MATH 123 Math Aspects of Data Analysis - Spring 2023
Tufts University, Department of Mathematics
Instructor: James M. Murphy
Practice Midterm 1

Instructions: This exam has four questions and is out of a total of 100 points. Each question is worth 25 points. No graphing calculators, books, or notes are allowed. Be sure to show all work for all problems. No credit will be given for answers without work shown. If you do not have enough room in the space provided you may use additional paper. Be sure to clearly label each problem and attach them to the exam. You have 75 minutes. Good luck! :-)

Your Printed Name: _____

Problem	Score
1	
2	
3	
4	
5	
Total	

Academic Honesty Certification:

I certify that I have taken this exam without the aid of unauthorized people or objects.

Signature: _____ Date: _____

Question 1

(a) Let $\{x_1, \ldots, x_n\} \subset \mathbb{R}^D$ be data with mean 0. Write the formula for the $D \times D$ covariance matrix Σ.

(b) Prove that for all $y \in \mathbb{R}^{D \times 1}, y^T \Sigma y \geq 0$.

(c) Is Σ necessarily invertible? Prove or give a counterexample.

Question 2

Suppose $\{x_1, \ldots, x_n\} \subset \mathbb{R}^D$ is formed into a data matrix $X \in \mathbb{R}^{n \times D}$ with the i^{th} row corresponding to x_i.

(a) Suppose X has singular value decomposition $X = U\Lambda V^T$, where $UU^T = I, VV^T = I$, and Λ is a diagonal matrix with diagonal entries $\sigma_1 > \sigma_2 > \cdots > \sigma_n \geq 0$ (note the strict inequality between successive singular values). In terms of this decomposition, what is the direction of maximum variance in the data? Explain.

(b) In terms of the singular value decomposition, what proportion of the variance in the data is retained when projecting onto the first principal component? Explain.

Question 3

Let $\{\theta_i\}_{i=0}^n$ be an equi-spaced partition of $[0, 2\pi]$, i.e. $\theta_i = i(2\pi/n)$.

(a) What visual shape do the data $\{(\cos(\theta_i), \sin(\theta_i))\}_{i=0}^n$ form when plotted in $\mathbb{R}^2$?

(b) Compute the covariance matrix of the data $\{(\cos(\theta_i), \sin(\theta_i)\}_{i=0}^n \subset \mathbb{R}^2$.

(c) Recall that for a continuous function $f(\theta)$ on $[0, 2\pi]$, the *Riemann integral* of $f(\theta)$ may be computed using *Riemann sums* as

$$\frac{1}{2\pi}\int_0^{2\pi} f(\theta)d\theta = \lim_{n\to\infty}\frac{1}{n}\sum_{i=0}^{n} f(\theta_i),$$

where $\{\theta_i\}_{i=0}^n$ is as above. Use this fact (without proving it) to argue that as $n \to \infty$, the covariance matrix in (b) becomes of the form $\begin{pmatrix}\alpha & 0\\ 0 & \alpha\end{pmatrix}$, for some $\alpha > 0$.

(d) Part (c) shows that, as $n \to \infty$, the covariance matrix has two eigenvectors with equal eigenvalues. Does this make sense, given your answer in (a)? Can you explain the conclusion of (c) geometrically?

Question 4

(a) Let $x_1, \ldots, x_n \subset \mathbb{R}^D$ be data to be clustered. Write down the functional that the K-means clustering algorithm attempts to minimize.

(b) Consider the "two moons" data. Will K-means with $K = 2$ learn the two moons exactly? Why or why not?