

**BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)**  
**May-June 2021**

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

**Introduction to Statistical Learning**

Date: Monday, 24 May 2021

Time: 09:00 to 11:30

Time Allowed: 2.5 hours

Upload Time Allowed: 30 minutes

**This paper has 5 Questions.**

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

Each sheet of paper should have your CID, Question Number and Page Number on the top.

Only use 1 side of the paper.

Allow margins for marking.

Any required additional material(s) will be provided.

Credit will be given for all questions attempted.

Each question carries equal weight.

**SUBMIT YOUR ANSWERS ONE PDF TO THE RELEVANT DROPBOX ON BLACKBOARD  
INCLUDING A COMPLETED COVERSHEET WITH YOUR CID NUMBER, QUESTION  
NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.**

1. Suppose in a regression problem we assume that we have a response vector  $Y = (Y_1, \dots, Y_n)$ , a  $n \times p$  full-rank data matrix  $X$ , an arbitrary (but fixed)  $p$ -dimensional vector of parameters  $\beta$  linked by the following model

$$Y = X\beta + \epsilon, \quad (1)$$

where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)$  is a vector of independent and identically distributed errors with mean zero and variance of  $\sigma^2$ . Let  $\hat{\beta}$  be the least squares estimator of  $\beta$ .

- (a) Define what is meant by an *unbiased linear estimator*,  $\tilde{\beta}$ , of  $\beta$ . (1 mark)
- (b) Suppose  $\check{\beta} = CY$  is an unbiased linear estimator of  $\beta$  with  $C = (X^T X)^{-1} X^T + D$ , where  $D$  is a non-zero  $p \times n$  matrix. You are given that  $\mathbb{E}(\check{\beta}) = (I_p + DX)\beta$  and so  $\check{\beta}$  is unbiased if and only if  $DX = 0$ .
- State and prove the Gauss-Markov theorem for unbiased linear estimators. (4 marks)
- (c) Explain why it might be sensible to consider using biased estimators in some circumstances. (1 mark)
- (d) This part of the question is about the ridge regression estimator:

$$\hat{\beta}^{\text{ridge}} = (X^T X + \lambda I_p)^{-1} X^T Y, \quad (2)$$

which can be obtained as the solution of a penalized least squares problem.

- (i) Give an example of a situation where you might use ridge regression. (1 mark)
- (ii) Briefly explain (about five sentences) the Bayesian formulation of ridge regression. State the relationship between the ridge parameter,  $\lambda$ , the prior variance,  $\tau^2$ , of  $\beta$  and the error variance  $\sigma^2$  (do not work out the estimators explicitly).  
Explain in at most a few sentences the main difference between the Bayesian formulation of ridge regression and that obtained by solving the ridge penalised least squares problem. (5 marks)
- (iii) By looking at second derivatives or otherwise, show that the ridge regression estimator in (2) is at a minimum of the ridge regression objective function. (3 marks)

*Question 1 continued next page ...*

- (e) The employment agency *HardWorkerz* decides to see whether the results of its applicant testing programme can predict future salaries. *HardWorkerz* gave each of 67 applicants four tests Test1, Test2, Test3 and Test4 and collected information on their salaries three years later. The correlation matrix (to two decimal places) between the four Test variables was

|       | Test1 | Test2 | Test3 | Test4 |
|-------|-------|-------|-------|-------|
| Test1 | 1.00  | 0.87  | -0.01 | 0.84  |
| Test2 |       | 1.00  | -0.07 | 0.80  |
| Test3 |       |       | 1.00  | -0.04 |
| Test4 |       |       |       | 1.00  |

The salary information was put into the variable `Salary` and, along with the variables `Test1`, `Test2`, `Test3` and `Test4`, were put into a data frame called `the.df`. They fitted the following linear model in R using the command

```
lm(Salary ~ Test1, data=the.df)
```

and obtained the following table of results

Coefficients:

|                | Estimate | Std. Error | t value  | Pr(> t )     |
|----------------|----------|------------|----------|--------------|
| (Intercept)    | 13.90529 | 3.08668    | 4.505    | 2.83e-05 *** |
| Test1          | 0.72600  | 0.05787    | 12.546   | < 2e-16 ***  |
| <hr/>          |          |            |          |              |
| Signif. codes: | 0 ‘***’  | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’     |
|                | 0.1 ‘ ’  | 1          |          |              |

