Some proof on Lecture 6-8

Lin Ning　linning51400@ruc.eud.cn

2022/4/23

> **Reference**
>
> **Source 1** Mathematics for Machine Learning by Marc Peter Deisenroth A Aldo Faisal Cheng Soon Ong, pp. 131-132, 326-331
>
> **Source 2** 机器学习，周志华，pp. 228-233
>
> **Source 3** www.stat.yale.edu/~lc436/papers/JCGS-mds.pdf

Transformation

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

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

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

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

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

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

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

and

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

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

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

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

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

The shifting of $\boldsymbol{z_i}$ won't affect the optimal function [2], i.e. won't help to avoid trivial solution, so the zero-mean constraint always be discarded.

Let

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

and

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

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

$$\begin{aligned} \boldsymbol{X}^* &:= \arg\min_{\tilde{\boldsymbol{X}}} \mathcal{L}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \arg\min_{\tilde{\boldsymbol{X}}} \sum_{i=1}^{N} \|\boldsymbol{x_i} - \tilde{\boldsymbol{x}}_{\boldsymbol{i}}\|_2^2 \\ &= \arg\min_{\tilde{\boldsymbol{X}}} \left\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\right\|_F^2, \ s.t. \ \ \mathrm{rank}(\tilde{\boldsymbol{X}}) \leqslant L \end{aligned}$$

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

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

and

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

and

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

Therefore

$$\tilde{\boldsymbol{X}} = \boldsymbol{BZ}$$

and

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

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

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

and

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

The optimal target can be transformed

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

Let

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

The optimal target is

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

or

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

therefore

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

To compress the data, we should use reversible 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

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

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

Kernel PCA

for kernel PCA, we apply PCA in kernel space

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

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

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

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

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

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

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

Let

$$
\boldsymbol{\Phi} = \mathrm{diag}(\boldsymbol{A 1_D}) - \boldsymbol{A}
$$

and

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

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 [4]

$$
\begin{aligned}
\boldsymbol{Z}^* &= \arg\min_{\boldsymbol{Z}} \mathrm{Strain}_d(\boldsymbol{Z}) = \arg\min_{\boldsymbol{Z}} \sum_{n,m=1}^{N} (k_{nm} - \boldsymbol{z_n^T z_m})^2 \\
&= \arg\min_{\boldsymbol{Z}} \left\| \boldsymbol{K} - \boldsymbol{Z^T Z} \right\|_F^2,\ s.t.\ \mathrm{rank}(\boldsymbol{Z}) \leqslant L
\end{aligned}
$$

where

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

and

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

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 $\boldsymbol{Z}$

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

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

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

and

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

The optimal problem can be converted to PCA

$$\boldsymbol{G^*} = \arg\min_{\boldsymbol{G}} \|\boldsymbol{K} - \boldsymbol{G}\|_F^2,\ s.t.\ \operatorname{rank}(\boldsymbol{G}) \leqslant L,\ \boldsymbol{G^T} = \boldsymbol{G},\ \boldsymbol{x^T G x} \geqslant 0$$

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

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

The restriction of $\boldsymbol{G}$ can be satisfied. [5]

Solution

Major Problem

It's easy to conclude that all the $\boldsymbol{\Phi}$ we defined is symmetric and positive semi-definite [6].

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

The next step is how to solve the optimal problem as

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

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

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

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

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

Let

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

and

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

Using the the bijection, we can simplify the optimal problem

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

and

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

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

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

Let

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

After analyze the constraints on a_n [7], we can prove that [8]

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

We finally conclude that

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

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

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

So the answer is

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

and

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

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

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

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

PCA

Apply SVD to $\boldsymbol{X}$

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

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 [8]

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

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

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

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

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

We can derive the following result in a winding way [8], in fact, we can take full use of SVD. 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 using matrix $\boldsymbol{K}$

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

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

ISOMap

Apply eigenvalue decompostion to $\boldsymbol{K}$

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

and the solution is

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

Consider the relationship between SVD and eigenvalue decomposition, we get

$$\boldsymbol{Z}^* = \boldsymbol{\Lambda}_{\boldsymbol{L}}^{\boldsymbol{1/2}}\boldsymbol{V}_{\boldsymbol{L}}^{\boldsymbol{T}}$$

