1 Transformation

1.1 LLE

In step 1 we choose k-nearest point of $\boldsymbol{x_i}$ and find their best linear representation in order to reduce the related point and get sparse weight matrix, the optimal target is

$$\boldsymbol{W}^* := \arg\min_{\boldsymbol{W}} \sum_{i=1}^{N} \left\| \boldsymbol{x_i} - \sum_{j \in \mathcal{N}_K(\boldsymbol{x_i})} w_{ij} \boldsymbol{x_j} \right\|_2^2 ,$$

$$s.t. \sum_{j=1}^{K} w_{ij} = \sum_{j=1}^{N} w_{ij} = 1 \Rightarrow \boldsymbol{W1_N} = \boldsymbol{1_N}$$

Define K-NN matrix $\boldsymbol{N_i}$ and their weight matrix of $\boldsymbol{x_i}$:

$$\boldsymbol{N_i} := \{\boldsymbol{x_j}\}_{j \in \mathcal{N}_K(\boldsymbol{x_i})} \in \mathbb{R}^{D \times K}$$

and the optimal solution is [1]

$$\boldsymbol{\omega_i} = \{w_{ij}\}^T_{j \in \mathcal{N}_K(\boldsymbol{x_n})} \in \mathbb{R}^{K \times 1}$$

The optimal problem is

$$\boldsymbol{\omega_i} := \arg\min_{\boldsymbol{\omega}} \|\boldsymbol{x_i} - \boldsymbol{N_i \omega}\|_2^2, \ s.t. \ \boldsymbol{1_K^T \omega} = 1, i = 1, 2, \cdots, N$$

Assume that $\boldsymbol{x_i}$ and its K nearest vectors are linearly independent, let

$$\boldsymbol{C} := (\boldsymbol{N_i} - \boldsymbol{x_i 1_K^T})^T (\boldsymbol{N_i} - \boldsymbol{x_i 1_K^T})$$

and

$$\boldsymbol{\omega}^* = \frac{\boldsymbol{C^{-1} 1_K}}{\boldsymbol{1_K^T C^{-1} 1_K}} = \mathrm{norm}(\boldsymbol{C^{-1} 1_K})$$

where norm is the function of normalization through sum all the elements of the vector

$$\text{norm}(\boldsymbol{x}) := \frac{1}{\sum_i x_i}\boldsymbol{x}$$

Using the parameters in $\boldsymbol{\omega_i}$ to fill the element w_{ij} of $\boldsymbol{W}$, step 1 end.

In step 2 we use the low-dimension vector $\boldsymbol{z_i}$ to get the embedding of weight matrix. In order to avoid trival solution, the optimal target is

$$\begin{aligned}\boldsymbol{Z}^* &:= \arg\min_{\boldsymbol{Z}} \sum_{i=1}^{N} \left\| \boldsymbol{z_i} - \sum_{i=1}^{N} w_{ij}\boldsymbol{z_i} \right\|_2^2 = \arg\min_{\boldsymbol{Z}} \left\| \boldsymbol{Z} - \boldsymbol{ZW^T} \right\|_F^2 \\ &= \arg\min_{\boldsymbol{Z}} \left\| \boldsymbol{Z^T} - \boldsymbol{WZ^T} \right\|_F^2,\ s.t.\ \boldsymbol{ZZ^T} = \boldsymbol{I_D},\ \boldsymbol{Z1_N} = \boldsymbol{0_N}\end{aligned}$$

The shifting of $\boldsymbol{z_i}$ won't affect the optimal function [2], i.e. won't help to avoid trival solution, so the restriction

$$\boldsymbol{Z1_N} = \boldsymbol{0_N}$$

always be discarded.

Let

$$\boldsymbol{\Phi} := (\boldsymbol{I} - \boldsymbol{W})^T(\boldsymbol{I} - \boldsymbol{W})$$

and

$$\boldsymbol{Z}^* := \arg\min_{\boldsymbol{Z}} \text{tr}(\boldsymbol{Z\Phi Z^T}),\ s.t.\ \boldsymbol{ZZ^T} = \boldsymbol{I_D}$$

1.2 PCA

Assuming that data matrix $\boldsymbol{X} \in \mathbb{R}^{D\times N}$.

The perspective of projection try to minimize the reconstruction error of in dimension reduction by setting the optimal target as follows

$$\boldsymbol{X}^* := \arg\min_{\tilde{\boldsymbol{X}}} \mathcal{L}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \arg\min_{\tilde{\boldsymbol{X}}} \sum_{i=1}^{N} \|\boldsymbol{x_i} - \tilde{\boldsymbol{x}}_{\boldsymbol{i}}\|_2^2,\ s.t.\ \text{rank}(\tilde{\boldsymbol{X}}) \leqslant L$$

or

$$\boldsymbol{X}^{*} = \arg\min_{\tilde{\boldsymbol{X}} \in \boldsymbol{\Omega}} \left\| \boldsymbol{X} - \tilde{\boldsymbol{X}} \right\|_F^2, \ s.t. \ \dim(\boldsymbol{\Omega}) \leqslant L$$

Select an unit orthogonal basis for L dimensional principle space

$$\tilde{\boldsymbol{x}}_{\boldsymbol{n}} := \sum_{i=1}^{L} z_{in} \boldsymbol{b}_{\boldsymbol{i}} = \boldsymbol{B} \boldsymbol{z}_{\boldsymbol{n}}$$

and

$$z_{in} = \langle \tilde{\boldsymbol{x}}_{\boldsymbol{n}}, \boldsymbol{b}_{\boldsymbol{i}} \rangle = \tilde{\boldsymbol{x}}_{\boldsymbol{n}}^{\boldsymbol{T}} \boldsymbol{b}_{\boldsymbol{i}} = \boldsymbol{b}_{\boldsymbol{i}}^{\boldsymbol{T}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}$$

and

$$\boldsymbol{z}_{\boldsymbol{n}} = \boldsymbol{B}^{\boldsymbol{T}} \tilde{\boldsymbol{x}}_{\boldsymbol{n}}$$

