P1

TO

Ref : [2011][E. Kokiopoulou, etal]

Trace optimization and eigen problems in DR methods.

Sec 2

Preliminaries :

Thm (Min-max thm) (Courant-Fischer-Weyl).

A : $n\times n$ self-adjoint (over $\mathbb{R}$ or $\mathbb{C}$) w/ e-val. : $\lambda_1 \le \cdots \le \lambda_k \le \cdots \le \lambda_n$ and

Then

$$\lambda_k = \min_{\substack{U: \\ \dim(U)=k}} \left(\max_{\substack{x\in U \\ \|x\|=1}} x^T A x \right), \text{ and}$$

x^*Ax if over $\mathbb{C}$

$$\lambda_k = \max_{\substack{U: \\ \dim(U)=n-k+1}} \left(\min_{\substack{x\in U \\ \|x\|=1}} x^T A x \right).$$

Thm (trace optimization)

A : $n\times n$ self-adjoint over $\mathbb{R}$, w/ e-val. : $\lambda_1 \le \cdots \le \lambda_n$.

Then

(i)
$$\max_{\substack{V\in\mathbb{R}^{n\times d}: \\ V^TV=I}} tr(V^TAV) = \lambda_n + \cdots + \lambda_{n-d+1} \quad (d\le n)$$

(ii)
$$\min_{\substack{V\in\mathbb{R}^{n\times d} \\ V^TV=I}} tr(V^TAV) = \lambda_1 + \cdots + \lambda_d. \quad (d\le n)$$

<Pf>

We prove (ii) ; (i) can be proved similarly.

Denote $V = [v_1 \cdots v_d]$ $n\times d$ (w/ $V^TV = I_d$).

Then $tr(V^TAV) = \sum_{i=1}^{d} v_i^T A v_i$.

WLOG, assume $v_1^TAv_1 \le v_2^TAv_2 \le \cdots \le v_d^TAv_d$.

claim : $v_i^TAv_i \le \lambda_i$, for $i=1,\cdots,d$.

(Thus, the minimization problem is solved by choosing $v_1,\cdots,v_d$ to be the first d e-vec. of A.)

$v_1^TAv_1 \le \lambda_1$ is clearly true.

We prove by induction.

Assume $v_i^TAv_i \le \lambda_i$, for $i=1,\cdots,k-1$. ($k\ge 2$)

Then

$$\lambda_k = \min_{\substack{U: \\ \dim U=k}} \left(\max_{\substack{x\in U \\ \|x\|=1}} x^TAx \right)$$

$$\le \min_{\substack{U: \\ \dim U=k \\ v_1,\cdots,v_{k-1}\in U}} \left(\max_{\substack{x\in U: \\ \|x\|=1 \\ x\perp v_i, \\ i=1,\cdots,k-1}} x^TAx \right)$$

$$= \min_{\substack{x\in\mathbb{R}^n \\ x\perp v_i, \\ i=1,\cdots,k-1}} x^TAx \le v_k^TAv_k.$$

Thus, by induction, the claim is proved. #

Sec 3

Nonlinear dimension reduction

LLE and Laplacian Eigenmaps are introduced here.

3.1 LLE : (Locally linear embedding).

$\mathcal{X} = \{x_j\}_{j=1}^{n} \subseteq \mathbb{R}^D$.

Algorithm :

(0) Construct a neighborhood graph on $\mathcal{X}$ using KNN (intuitively, choose $K=d+1$).

P2 TO

(1) For each i, solve

$$\underset{\substack{w_{ii_1},\cdots,w_{ii_K} \\ \sum_{j=1}^K w_{ii_j}=1}}{\operatorname{argmin}} \left\| x_i - \sum_{j=1}^K w_{ii_j} x_{i_j} \right\|_2^2 .$$

(Find best coefficients for neighbors of x_i to rep. x_i as a linear combination.)

(2) Define $w_{ij}=0$ if $j=i$ or x_j not a kNN of x_i.

For $Y=[y_1 \cdots y_n]$ $(d\times n)$, define

$$\mathcal{F}_{LLE}(Y) = \sum_i \|y_i - \sum_j w_{ij} y_j\|_2^2 .$$

$$\left(= tr\left(Y \underbrace{(I-W^T)(I-W)}_{\substack{\shortparallel \\ M}} Y^T \right)\right).$$

M, called the $\boxed{\text{LLE matrix}}$.

Solve

$$\underset{\substack{Y\in\mathbb{R}^{d\times n} \\ YY^T=I_d \\ y_1+\cdots+y_n=0}}{\operatorname{argmin}} \mathcal{F}_{LLE}(Y). \qquad (*)_{LLE}$$

The resulting Y is the output of LLE.

Rmk:

① $YY^T=I_d \iff$ rows of Y are orthonormal.

② The sol'n of $(*)_{LLE}$ is given by the 2nd to the $(d+1)$st bottom eigenvec. of M putting as the rows of Y.

<Pf>

$\because$ For all i, $\sum_j w_{ij}=1$

$\therefore W\mathbb{1}=\mathbb{1}. \Rightarrow (I-W)\mathbb{1}=0.$

Thus, $\mathbb{1}$ is an e-vec. of $(I-W^T)(I-W)$ w.r.t. e-val. 0.

