Imperial College
London

MATH70049

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2023

This paper is also taken for the relevant examination for the
Associateship of the Royal College of Science

Introduction to Statistical Learning

Date: 19 May 2023

Time: 14:00 – 16:30 (BST)

Time Allowed: 2.5hrs

This paper has 5 Questions.

Please Answer All Questions in 1 Answer Booklet

Candidates should start their answers to each question on a new sheet of paper.

Supplementary books may only be used after the relevant main book(s) are full.

Any required additional material(s) will be provided.

Allow margins for marking.

Credit will be given for all questions attempted.

Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

1. Let A be a set of objects that one is interested in studying.

(a) Let $x, y \in A$. Give the definition of a metric $d(x, y)$ for $x, y \in A$. (3 marks)

(b) Let $d_i(x, y)$ be a metric for each $i = 1, \ldots, q$, $x, y \in A$ for some integer $q > 0$. Show that $d_S(x, y) = \sum_{i=1}^{q} d_i(x, y)$ is a metric. (3 marks)

(c) Define the Hamming distance and show that it is a metric. (5 marks)

(d) Given two binary strings $x = 10011110$ and $y = 10010101$ compute the Hamming distance between x and y. (1 mark)

(e) Given two objects $x, y \in A$ we can count the number of presence/absence attributes that match or do not match and can put these in the following table.

Object		x	
		Present	Absent
y	Present	a	b
	Absent	c	d

Define the Jaccard distance $d_J(x, y)$ for objects x, y from the table. Why does the Jaccard distance not 'care' about d in the table above, i.e. what practical advantage does this confer and give an example? (2 marks)

(f) Given two strings of characters, the Levenshtein distance is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other.

For example, the Levenshtein distance from 'donut' to 'dog' is 3, which can be demonstrated by donut $\rightarrow$ donu $\rightarrow$ don $\rightarrow$ dog.

(i) What is the Levenshtein distance between 'cake' and 'cat'? Show the character edits that can be used to transform one to the other. (2 marks)

(ii) Let x, y be two text strings of length n_x and n_y respectively. State an upper bound on the Levenshtein distance between x and y in terms of n_x and n_y. (2 marks)

(iii) Is the Levenshtein distance symmetric? (2 marks)

(Total: 20 marks)

2. (a) Briefly explain the purpose of the bootstrap in statistics. (3 marks)

(b) Suppose X is a continuous random variable with distribution function F with some mean and variance, $-\infty < \mu < \infty, \sigma^2 < \infty$, all of which are unknown. Let x_i be an independent draw from X for $i = 1, \ldots, n$ for some integer $n > 0$. Let $\hat{\mu} = n^{-1} \sum_{i=1}^{n} X_i$ be the usual sample mean of $\{x_i\}_{i=1}^{n}$. Explain how you might use the bootstrap method to estimate the sampling distribution of $\hat{\mu} - \mu$. (2 marks)

(c) What does the abbreviation CART stand for? Briefly explain what a classification tree is. What are the pros and cons of the CART method? (8 marks)

(d) A materials scientist wishes to develop a method to identify the type of glass from its chemical composition. The `Glass` data consists of 214 observations on 10 variables. A description of the variables is given in the table:

Number	Abbreviation	RI or chemical content
1	RI	Refractive index
2	Na	Sodium
3	Mg	Magnesium
4	Al	Aluminium
5	Si	Silicon
6	K	Potassium
7	Ca	Calcium
8	Ba	Barium
9	Fe	Iron
10	Type	Type of glass (class attribute).

Figure 1 shows a boxplot of magnesium content split by glass type.

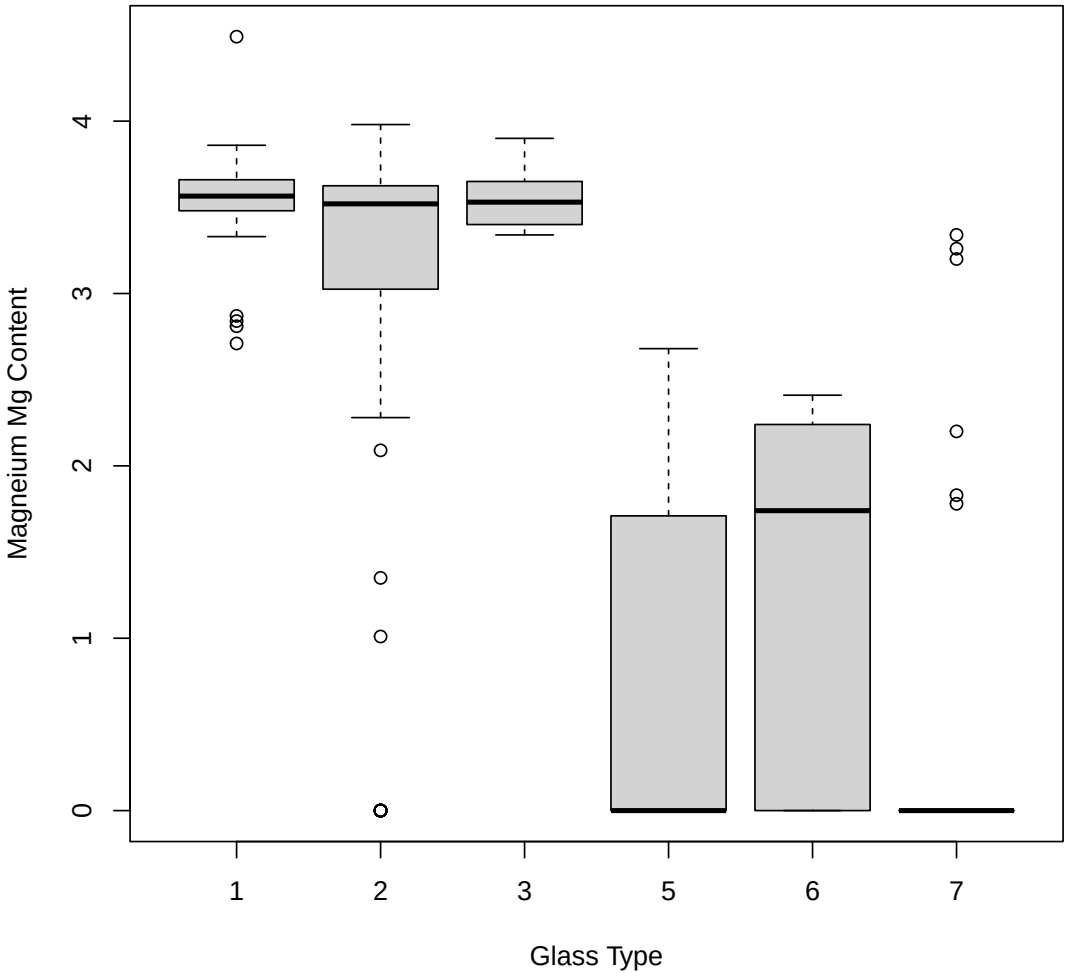

Figure 1: Magnesium content versus glass type boxplot.

