Introduction to ML

Anastasia Makarova and Mohammad Reza Karimi

ETH Zürich

22.04.2020

Tutorial plan

- Questions for Homework 3 with Anastasia
- Dimensionality Reduction via PCA with Mohammad

Note 1: Please open EduApp on your mobile phone or web browser (link in the chat)
Note 2: Please bring paper and pen for the second part

HW3 Q4: Valid kernel. Problem

Problem 2 (Kernels):

Use the basic rules for kernel decomposition discussed in class or otherwise and assuming that $k(x, y)$ kernel, letting $f : \mathbb{R} \to \mathbb{R}$ in a) and b), $g : \mathcal{X} \to \mathbb{R}_+$ for d), $f : \mathcal{X} \to \mathbb{R}$ for e) and f), and $\phi : \mathcal{X} \to \mathcal{X}'$.

4. Mark the following statements as True or False. Try to justify your answers to yourself.

   (a) $k_a(x, y) = f(k(x, y))$ is a valid kernel, if f is a polynomial with non-negative coefficients.

   (b) $k_b(x, y) = f(k(x, y))$ is a valid kernel, if f is any polynomial.

   (c) $k_c(x, y) = \exp(k(x, y))$ is a valid kernel.

   (d) $k_d(x, y) = g(x)k(x, y)g(y)$ is a valid kernel.

   (e) $k_e(x, y) = f(x)k(x, y)f(y)$ is a valid kernel.

   (f) $k_f(x, y) = k(\phi(x), \phi(y))$ is a valid kernel.

HW3 Q4: Valid kernel. Recap

Definition 1: Kernel

A kernel k is a two-argument real-valued function over $\mathcal{X} \times \mathcal{X}$, such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$k(x, y) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}}$$

for some inner-product space $\mathcal{F}$ such that $\forall \mathbf{x} \in \mathcal{X} \;\; \phi(\mathbf{x}) \in \mathcal{F}$.

How to show that k is a *valid* kernel?

- It is sufficient to present a mapping ϕ,
- Or show that $\mathbf{K}$ is PSD for finite subset of $\mathcal{X}$
  $\mathbf{K}$ denotes kernel matrix of corresponding kernel k.

HW3 Q4: Valid kernel. Recap

Definition 2: Kernel

A kernel k is a two-argument real-valued function over $\mathcal{X} \times \mathcal{X}$, such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$k(x, y) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}}$$

for some inner-product space $\mathcal{F}$ such that $\forall \mathbf{x} \in \mathcal{X} \;\; \phi(\mathbf{x}) \in \mathcal{F}$.

How to show that k is a *valid* kernel?

- It is sufficient to present a mapping ϕ,
- Or show that $\mathbf{K}$ is PSD for finite subset of $\mathcal{X}$
  $\mathbf{K}$ denotes kernel matrix of corresponding kernel k.

HW3 Q4: Valid kernel. Recap

Definition 3: Kernel

A kernel k is a two-argument real-valued function over $\mathcal{X} \times \mathcal{X}$, such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$:

$$k(x, y) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}}$$

for some inner-product space $\mathcal{F}$ such that $\forall \mathbf{x} \in \mathcal{X} \;\; \phi(\mathbf{x}) \in \mathcal{F}$.

How to show that k is a *valid* kernel?

- It is sufficient to present a mapping ϕ,
- Or show that $\mathbf{K}$ is PSD for finite subset of $\mathcal{X}$
  $\mathbf{K}$ denotes kernel matrix of corresponding kernel k.

