \in N_x,\ i<j, \\
(\mathbf{a}_k; -\mathbf{e}_j)(\mathbf{a}_k; -\mathbf{e}_j)^T \bullet \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix} &= \hat{d}_{kj}^2, \ \forall\ k,j \in N_a, \\
Y &= X^T X.
\end{aligned}
$$

SDP Relaxation [Biswas-Ye 2004]

Relax

$$Y = X^T X$$

to

$$Y \succeq X^T X;$$

or equivalently to

$$Z := \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix} \succeq \mathbf{0}.$$

SDP Standard Form [Biswas-Ye 2004]

Find a symmetric matrix $Z \in \mathcal{S}^{(n+2)}$ such that

$$\begin{aligned} Z_{1:2,1:2} &= I \\ (\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)(\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)^T \bullet Z &= d_{ij}^2, \ \forall\ i, j \in N_x, \ \ i < j, \\ (\mathbf{a}_k; -\mathbf{e}_j)(\mathbf{a}_k; -\mathbf{e}_j)^T \bullet Z &= \hat{d}_{kj}^2, \ \forall\ k, j \in N_a, \\ Z &\succeq \mathbf{0}. \end{aligned}$$

- This is an SDP problem,
- if every sensor point is connected, directly or indirectly, to an anchor point, then the solution set must be bounded,
- a solution matrix Z has rank at least 2,
- it's 2 if and only if $Y = X^T X$ and it solves the original problem.

SDP Dual Form [Biswas-Ye 2004]

$$
\begin{array}{ll}
\text{minimize} & I \bullet V + \sum_{i<j\in N_x} w_{ij} d_{ij}^2 + \sum_{k,j\in N_a} \hat{w}_{kj} \hat{d}_{kj}^2 \\
\text{subject to} & \begin{pmatrix} V & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \sum_{i<j\in N_x} w_{ij} (\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)(\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)^T \\
& + \sum_{k,j\in N_a} w_{kj} (\mathbf{a}_k; -\mathbf{e}_j)(\mathbf{a}_k; -\mathbf{e}_j)^T = S \succeq 0,
\end{array}
$$

where variable matrix $V \in \mathcal{S}^2$, variable w_{ij} is the (stress) weight on edge between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\hat{w}_{kj}$ is the (stress) weight on edge between $\mathbf{a}_k$ and $\mathbf{x}_j$.

- The dual is always feasible since $V = \mathbf{0}$ and all $w_{\cdot}$s equal 0 is a feasible solution.
- The rank of any optimal dual stress matrix S is less or equal to n.

Unique Localizability

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

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

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

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

Rank Condition for Universally Rigidity

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

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

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

- If we allow uncertainty with pairwise distance inequalities, then arbitrarily small rank solutions are possible [So-Ye-Zhang'2008].