Explain how this plot could help you discern glass type using magnesium content.

(2 marks)

(e) The scientist computes a classification tree on the first 205 observations, and then realizes that they accidentally omitted 9 observations. They then compute the tree on the full 214 observations. The two trees are shown in Figures 2 and 3.

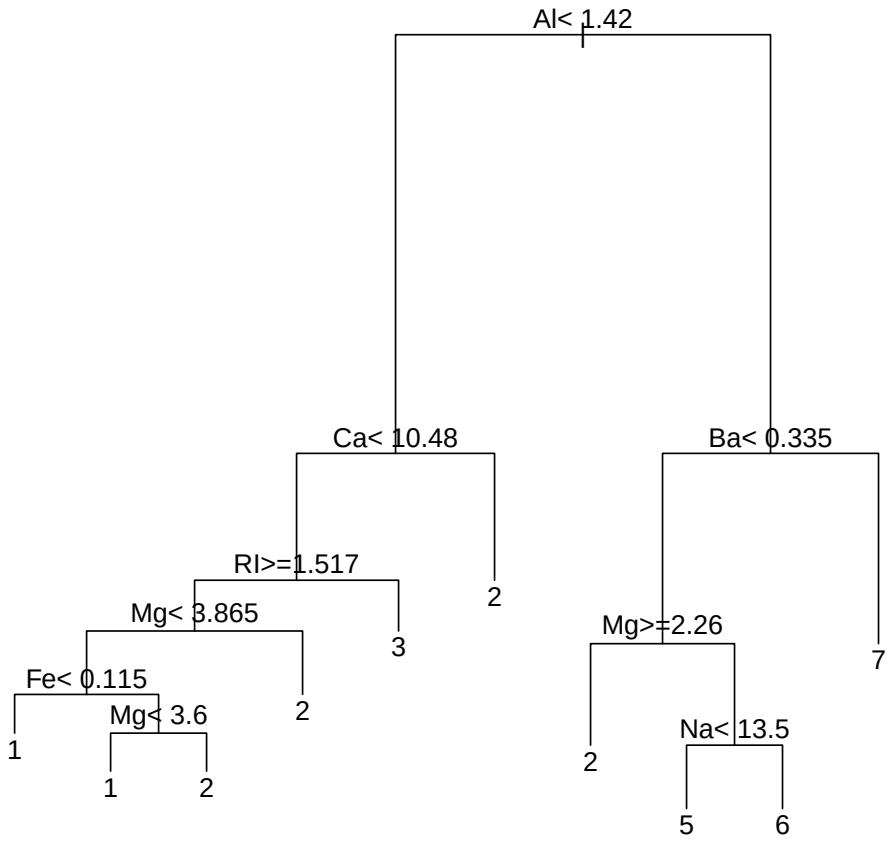

Figure 2: Classification tree of the first 205 observations of the glass data.

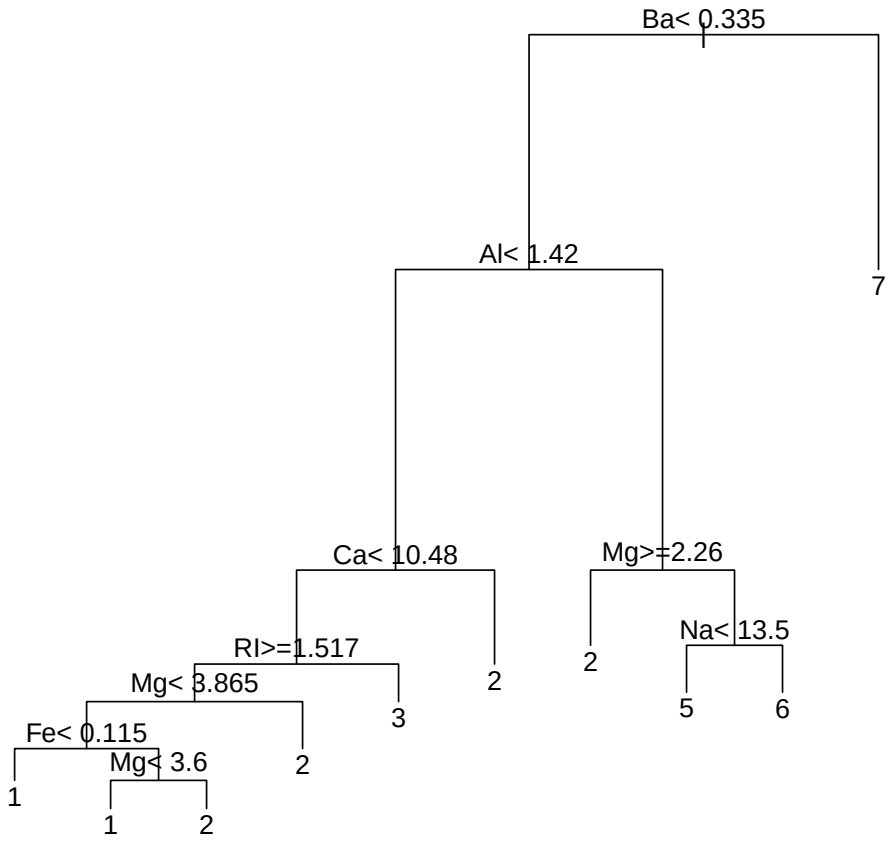

Figure 3: Classification tree for the full glass data.