$$K = \begin{bmatrix} k(x_1, x_1) & & \\ & \ddots & \\ & & \end{bmatrix}$$

$x_i \in \mathcal{X}$

K is PSD <=> eigenvalues are non negative

HW3 Q4: Valid kernel. Basics

1. $k(x, y) = k_1(x, y) + k_2(x, y)$
   $K = K_1 + K_2 \rightarrow$
2. $k(x, y) = \alpha k_1(x, y),\ \alpha \geq 0 \rightarrow$
3. $k(x, y) = k_1(x, y)k_2(x, y)$
4. $k_4(x, y) = f(x)f(y), f : \mathcal{X} \rightarrow R$
   By defining $\phi : x \rightarrow f(x)$ we show
   $\exists \phi : \mathcal{X} \rightarrow \mathcal{F} : k(x, y) = \langle \phi(x), \phi(y) \rangle$

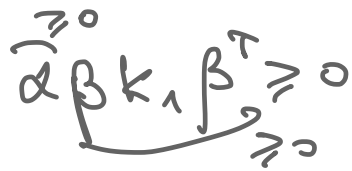
HW3 Q4: Valid kernel. Basics

1. $k(x, y) = k_1(x, y) + k_2(x, y)$
   $K = K_1 + K_2 \rightarrow$
   $\beta \in \mathbb{R}^n, \; K \in \mathbb{R}^{n \times n}$
   $\beta K \beta^T \geq 0$? why ?
   $\beta K \beta^T = \beta K_1 \beta^T + \beta K_2 \beta^T \geq 0$
2. $k(x, y) = \alpha k_1(x, y), \; \alpha \geq 0 \rightarrow$
3. $k(x, y) = k_1(x, y) k_2(x, y)$
4. $k_4(x, y) = f(x) f(y), f : \mathcal{X} \rightarrow R$
   By defining $\phi : x \rightarrow f(x)$ we show
   $\exists \phi : \mathcal{X} \rightarrow \mathcal{F} : k(x, y) = \langle \phi(x), \phi(y) \rangle$

HW3 Q4: Valid kernel. Basics

1. $k(x, y) = k_1(x, y) + k_2(x, y)$
   $K = K_1 + K_2 \rightarrow$

2. $k(x, y) = \alpha k_1(x, y),\ \alpha \geq 0 \rightarrow$



3. $k(x, y) = k_1(x, y)k_2(x, y)$

4. $k_4(x, y) = f(x)f(y), f : \mathcal{X} \rightarrow R$
   By defining $\phi : x \rightarrow f(x)$ we show
   $\exists \phi : \mathcal{X} \rightarrow \mathcal{F} : k(x, y) = \langle \phi(x), \phi(y) \rangle$

HW3 Q4: Valid kernel. Basics

1. $k(x, y) = k_1(x, y) + k_2(x, y)$
   $K = K_1 + K_2 \rightarrow$

2. $k(x, y) = \alpha k_1(x, y),\ \alpha \geq 0 \rightarrow$

3. $k(x, y) = k_1(x, y)k_2(x, y)$

eigenvalue decompositions

$K_1 = \sum_{i=1}^{n} \lambda_i v_i v_i^T$; $K_2 = \sum_{i=1}^{n} \beta_i u_i u_j^T$

$K = K_1 \circ K_2 = \sum_{i,j}^{n} \lambda_i \beta_j (v_i \circ u_j)(v_i \circ v_j)$ rank-1 matrices ≥ 0

4. $k_4(x, y) = f(x)f(y), f : \mathcal{X} \rightarrow R$
   By defining $\phi : x \rightarrow f(x)$ we show
   $\exists \phi : \mathcal{X} \rightarrow \mathcal{F} : k(x, y) = \langle \phi(x), \phi(y) \rangle$

HW3 Q4: Valid kernel. Basics

1. $k(x, y) = k_1(x, y) + k_2(x, y)$
   $K = K_1 + K_2 \rightarrow$
2. $k(x, y) = \alpha k_1(x, y),\ \alpha \geq 0 \rightarrow$
3. $k(x, y) = k_1(x, y)k_2(x, y)$
4. $k_4(x, y) = f(x)f(y), f : \mathcal{X} \rightarrow R$
   By defining $\phi : x \rightarrow f(x)$ we show
   $\exists \phi : \mathcal{X} \rightarrow \mathcal{F} : k(x, y) = \langle \phi(x), \phi(y) \rangle$