Therefore

$$\tilde{\boldsymbol{X}} = \boldsymbol{B} \boldsymbol{Z}$$

and

$$\boldsymbol{Z} = \boldsymbol{B}^{\boldsymbol{T}} \tilde{\boldsymbol{X}}$$

PCA actually define an orthogonal projection from the primal space to principal space [3], i.e.

$$\tilde{\boldsymbol{X}} = \boldsymbol{B} \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{X}$$

and

$$\boldsymbol{Z} = \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{B} \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{X} = \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{X}$$

The optimal target can be transformed

$$\begin{aligned}
\boldsymbol{X}^* &= \arg\min_{\boldsymbol{B}} \sum_{i=1}^{N} \|\boldsymbol{x_i} - \boldsymbol{B}\boldsymbol{z_n}\|_2^2 = \arg\min_{\boldsymbol{B}} \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{Z}\|_F^2 \\
&= \arg\min_{\boldsymbol{B}} \operatorname{tr}\big((\boldsymbol{X} - \boldsymbol{B}\boldsymbol{Z})^T(\boldsymbol{X} - \boldsymbol{B}\boldsymbol{Z})\big) \\
&= \arg\min_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{Z^T}\boldsymbol{Z}) - 2\operatorname{tr}(\boldsymbol{Z^T}\boldsymbol{B^T}\boldsymbol{X}) = \arg\min_{\boldsymbol{B}} -\operatorname{tr}(\boldsymbol{Z^T}\boldsymbol{Z}) \\
&= \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{Z^T}\boldsymbol{Z}) = \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{Z}\boldsymbol{Z^T}) \\
&= \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{B^T}\boldsymbol{X}\boldsymbol{X^T}\boldsymbol{B}),\ s.t.\ \boldsymbol{B^T}\boldsymbol{B} = \boldsymbol{I_L}
\end{aligned}$$

Let

$$\boldsymbol{\Phi} := \sum_{j=L+1}^{D} \boldsymbol{x_j}\boldsymbol{x_j^T} = \boldsymbol{X}\boldsymbol{X^T} \in \boldsymbol{R}^{N\times N}$$

The optimal target is

$$\boldsymbol{B}^* = \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{B^T}\boldsymbol{\Phi}\boldsymbol{B}),\ s.t.\ \boldsymbol{B^T}\boldsymbol{B} = \boldsymbol{I_L}$$

or

$$\boldsymbol{B}^* = \arg\min_{\boldsymbol{B}} -\operatorname{tr}(\boldsymbol{B^T}\boldsymbol{\Phi}\boldsymbol{B}),\ s.t.\ \boldsymbol{B^T}\boldsymbol{B} = \boldsymbol{I_L}$$

therefore

$$\boldsymbol{X}^* = \boldsymbol{B}^*\boldsymbol{B}^{*\boldsymbol{T}}\boldsymbol{X}$$

To compress the data, we should use linear maps to transform the data from the primal space to principal space. Since the change of base will preserve the information of data, using the orthogonal base of principle space, we represent data as

$$\boldsymbol{X_p} := \boldsymbol{Z}^* = \boldsymbol{B}^{*\boldsymbol{T}}\boldsymbol{X}$$

The $\boldsymbol{B}$ can be thought of encoder and the $\boldsymbol{B^T}$ can be thought of decoder.

1.3 Kernel PCA

for kernel PCA, we apply PCA in kernel space

$$\boldsymbol{X}^* := \arg\min_{\tilde{\boldsymbol{X}}} \mathcal{L}(\phi(\boldsymbol{X}), \tilde{\boldsymbol{X}}) = \arg\min_{\tilde{\boldsymbol{X}}} \sum_{i=1}^{N} \|\phi(\boldsymbol{x_i}) - \tilde{\boldsymbol{x}}_{\boldsymbol{i}}\|_2^2,\ s.t.\ \operatorname{rank}(\tilde{\boldsymbol{X}}) \leqslant L$$

make the substitution and we derive

$$\boldsymbol{\Phi} := \sum_{j=1}^{N} \phi(\boldsymbol{x_j})\phi(\boldsymbol{x_j})^T = \phi(\boldsymbol{X})\phi(\boldsymbol{X})^T \in \boldsymbol{R}^{N\times N}$$

and

$$\boldsymbol{B}^* = \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{B^T \Phi B}),\ s.t.\ \boldsymbol{B^T B} = \boldsymbol{I_L}$$

Which we pay more attention to is the representation in low-dimensional space

$$\boldsymbol{X_p} := \boldsymbol{Z}^* = \boldsymbol{B^{*T} X}$$

1.4 EigenMap

Define the similarity function and matrix

$$a(\boldsymbol{x_n}, \boldsymbol{x_m}) = a(\boldsymbol{x_m}, \boldsymbol{x_n}) \in [0, 1]$$

and

$$\boldsymbol{A} = \{a(\boldsymbol{x_n}, \boldsymbol{x_m})\}_{n,m=1}^{D}$$

For eigenmap, the optimal target is

$$\begin{aligned}
\boldsymbol{Z}^* &:= \arg\min_{\boldsymbol{Z}} \sum_{n,m=1}^{D} \|\boldsymbol{z_n} - \boldsymbol{z_m}\|_2^2 a_{nm} \\
&= \arg\min_{\boldsymbol{Z}} \sum_{n,m=1}^{D} (\boldsymbol{z_n^T z_n} - 2\boldsymbol{z_n^T z_m} + \boldsymbol{z_m^T z_m}) a_{nm} \\
&= \arg\min_{\boldsymbol{Z}} \left(\sum_{n=1}^{D} \boldsymbol{z_n^T z_n} \sum_{m=1}^{D} a_{nm} + \sum_{m=1}^{D} \boldsymbol{z_m^T z_m} \sum_{n=1}^{D} a_{nm} - 2 \sum_{n,m=1}^{D} \boldsymbol{z_n^T z_m} a_{nm} \right) \\
&= \arg\min_{\boldsymbol{Z}} \left(2\mathrm{tr}(\boldsymbol{Z^T} \mathrm{diag}(\boldsymbol{A1_N})\boldsymbol{Z}) - 2\mathrm{tr}(\boldsymbol{Z^T AZ}) \right),\ s.t.\ \boldsymbol{Z^T Z} = \boldsymbol{I_N}
\end{aligned}$$

