A38627

Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers



UNIVERSITY OF BIRMINGHAM

School of Computer Science

Fourth Year Undergraduate/Postgraduate

06-38968

LM Intelligent Data Analysis

Main Summer Examinations 2025

Time allowed: 2 hours

[Answer all questions]

Question 1

Table 1 corresponds to the coarse sampling over a normalized cylinder surface (See Figure 1) i.e. Radius = 1 and height = 1.

θ	ρ	Z	X	Y	Z
0	1	0	1	0	0
$\pi/4$	1	0	0.7071	0.7071	0
$\pi/2$	1	0	0.0000	1	0
$3\pi/4$	1	0	-0.7071	0.7071	0
π	1	0	-1	0	0
$5\pi/4$	1	0	-0.7071	-0.7071	0
$3\pi/2$	1	0	0	-1	0
$7\pi/4$	1	0	0.7071	-0.7071	0
0	1	1	1	0	1
$\pi/4$	1	1	0.7071	0.7071	1
$\pi/2$	1	1	0	1	1
$3\pi/4$	1	1	-0.7071	0.7071	1
π	1	1	-1	0	1
$5\pi/4$	1	1	-0.7071	-0.7071	1
$3\pi/2$	1	1	0	-1	1
$7\pi/4$	1	1	0.7071	-0.7071	1

Table 1: Sampling a cylinder surface. Left: Polar coordinates, Right: Cartesian coordinates.

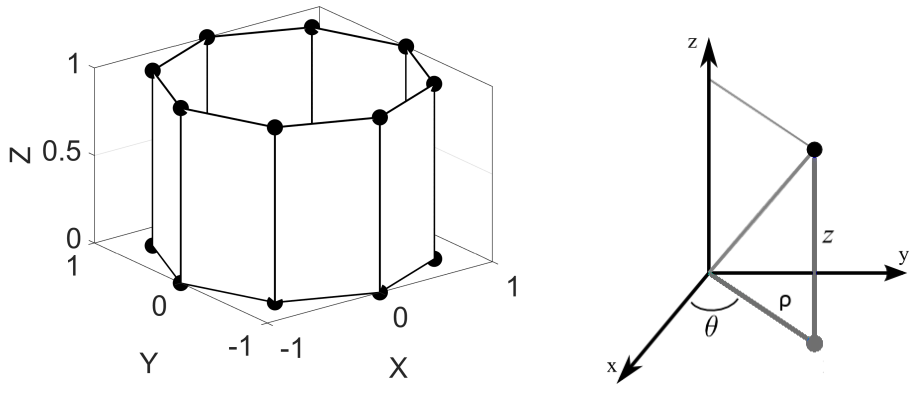


Figure 1: Left: Sampling a cylinder surface. Right: Cylindrical coordinates for reference.

(a) Report the covariance matrix. Show your intermediate calculations. **[9 marks]**

(b) Report the eigenvalues of the principal component analysis of the data, both raw and normalized. Show your intermediate calculations. **[9 marks]**

(c) Figure 2 shows the behaviour of the eigenvalues as we increase the sampling rate over the surface, whether in phase or in height yet maintaining the radius = 1 and height = 1. Discuss how altering the cylinder radius or height affects the output of PCA in terms of the changes in dimensionality of the cloud of points. **[8 marks]**

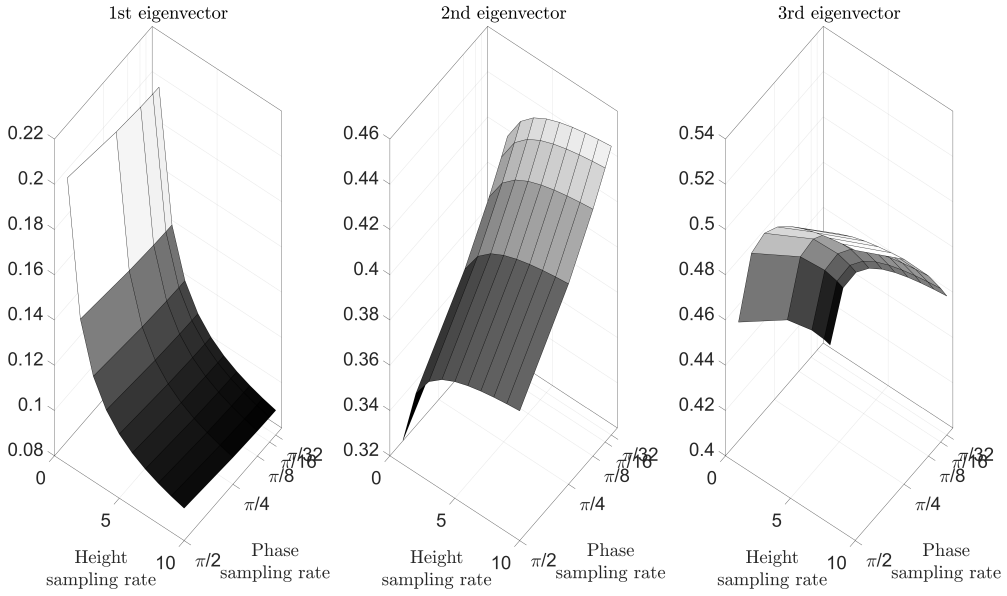


Figure 2: Normalized eigenvalues as a function of the phase and height sampling rate.