HW3 Q4: Valid kernel. Basics

1. $k(x, y) = k_1(x, y) + k_2(x, y)$
   $K = K_1 + K_2 \rightarrow$

2. $k(x, y) = \alpha k_1(x, y),\ \alpha \geq 0 \rightarrow$

3. $k(x, y) = k_1(x, y)k_2(x, y)$

4. $k_4(x, y) = f(x)f(y), f : \mathcal{X} \rightarrow R$
   By defining $\phi : x \rightarrow f(x)$ we show
   $\exists \phi : \mathcal{X} \rightarrow \mathcal{F} : k(x, y) = \langle \phi(x), \phi(y) \rangle$

HW3 Q4: Valid kernel. Solution

a. $k_a(x, y) = f(k(x, y))$ polynomial with non-negative coefficients

follows from (1), (2) and (3)

b. $k_b(x, y) = f(k(x, y))$ polynomial. Counter example

c. $k_b(x, y) = \exp(k(x, y))$

HW3 Q4: Valid kernel. Solution

a. $k_a(x, y) = f(k(x, y))$ polynomial with non-negative coefficients

b. $k_b(x, y) = f(k(x, y))$ polynomial. Counter example

$f(x) = -x$

$k(x, y) = x^T y$ $\longrightarrow$ $k_b(x, y) = -x^T y$ that will have negative eig-envalues

negation of a valid kernel

c. $k_b(x, y) = \exp(k(x, y))$

HW3 Q4: Valid kernel. Solution

a. $k_a(x, y) = f(k(x, y))$ polynomial with non-negative coefficients

b. $k_b(x, y) = f(k(x, y))$ polynomial. Counter example

c. $k_b(x, y) = \exp(k(x, y))$

$$e^x = 1 + x + \frac{x^2}{2!} + \dots$$

HW3 Q4: Valid kernel. Solution

$g: \mathcal{X} \to \mathbb{R}_+$
$f: \mathcal{X} \to \mathbb{R}$

→ 1. let's prove for a general case of f

2. $f(x)f(y)$ is a kernel (4)

3. product of kernel (3)

d. $k_d(x, y) = g(x)k(x, y)g(y)$

e. $k_e(x, y) = f(x)k(x, y)f(y) = k_4(x, y)k(x, y)$

f. $k_f = k(\phi(x), \phi(y))$ We want to show that $\mathbf{K}_f$ is PSD. As $k(.)$ is a kernel, when applied to any set of vectors $\{\phi(x_i)\}_{i=1}^N$ will result if PSD matrix.

HW3 Q4: Valid kernel. Solution

d. $k_d(x, y) = g(x)k(x, y)g(y)$

e. $k_e(x, y) = f(x)k(x, y)f(y) = k_4(x, y)k(x, y)$

f. $k_f = k(\phi(x), \phi(y))$ We want to show that $\mathbf{K}_f$ is PSD. As $k(.)$ is a kernel, when applied to any set of vectors $\{\phi(x_i)\}_{i=1}^N$ will result if PSD matrix.

HW3 Q4: Valid kernel. Solution

d. $k_d(x, y) = g(x)k(x, y)g(y)$

e. $k_e(x, y) = f(x)k(x, y)f(y) = k_4(x, y)k(x, y)$

f. $k_f = k(\phi(x), \phi(y))$ We want to show that $\mathbf{K}_f$ is PSD.
As $k(.)$ is a kernel, when applied to any set of vectors $\{\phi(x_i)\}_{i=1}^N$ will result if PSD matrix.