Let

$$\boldsymbol{\Phi} = \mathrm{diag}(\boldsymbol{A1_D}) - \boldsymbol{A}$$

and

$$\boldsymbol{Z}^* = \arg\min_{\boldsymbol{Z}} \mathrm{tr}(\boldsymbol{Z^T \Phi Z}),\ s.t.\ \boldsymbol{Z^T Z} = \boldsymbol{I_L}$$

1.5 ISOMap

Define the distance function and matrix

$$d(\boldsymbol{x_n}, \boldsymbol{x_m}) = d(\boldsymbol{x_m}, \boldsymbol{x_n}) \geqslant 0$$

and

$$\boldsymbol{D} = \{d(\boldsymbol{x_n}, \boldsymbol{x_m})\}_{n,m=1}^{N}$$

In order to find a low-dimensional representation in latent space whose Euclidean distance is close to the defined distance and get the best representation, the optimal target of Classic MDS is

$$\boldsymbol{Z}^* = \arg\min_{\boldsymbol{Z}} \frac{\sum_{n,m=1}^{N} (k_{nm} - \boldsymbol{z_n^T z_m})^2}{\sum_{n,m=1}^{N} k_{nm}^2},\ s.t.\ \mathrm{rank}(\boldsymbol{Z}) \leqslant L$$

where

$$\boldsymbol{K} = -\frac{1}{2}\boldsymbol{C}(\boldsymbol{D} \odot \boldsymbol{D})\boldsymbol{C}$$

and $\boldsymbol{C}$ is the centering matrix to set the sum of column or row to 0

$$\boldsymbol{C} = \boldsymbol{I_N} - \frac{1}{N}\boldsymbol{1_{N \times N}}$$

In ISOMAP, we always make assumption that $\boldsymbol{K}$ is positive semi-definite or force $\boldsymbol{K}$ to be positive semi-definite using the similar idea of ridge regression.

Define Gram matrix of $\boldsymbol{Z}$

$$\boldsymbol{G} := \{\boldsymbol{z_n^T z_m}\}_{n,m=1}^{N} = \boldsymbol{Z^T Z},\ \operatorname{rank}(\boldsymbol{G}) = \operatorname{rank}(\boldsymbol{Z})$$

Do eigenvalue decomposition to postive semi-definite positive $\boldsymbol{G}$, we get

$$\boldsymbol{G} = \boldsymbol{P\Lambda P^T}$$

and

$$\boldsymbol{Z} = \boldsymbol{\Lambda^{1/2} P^T}$$

and using the congruent transformation matrix, $\boldsymbol{K}$ can be similarly decompose to two matrixes. Due to the restriction of rank, the decomposition may not available for the solution. The optimal problem can be converted to PCA

$$\boldsymbol{G}^* = \arg\min_{\boldsymbol{G}} \sum_{n=1}^{N} \|\boldsymbol{k_n} - \boldsymbol{g_n}\|_2^2,\ s.t.\ \operatorname{rank}(\boldsymbol{G}) \leqslant L,\ \boldsymbol{G^T} = \boldsymbol{G},\ \boldsymbol{x^T G x} \geqslant 0$$

According to PCA, the $\boldsymbol{\Phi}$ is define as

$$\boldsymbol{\Phi} := \sum_{j=1}^{N} \boldsymbol{k_j k_j^T} = \boldsymbol{K K^T} \in \boldsymbol{R}^{N \times N}$$

and

$$\boldsymbol{B}^* = \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{B^T \Phi B}),\ s.t.\ \boldsymbol{B^T B} = \boldsymbol{I_N}$$

and

$$\boldsymbol{G}^* = \boldsymbol{B}_{\boldsymbol{L}}^* {\boldsymbol{B}_{\boldsymbol{L}}^*}^{\boldsymbol{T}} \boldsymbol{K}$$

The restriction of $\boldsymbol{G}$ can be satisfied. [4]

2 Solution

2.1 Major Problem

It's easy to conclude that all the $\boldsymbol{\Phi}$ we defined is symmetric and positive semi-definite [5].

Assume $\boldsymbol{Z}$ is a $D \times L$ matrix and suppose $D > L$.

The next step is how to solve the optimal problem as

$$\boldsymbol{Z}^* = \arg\min_{\boldsymbol{Z}} \operatorname{tr}(\boldsymbol{Z}^{\boldsymbol{T}} \boldsymbol{\Phi} \boldsymbol{Z}),\ s.t. \boldsymbol{Z}^{\boldsymbol{T}} \boldsymbol{Z} = \boldsymbol{I_D}$$

Apply eigenvalue decomposition to $\boldsymbol{\Phi}$ to simplify the problem

$$\boldsymbol{\Phi} = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^{\boldsymbol{T}},\ \boldsymbol{V}\boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{I_D}$$

Sort the eigenvalue to the satisfies following equation

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_D \geqslant 0$$

And rearrage the corresponding eigenvector

$$\boldsymbol{\Phi} \boldsymbol{v_i} = \lambda_i \boldsymbol{v_i},\ i = 1, 2, \cdots, D$$

Let

$$\boldsymbol{Y} = \boldsymbol{V}^{\boldsymbol{T}} \boldsymbol{Z},\ \boldsymbol{Y}^{\boldsymbol{T}} \boldsymbol{Y} = \boldsymbol{Z}^{\boldsymbol{T}} \boldsymbol{V} \boldsymbol{V}^{\boldsymbol{T}} \boldsymbol{Z} = \boldsymbol{I_D}$$

and

$$\boldsymbol{Z} = \boldsymbol{V} \boldsymbol{Y}$$

Using the the bijection, we can simplify the optimal problem

$$\boldsymbol{Y}^* = \arg\min_{\boldsymbol{Y}} \operatorname{tr}(\boldsymbol{Y}^{\boldsymbol{T}} \boldsymbol{\Lambda} \boldsymbol{Y}) = \arg\min_{\boldsymbol{Y}} \sum_{n=1}^{L} \boldsymbol{y}_{\boldsymbol{n}}^{\boldsymbol{T}} \boldsymbol{\Lambda} \boldsymbol{y_n},\ s.t. \boldsymbol{Y}^{\boldsymbol{T}} \boldsymbol{Y} = \boldsymbol{I_D}$$

