Both written exams for *MA500 Geometric Foundations in Data Analysis* will each consist of four questions: three from Graham and one from Emil. Students will be required to attempt all questions. The folloing are examples of the kinds of things that Graham could ask.

1 Least Squares Fitting

1. Find the best least squares straight line fit to the following measurements, and sketch your solution:

   $y = 2$ at $t = -1$,

   $y = 0$ at $t = 0$,

   $y = -3$ at $t = 1$,

   $y = -5$ at $t = 2$.

2. A middle-aged man was stretched on a rack to lengths $L = 5$, 6, and 7 feet under applied forces of $F = 1$, 2 and 4 tones. Assuming Hooke's Law $L = a + bF$, find his normal length a by least squares.

3. Let
$$y = b_0 + b_1 x_1 + \cdots + b_{p-1} x_{p-1} \tag{1}$$
denote the hyperplane in $\mathbb{R}^p$ that is the best least square hyperplane fit to a given collection of data points $(y_k, x_{k,1}, \ldots, x_{k,p-1}) \in \mathbb{R}^p$, $1 \leq k \leq n$.

   Either

   (a) Describe, in terms of partial derivatives, the normal equations that determine the constants $b_0, \ldots, b_n$.

   (b) Then use matrix notation to express these normal equations, making sure to define those matrices involved.

   Or (more easily and more preferably)

   Observe that on letting $||\ ||$ denote the Euclidean norm and writing
$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\ A = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ & \vdots & & \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix},\ b = \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix},$$
   our least squares requirement is simply that b should be chosen so that $||Ab - y||$ is as small as possible. In other words, the vector $Ab - y$ should be perpendicular to the plane spanned by the vectors of the form Au with $u \in \mathbb{R}^p$. In other words, for an

arbitrary vector $u \in \mathbb{R}^p$ we need $0 = u^t A^t(Ab - y)$ or, since u is arbitrary,

$$0 = A^t(Ab - y) \tag{2}$$

Convince yourself that (2) is the (correct way to think of the) system of normal equations.

4. Let $y = b_0 + b_1 x$ denote the best least squares straight line fit to given data points $(y_1, x_1), ..., (y_n, x_n)$.

   (a) Define the *fitted values* $\hat{y}_i$, *residuals* e_i, *sample mean* $\overline{y}$, *total sum of squares* $SSTO$, *error sum of squares* SSE, *regression sum of squares* SSR, and *coefficient of determination* R^2.

   (b) Prove that $\sum_{i=1}^n e_i = 0$.

   (c) Prove that $\sum_{i=1}^n \hat{y}_i e_i = 0$.

   (d) Prove that $SSTO = SSE + SSR$.

   (e) Prove that $1 \leq R^2 \leq 1$.

The theory of statistical inference can be applied to the output from a least squares fit. This topic is outside the main focus of this module. Nevertheless, the following two questions touch on the topic.

1. The theory of statistical inference can be applied to the output from a least squares fitting.

   Suppose given a random variable

   $$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots \beta_{p-1} x_{i,p-1} + \epsilon_i$$

   where

   - $i = 1, 2, \ldots, n$;
   - $x_{i,1}, \ldots, x_{i,p-1}$ are known constants;
   - $\beta_0, \ldots, \beta p - 1$ are unkown fixed parameters;
   - ϵ_i are independent random variables with common normal distribution $N(0, \sigma^2)$.

   Describe a criterion for choosing between the two hypotheses

   $\mathcal{C}_1$: $\beta_1 = \beta_2 = \cdots \beta_{p-1} = 0$

   $\mathcal{C}_2$: $\beta_i \neq 0$ for at least one $1 \leq i \leq p - 1$

   that controls Type I errors at level α.

2. In the context of the previous question, suppose that $q \leq p$ of the parameters β_k need to be estimated jointly. Describe the Bonferroni confidence intervals with family coefficient $1 - \alpha$ for these q parametrers.

2 Principal Component Analysis

1. Explain what is meant by *Principal Component Analysis*. Your explanation should include explanations of the terms: *geometric information*; *covariance matrix*; *orthogonal transformation*; *Spectral Theorem* and describe how the technique can be used to reduce dimensionality while retaining much geometric information.

2. Prove that any real symmetric matrix has at least one real eigenvector.

3. Use the fact that any real symmetric $n \times n$ matrix A has at least one eigenvector to prove that it has n linearly independent real eigenvectors.

4. Determine the maximum value of the function $f(x, y) = x^2 + 4xy + 4y^2$ on the unit sphere $\mathbb{S}^1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$. Also, find a linear homomorphism $\phi: \mathbb{R}^2 \to \mathbb{R}^2, (x, y) \mapsto (x', y')$ such that $f\phi(x, y) = \lambda_1 x'^2 + \lambda_2 y'^2$ for some $\lambda_1, \lambda_2 \in \mathbb{R}$.

5. A data set $S \subset \mathbb{R}^p$ consists of vectors whose first component is the number of kilometers that a salesperson has travelled during the last month. A principal component analysis is performed on S, the set S is then projected onto the three principal components with largest eigenvalues, and the projected points are visualized in $\mathbb{R}^3$. Would this visualization be any different if distance had been measured in miles? Justify your answer.

6. Suppose given a finite set S of data points in $\mathbb{R}^3$ and that a visual inspection suggests that all points look to lie close to some 2-dimensional plane containing the origin. We could construct a plane by regarding the first coordinate as a dependent variable and taking a least squares fit. Alternatively, we could construct a plane using Principal Component Analysis and taking the span of the eigenvectors corresponding to the two larger eignevalues. In general, would the two constructed planes differ? If so, in what way?

3 Clustering and Persistence

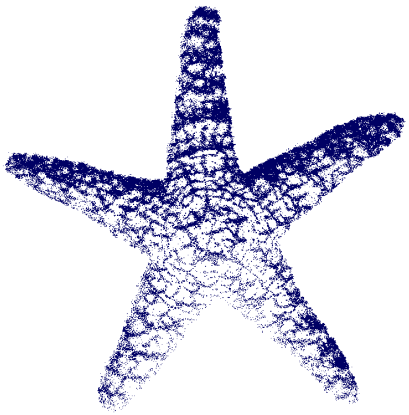
1. Describe an algorithm that inputs an $n \times n$ distance (or dissimilarity) matrix for n items, applies single-linkage hierarchical clustering, and returns the corrsponding barcode. Determine a worst-case time estimate for the algorithm.

2. Describe the Smith-Waterman algorithm for determining the optimal score of a local alignment of two sequences of letters. Include an explanation of the terms *local alignment* and *optimal score*.

3. Explain how cluster analysis and barcodes can be used to estimate the number of objects in the digital photograph.

Explain how cluster analysis can also be used to estimate the number of objects with holes.

4. Explain how cluster analysis and barcodes can be used to estimate the number of 'limbs' of an object such as a starfish from a digital image of the object.



5. A property $I(S)$ of a set $S \subset \mathbb{R}^n$ is said to be a *geometric invariant* if its value does not change when S undergoes an orthogonal tranaformation. Explain why the barcodes in the preceding questions (3) and (4) are geometric invariants.

4 Factor Analysis