HW3 Q4: Valid kernel. Solution

d. $k_d(x, y) = g(x)k(x, y)g(y)$

e. $k_e(x, y) = f(x)k(x, y)f(y) = k_4(x, y)k(x, y)$

f. $k_f = k(\phi(x), \phi(y))$ We want to show that $\mathbf{K}_f$ is PSD. As $k(.)$ is a kernel, when applied to any set of vectors $\{\phi(x_i)\}_{i=1}^N$ will result if PSD matrix.

HW3 Q4: Valid kernel. Eduapp question

Please open eduapp

HW3 Q4: Valid kernel. Eduapp question

Which of the following functions are kernels?

- ✗ $k_1(x, y) = -\frac{<x,y>}{||x||_2||y||_2}$ negation of a valid kernel
- ✗ $k_2(x, y) = \sqrt{(||x - y||_2^2 + 1)}$ counter example unit vectors $x=(1,0)$ $y=(0,1)$

$K_2 = \begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & 1 \end{pmatrix} \rightarrow$

$\det \begin{pmatrix} 1-\lambda & \sqrt{3} \\ \sqrt{3} & 1-\lambda \end{pmatrix} = 0 \rightarrow (1-\lambda)^2 - \sqrt{3}^2 = 0$

$(1-\lambda-\sqrt{3})(1-\lambda+\sqrt{3}) = 0 \rightarrow$ K_2 has negative eigenvalue

- Valid $k_3(x, y) = \prod_{i=1}^{D} \underbrace{\exp(\frac{x_i^2 - c}{2})}_{f(x_i)} \underbrace{\exp(\frac{y_i^2 - c}{2})}_{f(y_i)}$

can be rewritten as $k_3(x,y) = f(x)f(y)$

$\rightarrow$ according to (4) is valid

HW4 Q3: SVM update rule via SGD

The exercise suggests to train an SVM, where we penalise the margin violation given by $(1 - y\mathbf{w}^T\mathbf{x})_+ = \max(1 - y\mathbf{w}^T\mathbf{x}, 0)$, not linearly but with the square root instead. Correspondingly, our modified SVM seeks to optimise the following objective

$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(1 - y\mathbf{w}^T\mathbf{x})_+} + \lambda||\mathbf{w}||_2$$

Dimensionality Reduction via PCA

1) Review of Some Linear algebra

Spectral Decomp $X \in \mathbb{R}^{n \times n}$ symmetric $\quad X = V \Lambda V^T = \sum \lambda_i v_i v_i^T$

V is orthogonal (eigenvectors)

if all $\lambda_i \geq 0 \longrightarrow$ PSD Matrix $\longrightarrow$ Cholskey Dccomp.

$$X = V \Lambda^{1/2} \Lambda^{1/2} V^T = (V \Lambda^{1/2})(V \Lambda^{1/2})^T = Q Q^T .$$

Trace For square matrices $X \in \mathbb{R}^{n \times n}$,

$$tr(X) = \sum_{i=1}^{n} X_{ii} = \text{sum of diagonal entries.}$$

* Properties of Trace (Exercise):

- $tr(AB) = tr(BA) = \sum_{i,j=1}^{n} A_{ij} B_{ij}$.
- $tr(ABC) = tr(BCA) = tr(CAB)$.
- $tr(PAP^{-1}) = tr(A)$.

$\Rightarrow$ trace of a "linear transformation" is well difined as it does not depend on the matrix repr. of the linear operation.

Frobenius norm of a matrix A is

$$\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = tr(AA^T) .$$

Also, if A's columns are $a_1, \ldots, a_n$,

$$\|A\|_F^2 = \|a_1\|_2^2 + \cdots + \|a_n\|_2^2$$

An Optimization Problem

We are interested in $\max_{A \in \mathbb{R}^{n\times d},\ A^TA=I} \operatorname{tr}(A^TXA)$ ($X \in \mathbb{R}^{n\times n}$ symmetric).

We claim that the optimal A is indeed the set of eigenvectors of X corresponding to the d largest eigenvalues.