and

$$\boldsymbol{Z^*} = \boldsymbol{VY^*}$$

The optimal problem may difficult to directly solve for the complex constrains for us, so we should make full use of its excellent algebra property. Let

$$\boldsymbol{Y^T} = (\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \cdots, \boldsymbol{\gamma_D}),\ \boldsymbol{\gamma_i'} \in \mathbb{R}^{L\times 1}$$

and

$$\begin{aligned}
\boldsymbol{Y^*} &= \arg\min_{\boldsymbol{Y}} \operatorname{tr}(\boldsymbol{Y^T\Lambda Y}) \\
&= \arg\min_{\boldsymbol{Y}} \operatorname{tr}\left(\sum_{n=1}^{D} \lambda_n \boldsymbol{\gamma_n}\boldsymbol{\gamma_n^T}\right) \\
&= \arg\min_{\boldsymbol{Y}} \sum_{n=1}^{D} \lambda_n \operatorname{tr}(\boldsymbol{\gamma_n}\boldsymbol{\gamma_n^T})
\end{aligned}$$

Notice

$$\sum_{n=1}^{D} \operatorname{tr}(\boldsymbol{\gamma_n}\boldsymbol{\gamma_n^T}) = \operatorname{tr}\left(\sum_{n=1}^{D} \boldsymbol{\gamma_n}\boldsymbol{\gamma_n^T}\right) = \operatorname{tr}(\boldsymbol{YY^T}) = \operatorname{tr}(\boldsymbol{Y^TY}) = L$$

and do Gram-Schmidt Orthogonalization to the column vectors of $\boldsymbol{Y}$

$$(\boldsymbol{Y}, \boldsymbol{Y'}) \in \mathbb{R}^{D\times D},\ (\boldsymbol{Y}, \boldsymbol{Y'})^T(\boldsymbol{Y}, \boldsymbol{Y'}) = (\boldsymbol{Y}, \boldsymbol{Y'})(\boldsymbol{Y}, \boldsymbol{Y'})^T = \boldsymbol{I_D}$$

and

$$\boldsymbol{Y'^T} = (\boldsymbol{\gamma_1'}, \boldsymbol{\gamma_2'}, \cdots, \boldsymbol{\gamma_D'}),\ \boldsymbol{\gamma_i'} \in \mathbb{R}^{(D-L)\times 1}$$

Therefore

$$\begin{aligned}
&(\boldsymbol{Y}, \boldsymbol{Y'})(\boldsymbol{Y}, \boldsymbol{Y'})^T = \boldsymbol{I_D} \\
&\Rightarrow (\boldsymbol{\gamma_i^T}, \boldsymbol{\gamma_i'^T})(\boldsymbol{\gamma_i^T}, \boldsymbol{\gamma_i'^T})^T = \boldsymbol{\gamma_i^T\gamma_i} + \boldsymbol{\gamma'^T_i\gamma_i'} = 1 \\
&\Rightarrow 0 \leqslant \boldsymbol{\gamma_i^T\gamma_i} = \operatorname{tr}(\boldsymbol{\gamma_i^T\gamma_i}) = \operatorname{tr}(\boldsymbol{\gamma_i\gamma_i^T}) \leqslant 1,\ i = 1, 2, \cdots, D
\end{aligned}$$

Let

$$a_n = \operatorname{tr}(\boldsymbol{\gamma_n}\boldsymbol{\gamma_n^T})$$

It's simple to guess that

$$\sum_{n=1}^{D} \lambda_n a_n \geqslant \sum_{n=D-L+1}^{D} \lambda_n,\ s.t.\ \sum_{n=1}^{D} = a_n = L,\ 0 \leqslant a_i \leqslant 1,\ i = 1, 2, \cdots, D$$

If not, there must exists a smaller combination

$$\exists s \geqslant D - L + 1, t < D - L + 1,\ a_s < 1, a_t < 1$$

The solution is not the smallest solution for moving the weight and we can get

$$\sum_{n \neq s,t}^{D} \lambda_n a_n + \lambda_s(a_s + \varepsilon) + \lambda_t(a_t - \varepsilon) < \sum_{n=1}^{D} \lambda_n a_n$$

where

$$0 \leqslant a_s + \varepsilon \leqslant 1, 0 \leqslant a_t - \varepsilon \leqslant 1$$

We finally conclude that

$$\min_{\boldsymbol{Z}} \operatorname{tr}(\boldsymbol{Z^T \Phi Z}) = \min_{\boldsymbol{Y}} \operatorname{tr}(\boldsymbol{Y^T \Lambda Y}) = \min_{\boldsymbol{Y}} \sum_{n=1}^{D} \lambda_n \operatorname{tr}(\boldsymbol{\gamma_n \gamma_n^T}) \geqslant \sum_{n=D-L+1}^{D} \lambda_n$$

Select eigenvectors corrsponding to L th smallest eigenvalues to construct $\boldsymbol{Z}$

$$\boldsymbol{Z} = (\boldsymbol{v_{D-L+1}}, \boldsymbol{v_{D-L+2}}, \cdots, \boldsymbol{v_D})$$

and

$$\operatorname{tr}(\boldsymbol{Z^T \Phi Z}) = \sum_{n=1}^{L} \boldsymbol{z_n^T \Phi z_n} = \sum_{n=1}^{D} \lambda_{D-L+n} \boldsymbol{z_n^T z_n} = \sum_{n=D-L+1}^{D} \lambda_n$$

So the answer is

$$\boldsymbol{Z^*} = (\boldsymbol{v_{D-L+1}}, \boldsymbol{v_{D-L+2}}, \cdots, \boldsymbol{v_D}) = \boldsymbol{V_{-L}}$$

and

$$\min_{\boldsymbol{Z}} \operatorname{tr}(\boldsymbol{Z^T \Phi Z}) = \sum_{n=D-L+1}^{D} \lambda_n$$

If we change the minimizing problem into maximizing problem, through similarly derivation, the solution is

$$\boldsymbol{Z}^* = (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_L) = \boldsymbol{V}_L$$

and

$$\max_{\boldsymbol{Z}} \operatorname{tr}(\boldsymbol{Z}^T \boldsymbol{\Phi} \boldsymbol{Z}) = \sum_{n=1}^{L} \lambda_n$$

The solution is interesting for PCA, kernel PCA, and ISOMap when combined with SVD.

