COMP/ENGN8535 Homework 3
Solutions

Problem 1

a) Mean vector: $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ so $x_j - \mu = \hat{x}_j = x_j$

Centered data matrix: $X = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$

Covariance matrix: $C = \frac{1}{4} X X^T$

$$= \frac{1}{4} \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

Eigenvalues of C: $\frac{1}{2}$ & $\frac{1}{2}$ (diagonal elements)

Eigenvectors of C: $\begin{bmatrix}1\\0\end{bmatrix}$, $\begin{bmatrix}0\\1\end{bmatrix}$

The principle components are $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ with both of these being the "first" principle component because they describe the same amount of variance in the data (ie they both share the same e-value).

b) Kernel $K(x_i, x_j) = (x_i^T x_j)^{10}$

Gram matrix with $K_{ij} = (x_i^T x_j)^{10}$:

$$K = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Mean-subtracted Gram matrix:

$\tilde{K} = K - 1_n K - K 1_n + 1_n K 1_n$ where $n = 4$ so $1_n = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Now, solve the eigen problem

$$\tilde{K} a_i = \lambda_i n a_i \quad (1)$$

for vectors a_i and scalars λ_i with $n = 4$.

Two possible approaches,

Approach 1) Note that eigenvalues d_i and eigenvectors v_i of $\tilde{K}$ by definition satisfy the equation

↓ Compute λ_i & a_i by scaling e-values & e-vectors of $\tilde{K}$.

$$\tilde{K} v_i = d_i v_i \quad (2)$$

By inspection of (1) and (2) we see that $v_i = a_i$ (up to scale) and $d_i = \lambda_i n = 4\lambda_i$.

Eigenvalues d_i of $\tilde{K}$ are $0, 0, 0, 2$ so $\lambda_i = 0, 0, 0, \frac{1}{2}$

We need $y_i \in \mathbb{R}$ so finding eigenvector v_i of $\tilde{K}$ corresponding to largest $\lambda_i = \frac{1}{2}$ (or $d_i = 2$) is

$$a_i = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$$

Note that $\|a_i\|^2 = 1 = a_i^T a_i$ and that we require

$$\|a_i\|^2 = \frac{1}{\lambda_i n} = \frac{1}{2} = a_i^T a_i$$

so we must scale a_i by $\frac{1}{\sqrt{2}}$ which gives that

$$a_i = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{4} \\ -\frac{\sqrt{2}}{4} \\ \frac{\sqrt{2}}{4} \\ -\frac{\sqrt{2}}{4} \end{bmatrix}$$

Output Points are then

$$y_{ij} = \sum_{i=1}^{4} \tilde{K}_{jl} a_{il} \quad \text{with } i = 1$$

then $y_1 = y_3 = \frac{\sqrt{2}}{2}$

$$y_2 = y_4 = -\frac{\sqrt{2}}{2}.$$

Approach 2)

Solve $\tilde{K} a_i = \lambda_{min} a_i$ via first principles by noting that it implies that

$$(\tilde{K} - \lambda_{min} I) a_i = 0$$

which is solved by λ_i satisfying

$$|\tilde{K} - \lambda_{min} I| = 0 = |\tilde{K} - 4\lambda_i I|$$

This yields $\lambda_i = 0$ or $\frac{1}{2}$.

Choosing $\lambda_i = \frac{1}{2}$ then solving

$$(\tilde{K} - \tfrac{1}{2} I) a_i = 0 \text{ for } a_i$$

yields $a_i = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$. Need $\|a_i\|^2 = \frac{1}{2}$ so

scaling a_i gives same answers as above.

Problem 2

a) Show

$$\max_{\{u : \|u\|_2 = 1\}} u^T C u^T = \max_{\{v : \|v\|_2 = 1\}} v^T K v$$

LHS RHS.

Replace constraints $\|u\|_2 = 1$ and $\|v\|_2 = 1$ in both problems with equivalent constraints $\|u\|_2^2 = 1$ and $\|v\|_2^2 = 1$.