Case $d=1$: $\max_{a\in\mathbb{R}^n,\ \|a\|=1} a^TXa$

$X = V\Lambda V^T$

$\Rightarrow \quad a^TXa = a^TV\Lambda V^Ta = (V^Ta)^T\Lambda(V^Ta)$

$= u^T\Lambda u$

and V ortho $\Rightarrow \|u\| = \|a\|$ + V one-to-one

$\Rightarrow$ OPT $= \max_{\|u\|=1} u^T\Lambda u$.

But $u^T\Lambda u = \sum_{i=1}^n u_i^2\lambda_i$, $\|u\|^2 = \sum_{i=1}^n u_i^2 = 1$

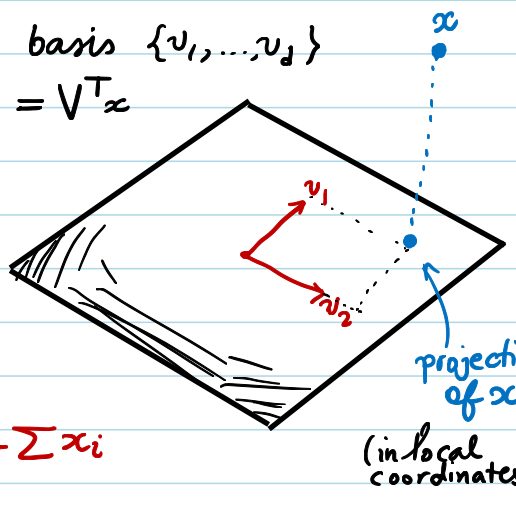
max is when $u_1 = 1$, $u_i = 0$ $(i>1)$. ▨

2) Variance point-of-view to PCA

Goal: find orthonormal basis $v_1, \dots, v_d$ such that the projections of $x_1, \dots, x_n$ on span of $\{v_1, \dots, v_d\}$ has the most variance.

(Variance above should be read "sample variance"...)

The projection of x_i, expressed in the basis $\{v_1, \dots, v_d\}$ has coordinates $(v_1^Tx_i, \dots, v_d^Tx_i) = V^Tx$



Hence, the variance of the Projection is

$$\sum_{i=1}^n \left\| V^Tx_i - V^T\bar{x} \right\|^2$$

$V = [v_1 | \dots | v_d]$

mean $= \frac{1}{n}\sum x_i$

The goal of PCA is then to find

$$\max_{V^TV=I} \sum_{i=1}^{n} \|V^Tx_i - V^T\bar{x}\|^2 .$$

Now some algebraic tricks:

$$\|V^Tx_i - V^T\bar{x}\|^2 = \|V^T(x_i - \bar{x})\|^2$$

$$= (x_i - \bar{x})^T V V^T (x_i - \bar{x})$$

$$= tr\{(x_i - \bar{x})^T V V^T (x_i - \bar{x})\}$$

$$= tr\{V^T(x_i - \bar{x})(x_i - \bar{x})^T V\}$$

$$\Rightarrow \sum_{i=1}^{n} \|V^Tx_i - V^T\bar{x}\|^2$$

$$= \sum tr\{V^T(x_i - \bar{x})(x_i - \bar{x})^T V\}$$

$$= tr\{\sum V^T(x_i - \bar{x})(x_i - \bar{x})^T V\}$$

$$= tr\{V^T[\Sigma(x_i - \bar{x})(x_i - \bar{x})^T]V\}$$

$$= (n-1)\, tr\, V^T \Sigma_n V$$

↖ the covariance matrix!

... and we know the answer!

3) PCA in High dimensions

Assume that the data points $x_1, \dots, x_n$ are drawn i.i.d from some distribution.

What PCA computes $\xrightarrow[\text{on}]{\text{depends}}$ Spectrum of Σ_n, the sample covariance matrix

What we "Hopefully" want $\xrightarrow[\text{on}]{\text{depends}}$ Spectrum of Σ, the covariance matrix of the distribution

if the data is "really" having a low-dimensional structure, does PCA see it?