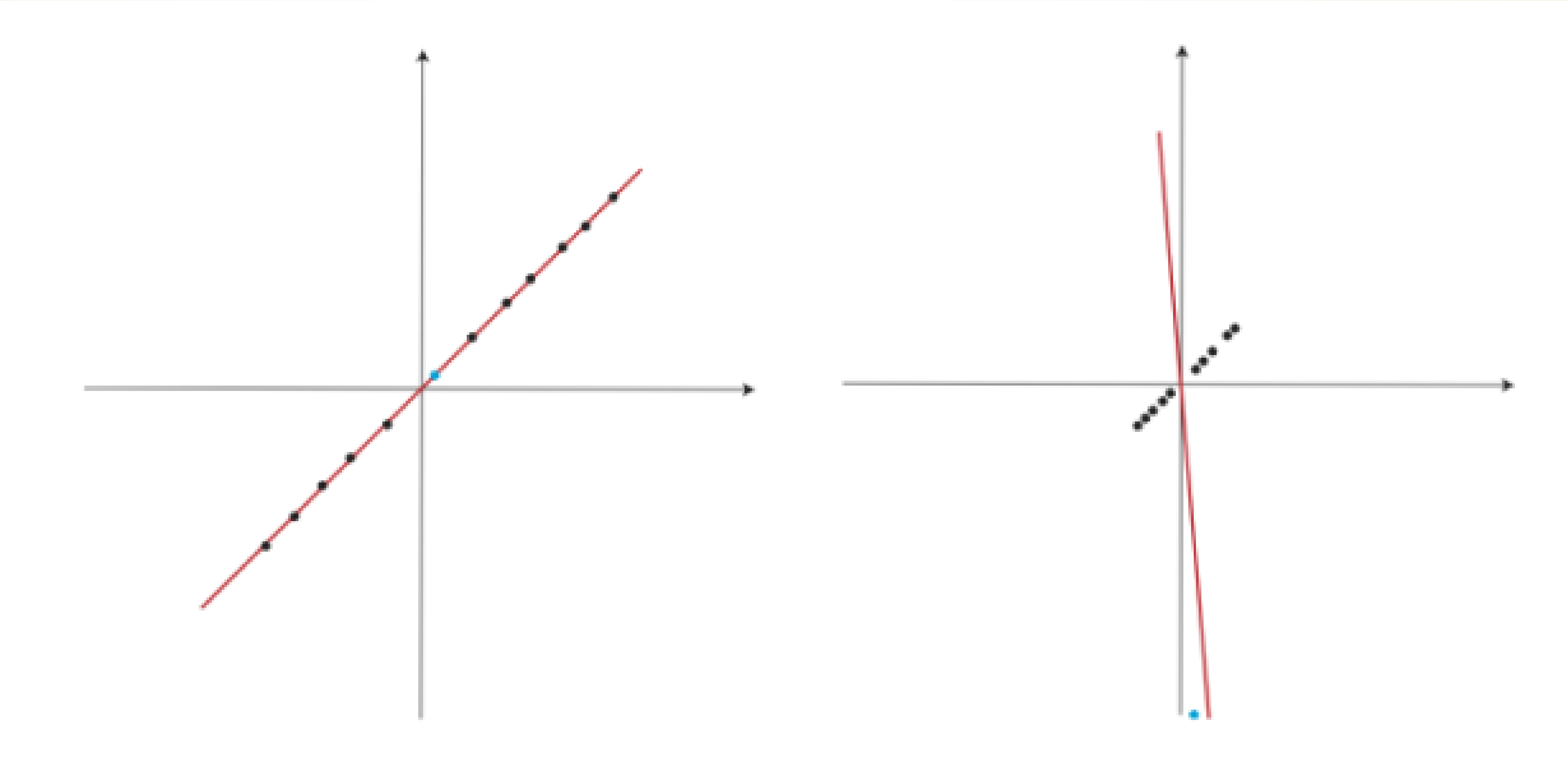
Reference

- Schoenberg, I. J. (1938a) Metric Spaces and Completely Monotone Functions. The Annals of Mathematics 39, 811–841
- Schoenberg, I. J. (1938b) Metric Spaces and Positive Definite Functions. Transactions of the American Mathematical Society 44, 522–536
- Francois Bavaud (2010) On the Schoenberg Transformations in Data Analysis: Theory and Illustrations, arXiv:1004.0089v2
- T. F. Cox and M. A. A. Cox, Multidimensional Scaling. New York: Chapman & Hall, 1994.
- Yinyu Ye, Semidefinite Programming and its Low-Rank Theorems, ICCM 2010

Robust PCA

A sad fact is that PCA is sensitive to outliers…

PCA is sensitive to outliers



PCA as Low-Rank Approximation

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

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

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

$$X = L + E$$

- Here error term E is small in Frobenius norm or matrix spectral norm, sensitive to outliers

Robust PCA

Robust PCA looks for the following decomposition instead

$$X = L + S$$

where

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

Example: Rank-1 Signal Model

$$Y = X + \varepsilon,$$

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

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

Example

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

Example: Gaussian Graphical Models

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

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

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

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

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

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

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

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

A Convex Relaxation Approach: SDP

$$\begin{array}{ll} \min & \|X - L\|_0 \\ s.t. & \operatorname{rank}(L) \le k \end{array}$$

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

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

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

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

When does it work?

It is necessary to assume that

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

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

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

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

and

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

Exact Recovery Theory (Candes-Recht'09, Candes-Tao'10)

Theorem 3.1. Assume the following holds,

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

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

- It can be generalized to incomplete measurements

Sparse PCA

How about the principle components are sparse, i.e. with many zero elements?

PCA as Maximal Variance Projection

Recall that classical PCA is to solve

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

which gives the maximal variation direction of covariance matrix Σ.

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

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

A SDP approach to Sparse PCA

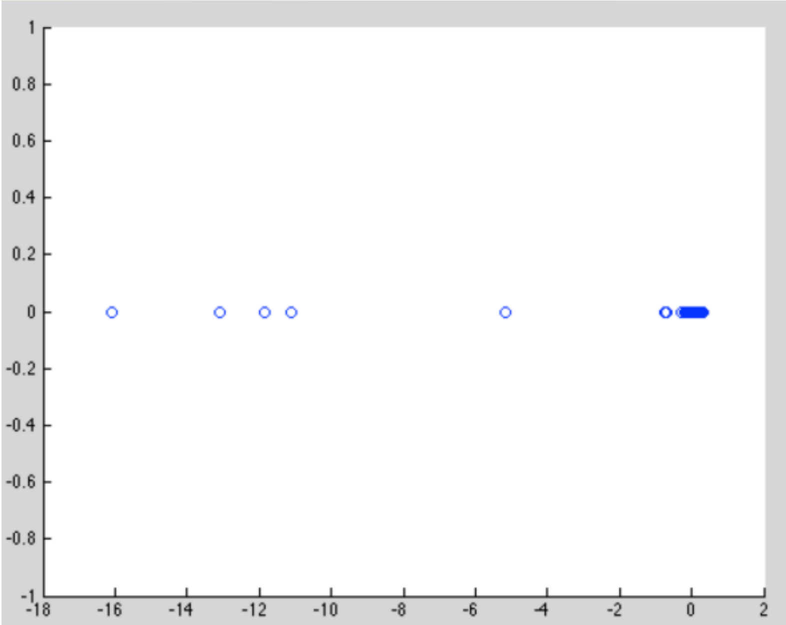
Now we are looking for sparse principal components, i.e. $\#\{X_{ij} \neq 0\}$ are small. Using 1-norm convexification, we have the following SDP formulation [**dGJL07**] for Sparse PCA

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

Example: A Journey to the West

(provided by Song Mei, Fangyuan Li, Yiding Lu, PKU course 2014.)

- Data contains a 302-by-408 (p-by-n) character-scene cooccurance matrix
- Figure below: (Left) first principle component; (middle) scores of the top 5 characters; (right) sparse PCA



人物	主成分得分
孙悟空	16.0893
唐僧	11.8656
猪八戒	13.0722
沙僧	11.1315
白龙马	5.1673

5.1.3 $\lambda = 10$

人物	得分
孙悟空	0.606
唐僧	0.441
猪八戒	0.487
沙僧	0.411
白龙马	0.181
观音菩萨	0.011
土地神	0.009

Summary

- Random projections
  - Universal basis to achieve almost isometry (Johnson-Lindenstrauss Lemma)
  - Applications in compressed sensing (RIP)
- SDP extensions of MDS/PCA
  - Incomplete MDS (graph realization/sensor network localization)
  - Robust PCA
  - Sparse PCA