Apply Lagrange multiplier method to LHS giving

$$\mathcal{L}(u, \lambda) = u^T C u + \lambda(1 - u^T u)$$

$$\therefore \frac{\partial \mathcal{L}}{\partial u} = 2Cu - 2\lambda u = 0 \Rightarrow Cu = \lambda u$$

So solutions (u^*, λ^*) to LHS must be e-vectors, e-values of C.

Substituting $Cu^* = \lambda^* u^*$ into LHS cost gives

$$\max_{\{u : \|u\|_2^2 = 1\}} u^T C u = u^{*T} C u^* = u^{*T} \lambda u^* = \underbrace{(u^{*T} u)}_{=1} \lambda^*$$

So optimal λ^* must be largest e-value of C.

Apply Lagrange multiplier method to RHS giving

$$\mathcal{L}(v, \bar{\lambda}) = v^T K v + \bar{\lambda}(1 - v^T v)$$

$$\therefore \frac{\partial \mathcal{L}}{\partial v} = 2Kv - 2\bar{\lambda} v = 0 \Rightarrow Kv = \bar{\lambda} v$$

So solutions $(v^*, \bar{\lambda}^*)$ to RHS must be e-vectors, e-values of K.

Substituting $Kv^* = \lambda^* v^*$ into RHS gives

$$\max_{\{v : \|v\|_2^2 = 1\}} v^T K v = v^{*T} K v^* = v^{*T} \lambda^* v^*$$

$$= \lambda^* \underbrace{\|v^*\|_2^2}_{=1} = \lambda^*$$

So optimal λ^* must be largest e-value of K.

Must now show that largest e-values of C & K are the same.

Note that the e-values & e-vectors of $C = \frac{1}{n} X X^T$ satisfy

$$Cu = \lambda u$$

$$\Rightarrow \quad \frac{1}{n} X X^T u = \lambda u$$

multiply both sides by X^T gives

$$\frac{1}{n} X^T X X^T u = \lambda X^T u$$

$$\Rightarrow \quad K X^T u = \lambda X^T u \qquad \text{noting } K = \frac{1}{n} X^T X.$$

Letting $v = X^T u$ we have

$$Kv = \lambda v$$

which is the eigenvector/eigenvalue problem for the gram matrix K thus C & K share the same eigenvectors/eigenvalues (provided they are non-zero).

$\therefore$ LHS solved by u^* corresponding to largest e-value λ^* & RHS solved by $v^* = X^T u^*$ with $\lambda^* = \lambda^*$.

Alternative Approach

Recall economic or compact SVD of $X = UDV^T$

where $U \in \mathbb{R}^{d \times r}$, $D \in \mathbb{R}^{r \times r}$, $V \in \mathbb{R}^{n \times r}$ where X has r non-zero singular values. Then

$$C = \frac{1}{n} XX^T = \frac{1}{n} (UDV^T)(VDU^T) = \frac{1}{n} UD^2U^T$$

$\therefore$ LHS optimisation cost is

$$u^T C u = \frac{1}{n} u^T (UD^2U^T) u$$

$$= \frac{1}{n} \tilde{u}^T D^2 \tilde{u}$$

where $\tilde{u} \triangleq U^T u \in \mathbb{R}^r$ and note $\|\tilde{u}\| = \|u\|$ since U orthogonal

LHS can be written as the optimisation

$$\max_{\{u \in \mathbb{R}^d : \|u\|_2^2 = 1\}} u^T C u = \max_{\{\tilde{u} \in \mathbb{R}^r : \|\tilde{u}\|_2^2 = 1\}} \frac{1}{n} \tilde{u}^T D^2 \tilde{u} \quad (3)$$