2.2 PCA

Apply SVD to $\boldsymbol{X}$

$$\boldsymbol{X} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^T$$

The optimal solution and representation is

$$\boldsymbol{X}^* = \boldsymbol{U}_L \boldsymbol{U}_L^T \boldsymbol{X} = \boldsymbol{U}_L \boldsymbol{\Sigma}_L \boldsymbol{V}_L^T$$

and

$$\boldsymbol{X}_p = \boldsymbol{U}_L^T \boldsymbol{X} = \boldsymbol{\Sigma}_L \boldsymbol{V}_L^T$$

We can also get the optimal reconstruct error

$$\min_{\tilde{\boldsymbol{X}}} \mathcal{L}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \sum_{n=1}^{L} \sigma_n$$

It can be proved that the solution is also the optimal one in the following optimal problem [6]

$$\boldsymbol{X}^* = \arg\min_{\tilde{\boldsymbol{X}} \in \boldsymbol{\Omega}} \left\| \boldsymbol{X} - \tilde{\boldsymbol{X}} \right\|_2, \ s.t. \ \dim(\boldsymbol{\Omega}) \leqslant L$$

where the matrix norm $\| . \|_2$ is spectrual norm defined as

$$\|A\|_2 := \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\| \boldsymbol{x} \|_2}$$

2.3 Kernel PCA

Apply SVD to $\phi(\boldsymbol{X})$ and we get

$$\phi(\boldsymbol{X}) = \boldsymbol{U\Sigma V^T}$$

and the representation

$$\boldsymbol{X_p} = \boldsymbol{U_L^T}\phi(\boldsymbol{X})$$

In fact, the specific form of ϕ is always unknown so we define kernel function and its matrix

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x_j})^T\phi(\boldsymbol{x_j})$$

and

$$\boldsymbol{K} = \{\kappa(\boldsymbol{x_n}, \boldsymbol{x_m})\}_{n,m=1}^N = \phi(\boldsymbol{X})^T\phi(\boldsymbol{X})$$

It will be shown in the derivation of the optimal problem

$$\boldsymbol{B}^* = \arg\max_{\boldsymbol{B}} \operatorname{tr}(\boldsymbol{B^T\Phi B}),\ s.t.\ \boldsymbol{B^TB} = \boldsymbol{I_L}$$

The optimal solution satisfies

$$\boldsymbol{\Phi b_i} = \phi(\boldsymbol{X})\phi(\boldsymbol{X})^T\boldsymbol{b_i} = \lambda_i\boldsymbol{b_i}$$

Notice

$$\boldsymbol{b_i} = \frac{1}{\lambda_i}\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T\boldsymbol{b_i}$$

Let

$$\boldsymbol{c_i} = \frac{1}{\lambda_i}\phi(\boldsymbol{X})^T\boldsymbol{b_i} \Rightarrow \boldsymbol{C} = \phi(\boldsymbol{X})^T\boldsymbol{B\Lambda^{-1}}$$

and

$$\boldsymbol{b_i} = \phi(\boldsymbol{X})\boldsymbol{c_i} \Rightarrow \boldsymbol{B} = \phi(\boldsymbol{X})\boldsymbol{C}$$

Therefore

$$\boldsymbol{\Phi b_i} = \phi(\boldsymbol{X})\boldsymbol{Kc_i} = \phi(\boldsymbol{X})\lambda_i\boldsymbol{c_i} \Leftarrow \boldsymbol{Kc_i} = \lambda_i\boldsymbol{c_i}$$

The equation must have solution for $\boldsymbol{K}$ and $\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T$ have the same non-zero eigenvalue [7]. Consider the relationship between SVD and eigenvalue decomposition, $\boldsymbol{K}$, we get

$$\boldsymbol{K} = \boldsymbol{V\Lambda V^T}$$

So we can solve the problem by derive the eigenvector of $\boldsymbol{K}$ to get $\boldsymbol{b_i}$, the optimal representation is

$$\boldsymbol{X_p} = \boldsymbol{B^{*T}}\phi(\boldsymbol{X}) = \boldsymbol{C^{*T}K} = \boldsymbol{V_L^T K} = \boldsymbol{\Lambda_L V_L^T}$$

although there is no explicit solution to $\boldsymbol{X^*}$

2.4 ISOMap

Apply eigenvalue decompostion to $\boldsymbol{K}$

$$\boldsymbol{K} = \boldsymbol{V\Lambda V^T}$$

and the solution is

$$\boldsymbol{G^*} = \boldsymbol{V_L V_L^T K} = \boldsymbol{V_L \Lambda_L V_L^T}$$

Consider the relationship between SVD and eigenvalue decomposition, we get

$$\boldsymbol{Z^*} = \boldsymbol{\Lambda_L^{1/2} V_L^T}$$

3 Appendix

3.1 [1]

According to the constraints

$$\begin{aligned}\boldsymbol{\omega_i} &= \arg\min_{\boldsymbol{\omega}} \left\|\boldsymbol{x_i} \mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}} \boldsymbol{\omega} - \boldsymbol{N_i}\boldsymbol{\omega}\right\|_2^2 \\ &= \arg\min_{\boldsymbol{\omega}} \left\|(\boldsymbol{x_i}\mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}} - \boldsymbol{N_i})\boldsymbol{\omega}\right\|_2^2,\ s.t.\ \mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}}\boldsymbol{\omega} = 1\end{aligned}$$

Let

$$\boldsymbol{Y} := \boldsymbol{x_i}\mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}} - \boldsymbol{N_i}$$

and

$$\boldsymbol{\omega_i} = \arg\min_{\boldsymbol{\omega}} \|\boldsymbol{Y}\boldsymbol{\omega}\|_2^2,\ s.t.\ \mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}}\boldsymbol{\omega} = 1$$