They then fitted the model

```
Salary ~ Test1 + Test2 + Test3 + Test4
```

and obtained the following results on the next page

*Question 1 continued next page ...*

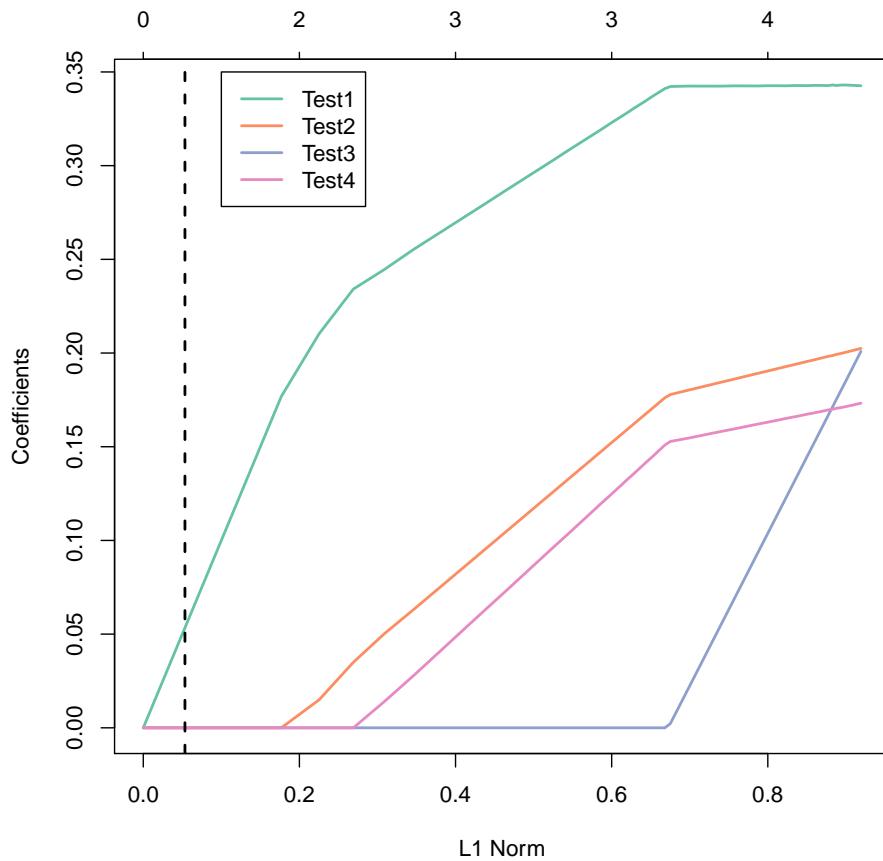


Figure 1: Lasso plot from *Hardworkerz* data set.

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )   |
|-------------|----------|------------|---------|------------|
| (Intercept) | 3.59968  | 11.34036   | 0.317   | 0.75199    |
| Test1       | 0.34228  | 0.12610    | 2.714   | 0.00859 ** |
| Test2       | 0.20430  | 0.09458    | 2.160   | 0.03464 *  |
| Test3       | 0.21527  | 0.22359    | 0.963   | 0.33940    |
| Test4       | 0.17494  | 0.08778    | 1.993   | 0.05068 .  |

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The graphical results from using the `glmnet` function with the lasso option, and all other arguments set to defaults, based on the possibility of including any of the four explanatory variables, are shown in Figure 1. Interpret the results of these analyses. (5 marks)

(Total: 20 marks)