Can PCA trick us to "see" something which is not true?

MAIN QUESTION

→ requires that spectra of Σ_n is "close" to Σ ...

ANSWER: maybe! yes! So, be careful!

BABY example: Gaussian data

Suppose that $x_1, \dots, x_n \sim N(0, \Sigma)$ multivariate in p dimensions

Construct $\Sigma_n = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$

law of large numbers says for fixed p, $\Sigma_n \to \Sigma$ as $n \to \infty$ ...

In practice, p often is of order of n ... cannot use LLN!

Assume that $\Sigma = I$.

[Show the notebook ...]

Lots of eigenvalues > 1 !

PCA with small d gets fooled ... but the original distribution is rotation-invariant (isotropic)!

BABY example 2 : Spike model

The data distribution really has low-dimensional structure. Does PCA find it ?

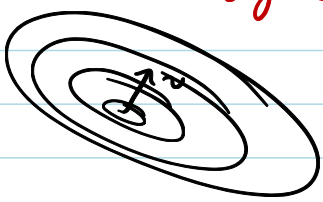
$$\Sigma = I + \beta v v^T \qquad \beta > 0, \ \|v\| = 1.$$

if $x \sim \mathcal{N}(0, \Sigma)$, $\quad x = \sqrt{\beta}\, g_0\, v + \varepsilon$ → noise

signal

$$\varepsilon \sim \mathcal{N}(0, I), \ g_0 \sim \mathcal{N}(0,1)$$

β : How powerful is the signal (SNR).

Quiz :  or 

Reformulated Question: How large β should be so that PCA "sees" the direction v ?

Intuition : Something should be visible in spectra of Σ_n ...

[show notebook ...]

BBP Transition Suppose $\frac{p}{n} = \gamma$ and let $n, p \to \infty$.

then if $\beta > \sqrt{\gamma}$, $\lambda_{max}(\Sigma_n)$ gets out of MP curve
if $\beta \leq \sqrt{\gamma}$, $\lambda_{max}(\Sigma_n)$ stays inside ...

Beautiful theory, but little time ...

Kernel k-means & Spectral Clustering

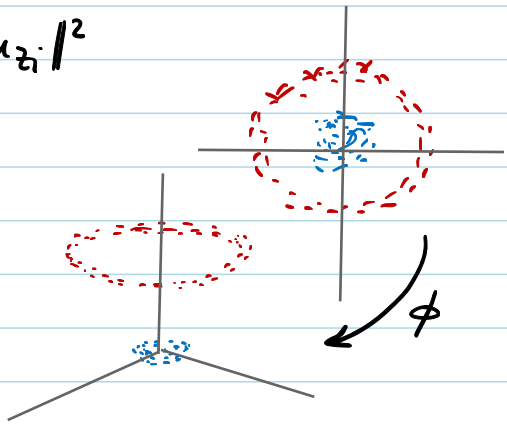
Setup a set of points $x_1, \dots, x_n$
k-means wish to find clusters with means $\{\mu_1, \dots, \mu_k\} = \mu$

$$\text{s.t.} \quad C(z, \mu) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2 \quad \text{is minimized.}$$

(z_i: the cluster assigned to point x_i)

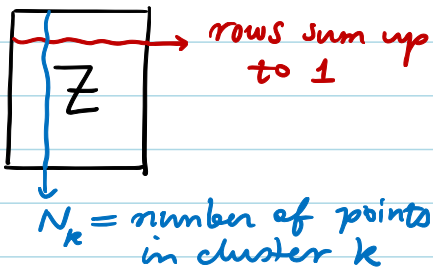
Now, assume that before clustering, we first apply a feature map ϕ to our data, and then perform clustering.

$$\text{New obj: } C(z, \mu) = \sum_{i=1}^{n} \|\phi(x_i) - \mu_{z_i}\|^2$$



To write the objective nicer, we define a "cluster assignment matrix" $Z_{n \times K}$ as

$$Z_{ik} = \begin{cases} 1 & \text{if } z_i = k \\ 0 & \text{otherwise} \end{cases}$$



Define $L = \text{diag}\left(\frac{1}{N_1}, \dots, \frac{1}{N_K}\right)$.

and $\Phi = \left[\phi(x_1) \mid \dots \mid \phi(x_n)\right]$.