The scientist is puzzled because the first tree contains over 95% of the data, but the trees look different. Why are the trees different even though they were constructed from datasets that shared 95% of the same observations? (1 mark)

(f) The scientist acquires a new piece of glass for testing and carries out a scientific analysis and arrives at the following values on the variables for the new glass: RI$=1.51756$, Na$=13.15$, Mg$=3.61$, Al$=1.05$, Si$=73.24$, K$=0.57$, Ca$=8.24$, Ba$=0$, Fe$=0$. The scientist decides to compute a bagged estimate for the glass using the two trees in Figures 2 and 3. Explain how the bagged estimate is obtained and write down what it is. (3 marks)

(g) Explain what a random forest is and how it can ameliorate the problem described in part(e) where small changes in input data can result in quite different trees. (1 mark)

(Total: 20 marks)

3. (a) Define $L_2(\mathbb{R})$, the space of square-integrable functions on $\mathbb{R}$. (1 mark)

(b) Provide the mathematical definition of a multiresolution analysis of $L^2(\mathbb{R})$. (5 marks)

(c) State and prove both Parseval's and Plancheral's theorems. (3 marks)

(d) Let

$$\psi(x) = \begin{cases} 1, & x \in (0, 1/2), \\ -1, & x \in (1/2, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

be the Haar wavelet. Let $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ for $j, k \in \mathbb{Z}$. Show that $\psi_{1,0}(x)$ and $\psi_{1,1}(x)$ are both orthogonal to $\psi(x)$. Show that $||\psi_{j,k}|| = 1$ for all $j, k \in \mathbb{Z}$. (4 marks)

(e) Suppose $f(x) \in L^2(0, 1)$ is a function and define $f_i = f(i/n)$ for $i = 1, \ldots, n$ and some integer $n > 0$. Let $\{\epsilon_i\}_{i=1}^n$ be a set of independent and identically normally-distributed random variables with mean zero and variance of σ^2. Let $y_i = f_i + \epsilon_i$ for $i = 1, \ldots, n$. Let a discrete wavelet transform be represented by the orthogonal $n \times n$ matrix W. Let $w = Wy$, $d = Wf$ and $e = W\epsilon$, where y, d, ϵ are n-vectors consisting of the $\{y_i\}, \{f_i\}, \{\epsilon_i\}$ entries.

Prove that the e vector has zero mean and covariance matrix of $\sigma^2 I_n$. Explain that, with reference to Plancheral's theorem, why the signal-to-noise ratio in the wavelet-transformed sequence, w, is typically greater than that in the original sequence y. (4 marks)

(f) In density estimation the Epanechnikov (parabolic) kernel is given by

$$K(u) = \begin{cases} \frac{3}{4}(1 - u^2) & u \in (-1, 1), \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Figure 4 shows $\hat{f}(x)$, the Epanechnikov kernel density estimate of two data points $x_1 = 0, x_2 = 1.5$ realized from some continuous probability density function $f(x)$. The bandwidth of the density estimate is $h = 1$. What statistical task would the Epanechnikov kernel density estimate be unsuitable for? (3 marks)

(Total: 20 marks)

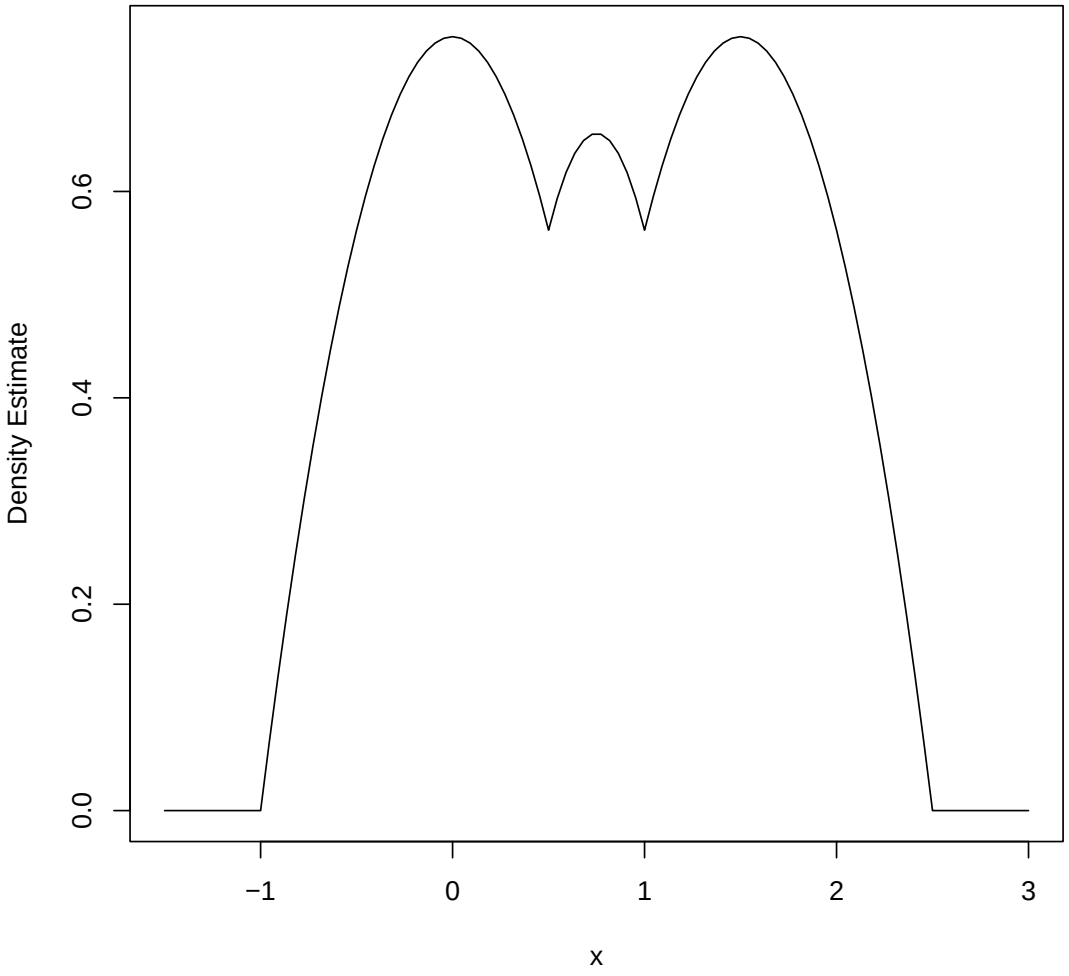

Figure 4: Epanechnikov density estimate formed from two data points $x_1 = 0, x_2 = 0$ with bandwidth $h = 1$.