Furthermore

$$K = \frac{1}{n} X^T X = \frac{1}{n} (VDU^T)(UDV^T) = \frac{1}{n} VD^2V^T$$

$\therefore$ RHS optimisation cost is

$$v^T K v = \frac{1}{n} v^T (VD^2V^T) v$$

$$= \frac{1}{n} \tilde{v}^T D^2 \tilde{v}$$

where $\tilde{v} \triangleq V^T v \in \mathbb{R}^r$ and note $\|\tilde{v}\| = \|v\|$ since V orthogonal

RHS can be written as the optimisation proble

$$\max_{\{v \in \mathbb{R}^n : \|v\|_2^2 = 1\}} v^T K v = \max_{\{\tilde{v} \in \mathbb{R}^r : \|\tilde{v}\|_2^2 = 1\}} \frac{1}{n} \tilde{v}^T D^2 \tilde{v} \quad (4)$$

Clearly the right-hand sides of (3) and (4) are equivalent since they are both optimising over vectors in $\mathbb{R}^r$ with the same cost and constraint.

b) In PCA we select k principle components from the eigenvectors of the covariance matrix C, and then perform the projection $y_j = A^T x_j$ where $A = [u_1 \; u_2 \; \dots \; u_k]$ for (centered data) $\{x_j : 1 \le j \le n\}$.

Under Approach 1 to Part (a)

In Part (a) it is shown that the covariance matrix C & Gram matrix K share the same (non-zero) eigenvalues and their eigenvectors are related via $v = X^T u$. Thus, to perform PCA using the Gram matrix, we can identify the k-largest eigenvalues of K ($\& C$) and the corresponding e-vectors of K ($\& C$) as

$$v_1 = X^T u_1, \; v_2 = X^T u_2, \; \dots \; v_k = X^T u_k$$

Note then that properties of the transpose give

$$v_1^T = u_1^T X, \; v_2^T = u_2^T X, \; \dots, \; v_k = u_k^T X$$

and recall that the output points are

$$y_j = \begin{bmatrix} u_1^T \\ \vdots \\ u_n^T \end{bmatrix} x_j$$

for $j = 1, \ldots, n$ and so noting $X = [x_1, \ldots, x_n]$ we have that

$$[y_1 \quad y_2 \quad \cdots \quad y_n] = \begin{bmatrix} u_1^T X \\ u_2^T X \\ \vdots \\ u_n^T X \end{bmatrix} = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_n^T \end{bmatrix}$$

The output points can thus be computed as the columns of $\begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$.

Alternatively, once we have v_i, we could try solving $\underset{n\times 1}{v_i} = \underset{n\times d}{X^T} \underset{d\times 1}{u_i}$ for u_i with $\|u_i\| = 1$ then compute output points as usual.

Under Approach 2 to Part (a)

In Part (a) it is shown that

$$C = \frac{1}{n} U D^2 U^T \quad \text{and} \quad K = \frac{1}{n} V D^2 V^T$$

Thus C & K share the same (non-zero) eigenvalues corresponding to the elements of $\frac{1}{n} D^2$.

Thus $C = U \Sigma U^T$ and $K = V \Sigma V^T$

with $\Sigma = \frac{1}{n} D^2$.

Recall that the principle components are the eigenvectors of C, and since $C = U \Sigma U^T$ is the eigendecomposition of C the principle components are the columns of U.

Note also that the output is

$$y_j = A^T x_j \quad \text{for } j = 1, \ldots, n$$

with $A = [u_1 \; u_2 \; \ldots \; u_n]$.

Then without loss of generality consider $k = r = n$, we have then that $A = U$ and so

$$y_j = U^T x_j$$

$$\Rightarrow [y_1 \; y_2 \; \ldots \; y_n] = U^T X$$

Recalling the SVD of X gives $X = UDV^T$

$\Rightarrow U^T X = DV^T$ since U orthogonal

$\therefore \; [y_1 \; y_2 \; \ldots \; y_n] = DV^T$ (with D here because v_i in V normalised).

