Data Science Assignment 4

Nike Marie Pulow – Henri Paul Heyden
stu239549 – stu240825

1 PCA, SVD

1. We search for matrices U, Σ, V with $X = U\Sigma V^T$ so that the columns of U and V are eigenvectors of XX^T and X^TX respectively.
We'll begin by computing XX^T:

$$\begin{aligned} XX^T &= \begin{bmatrix} 4 & 2 \\ \sqrt{2} & 2\sqrt{2} \\ -\sqrt{2} & -2\sqrt{2} \end{bmatrix} \times \begin{bmatrix} 4 & \sqrt{2} & -\sqrt{2} \\ 2 & 2\sqrt{2} & -2\sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} \end{aligned}$$

Now we have to solve $\det(X - \lambda I_{3\times 3}) = 0$ for the eigenvalues of XX^T.
An expanded form of this equation is:

$$\begin{aligned} 0 &= \det(X - \lambda I_{3\times 3}) \\ &= \det\left(\begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right) \\ &= \det\left(\begin{bmatrix} 20-\lambda & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10-\lambda & -10 \\ -8\sqrt{2} & -10 & 10-\lambda \end{bmatrix} \right) \\ &= (20-\lambda)(10-\lambda)(10-\lambda) + (8\sqrt{2})(-10)(-8\sqrt{2}) + (-8\sqrt{2})(8\sqrt{2})(-10) \\ &\quad -(-8\sqrt{2})(10-\lambda)(-8\sqrt{2}) - (8\sqrt{2})(8\sqrt{2})(10-\lambda) - (20-\lambda)(-10)(-10) \end{aligned}$$

This is a third degree polynomial equation with expanded form $-x^3 + 40x^2 - 144x = 0$ [1]
We get the solutions:

$$\lambda_1 = 36 \wedge \lambda_2 = 4 \wedge \lambda_3 = 0$$

With this we have

$$\Sigma = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

With the eigenvalues of XX^T being calculated, we can now solve for the eigenvectors of XX^T:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} &= \left(\begin{bmatrix} 20 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & 10 & -10 \\ -8\sqrt{2} & -10 & 10 \end{bmatrix} - \begin{bmatrix} 36 & 0 & 0 \\ 0 & 36 & 0 \\ 0 & 0 & 36 \end{bmatrix} \right) \times \begin{bmatrix} v_{1_1} \\ v_{1_2} \\ v_{1_3} \end{bmatrix} \\ &= \begin{bmatrix} -16 & 8\sqrt{2} & -8\sqrt{2} \\ 8\sqrt{2} & -26 & -10 \\ -8\sqrt{2} & -10 & -26 \end{bmatrix} \times \begin{bmatrix} v_{1_1} \\ v_{1_2} \\ v_{1_3} \end{bmatrix} \\ &= \begin{bmatrix} -16v_{1_1} + 8\sqrt{2}v_{1_2} - 8\sqrt{2}v_{1_3} \\ 8\sqrt{2}v_{1_1} - 26v_{1_2} - 10v_{1_3} \\ -8\sqrt{2}v_{1_1} - 10v_{1_2} - 26v_{1_3} \end{bmatrix} \end{aligned}$$

[1] Actually writing each step of this down would take too much space.

When numbering the equations in collumns **I**, **II** and **III** we can see that $\mathbf{I} \Rightarrow v_{1_1} = \frac{\sqrt{2}v_{1_2}}{2} - \frac{\sqrt{2}v_{1_3}}{2}$ and $\mathbf{II} + \mathbf{III} \Rightarrow v_{1_2} = -v_{1_3}$ which results in:

$$v_{1_1} = \sqrt{2}v_{1_2}$$

and respectively:

$$v_{1_2} = \frac{v_{1_1}}{\sqrt{2}} \wedge v_{1_3} = -\frac{v_{1_1}}{\sqrt{2}}$$

There are infinite solutions for this, with one being:

$$v_{1_1} = 1 \wedge v_{1_2} = \sqrt{2}^{-1} \wedge v_{1_3} = -\sqrt{2}^{-1}$$

So we get one eigenvector:

$$v_1 = \begin{bmatrix} 1 \\ \sqrt{2}^{-1} \\ -\sqrt{2}^{-1} \end{bmatrix}$$

Because writing the processes down takes a lot of space, the following linear equation systems will not be solved, but solutions provided. We have already demonstrated that we can solve one and the probability of minor mistakes when writing it down ruining the whole task is too great.
The second eigenvector we get by this time subtracting the diagonal matrix times $\mathbf{4}$ instead of $\mathbf{36}$ and we get a solution:

$$v_2 = \begin{bmatrix} 1 \\ -\sqrt{2}^{-1} \\ \sqrt{2}^{-1} \end{bmatrix}$$

And the third eigenvector is quite trivial:

$$v_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}$$

Something that wasn't specified in the slides was that the eigenvectors must be normalized so that U and V are orthonormal.
For that we have to divide all these vectors by $\sqrt{2}$.
We get:

$$U = \begin{bmatrix} \sqrt{2}^{-1} & \sqrt{2}^{-1} & 0 \\ 1/2 & -1/2 & \sqrt{2}^{-1} \\ -1/2 & 1/2 & \sqrt{2}^{-1} \end{bmatrix}$$

For a quick verification, we can compute UU^T and yes we get $I_{3\times3}$, so U is indeed orthonormal.
Now to V
As the lecture specifies, the eigenvalues of XX^T are the same like those of X^TX.
This means we „just" have to solve the linear equations but for X^TX and not XX^T.
We'll begin by computing X^TX:

$$\begin{aligned} X^TX &= \begin{bmatrix} 4 & \sqrt{2} & -\sqrt{2} \\ 2 & 2\sqrt{2} & -2\sqrt{2} \end{bmatrix} \times \begin{bmatrix} 4 & 2 \\ \sqrt{2} & 2\sqrt{2} \\ -\sqrt{2} & -2\sqrt{2} \end{bmatrix} \\ &= \begin{bmatrix} 20 & 16 \\ 16 & 20 \end{bmatrix} \end{aligned}$$

Thank god that this time it's small and regular.
We get the following systems of linear equations: With eigenvalue $\mathbf{36}$ we get:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} -16 & 16 \\ 16 & -16 \end{bmatrix} \times \begin{bmatrix} v_{1_1} \\ v_{1_2} \end{bmatrix} \\ &= \begin{bmatrix} 16v_{1_1} - 16v_{1_2} \\ 16v_{1_1} - 16v_{1_2} \end{bmatrix} \end{aligned}$$

With one solution being:

$$v_1 = \begin{bmatrix} \sqrt{2}^{-1} \\ \sqrt{2}^{-1} \end{bmatrix}$$

With eigenvalue 4 we get:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} 16 & 16 \\ 16 & 16 \end{bmatrix} \times \begin{bmatrix} v_{2_1} \\ v_{2_2} \end{bmatrix} \\ &= \begin{bmatrix} 16v_{2_1} + 16v_{2_2} \\ 16v_{2_1} + 16v_{2_2} \end{bmatrix} \end{aligned}$$

With one solution being

$$v_2 = \begin{bmatrix} \sqrt{2}^{-1} \\ -\sqrt{2}^{-1} \end{bmatrix}$$

For the last eigenvalue 0 we get $v_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, which formally isn't an eigenvector, since it would be an eigenvector of every matrix, but it is the only solution to the corresponding eigenvalue.
With that we get:

$$V^T = \begin{bmatrix} \sqrt{2}^{-1} & \sqrt{2}^{-1} \\ \sqrt{2}^{-1} & -\sqrt{2}^{-1} \\ 0 & 0 \end{bmatrix}$$

And now we can calculate the product $U\Sigma V^T$ to verify that it is actually right, which I did, but to be honest, writing all that down with each addition and so on would take a lot of time. □

2. We firstly have to center the points, since to calculate the eigenvectors of the covariance Matrix, we can't assume that the mean x and y components are 0. We get:

$$\overline{x} = 4/3 \wedge \overline{y} = 2/3$$

With this we can center X:

$$X_C = \begin{bmatrix} 4 - 4/3 & 2 - 2/3 \\ \sqrt{2} - 4/3 & 2\sqrt{2} - 2/3 \\ -\sqrt{2} - 4/3 & -2\sqrt{2} - 2/3 \end{bmatrix}$$

And get the covariance Matrix:

$$\begin{aligned} C &= 1/3 \cdot X_C^T X_C \\ &= 1/3 \cdot \begin{bmatrix} 44/3 & 40/3 \\ 40/3 & 56/3 \end{bmatrix} \end{aligned}$$

The actual expanded form of this with appropriate steps in between would be way too long to fit on one page.
We actually don't care for the scalar $1/9$, since the eigenvalues are independent of it since the characteristic polynomial would just be scaled by a scalar, so the zeroes of it are the same.
With this, to get the eigenpairs of the covariance matrix C, we can just get the eigenpairs of $9 \cdot C$, since it does not matter, since we only need the eigenvectors and they will be normalized later anyways.
We have to solve:

$$\begin{aligned} 0 &= \det(9 \cdot C - \lambda I_{2\times 2}) \\ &= \det\left(\begin{bmatrix} 44 - \lambda & 40 \\ 40 & 56 - \lambda \end{bmatrix}\right) \\ &= (44 - x)(56 - x) - 40 \cdot 40 \\ &= x^2 - 100x + 864 \end{aligned}$$