4. Let X be an $n \times p$ data matrix where n, p are positive integers.

(a) Let $S = n^{-1}X^TX$. Define the principal components $\{v_j\}_{j=1}^p$ of S. Briefly explain what the first principal component is in terms of the point cloud X. (2 marks)

(b) Compare and contrast principal components regression and ridge regression. (1 mark)

(c) Suppose X is a 896610×9 data matrix, which contains the results of remote sensing an area of ground in the UK at nine difference frequency bands ranging from violet light to thermal infra-red. The correlation matrix of the remote sensing data is

Channel	2	3	4	5	6	8	9	10	11
2	1.00								
3	0.98	1.00							
4	0.98	0.99	1.00						
5	0.97	0.99	0.99	1.00					
6	0.36	0.45	0.34	0.43	1.00				
8	0.33	0.42	0.32	0.39	0.91	1.00			
9	0.79	0.85	0.80	0.83	0.65	0.70	1.00		
10	0.89	0.92	0.89	0.90	0.51	0.53	0.96	1.00	
11	0.75	0.79	0.75	0.77	0.46	0.46	0.87	0.89	1.00

A principal components analysis was carried out on the correlation matrix. The eigenvalues and percentage variance explained by each one is shown in the following table

No.	Eigenvalue	% Variance	No.	Eigenvalue	% Variance
1	6.88	76	6	0.032	0.36
2	1.50	17	7	0.014	0.15
3	0.387	4.3	8	0.006	0.07
4	0.130	1.4	9	0.001	0.02
5	0.057	0.63			

(i) Describe two interesting aspects of the correlation matrix. (1 mark)

(ii) Colour images are usually comprised of three colours, which corresponds to three projected dimensions. Looking at the table of eigenvalues, give your opinion over whether three dimensions would be enough to reasonably encapsulate the information present in the original nine-dimensional remote sensed image. (1 mark)

(iii) The first principal component vector corresponding to the correlation matrix above is

$$-(0.35, 0.37, 0.35, 0.36, 0.23, 0.23, 0.36, 0.37, 0.33)^T. \tag{3}$$

Give a practical interpretation for the first principal component. (Note: $1/\sqrt{9} = 1/3$). (2 marks)

(d) What is the difference between principal components analysis and exploratory projection pursuit? (1 mark)

(e) Define the F-divergence from p to q by

$$\mathcal{F}_F(p|q) = \int q(x) F\{p(x)/q(x)\}\, dx, \tag{4}$$

where p, q are probability density functions on $\mathbb{R}$ and $F : [0, \infty) \to \mathbb{R}$ is some strictly convex function. It can be shown that $\mathcal{F}_F(p|q) \geq F(1)$ for all densities p, q with equality if and only if $p = q$.

(i) If F is strictly convex, then so is $\tilde{F}(u) = uF(1/u)$ for all $u \in (0, \infty)$. Show that $\mathcal{F}_{\tilde{F}}(p|q) = \mathcal{F}_F(q|p)$. (2 marks)

(ii) Define the entropy $H(g)$ for a probability density function g. (1 mark)

(iii) Rewrite $H(g)$ as an F-divergence. Show that the standard normal distribution maximises the entropy amongst all probability densities with zero mean and unit variance (sphered). (7 marks)

(f) Let $t_3^s(x)$ be Student's t-distribution on three degrees of freedom that is scaled to have unit variance. Let $f(x)$ be a probability density function with mean zero and unit variance. Define the projection index

$$I_{t_3}(f) = \int_{-\infty}^{\infty} \{f(x) - t_3^s(x)\}^2\, dx, \tag{5}$$

and the projection index I_ϕ given by

$$I_\phi(f) = \int_{-\infty}^{\infty} \{f(x) - \phi(x)\}^2\, dx. \tag{6}$$

What kinds of projections does $I_{t_3}(f)$ favour compared to $I_\phi(f)$? (2 marks)

(Total: 20 marks)

5. Suppose we have a linear model $Y = X\beta + \epsilon$, where X is known, fixed, of full rank and dimension $n \times p$, Y is an observed n-vector, β is a p-dimensional vector of unknown parameters and ϵ is an n-vector of unknown, but assumed independent and identically distributed errors ϵ_i with mean zero and variance of σ^2.

The ridge regression estimator of β is given by

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T Y, \tag{7}$$

for smoothing parameter $\lambda > 0$.

(a) Show that the bias of $\hat{\beta}_{\text{ridge}}$ is given by

$$\begin{aligned} \text{bias}(\hat{\beta}_{\text{ridge}}) &= \mathbb{E}(\hat{\beta}_{\text{ridge}}) - \beta & (8) \\ &= \left\{(X^T X + \lambda I_p)^{-1} - (X^T X)^{-1}\right\} X^T X \beta. & (9) \end{aligned}$$

Prove that the bias is zero if and only if $\lambda = 0$. (3 marks)

(b) Prove that the ridge estimator can be written as a function of the least-squares estimator by

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T X \hat{\beta}, \tag{10}$$

where $\hat{\beta}$ is the usual least-squares estimator.

You are given (reminded) that $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$.

Hence, prove that the variance of the ridge estimator is given by

$$\text{var}(\hat{\beta}_{\text{ridge}}) = \sigma^2 (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1}. \tag{11}$$

(4 marks)

(c) Define $W = X^T X (X^T X + \lambda I_p)^{-1}$.