Construct Lagrangian

$$\mathcal{L}(\boldsymbol{\omega}, \lambda) := \frac{1}{2}\|\boldsymbol{Y}\boldsymbol{\omega}\|_2^2 + \lambda(1 - \mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}}\boldsymbol{\omega})$$

Take derivative

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\omega}} = \boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y}\boldsymbol{\omega} - \lambda\mathbf{1}_{\boldsymbol{K}}$$

and

$$\frac{\partial\mathcal{L}}{\partial\lambda} = 1 - \mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}}\boldsymbol{\omega} = 0$$

Assume that $\boldsymbol{x_i}$ and its K nearest vectors are linearly independent

$$\text{rank}(\boldsymbol{N_i}) = \text{rank}(\boldsymbol{N_i^T}\boldsymbol{N_i}) = K$$

The equation about rank is easy to prove. The column vector of $\boldsymbol{Y}$ are linearly independent, if not, we can get

$$\exists \lambda_i \neq 0,\ i = 1, 2, \cdots, K,\ \sum_{j=1}^{K} \lambda_j (\boldsymbol{x_i} - \boldsymbol{n_j}) = 0$$

and

$$\boldsymbol{x_i} \sum_{j=1}^{K} \lambda_j - \sum_{j=1}^{K} \lambda_j \boldsymbol{n_j} = 0$$

It shows that $\boldsymbol{x_i}$ and its K nearest vectors will not be linearly independent which leads to a contradiction. Therefore

$$\operatorname{rank}(\boldsymbol{Y}) = \operatorname{rank}(\boldsymbol{Y}^T \boldsymbol{Y}) = K$$

If unluckily $\boldsymbol{x_i}$ and its K nearest vectors not fit our assumption, we can always force our matrix to be reversible using the idea of ridge regression. For the primal problem, the optimal solution is

$$\boldsymbol{\omega^*} = (\boldsymbol{N_i^T N_i})^{-1} (\lambda \boldsymbol{1_K} + \boldsymbol{N_i^T x_i})$$

and for our transformed problem

$$\boldsymbol{\omega^*} = \lambda (\boldsymbol{Y^T Y})^{-1} \boldsymbol{1_K}$$

Let

$$\boldsymbol{C} := \boldsymbol{Y^T Y} = (\boldsymbol{N}_i - \boldsymbol{x}_i \boldsymbol{1_K^T})^T (\boldsymbol{N}_i - \boldsymbol{x}_i \boldsymbol{1_K^T})$$

Using the constraints

$$\boldsymbol{1_K^T \omega^*} = \lambda \boldsymbol{1_K^T C^{-1} 1_K} = 1$$

And

$$\boldsymbol{\omega^*} = \frac{\boldsymbol{C^{-1} 1_K}}{\boldsymbol{1_K^T C^{-1} 1_K}} = \operatorname{norm}(\boldsymbol{C^{-1} 1_K})$$

where norm is the function of normalization through sum all the elements of the vector

$$\operatorname{norm}(\boldsymbol{x}) := \frac{1}{\sum_i x_i} \boldsymbol{x}$$

3.2 [2]

Notice

$$\sum_{j=1}^{N} w_{ij} = 1$$

and

$$\begin{aligned}\forall \boldsymbol{b} \in \mathbb{R}^{D\times 1},\ \min_{\boldsymbol{Z}} \sum_{i=1}^{N} \left\| (\boldsymbol{z_i} + \boldsymbol{b}) - \sum_{i=1}^{N} w_{ij}(\boldsymbol{z_i} + \boldsymbol{b}) \right\|_2^2 \\ = \min_{\boldsymbol{Z}} \sum_{i=1}^{N} \left\| \boldsymbol{z_i} - \sum_{i=1}^{N} w_{ij}\boldsymbol{z_i} \right\|_2^2\end{aligned}$$

3.3 [3]

select another unit orthogonal basis for the orthogonal complement

$$\boldsymbol{x_n} := \sum_{i=1}^{L} \zeta_{in}\boldsymbol{b_i} + \sum_{j=L+1}^{D} \zeta_{jn}\boldsymbol{b_j}$$

and the two component of $\boldsymbol{x_n}$ can be defined as

$$\boldsymbol{x_n^{(1)}} := \sum_{i=1}^{L} \zeta_{in}\boldsymbol{b_i}$$

and

$$\boldsymbol{x_n^{(2)}} := \sum_{j=L+1}^{D} \zeta_{jn}\boldsymbol{b_j}$$

Notice

$$z_{in} = \langle \boldsymbol{\tilde{x}_n}, \boldsymbol{b_i} \rangle = \boldsymbol{\tilde{x}_n^T}\boldsymbol{b_i} = \boldsymbol{b_i^T}\boldsymbol{\tilde{x}_n}$$

and

$$\zeta_{in} = \langle \boldsymbol{x_n}, \boldsymbol{b_i} \rangle = \boldsymbol{x_n^T b_i} = \boldsymbol{b_i^T x_n}$$

Therefore

$$\frac{\partial \mathcal{L}}{\partial z_{in}} = \frac{\partial \mathcal{L}}{\partial \tilde{\boldsymbol{x}}_n}\frac{\partial \tilde{\boldsymbol{x}}_n}{\partial z_{in}} = 2(\tilde{\boldsymbol{x}}_{\boldsymbol{n}} - \boldsymbol{x_n})^T \boldsymbol{b_i} = 2(z_{in} - \zeta_{in})$$