Appendix

[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

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

and

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

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

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

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

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

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

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

and for our transformed problem

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

Let

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

Using the constraints

$$\mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}}\boldsymbol{\omega}^{*}=\lambda\mathbf{1}_{\boldsymbol{K}}^{\boldsymbol{T}}\boldsymbol{C}^{-1}\mathbf{1}_{\boldsymbol{K}}=1$$

And

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

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

$$\mathrm{norm}(\boldsymbol{x}):=\frac{1}{\sum_{i}x_{i}}\boldsymbol{x}$$

[2]

Notice

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

and

$$\begin{aligned}\forall\boldsymbol{b}\in\mathbb{R}^{D\times1},\ \min_{\boldsymbol{Z}}\sum_{i=1}^{N}\left\|(\boldsymbol{z_i}+\boldsymbol{b})-\sum_{i=1}^{N}w_{ij}(\boldsymbol{z_i}+\boldsymbol{b})\right\|_2^2\\=\min_{\boldsymbol{Z}}\sum_{i=1}^{N}\left\|\boldsymbol{z_i}-\sum_{i=1}^{N}w_{ij}\boldsymbol{z_i}\right\|_2^2\end{aligned}$$

[3]

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

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

and

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

Therefore

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

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

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

and

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

and optimal target is

$$\min_{\boldsymbol{B}} \sum_{i=1}^{N} \left\| \boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{2})} \right\|_2^2 = \min_{\boldsymbol{B}} \sum_{i=1}^{N} \sum_{j=L+1}^{D} \zeta_{ji}^2 = \min_{\boldsymbol{B}} \sum_{i=1}^{N} \sum_{j=L+1}^{D} \boldsymbol{b}_{\boldsymbol{j}}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{b}_{\boldsymbol{j}}$$

$$= \min_{\boldsymbol{B}} \sum_{j=L+1}^{D} \boldsymbol{b}_{\boldsymbol{j}}^{\boldsymbol{T}} \left(\sum_{i=1}^{N} \boldsymbol{x}_{\boldsymbol{j}} \boldsymbol{x}_{\boldsymbol{j}}^{\boldsymbol{T}} \right) \boldsymbol{b}_{\boldsymbol{j}}, \ s.t. \ \boldsymbol{B}^{\boldsymbol{T}} \boldsymbol{B} = \boldsymbol{I}_{\boldsymbol{D-L}}$$

corresponding to the perspective of maximize variance.

[4]

The low-dimensional representation in latent space whose Euclidean distance is close to the defined distance

$$d_{ij}^2 \approx \left\| \boldsymbol{z}_{\boldsymbol{i}} - \boldsymbol{z}_{\boldsymbol{j}} \right\|_2^2 = \left\| \boldsymbol{z}_{\boldsymbol{i}} \right\|_2^2 + \left\| \boldsymbol{z}_{\boldsymbol{j}} \right\|_2^2 - 2 \left\langle \boldsymbol{z}_{\boldsymbol{i}}, \boldsymbol{z}_{\boldsymbol{j}} \right\rangle$$

Define Gram matrix of $\boldsymbol{Z}$

$$\boldsymbol{G} := \{ \boldsymbol{z}_{\boldsymbol{n}}^{\boldsymbol{T}} \boldsymbol{z}_{\boldsymbol{m}} \}_{n,m=1}^{N} = \boldsymbol{Z}^{\boldsymbol{T}} \boldsymbol{Z}$$

and we can obtain

$$g_{ij} \approx -\frac{1}{2} (d_{ij}^2 - \left\| \boldsymbol{z}_{\boldsymbol{i}} \right\|_2^2 - \left\| \boldsymbol{z}_{\boldsymbol{j}} \right\|_2^2) = -\frac{1}{2} (d_{ij}^2 - g_{ii} - g_{jj})$$

To simplify the problem, in classical MDS, we often assume that $\boldsymbol{Z}$ has zero mean, i.e.

$$\boldsymbol{Z} \boldsymbol{1}_{\boldsymbol{N}} = \boldsymbol{0}_{\boldsymbol{N}}$$