Now check that $M = \Phi Z L Z^T$ has n columns, each column is the (copy of) cluster center associated to that point via Z.

$$\Rightarrow \text{K-means cost } C = \|\Phi - M\|_F^2$$

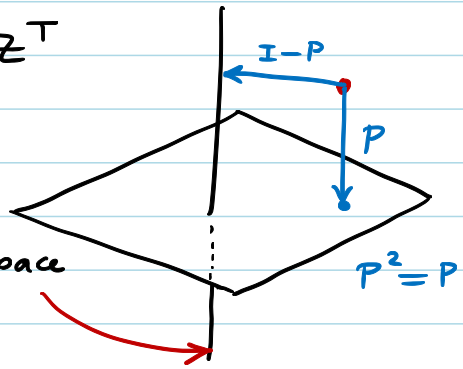
$$= \text{tr}\left\{(\Phi - M)(\Phi - M)^T\right\}.$$

Some facts: • $Z^T Z = L^{-1}$

• $(ZLZ^T)^2 = ZLZ^T$

$\Rightarrow$ ZLZ^T is a projection

$\Rightarrow$ $I - ZLZ^T$ is projection on the complement space



$\Phi - M = \Phi(I - ZLZ^T)$

Now we simplify the objective:

$$C = tr\{(\Phi - M)(\Phi - M)^T\}$$

$$= tr\{\Phi(I - ZLZ^T)^2 \Phi^T\}$$

$$= tr\{\Phi(I - ZLZ^T)\Phi^T\}$$

$$= tr\{\Phi\Phi^T\} - tr\{\Phi ZLZ^T\Phi^T\}$$

$$= \underbrace{tr\{K\}}_{\text{does not depend on clustering } Z} - tr\{L^{1/2} Z^T K Z L^{1/2}\} \quad (K = \Phi^T\Phi)$$

$\Rightarrow$ new objective: $\max tr\{L^{1/2} Z^T K Z L^{1/2}\}$

s.t. Z is a binary clustering matrix.

"Binary" makes it hard $\longrightarrow$ relax!

$Z^T Z = L^{-1} \Rightarrow L^{1/2} Z^T \underbrace{Z L^{1/2}}_{H} = I \Rightarrow H^T H = I$

relaxed problem $\max tr\{H^T K H\}$

s.t. $H^T H = I$

... our friend in PCA!

Hence, optimal H is obtaind via taking top k eigenvectors of K.

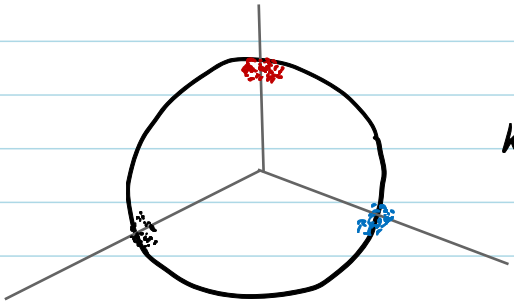
Last part: How to find the clusters? (Z)

<u>proposal</u>: H approximates Z ....
each row of Z summed up to 1 ....
$\Rightarrow$ for each row of H, set max to 1 and the rest to 0 (rounding)

<u>Better</u>: Normalize each row of H by its norm.
$\longrightarrow$ rows of H are now points on the unit sphere.

Run k_means and extract clusters!

dream :



$k = 3$

Called 'Spectral Clustering'.

There are other approaches as well, but we don't discuss them here.