(i) Show that $\text{var}(\hat{\beta}_{\text{ridge}})$ can be written as

$$\text{var}(\hat{\beta}_{\text{ridge}}) = \sigma^2 W^T (X^T X)^{-1} W. \tag{12}$$

(1 mark)

(ii) Let DD be the difference between the variances of the least-squares estimator $\hat{\beta}$ and ridge estimator $\hat{\beta}_{\text{ridge}}$. Show that

$$\begin{aligned} DD &= \text{var}(\hat{\beta}) - \text{var}(\hat{\beta}_{\text{ridge}}) & (13) \\ &= \sigma^2 W^T \{(W^T)^{-1} (X^T X)^{-1} W^{-1} - (X^T X)^{-1}\} W. \end{aligned}$$

(2 marks)

(iii) Then, substituting for W (twice) in the internal brackets, show that

$$\begin{aligned} DD &= \text{var}(\hat{\beta}) - \text{var}(\hat{\beta}_{\text{ridge}}) && (14) \\ &= \sigma^2 W^T \{2\lambda (X^T X)^{-2} + \lambda^2 (X^T X)^{-3}\} W. && (15) \end{aligned}$$

(3 marks)

(iv) Now substituting for W twice in (15) show that

$$\begin{aligned} DD &= \text{var}(\hat{\beta}) - \text{var}(\hat{\beta}_{\text{ridge}}) && (16) \\ &= \sigma^2 (X^T X + \lambda I_p)^{-1} \{2\lambda I_p + \lambda^2 (X^T X)^{-1}\} (X^T X + \lambda I_p)^{-1}. \end{aligned}$$

Recall that X is of full rank and hence show that DD is a positive definite matrix. What implications does DD being positive definite have for ridge regression compared to least-squares regression? (5 marks)

(d) Suppose now that $p = 1$ and that the aim is to fit the univariate regression model $y_i = x_i \beta_1 + \epsilon_i$, for $i = 1, \ldots, n$ with all other assumptions as above. Briefly discuss whether it is (a) possible or (b) advisable to fit a ridge regression in this situation? (2 marks)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2023

This paper is also taken for the relevant examination for the Associateship.

MATH60049/70049

Intro. to Statistical Learning (Solns)

Setter's signature	Checker's signature	Editor's signature
.....	.....	.....

1. (a) A metric d satisfies the following

seen ⇓

Non-negativity $d(x,y) \geq 0$ and $d(x,y) = 0$ if $y = x$.
Symmetry $d(x,y) = d(y,x)$
Triangle inequality $d(x,y) + d(y,z) \geq d(x,z)$

3, A

(b) We show it for $q = 2$, for arbitrary n it just follows. Clearly $d(x,y) = d_1(x,y) + d_2(x,y) \geq 0$ as both d_1, d_2 are metrics for all x, y. Ditto:

seen ⇓

$$d(x,y) = d_1(x,y) + d_2(x,y) = d_1(y,x) + d_2(y,x) = d(y,x), \tag{1}$$

for all x, y as d_1, d_2 are symmetric. For the triangle inequality

$$\begin{aligned} d(x,y) + d(y,z) &= d_1(x,y) + d_2(x,y) + d_1(y,z) + d_2(y,z) && (2)\\ &= d_1(x,y) + d_1(y,z) + d_2(x,y) + d_2(y,z) && (3)\\ &\geq d_1(x,z) + d_2(x,z) = d(x,z), && (4) \end{aligned}$$

since both d_1, d_2 satisfy the triangle inequality.

3, A

seen ⇓

(c) The Hamming distance merely counts the number of mismatches in two binary strings.

1, A

Define $d_i(x,y) = 1$ if $x_i \neq y_i$ or 0 otherwise, where x_i, y_i are the ith digits of two binary strings x, y. Then the Hamming distance for two binary strings of length n is $d(x,y) = \sum_{i=1}^{n} d_i(x,y)$. Due to the result in part (b) Hamming is a metric if d_i is. So, we check the properties of a metric for d_i. Clearly $d_i(x,y) \geq 0$ for all x, y and $d_i(x,y) = d_i(y,x)$. For the triangle inequality, we can use this table:

x_i	y_i	z_i	$d_i(x,y)$	$d_i(y,z)$	Sum	$d_i(x,z)$
0	0	0	0	0	0	0
0	0	1	0	1	1	1
0	1	0	1	1	2	0
0	1	1	1	0	1	1
1	0	0	1	0	1	1
1	0	1	1	1	2	0
1	1	0	0	1	1	1
1	1	1	0	0	0	0

which enumerates all possibilities.

4, A

unseen ⇓

(d) The Hamming distance between $x = 10011110$ and $y = 10010101$ is 3.

1, A

sim. seen ⇓

(e) The Jaccard distance $d_J(x,y) = (b+c)/(a+b+c)$. The Jaccard distance does not take notice of d — this is because it is often irrelevant to known that both objects do not possess an attribute. For example, when comparing two archeological sites you will not be interested when both sites do not contain an Apple iPhone.

1, A

1, B

unseen ⇓

(f) (i) 'cake' → 'cak' → 'cat'. So, the Levenshtein distance is 2.

2, B

(ii) An upper bound is $\max(n_x, n_y)$ (as the smaller string, say, x, can overwrite the first n_x entries of y, and then delete the remaining $n_y - n_x$).

2, C

(iii) Yes. (E.g. just reverse the operations. If a smaller number of edits gets you there, then just reverse those to beat the initial set, but they can't be beaten, by definition).

2, D

2. (a) The bootstrap acquires an estimate of the sampling distribution of a statistic by computing that statistic on a resampled (with replacement) sample from the dataset and repeating this procedure many, B, times. The entire B resampled statistic values can be used to estimate the statistic's sampling distribution.

seen ⇓

3, A

(b) Let a single bootstrap sample B_b be a simple random sample, with replacement, from $x_1, \ldots, x_n$ of size n. Define the bth sample mean bootstrap estimate by $\hat{\mu}^{(b)} = |B_b|^{-1} \sum_{x_i \in B_b} x_i$. Then, we can estimate the sampling distribution of $\hat{\mu} - \mu$ by the sample (or CDF of) $\hat{\mu}^{(b)} - \hat{\mu}$, over $b = 1, \ldots, B$, where B is the number of bootstrap simulations.

seen/sim.seen ⇓

2, B

(c) CART stands for 'Classification and Regression Trees'. A classification tree is a tree constructed from data that enables the class membership of a new observation to be determined by dropping the observation through the tree and that observation following the left or right forks by a left/right decision based on comparing the new observation's data on a particular variable to a threshold. Those thresholds are computed in advance by an algorithm that recursively works out the (empirically) best splits on a sequence of variables. CART partitions feature space into a set of rectangles, and assigns a constant (or class membership) on each of those rectangles.

seen/sim.seen ⇓

2, A

2, B

Pros: simple and easily explained. Fast to compute.

2, A

Cons: Lack of continuity. Small changes in input data can result in very different trees. Inefficient sometimes (e.g. if class divides are not parallel to measured variable axes.