We can therefore perform PCA using the eigenvalues and eigenvectors of the gram matrix K (namely $\Sigma = \frac{1}{n} D^2$ & V).

c) Choosing the kernel $K(x_i, x_j) = x_i^T x_j$ in Kernel PCA yields the Gram matrix $K = X^T X = \tilde{K}$ since X is already mean-subtracted.

Kernel PCA involves solving for vectors a_i by solving the eigen problem

$$\tilde{K} a_i = \lambda_i n a_i \quad \text{with} \quad \|a_i\|^2 = \frac{1}{(\lambda_i n)}.$$

This problem is the same as solving (divide by n)

$$\frac{1}{n}\tilde{K} a_i = \lambda_i a_i \quad \text{with} \quad \|a_i\|^2 = \frac{1}{(\lambda_i n)} \quad (5)$$

where

$\frac{1}{n}\tilde{K} = \frac{1}{n} X^T X$ is the Gram matrix studied in Parts (a) & (b).

$\therefore$ The a_i solving (5) are eigenvectors of $\frac{1}{n} X^T X$ (v_i in Parts (a) & (b)).

The output of Kernel PCA is $y_{ij} = \sum_{i=1}^{n} \tilde{K}_{jl} a_{il}$

so $[y_1 \cdots y_n] = \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix}$ showing that

Kernel PCA with kernel $K(x_i, x_j) = x_i^T x_j$ provides same output as standard PCA.

Problem 3

a) $$D = \begin{bmatrix} 0 & 2 & 4 & 2 \\ 2 & 0 & 2 & 4 \\ 4 & 2 & 0 & 2 \\ 2 & 4 & 2 & 0 \end{bmatrix}$$

b) (1) $$S_i = \sum_{j=1}^{n} D_{ij} \Rightarrow S_1 = S_2 = S_3 = S_4 = 8$$

$$S = \sum_i S_i = 32$$

(2) $$B_{ij} = -\frac{1}{2}\left(D_{ij} - \frac{1}{n}S_i - \frac{1}{n}S_j + \frac{1}{n^2}S\right)$$

$$= -\frac{1}{2}\left(D_{ij} - \frac{1}{4}8 - \frac{1}{4}8 + \frac{1}{16}32\right)$$

$$\therefore B = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}$$

(3) eigen decomposition of B is

$$B = U \Lambda U^T$$

with e-values

$$\Lambda = \begin{bmatrix} 2 & & & \\ & 2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

and e-vectors $U = [u_1\ u_2\ u_3\ u_4]$

$$u_1 = \frac{\sqrt{2}}{2}\begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad u_2 = \frac{\sqrt{2}}{2}\begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} \quad u_3 = \frac{\sqrt{2}}{2}\begin{bmatrix} 0 \\ -1 \\ 0 \\ -1 \end{bmatrix} \quad u_4 = \frac{\sqrt{2}}{2}\begin{bmatrix} -1 \\ 0 \\ -1 \\ 0 \end{bmatrix}.$$

(Unnormalised e-vectors without factor of $\frac{\sqrt{2}}{2}$ also work in this case.)

Output points thus either

$$Y = \sqrt{2}\,u_1 = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{or} \quad Y = \sqrt{2}\,u_2 = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

since two large e-values equal to 2.

Using unnormalised e-vectors gives points

$$Y = \begin{bmatrix} -\sqrt{2} \\ 0 \\ \sqrt{2} \\ 0 \end{bmatrix} \text{ or } \quad Y = \begin{bmatrix} 0 \\ -\sqrt{2} \\ 0 \\ \sqrt{2} \end{bmatrix}$$

which are valid points in this case.
Safest to always normalise e-vectors through.

Problem 4

- LLE avoids computation of pairwise geodesic distances as required by IsoMap, leading to significantly less computational effort.
- LLE also avoids eigendecompositions involved in IsoMap (due to underlying MDS), so may be less computationally & memory intensive than IsoMap for larger datasets.
- Matrix of geodesic distances in IsoMap much denser than the weight matrix in LLE since distance matrix in IsoMap is between all pairs whilst weight matrix only via neighbours.