Therefore, the rest of the e-sp. are orthogonal to $\mathbb{1}$.

Notice that $y_1+\cdots+y_n=0 \iff (\text{rows of } Y)\perp \mathbb{1}$

Thus, $(*)$ can be rewritten as

$$\underset{\substack{Y\in\mathbb{R}^{d\times n} \\ (\text{rows of } Y)\in\langle \mathbb{1}\rangle^{\perp} \\ (\text{rows of } Y) \text{ are o.n.}}}{\operatorname{argmin}} \mathcal{F}_{LLE}(Y).$$

By the trace optimization thm, the result follows. #

3.2 Laplacian Eigenmaps

$$\mathcal{X}=\{x_j\}_{j=1}^n \subseteq \mathbb{R}^D.$$

(0) Construct a connected graph G w/ vertex set $\mathcal{X}$ using ε-nbd or kNN.

(1) Assign weights on edges of G by

$$w_{ij}=\begin{cases} \exp(-\|x_i-x_j\|^2/t) & \text{if } x_ix_j\in E_G \\ 0 & \text{o.w.} \end{cases}$$

or

$$w_{ij}=\begin{cases} 1 & \text{if } x_ix_j\in E_G \\ 0 & \text{o.w.} \end{cases}$$

(2) Define $W=[w_{ij}]$. (Rmk: $w_{ii}:=0$).

$D:=\operatorname{diag}(D_{11},\cdots,D_{nn})$, where $D_{ii}=\sum_j w_{ij}$.

Define $L:=D-W$, called the graph Laplacian of G.

(3) For $Y\in\mathbb{R}^{d\times n}$, define

$$\mathcal{F}_{\underset{\substack{\shortparallel \\ \text{EigenMap}}}{EM}}(Y)=\sum_{i,j=1}^n w_{ij}\cdot\|y_i-y_j\|_2^2 .$$

$$(=2tr(YLY^T)).$$

Solve

$$\underset{\substack{\{Y\in\mathbb{R}^{d\times n}: \\ YDY^T=I_d \\ \sum_{i=1}^n \sqrt{D_{ii}}\, y_i = 0\}}}{\operatorname{argmin}} \mathcal{F}_{EM}(Y). \qquad (*)_{EM}$$

Rmk:

① Setting $\hat{Y} = YD^{1/2}$, $\hat{W} = D^{-1/2} W D^{-1/2}$, and

$\hat{L} = I - \hat{W}$, ← called the $\boxed{\text{normalized Laplacian}}$.

then $(*)_{EM}$ can be rewritten as

$$\underset{\substack{\hat{Y}\in\mathbb{R}^{d\times n} \\ \hat{Y}\hat{Y}^T=I_d \\ \hat{y}_1+\cdots+\hat{y}_n=0}}{\operatorname{argmin}} \operatorname{tr}\left(\hat{Y}(I-\hat{W})\hat{Y}^T\right).$$

Thus, the sol'n is

$$\hat{Y} = [\hat{u}_2, \cdots, \hat{u}_{d+1}]^T$$

(or, equivalently, $Y = [\hat{u}_2, \cdots, \hat{u}_{d+1}]^T D^{1/2}$).

Sec 4 Linear dimension reduction

Idea: Find $V \in \mathbb{R}^{D\times d}$ s.t. $\underset{(d\times D)}{Y = V^T}\underset{(D\times n)}{X}$ "preserves some information the best".

4.1 PCA (Principal Component Analysis).

(0)

$$\mathcal{F}_{PCA}(Y) := \left(\sum_{i=1}^n \Big\| y_i - \underbrace{\frac{1}{n}\sum_{j=1}^n y_j}_{\text{(mean)}} \Big\|_2^2\right) \leftarrow \boxed{\text{variance}}$$

$$\boxed{Y=V^TX} \rightarrow \left(= \operatorname{tr}\Big(V^T X \underbrace{\left(I - \tfrac{1}{n}11^T\right)}_{\text{called the centering matrix}} X^T V\Big)\right)$$

(1) Solve

$$\underset{\substack{V\in\mathbb{R}^{D\times d} \\ V^TV=I}}{\operatorname{argmax}} \mathcal{F}_{PCA}(Y). \qquad (*)_{PCA}$$

i.e. PCA seeks orthogonal projection that preserves maximal variance.

Rmk:

① $\left(I - \frac{1}{n}11^T\right)^2 = I - \frac{1}{n}11^T$.

Thus, writing $\overline{X} := X\left(I - \frac{1}{n}11^T\right)$, $(*)_{PCA}$ can be rewritten as

$$\underset{\substack{V\in\mathbb{R}^{D\times d} \\ V^TV=I}}{\operatorname{argmax}} \operatorname{tr}\left(V^T \overline{X}\,\overline{X}^T V\right).$$

② The sol'n is given by putting the bottom d e-vec. of $\underset{(D\times D)}{\overline{X}\,\overline{X}^T}$ to the columns of V. ← for relation to SVD of $\overline{X}\,\overline{X}^T$, see rmks in MDS below.