Let the value of derivative function to zero and the optimal solution for dimension reduction satisfies

$$\begin{aligned}
\tilde{\boldsymbol{x}}_{\boldsymbol{n}} &= \sum_{i=1}^{L} z_{in}\boldsymbol{b_i} = \sum_{i=1}^{L} \zeta_{in}\boldsymbol{b_i} \\
&= \boldsymbol{x}_{\boldsymbol{n}}^{(\boldsymbol{1})} = \pi_{\text{span}(\boldsymbol{b_1},\boldsymbol{b_2},\cdots,\boldsymbol{b}_L)}(\boldsymbol{x_n}) \\
&= \left(\sum_{i=1}^{L} \boldsymbol{b_i} \boldsymbol{b}_{\boldsymbol{i}}^{T}\right)\boldsymbol{x_n} = \boldsymbol{BB^T x_n}
\end{aligned}$$

and

$$\tilde{\boldsymbol{X}} = \boldsymbol{BB^T X}$$

and optimal target is

$$\begin{aligned}
\min_{\boldsymbol{B}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{2})} \right\|_2^2 &= \min_{\boldsymbol{B}} \sum_{i=1}^{N} \sum_{j=L+1}^{D} \zeta_{ji}^2 = \min_{\boldsymbol{B}} \sum_{i=1}^{N} \sum_{j=L+1}^{D} \boldsymbol{b_j^T} \boldsymbol{x}_i \boldsymbol{x_i^T b_j} \\
&= \min_{\boldsymbol{B}} \sum_{j=L+1}^{D} \boldsymbol{b_j^T} \left(\sum_{i=1}^{N} \boldsymbol{x}_j \boldsymbol{x_j^T} \right) \boldsymbol{b_j}, \ s.t. \ \boldsymbol{B^T B} = \boldsymbol{I_{D-L}}
\end{aligned}$$

corresponding to the perspective of maximize variance.

3.4 [4]

Notice

$$\boldsymbol{G}^* = \boldsymbol{B_L^* B_L^*}^{\boldsymbol{T}} \boldsymbol{K} \boldsymbol{B_L^*}^{\boldsymbol{T}} \boldsymbol{B_L^*}$$

It's easy to conclude that

$$\boldsymbol{G^*}^{\boldsymbol{T}} = \boldsymbol{G^*}$$

and

$$\boldsymbol{x^T G x} = (\boldsymbol{B_L^*}^{\boldsymbol{T}} \boldsymbol{B_L^* x})^{\boldsymbol{T}} \boldsymbol{K} (\boldsymbol{B_L^*}^{\boldsymbol{T}} \boldsymbol{B_L^* x}) \geqslant 0$$

for $\boldsymbol{K}$ is positive semi-definite.

3.5 [5]

It's easy to conclude that all the $\boldsymbol{\Phi}$ we defined is symmetric and positive semi-definite, mostly for the reason

$$\forall \boldsymbol{M} \in \mathbb{R}^{n\times m}, f(\boldsymbol{x}) = \boldsymbol{x^T M M^T x} = \left\| \boldsymbol{M^T x} \right\|_2^2 \geqslant 0$$

For LLE

$$\boldsymbol{\Phi} := (\boldsymbol{I} - \boldsymbol{W})^T (\boldsymbol{I} - \boldsymbol{W}) For constraints$$

$$\boldsymbol{W 1_N} = \boldsymbol{1_N}$$

We can get

$$(\boldsymbol{I} - \boldsymbol{W})^T (\boldsymbol{I} - \boldsymbol{W}) \boldsymbol{1_N} = (\boldsymbol{I} - \boldsymbol{W})^T (\boldsymbol{1_N} - \boldsymbol{W 1_N}) = \boldsymbol{0_N} = 0 \boldsymbol{1_N}$$

$\boldsymbol{\Phi}$ has a eigenvalue 0 and its corresponding eigenvector. In fact, the eigenvalue and corresponding eigenvector are always discarded for it is trivial for the problem.

For PCA

$$\boldsymbol{\Phi} := \sum_{j=L+1}^{D} \boldsymbol{x_j x_j^T} \in \boldsymbol{R}^{N\times N}$$

For Kernel PCA

$$\boldsymbol{\Phi} := \sum_{j=L+1}^{D} \phi(\boldsymbol{x_j}) \phi(\boldsymbol{x_j})^T \in \boldsymbol{R}^{N\times N}$$

For ISOMap

$$\boldsymbol{\Phi} := \sum_{j=L+1}^{D} \boldsymbol{k}_j \boldsymbol{k}_j^{\boldsymbol{T}} \in \boldsymbol{R}^{N\times N}$$

For Eigenmap

We just need to proove that for Eigenmap

$$\boldsymbol{\Phi} = \operatorname{diag}(\boldsymbol{A}\mathbf{1}_{\boldsymbol{N}}) - \boldsymbol{A}$$

is semi-definite.

Notice

$$\begin{aligned} f(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\Phi}\boldsymbol{x} &= \sum_{i,j=1}^{n} a_{ij}x_i^2 - \sum_{i,j=1}^{n} a_{ij}x_ix_j \\ &= \sum_{i,j=1}^{n} a_{ij}\left(\frac{x_i^2+x_j^2}{2} - x_ix_j\right) \geqslant 0 \end{aligned}$$

It's easy to prove that $\boldsymbol{\Phi}$ has a eigenvalue 0 and its corresponding eigenvector

$$\boldsymbol{\Phi}\mathbf{1}_{\boldsymbol{N}} = (\operatorname{diag}(\boldsymbol{A}\mathbf{1}_{\boldsymbol{N}}) - \boldsymbol{A})\mathbf{1}_{\boldsymbol{N}} = \mathbf{0}_{\boldsymbol{N}} = 0\mathbf{1}_{\boldsymbol{N}}$$

In fact, the eigenvalue and corresponding eigenvector are always discarded for it is trivial for the problem.

3.6 [6]

Let

$$\hat{\boldsymbol{A}}(\boldsymbol{L}) := \sum_{n=1}^{L} \sigma_i \boldsymbol{u_i}\boldsymbol{v}_{\boldsymbol{i}}^{\boldsymbol{T}} = \boldsymbol{U}_{\boldsymbol{L}}\boldsymbol{\Sigma}_{\boldsymbol{L}}\boldsymbol{V}_{\boldsymbol{L}}^{\boldsymbol{T}}$$

and

$$\boldsymbol{\Sigma}'_{\boldsymbol{L}} := \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{L}} & \\ & \boldsymbol{O} \end{pmatrix}$$

For $\hat{\boldsymbol{A}}(\boldsymbol{L})$ can be decomposited as

$$\hat{\boldsymbol{A}}(\boldsymbol{L}) = \boldsymbol{U}_L \boldsymbol{\Sigma}_L \boldsymbol{V}_L^T = \boldsymbol{U} \boldsymbol{\Sigma}_L' \boldsymbol{V}^T$$

which is the expression of SVD of $\hat{\boldsymbol{A}}(\boldsymbol{L})$. so

$$\operatorname{rank}(\hat{\boldsymbol{A}}(\boldsymbol{L})) = \operatorname{rank}(\boldsymbol{\Sigma}_L') = L$$