(d) Finally, imagine that we have a cylinder but this time not normalized, e.g. either radius and/or height different from 1. How would this affect the output of PCA in terms of the appreciation of changes in dimensionality of the cloud of points. **[8 marks]**

Question 2

Somewhat inspired by true events. The regional secretary of transport and infrastructure is seeking to preserve a old Roman bridge that even up to this day was sustaining quite heavy transit. The nodes of graph $G = (V, E)$ represent towns in the county connected by roads (links in the graph), including the Roman bridge (See Figure 3).

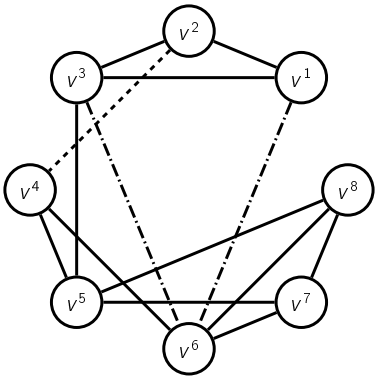


Figure 3: Depiction of towns in the county, with the road traversing the Roman bridge (dashed), and the two new potential location for additional links (dash-dotted).

In order to divert some traffic to alleviate the Roman bridge, the regional government has come to two final potential alternatives (dash-dotted links) where to build an additional link between the south and central part of the county. Only the one that minimizes distances travelled will be built.

As an expert on data analysis, you are asked to provide your informed advice to the government as to which one to build.

Do not simply answer with the final outcome. Explain and justify the steps taken as well as show the intermediate calculations if any. Remember that in unweighted graphs distances are proportional to the number of edges traversed.

The new road (and bridge) will depart from v^6. Which of the two alternative endpoints v^1 or v^3 (or no difference) would you advise as the best option for building the additonal road? **[33 marks]**

Question 3

Algorithm 1 encodes the PageRank algorithm in pseudocode.

Algorithm 1: PageRank

```
Data: G: A directed graph
Data: d ∈ [0, 1] ⊂ ℝ: A dumping parameter
Data: e : e^i ∈ ℝ: A rank sources vector
Data: p0 : p0^i ∈ ℝ: PageRank scores initialization vector (initial prestiges)
Data: maxIter ∈ ℕ: Number of maximum iterations
Data: tol ∈ ℝ: Tolerance error
Data: Dangling: Outedges to be assigned to any dangling nodes
Result: p : p^i ∈ ℝ: PageRank scores (final prestiges)
1  /* Initialization:                                                          */
2  /* - Set up the various parameters and build the normalized adjacency matrix */
3  A ← adjacency(G) /* Retrieve adjacency matrix                              */
4  if A = ∅ then
5  |   p ← ∅
6  end if
7  A^norm ← A
8  a_ij^norm ← a_ij^norm / Σ_j a_ij^norm /* Normalize                         */
9  if Σ_i p0^i ≠ 1 then
10 |   p0^i ← p0^i / Σ_i p0^i
11 end if
12 if Σ_i e^i ≠ 1 then
13 |   e^i ← e^i / Σ_i e^i
14 end if
15 Dangling_ij ← Dangling_ij / Σ_j Dangling_ij /* Normalize dangling          */
16 DanglingNodes ← v^i : Σ_j a_ij = 0
17 for v^i ∈ DanglingNodes do
18 |   a_ij^norm ← Dangling_ij
19 end for
20 /* Main loop:                                                               */
21 p ← p0
22 for iter = 1 : maxIter do
23 |   p^last ← p
24 |   p ← d(A^normT * p^last) + (1 − d)e
25 |   /* Check convergence, using L1 norm                                     */
26 |   if ‖p − p^last‖_1 < tol then
27 |   |   break
28 |   end if
29 end for
```

(a) PageRank imitates a random surfer where all links available in a page have the same probability of being visited i.e. alike a uniform probability distribution of clicks over the links. In practice, links at the top of the page have a slightly higher probability of being visited than those further down, specially in long pages. Explain how the random surfer behaviour is encoded in PageRank and refer to Algorithm 1's line numbers if you need to point where it happens. **[11 marks]**

(b) Modify the above algorithm to allow a non-uniform distribution of clicks over the links. Do NOT concern with checking the order of the links as such, just ensure that you permit the manipulation of the probability of links to be clicked. There is NO need to copy the whole pseudo code for your new version, just focus on the

lines that need to be altered. Algorithm 1 and the new pseudocode may or may not differ in the number of lines of code. If your new version differs e.g. one original line now necessitates more than one, or several original lines can be collapsed, please indicate both line countings using the original as the reference (i.e. do not bother considering any shifting in the line counting); e.g.

Original	Modified	code
14	14	code for new line 14
-	15	code for new line 15
21	21	code for new line 21 (replacing 21 and 22)
22	-	

[11 marks]

(c) Explain the rationale of your solution above. How does your new code achieve a non-uniform clicking? **[11 marks]**

This page intentionally left blank.

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.