This polynomial has solutions:

$$\lambda_1 = 50 + 2\sqrt{409} \wedge \lambda_2 = 50 - 2\sqrt{409}$$

With λ_1 being the bigger eigenvalue, we know that the direction of biggest covariance is the eigenvector corresponding to λ_1.
On a side note, the eigenvector corresponding to λ_2 must be orthogonal to the one corresponding to λ_1, since C is symmetric, this also makes more sense regarding what were calculating.
To get λ_1's eigenvector, we have to solve:

$$\begin{aligned}\begin{bmatrix}0\\0\end{bmatrix} &= \begin{bmatrix}44-\lambda_1 & 40\\ 40 & 56-\lambda_1\end{bmatrix} \times \begin{bmatrix}v_x\\v_y\end{bmatrix}\\ &= \begin{bmatrix}-(6+2\sqrt{409})v_x + 40v_y\\ 40v_x + (6-2\sqrt{409})v_y\end{bmatrix}\end{aligned}$$

Which has infinite solutions, one being:

$$v = \begin{bmatrix}1\\ 1/20(\sqrt{409}-3)\end{bmatrix}$$

Which is our eigenvector. And with that being the direction of maximal covariance in X.

On a side note, you see all calculation in this subtask thus far would have been trivial, if X would have been centered already to begin with, since then we could just have took the first eigenvector in V from 1.1,this whole ordeal makes the hint „you may use your results from one subtask for the other“ extremely misleading.

Now we have[2] to normalize v. The length of v is $s := \sqrt{1 + 1/400 \cdot (\sqrt{409}-3)^2} \approx 1.31971352$, which has no further simplification.
Finally we get:

$$v_n = \frac{1}{s} \cdot \begin{bmatrix}1\\ 1/20(\sqrt{409}-3)\end{bmatrix}$$

As the normalized version.
According to the lecture, the transformation matrix that projects any point onto the corresponding line is $v_n \times v_n^T = s^{-2} \cdot v \times v^T$ which is:

$$\begin{bmatrix}\frac{1}{2} + \frac{3}{2\sqrt{409}} & \frac{10}{\sqrt{409}}\\ \frac{10}{\sqrt{409}} & \frac{1}{2} - \frac{3}{2\sqrt{409}}\end{bmatrix}$$

Which is approximately:

$$\begin{bmatrix}0.57417 & 0.49447\\ 0.49447 & 0.42583\end{bmatrix}$$

□

[2] I think we don't have to, since for $s \in \mathbb{R}$ we have $(s \cdot u) \times (s \cdot u^T) = s^2 \cdot u \times u^T$, which means that the projected point would be still on the line, but just with different distance to (0,0), but the lecture suggests that we should.

2 TF-IDF

1. Vocabulary: [”fast”, ”car”, ”highway”, ”road”, ”bike”, ”wheel”]

2. term frequencies:

t	$d=1$	$d=2$	$d=3$	$d=4$
fast	0.2	0.4	0.2	0
car	0.4	0.4	0	0.25
highway	0.2	0	0.2	0
road	0.2	0	0.4	0
bike	0	0.2	0	0.25
wheel	0	0	0.2	0.5

3. document frequency:

t	$\#d$ containing t	IDF
fast	3	$\log \frac{4}{3}$
car	3	$\log \frac{4}{3}$
highway	2	$\log \frac{4}{2} = \log 2$
road	2	$\log(2)$
bike	2	$\log(2)$
wheel	2	$\log(2)$

4. calculating TF-IDF vectors:

t	$d=1$	$d=2$	$d=3$	$d=4$
fast	0.025	0.05	0.025	0
car	0.05	0.05	0	0.031
highway	0.06	0	0.06	0
road	0.06	0	0.120	0
bike	0	0.06	0	0.075
wheel	0	0	0.06	0.151

Which gives us the following vectors for each d:

$$r(d=1) = (0.025, 0.05, 0.06, 0.06, 0, 0)$$
$$r(d=2) = (0.05, 0.05, 0, 0, 0.06, 0)$$
$$r(d=3) = (0.025, 0, 0.06, 0.12, 0, 0.06)$$
$$r(d=4) = (0, 0.031, 00, 0.075, 0.151)$$