2, A

(d) There is a clear split between glass types $\{1, 2, 3\}$ and $\{5, 6, 7\}$ with the latter having considerably lower values on Magnesium content. This could be used in a tree to split those glass types.

unseen ⇓

2, C

(e) This is the lack of continuity property of regression trees. Tree construction is highly sensitive to input values. In deciding splitting criteria the algorithm works on a simple sum-of-squares formula and if two variables are ranked quite closely then small changes in the input data can choose on or the other to be selected — this causes quite a big change in the tree, where a decision node can be completely replaced by another variable.

seen/sim.seen ⇓

1, B

(f) Let's follow both trees with this explanatory variable values. Underlined value is the value on the variable of the new sample.

meth seen ⇓

Tree in figure 2 of the paper: Al: $\underline{1.05} < 1.42$ (go left); Ca: $\underline{8.24} < 10.48$ (go left); RI: $\underline{1.51756} > 1.517$ (go right) $\to$ 3.

1, B

Tree in figure 3 of paper: Ba: $0 < 0.335$ (go left); Al: $\underline{1.05} < 1.42$ (go left); Ca: $\underline{8.24} < 10.48$ (go left); RI: $\underline{1.51756} > 1.517$ (go right) $\to$ 3.

1, B

Both trees assign this new sample to type 3. So, the predicted bagged type is 3.

1, C

(g) Random forest is simply generate many trees each based on a bootstrap sample, but we also randomly select a set of variables to split on. Then the results are based on the aggregate of the modified bootstrap trees. This combination of information across many trees makes one single tree much less influential and, with many trees in the forest, similar decisions are made by many trees, especially if their underlying variables are important (or similar).

seen/sim.seen ⇓

1, B

3. (a) The space $L_2(\mathbb{R}) = \{f(x) : \int f^2(x)\,dx < \infty\}$. seen ⇓

1, A

(b) Let $\{V_j\}_{j\in\mathbb{Z}}$ be a collection of spaces satisfying: seen ⇓

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \tag{5}$$

and

$$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L_R(\mathbb{R}), \tag{6}$$

and

$$\bigcap_{j\in\mathbb{Z}} V_j = \{0\}, \tag{7}$$

and

$$f(x) \in V_j \implies f(2x) \in V_{j+1}, \forall j \in \mathbb{Z}, \tag{8}$$

and we also need

$$f(x) \in V_0 \implies f(x-k) \in V_0, \forall k \in \mathbb{Z}, \tag{9}$$

integer translates of a function in V_0 are also in V_0. 5, A

(c) (from homework) Let $f(x), g(x)$ be two functions with orthogonal series expansions of $f(x) = \sum_\nu f_\nu \xi_\nu(x)$ and $g(x) = \sum_\nu g_\nu \xi_\nu(x)$, where $\{\xi_\nu\}$ is some orthogonal basis for the space of functions we're considering. Define $F = \{f_\nu\}_\nu$ and similarly for G and the inner products on the original and transformed coefficients are $< f, g >= \int f(x)\overline{g(x)}\,dx$ and $< F, G >= \sum_\nu f_\nu \overline{g_\nu}$, respectively. seen ⇓

Then

$$\begin{aligned}
< f, g > &= \int f(x)\overline{g(x)}\,dx && (10)\\
&= \int \sum_\nu f_\nu \xi_\nu(x) \sum_\mu \overline{g_\mu}\,\overline{\xi_\mu(x)}\,dx && (11)\\
&= \sum_\nu \sum_\mu f_\nu \overline{g_\mu} \int \xi_\nu(x)\overline{\xi_\mu(x)}\,dx && (12)\\
&= \sum_\nu \sum_\mu f_\nu \overline{g_\mu}\delta_{\nu,\mu} && (13)\\
&= \sum_\nu f_\nu \overline{g_\nu} && (14)\\
&= < F, G >, && (15)
\end{aligned}$$

where $F = \{f_\nu\}_\nu$ is the coefficient set of all the f_ν, and similarly for G. This is Parseval's relation. Plancherel's theorem is 2, B

$$||f||^2 =< f, f >=< F, F >= ||F||_\nu^2, \tag{16}$$

where the first norm is a norm on functions and the second is a norm on the sequence space. 1, B

(d) For the first question

unseen ⇓

$$< \psi, \psi_{1,0} > = \int \psi(x)\psi_{1,0}(x)\,dx \tag{17}$$

$$= \int \psi(x)2^{1/2}\psi(2x)\,dx = 0, \tag{18}$$

because $\psi(2x)$ is constant on $(0, 1/4)$ and precisely negative that constant on $(1/4, 1/2)$ multiplied by the constant value of $\psi(x)$ on $(0, 1/2)$, so the two parts cancel.

2, C

For $\psi_{1,1}(x)$ the same is true, but the smaller scaled wavelet carries out its action on $(1/2, 1)$ instead.

1, C

For the norm

$$||\psi||^2 = \int \psi^2(x)\,dx \tag{19}$$

$$= \int_0^1 1\,dx = 1. \tag{20}$$

1, B

(e) Observe that $\mathbb{E}(e) = \mathbb{E}(W\epsilon) = W\mathbb{E}(\epsilon) = 0$. For the covariance $\mathrm{var}(e) = \mathrm{var}(W\epsilon) = W\,\mathrm{var}(\epsilon)W^T = \sigma^2 W I_n W^T = \sigma^2 W W^T = \sigma^2 I_n$.

seen/sim.seen ⇓

For the second part, wavelets are known sparsifiers. The vector d will, in general, be considerably more sparse than the original truth vector f. However, because of Parseval's theorem we also have $||d|| = ||f||$, which means, in general, that many of the wavelet 'signal' coefficients should be considerably larger than those in f. The Gaussian IID noise is invariant to orthogonal transformation and so the values in e and ϵ should be commensurate. Hence, the signal-to-noise ratio is improved after wavelet transformation.

2, B

2, D

(f) Detecting modes or 'bump hunting'.

unseen ⇓

3, D

4. (a) The eigendecomposition of $S = VDV^T$, where V are orthogonal and D is a diagonal matrix. The principal components are the eigenvectors, i.e. v_j is the jth column of V.

seen ⇓

1, A

The first principal component is the direction which maximises the projected variance of the point cloud over all directions.