First we should derive the value of spectual norm of a arbitrary matrix

$$\|A\|_2 := \max_{\boldsymbol{x}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$$

The optimal problem can be transformed to our familiar form

$$\max_{\boldsymbol{x}} \boldsymbol{x}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \max_{\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}),\ s.t.\ \boldsymbol{x}^T\boldsymbol{x} = 1$$

We have already solve the problem, so the solution is

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}) = \sigma_1$$

Therefore, consider the SVD of $\hat{\boldsymbol{A}}(\boldsymbol{L})$, we can get

$$\left\|\boldsymbol{A} - \hat{\boldsymbol{A}}(\boldsymbol{L})\right\|_2 = \sigma_{L+1}$$

Suppose

$$\exists \boldsymbol{B},\ \operatorname{rank}(\boldsymbol{B}) \leqslant L,\ \|\boldsymbol{A} - \boldsymbol{B}\|_2 < \left\|\boldsymbol{A} - \hat{\boldsymbol{A}}(\boldsymbol{L})\right\|_2$$

which implies

$$\forall \boldsymbol{x},\ \|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{A} - \boldsymbol{B}\|_2 \|\boldsymbol{x}\|_2 < \sigma_{L+1} \|\boldsymbol{x}\|_2$$

Think of the rank equation

$$\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{Col}(\boldsymbol{A})) = n - \dim(\operatorname{Null}(\boldsymbol{A}))$$

We know that

$$\dim(\operatorname{Null}(\boldsymbol{B})) = n - \operatorname{rank}(\boldsymbol{B}) \geqslant N - L$$

and

$$\forall \boldsymbol{x} \in \text{Null}(\boldsymbol{B}),\ \boldsymbol{x} \neq \boldsymbol{0},\ \boldsymbol{Bx} = \boldsymbol{0}$$

We obtain

$$\|\boldsymbol{Ax}\|_2 = \|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2 \leqslant \|(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x}\|_2 < \sigma_{L+1} \parallel \boldsymbol{x} \parallel_2$$

Think of the vector in the space spanned by orthogonal $L+1$ right sigular eigenvectors corresponding to the biggest $L+1$ singular value

$$\forall \boldsymbol{x} \in \text{span}(\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_{L+1}}),\ \boldsymbol{x} = \sum_{n=1}^{L+1} z_n \boldsymbol{v_n},\ \parallel \boldsymbol{x} \parallel_2^2 = \sum_{n=1}^{L+1} z_n^2$$

It will satisfies

$$\|\boldsymbol{Ax}\|_2^2 = \boldsymbol{x^T}(\boldsymbol{A^T}\boldsymbol{Ax}) = \sum_{n=1}^{L+1} \sigma_n^2 z_n^2 \geqslant \sigma_{L+1}^2 \parallel \boldsymbol{x} \parallel_2^2 \Rightarrow \|\boldsymbol{Ax}\|_2 \geqslant \sigma_{L+1} \parallel \boldsymbol{x} \parallel_2$$

Denote

$$V_1 = \text{Null}(\boldsymbol{B}),\ V_2 = \text{span}(\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_{L+1}})$$

and

$$V_1 \cap V_2 = \{\boldsymbol{0}\},\ \dim(V_1) \geqslant N - L,\ \dim(V_2) = L + 1$$

Thick of dimension equation

$$\dim(V_1) + \dim(V_2) - \dim(V_1 \cap V_2) = \dim(V_1 + V_2) \geqslant N + 1$$

It will lead to contradictory for $V_1, V_2 \subset \mathbb{R}^N$.

3.7 [7]

For $\lambda \geqslant 0$, we should verify that

$$\exists k \neq 0, f_{\boldsymbol{AA}^T}(\lambda) := \left|\boldsymbol{AA}^T - \lambda \boldsymbol{I}\right| = k f_{\boldsymbol{A}^T\boldsymbol{A}}(\lambda) := k\left|\boldsymbol{A}^T\boldsymbol{A} - \lambda \boldsymbol{I}\right|$$

Think of block matrices, assume $\boldsymbol{A} \in \mathbb{R}^{m \times n}$

$$\begin{pmatrix} \sigma \boldsymbol{I}_n & \boldsymbol{A}^T \\ \boldsymbol{A} & \sigma \boldsymbol{I}_m \end{pmatrix} \rightarrow \begin{pmatrix} \sigma \boldsymbol{I}_n & \boldsymbol{A}^T \\ \boldsymbol{O} & \sigma \boldsymbol{I}_m - \frac{1}{\sigma} \boldsymbol{A}\boldsymbol{A}^T \end{pmatrix}$$

和

$$\begin{pmatrix} \sigma \boldsymbol{I}_n & \boldsymbol{A}^T \\ \boldsymbol{A} & \sigma \boldsymbol{I}_m \end{pmatrix} \rightarrow \begin{pmatrix} \sigma \boldsymbol{I}_n - \frac{1}{\sigma} \boldsymbol{A}^T\boldsymbol{A} & \boldsymbol{O} \\ \boldsymbol{A} & \sigma \boldsymbol{I}_m \end{pmatrix}$$

where $\sigma \neq 0$.

For elementary transformation of partitioned matrices will not change the value of determinant, so

$$\sigma^n \left| \sigma \boldsymbol{I}_m - \frac{1}{\sigma} \boldsymbol{A}\boldsymbol{A}^T \right| = \sigma^m \left| \sigma \boldsymbol{I}_n - \frac{1}{\sigma} \boldsymbol{A}^T\boldsymbol{A} \right|$$

let $\lambda = \sigma^2$, finally we derive the result

$$(-1)^m \sigma^{n-m} \left| \boldsymbol{A}\boldsymbol{A}^T - \sigma^2 \boldsymbol{I} \right| = (-1)^n \sigma^{m-n} \left| \boldsymbol{A}\boldsymbol{A}^T - \sigma^2 \boldsymbol{I} \right|$$