And we also constraint the rows and columns of distance matrix to zero mean, define the margin min

$$\bar{d}_{i\cdot}^2 = \frac{1}{N} \sum_{j=1}^{N} d_{ij}^2 = \left\| \boldsymbol{z}_{\boldsymbol{i}} \right\|_2^2 + \frac{1}{N} \sum_{j=1}^{N} \left\| \boldsymbol{z}_{\boldsymbol{j}} \right\|_2^2 - \frac{2}{N} \left\langle \boldsymbol{z}_{\boldsymbol{i}}, \boldsymbol{Z} \boldsymbol{1}_{\boldsymbol{N}} \right\rangle$$

$$= \left\| \boldsymbol{z}_{\boldsymbol{i}} \right\|_2^2 + \frac{1}{N} \sum_{j=1}^{N} \left\| \boldsymbol{z}_{\boldsymbol{j}} \right\|_2^2 = \left\| \boldsymbol{z}_{\boldsymbol{i}} \right\|_2^2 + \frac{1}{N} \mathrm{tr}(\boldsymbol{G})$$

and

$$\begin{aligned}\bar{d}_{\cdot j}^2 &= \frac{1}{N}\sum_{i=1}^{N} d_{ij}^2 = \|\boldsymbol{z_j}\|_2^2 + \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{z_i}\|_2^2 - \frac{2}{N}\langle \boldsymbol{Z}\mathbf{1_N}, \boldsymbol{z_j}\rangle \\ &= \|\boldsymbol{z_j}\|_2^2 + \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{z_i}\|_2^2 = \|\boldsymbol{z_j}\|_2^2 + \frac{1}{N}\mathrm{tr}(\boldsymbol{G})\end{aligned}$$

and

$$\bar{d}_{\cdot\cdot}^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}^2 = \frac{1}{N}\sum_{i=1}^{N}\bar{d}_{i\cdot}^2 = \frac{2}{N}\mathrm{tr}(\boldsymbol{G})$$

so

$$\begin{aligned}g_{nn} &= \|\boldsymbol{z_n}\|_2^2 \\ &= \bar{d}_{n\cdot}^2 - \frac{1}{N}\mathrm{tr}(\boldsymbol{G}) = \bar{d}_{n\cdot}^2 - \frac{1}{2}\bar{d}_{\cdot\cdot}^2 \\ &= \bar{d}_{\cdot n}^2 - \frac{1}{N}\mathrm{tr}(\boldsymbol{G}) = \bar{d}_{\cdot n}^2 - \frac{1}{2}\bar{d}_{\cdot\cdot}^2\end{aligned}$$

and

$$\begin{aligned}g_{ij} &= -\frac{1}{2}(d_{ij}^2 - g_{ii} - g_{jj}) \\ &= -\frac{1}{2}(d_{ij}^2 - \bar{d}_{i\cdot}^2 - \bar{d}_{\cdot j}^2 + \bar{d}_{\cdot\cdot}^2)\end{aligned}$$

Using centering matrix $\boldsymbol{C}$ to remove the margin mean of our defined distance, i.e. to set the sum of column or row to 0, then set the coefficient

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

and $\boldsymbol{C}$ is centering matrix

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

and we final get

$$g_{ij} = \boldsymbol{z_i^T z_j} \approx k_{ij}$$

Therefore, we construct the strain loss to force g_{ij} close to k_{ij}

$$\mathrm{Strain}_d(\boldsymbol{Z}) = \left(\frac{\sum_{n,m=1}^{N} (k_{nm} - \boldsymbol{z_n^T z_m})^2}{\sum_{n,m=1}^{N} k_{nm}^2} \right)^{1/2}$$

[5]

Notice

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

It's easy to conclude that

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

and

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

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

[6]

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

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

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

We can get

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

$\boldsymbol{\Phi}$ 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

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

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

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

For Eigenmap
We just need to proove that for Eigenmap

$$\boldsymbol{\Phi} = \operatorname{diag}(\boldsymbol{A1_N}) - \boldsymbol{A}$$

is semi-definite.

Notice

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

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

$$\boldsymbol{\Phi 1_N} = (\operatorname{diag}(\boldsymbol{A1_N}) - \boldsymbol{A}) \boldsymbol{1_N} = \boldsymbol{0_N} = 0 \boldsymbol{1_N}$$

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

[7]

Notice

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

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

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

So

$$\sum_{n=1}^{D}a_n=L,\ 0\leqslant a_n\leqslant 1,\ n=1,2,\cdots,D$$

[8]

If not, there must exists a smaller combination

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

and

$$\exists t<D-L+1,\ a_t>0$$

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

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

[9]

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

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

The equation must have solution for $\boldsymbol{K}$ and $\phi(\boldsymbol{X})\phi(\boldsymbol{X})^T$ have the same non-zero eigenvalue, because for $\lambda\geqslant 0$, we should verify that

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

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

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

和

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

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