1, A

(b) Ridge regression can be seen to shrink the contribution of principal directions, where the eigenvalues are small, whereas principal components regression omits contributions from principal directions where the eigenvalues are small. (mark positively anything reasonable. There might be more than one answer)

seen ⇓

1, B

unseen ⇓

(c) (i) There are many extremely high correlations. Some variables appear to be grouped. E.g. $\{2, 3, 4, 5, 10\}$ and $\{6, 8\}$ and possibly 9 and 11 less so. (mark positively anything reasonable. There might be more than one answer)

1, B

(ii) Yes. Even two dimensions encapsulates 93% of the variation and three dimensions over 97%, so three dimensions would be adequate.

seen/sim.seen ⇓

1, B

(iii) This vector weights all original variables roughly equally. All of the values are close to $1/3$. Practically, this corresponds to a 'brightness' variable, which is contributed to by all variables.

unseen ⇓

2, C

(d) Principal components selects ordered orthogonal directions based on maximising variance, whereas exploratory project pursuit maximises entropy or information.

seen ⇓

1, A

(e) (i) For the first part:

unseen ⇓

$$\begin{aligned}
\mathcal{F}_{\tilde{F}}(p|q) &= \int q(x)\tilde{F}\{p(x)/q(x)\}\,dx && (21)\\
&= \int \cancel{q(x)}^{1}\{p(x)/\cancel{q(x)}^{1}\}F\{q(x)/p(x)\}\,dx && (22)\\
&= \int p(x)F\{q(x)/p(x)\}\,dx && (23)\\
&= \mathcal{F}_F(q|p). && (24)
\end{aligned}$$

2, C

(ii) The entropy, $H(g)$, of density g is defined as

$$H(g) = -\int_{\mathbb{R}} g(x)\log\{g(x)\}\,dx \qquad (25)$$

1, A

(iii) For the third part, choose $F(u) = u\log(u)$, $p = f$ and $q = \phi$, the standard normal density. Then, [2, D]

$$\begin{aligned}
\mathcal{F}_F(p|q) &= \mathcal{F}_{u\log u}(f|\phi) && (26)\\
&= \int \phi(x) F\{f(x)/\phi(x)\}\, dx && (27)\\
&= \int \phi(x) \frac{f(x)}{\phi(x)} \log\{f(x)/\phi(x)\}\, dx && (28)\\
&= \int f(x)\log f(x)\, dx - \int f(x)\log\phi(x)\, dx && (29)\\
&= \int f(x)\log f(x)\, dx + \frac{1}{2}\log(2\pi) + \frac{1}{2}\int x^2 f(x)\, dx && (30)\\
&= \int f(x)\log f(x)\, dx + \frac{1}{2}\log(2\pi e). && (31)
\end{aligned}$$

Due to the properties of F-divergence it is the case that $\phi(x)$ minimises (26) and hence maximises the entropy. [5, D]

(f) The I_{t_3} index is measuring divergence of $f(x)$ from a heavier-tailed distribution when compared to I_ϕ and so the t-index will prefer heavy-tailed solutions less often than I_ϕ and, in that sense, the t-index is more robust than I_ϕ. [unseen ⇓] [2, D]

5. The ridge regression estimator in this setting is

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T Y. \tag{32}$$

It's assumed X is a fixed design. seen/sim.seen ⇓

(a) The expectation of $\hat{\beta}_{\text{ridge}}$ is

$$\mathbb{E}(\hat{\beta}_{\text{ridge}}) = (X^T X + \lambda I_p)^{-1} X^T X \beta. \tag{33}$$

Hence the bias of $\hat{\beta}_{\text{ridge}}$ is 2, M

$$\mathbb{E}(\hat{\beta}_{\text{ridge}}) - \beta = \left\{(X^T X + \lambda I_p)^{-1} - (X^T X)^{-1}\right\} X^T X \beta. \tag{34}$$

seen ⇓

Clearly, the bias is only zero if and only if $\lambda = 0$, as we return to the least-squares estimator. 1, M

(b) First write the ridge estimator as a function of the least-squares estimator: unseen ⇓

$$\begin{aligned} \hat{\beta}_{\text{ridge}} &= (X^T X + \lambda I_p)^{-1} X^T Y & (35)\\ &= (X^T X + \lambda I_p)^{-1} (X^T X)(X^T X)^{-1} X^T Y & (36)\\ &= (X^T X + \lambda I_p)^{-1} X^T X \hat{\beta}, & (37) \end{aligned}$$

where $\hat{\beta}$ is the usual least-squares estimator. 2, M

We are reminded in the question that $\text{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$. seen/sim.seen ⇓

Thus

$$\begin{aligned} \text{var}(\hat{\beta}_{\text{ridge}}) &= \text{var}\left\{(X^T X + \lambda I_p)^{-1} X^T X \hat{\beta}\right\} & (38)\\ &= (X^T X + \lambda I_p)^{-1} X^T X \,\text{var}(\hat{\beta}) \{(X^T X + \lambda I_p)^{-1} X^T X\}^T & (39)\\ &= (X^T X + \lambda I_p)^{-1} X^T X \sigma^2 (X^T X)^{-1} X^T X (X^T X + \lambda I_p)^{-1} & (40)\\ &= \sigma^2 (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1}. & (41) \end{aligned}$$

2, M

(c) Let $W = X^T X (X^T X + \lambda I_p)^{-1}$. seen/sim.seen ⇓

(i) Then

$$\begin{aligned} \text{var}(\hat{\beta}_{\text{ridge}}) &= \sigma^2 (X^T X + \lambda I_p)^{-1} (X^T X)(X^T X)^{-1} X^T X (X^T X + \lambda I_p)^{-1}\\ &= \sigma^2 W^T (X^T X)^{-1} W. & (42) \end{aligned}$$

1, M

(ii) The difference between the variance of the least squares and ridge estimators is given by

unseen ⇓

$$\begin{aligned} DD &= \mathrm{var}(\hat{\beta}) - \mathrm{var}(\hat{\beta}_{\mathsf{ridge}}) && (43)\\ &= \sigma^2(X^TX)^{-1} - \sigma^2 W^T(X^TX)^{-1}W && (44)\\ &= \sigma^2\left\{W^T(W^T)^{-1}(X^TX)^{-1}W^{-1}W - W^T(X^TX)^{-1}W\right\} && (45)\\ &= \sigma^2 W^T\left\{(W^T)^{-1}(X^TX)^{-1}W^{-1} - (X^TX)^{-1}\right\}W. && (46)\end{aligned}$$