③ It turns out that

max (variance) $\Longleftrightarrow$ min (projected error)

$\boxed{\text{proj. error}} := \|\overline{X} - VV^T\overline{X}\|_F^2$.

The pts Vy_i, $i=1,\cdots,n$, are called $\boxed{\text{reconstructed pts}}$.

4.2 MDS and ISOMAP.

MDS (metric MDS) ← multidimensional scaling.

Now assume (the data matrix) is centered (at 0), denoted by $\overline{X} = \left[\overline{x}_1 \cdots \overline{x}_n\right]$.

$G := \left[\langle \overline{x}_i, \overline{x}_j\rangle\right]_{n\times n} = [g_{ij}]$. $(= \overline{X}^T\overline{X})$ is called the Grammian of $\overline{X}$.

Rmk:

① $s_{ij} := \|x_i - x_j\|_2^2 = g_{ii} + g_{jj} - 2g_{ij}$. ← squared distance

$S := [s_{ij}]$

② $G = [g_{ij}] = -\frac{1}{2}\left[I - \frac{1}{n}11^T\right] S \left[I - \frac{1}{n}11^T\right]$

i.e. $g_{ij} = \frac{1}{2}\left[\frac{1}{n}\sum_k (s_{ik} + s_{jk}) - s_{ij} - \frac{1}{n^2}\sum_{k,l} s_{kl}\right]$.

MDS seeks to find the sol'n of

$$\underset{Y\in\mathbb{R}^{d\times n}}{\operatorname{argmin}} \|G - Y^TY\|_F^2 . \quad (*)_{MDS}$$

i.e. low dim. rep. Y whose Grammian matrix is closest to Grammian of $\bar{X}$.

Rmk:

① A sol'n to $(*)_{MDS}$ is $\underbrace{Y}_{d\times n} = \Lambda_d^{1/2} Z_d^T$ — output of MDS

where $G = Z\Lambda Z^T$ is the spectral decomp. of G, $\Lambda = \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}$, $\lambda_1 \geq \cdots \geq \lambda_n$,

$\underbrace{Z}_{n\times n} = \begin{pmatrix} z_1 & \cdots & z_n \end{pmatrix}$, $\underbrace{\Lambda_d}_{d\times d} = \begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_d\end{pmatrix}$, $\underbrace{Z_d}_{n\times d} = \begin{pmatrix} z_1 & \cdots & z_d \end{pmatrix}$.

② Recall the sol'n of PCA is $Y_{PCA} = U_d^T \bar{X}$ where $\underbrace{U}_{D\times D} = \begin{pmatrix} u_1 & \cdots & u_D \end{pmatrix}$, $u_1, \cdots, u_D$: top e-vec. of $\bar{X}\bar{X}^T$.

Notice that this U also shows up in SVD of $\bar{X}$: $\bar{X} = U\Sigma Z^T$, where Z is as in ①.

Thus, $Y_{PCA} = U_d^T \bar{X} = U_d^T U \Sigma Z^T = \Lambda_d^{1/2} Z_d^T$.

Surprisingly, it coincides w/ the output of MDS.

③ The sol'n of $(*)_{MDS}$ is only unique up to orthogonal transformation.

ISOMAP

(0) Construct affinity graph G via ε-nbd ~~of~~ or kNN. Assign weights on edges using Euclidean distances.

(1) Apply Dijkstra algorithm to obtain $d_G(x_i, x_j)$, $\forall\, x_i, x_j \in \mathcal{X}$.

$\hat{S} := [\, d_G^2(x_i, x_j) \,]$.

(2) $\hat{G} := -\frac{1}{2} J \hat{S} J$, where $J = I - \frac{1}{n} 1 1^T$ is the centering matrix.

Apply MDS on $\hat{G}$.

i.e. solve

$$\underset{Y\in\mathbb{R}^{d\times n}}{\operatorname{argmin}} \|\hat{G} - Y^TY\|_F^2 . \quad (*)_{ISOMAP}$$

4.3 LPP.

(= Locality Preserving Projections).

This is basically Laplacian Eigenmap w/ $y_i = V^T x_i$, $i = 1, \cdots, n$.

(0) $\mathcal{F}_{LPP}(Y) = \sum_{i,j=1}^{n} W_{ij} \|y_i - y_j\|^2$, $Y = V^T X$.

(1) Solve

$$\underset{\substack{V\in\mathbb{R}^{m\times d} \\ V^T(XDX^T)V = I \\ Y = V^TX}}{\operatorname{argmin}} \mathcal{F}_{LPP}(Y) = \operatorname{tr}\left[V^T X (D - W) X^T V\right].$$

where D is as in Lap. Eig.

4.4 ONPP (= Orthogonal Neighborhood Preserving Projection).

Idea: To seek an orthogonal mapping to best preserve the same affinity graph as LLE.

The optimizat'n we are solving is

$$\underset{\substack{V\in\mathbb{R}^{D\times d} \\ V^TV = I}}{\operatorname{argmin}} \operatorname{tr}\left[V^T X (I - W^T)(I - W) X^T V\right].$$