2, M

(iii) Then for the next part substituting in for the value of W in the internal brackets gives

unseen ⇓

$$\begin{aligned} DD &= \mathrm{var}(\hat{\beta}) - \mathrm{var}(\hat{\beta}_{\mathsf{ridge}})\\ &= \sigma^2 W^T\left\{(X^TX)^{-1}(X^TX+\lambda I_p)\times\right.\\ &\quad \left.(X^TX)^{-1}(X^TX+\lambda I_p)(X^TX)^{-1} - (X^TX)^{-1}\right\}W && (47)\\ &= \sigma^2 W^T\left[\{I_p+\lambda(X^TX)^{-1}\}(X^TX)^{-1}\{I_p+\lambda(X^TX)^{-1}\}\right.\\ &\quad \left.-(X^TX)^{-1}\right]W && (48)\\ &= \sigma^2 W^T\left[\{(X^TX)^{-1}+\lambda(X^TX)^{-2}\}\{I_p+\lambda(X^TX)^{-1}\}\right.\\ &\quad \left.-(X^TX)^{-1}\right]W && (49)\\ &= \sigma^2 W^T\left\{(X^TX)^{-1}+\lambda(X^TX)^{-2}+\lambda(X^TX)^{-2}\right.\\ &\quad \left.+\lambda^2(X^TX)^{-3} - (X^TX)^{-1}\right\}W && (50)\\ &= \sigma^2 W^T\left\{2\lambda(X^TX)^{-2}+\lambda^2(X^TX)^{-3}\right\}W. && (51)\end{aligned}$$

3, M

(iv) Then expanding the W^T and W at the beginning and end gives

unseen ⇓

$$\begin{aligned} DD &= \mathrm{var}(\hat{\beta}) - \mathrm{var}(\hat{\beta}_{\mathsf{ridge}})\\ &= \sigma^2(X^TX+\lambda I_p)^{-1}X^TX\left\{2\lambda(X^TX)^{-2}+\lambda^2(X^TX)^{-3}\right\}\times\\ &\quad X^TX(X^TX+\lambda I_p)^{-1} && (52)\\ &= \sigma^2(X^TX+\lambda I_p)^{-1}\{2\lambda I_p+\lambda^2(X^TX)^{-1}\}(X^TX+\lambda I_p)^{-1}. && (53)\end{aligned}$$

If $\lambda > 0$ the last matrix is positive definite because for any $v \neq 0$ we have

2, M

$$z = (X^TX+\lambda I_p)^{-1}v \neq 0, \qquad (54)$$

as everything is of full rank. Thus

$$\begin{aligned} v^TDDv &= \sigma^2 z^T\{2\lambda I_p+\lambda^2(X^TX)^{-1}\}z && (55)\\ &= 2\sigma^2\lambda z^Tz + \sigma^2\lambda^2 z^T(X^TX)^{-1}z > 0. && (56)\end{aligned}$$

because X^TX and its inverse are positive definite. The implication is that the variance of ridge is always less than, or equal to, least squares regression in the sense that the variance of any one-dimensional projection is always less.

2, M

1, M

(d) It is possible — in that the equations can be computed and the computer can be used to compute everything (although, actually several implementations won't do it). There's not much point, though. This is because ridge regression is used typically when there is suspected co-linearity between variables and this cannot occur when only one variable is used.

unseen ⇓

2, M

Review of mark distribution:
Total A marks: 32 of 32 marks
Total B marks: 20 of 20 marks
Total C marks: 12 of 12 marks
Total D marks: 16 of 16 marks
Total marks: 100 of 80 marks
Total Mastery marks: 20 of 20 marks

If your module is taught across multiple year levels, you might have received this form for each level of the module. You are only required to fill this out once for each question.		
ExamModuleCode	**QuestionNumber**	**Comments for Students**
MATH60049/70049	1	Most students performed well or reasonably well.A surprising minority of students could not define the Hamming distance, which is particularly simple (part. c). Most students defined the Jaccard distance (part e) but often had difficulty in proposing a practice advantage.Part f was answered extremely well, in the main.
MATH60049/70049	2	Students found it tricky to explain the purpose of the bootstrap. Some explained the bootstrap itself, but not its purpose (part a)Probably about 30% of students answered this question well.A truly amazing number of students did not know what the abbreviation CART stood for (possibly 70%, a big surprise).Part d was answered very well - and there were many ways to answer it.Part e was mostly answered well.Most students understood what to do in part f, However, several students got caught out by the "RI >=1.517" condition, which is the opposite way around to the others. In fact, the solution was wrong in this respect too. However, marks were given if this mistake was made and then the student correctly worked out the bagged estimate from the incorrect answers from the trees.Most students had an idea of what a random forest was, and its purpose. However, few got into the specifics of how it was done, which attracted full marks.
MATH60049/70049	3	Most students could answer a, b and d relatively well.Only a few students seemed to answer part bMany students answered the first two bits of part (e) well, and most attempted the remainder of the part and sometimes did ok.Few students fully grasped and answered part (f), but a mark was usually identified if the student mentioned modes.
MATH60049/70049	4	Most students answered parts (a) and (b) reasonably well.In (c) part (i) the question was really after what data-related interesting aspects were present in the matrix. Correlation matrices are always symmetric and have 1s on the diagonal, so no marks were given for this.(ii) Many people answered this correctly, which was pleasing.(iii) A minority answered this question well and few understood the PC to be a `brightness' or overall level component.(d) was answered reasonably well by mosts(e) Parts (i) and (ii) were answered well by mostPart (iii) was partly answered well by many people, but details were left out or some people stated that 1 was a density (incorrect) just to try to make it work. Many students relied on the standard proof from lectures, rather than using the F-divergence itself.Few students understood this, or that the projection index was looking for projections AWAY from data with heavy tails (discourages solutions with outliers). Some students got a mark for identifying heavy-tailed.
MATH70049	5	Most students did parts (a), (b), (c) (I-iii) well.Students struggled a little on (c-iv) in showing the positive definiteness. Many students answered part (d) reasonably well, which was pleasing [certain on the `possible' bit]