End of Question 1

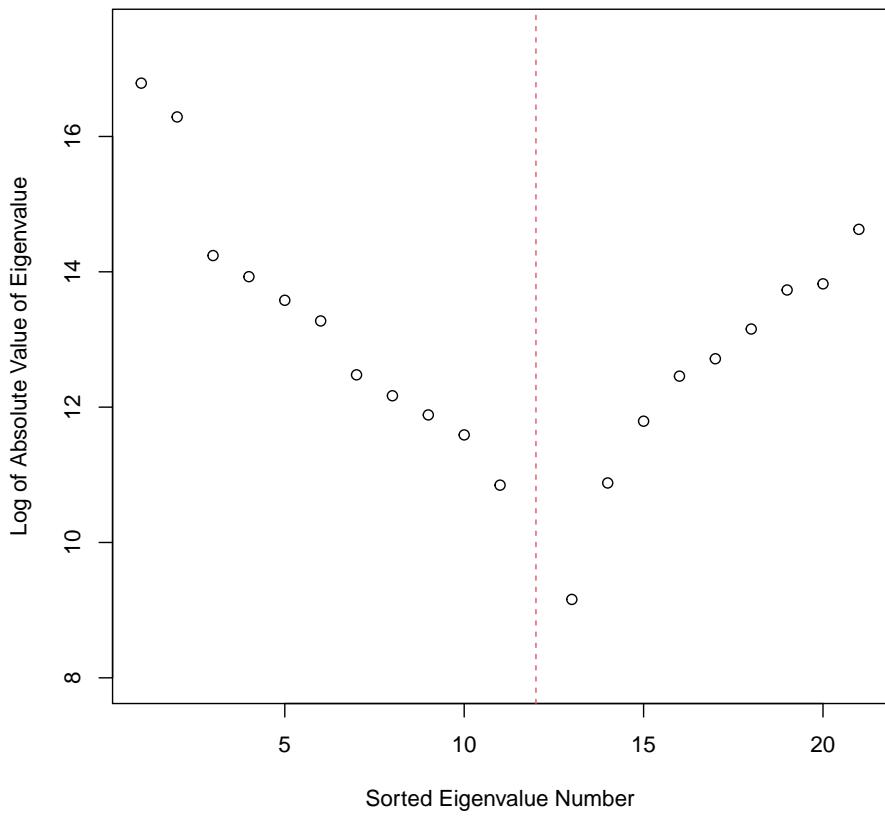


Figure 2: Eigenvalues from `cmdscale()` function on `eurodist`.

2. (a) (i) Explain how the Miles algorithm for least squares monotone linear regression works. (6 marks)
- (ii) Apply the Miles algorithm to the following sequence: 3, 9, 17, 56, 8, 17, 7, 22, 10 and show your working. (2 marks)
- (b) The eurodist dataset in R contains the distances between 21 European cities in km. The first four cities and their distances are shown:

|           | Athens | Barcelona | Brussels | Calais |
|-----------|--------|-----------|----------|--------|
| Barcelona |        | 3313      |          |        |
| Brussels  |        | 2963      | 1318     |        |
| Calais    |        | 3175      | 1326     | 204    |
| ...       |        |           |          |        |

Classical multidimensional scaling was carried out using the `cmdscale()` function in R and the corresponding eigenvalue plot was produced as in Figure 2.

*Question 2 continued next page ...*

- (i) For the general case, show how the eigenvalues, such as those in Figure 2, arise from  $E$ , a matrix of Euclidean distances [Hint: you are given that the Euclidean distances arise from some (unknown)  $n \times p$  centred configuration  $X$ , with corresponding inner product matrix  $B = XX^T$ , the recovered configuration will be centred, and the formula  $e_{m,\ell} = b_{m,m} + b_{\ell,\ell} - 2b_{m,\ell}$ .] (6 marks)
- (ii) The red dotted line in Figure 2 indicates where one of the eigenvalues is precisely zero. In classical scaling, why is (at least) one of the eigenvalues always zero? (1 mark)
- (iii) Using the eigenvalue plot in Figure 2 what dimensionality would you recommend that the scaling solution should be presented in? (1 mark)
- (c) *Jenks' natural breaks optimization* is the one-dimensional version of  $k$ -means clustering. Suppose you have four one-dimensional datapoints  $x_1 = 1, x_2 = 2 + \epsilon, x_3 = 3$  and  $x_4 = 4$ , where  $\epsilon > 0$  and  $\epsilon$  is considerably smaller than 1. You are told that the true number of clusters  $k = 2$ . Assume that one cluster centre is  $m_1 = x_1 = 1$  and demonstrate that, depending on what points are initially used as the second cluster centre, Jenks' algorithm can only produce two possible different allocations of points to clusters. List the possible clusters and the associated cluster means for each of these two different allocations. (4 marks)

(Total: 20 marks)

End of Question 2

3. (a) Suppose we have a sample of  $n$  identically and independently-distributed observations,  $X_1, \dots, X_n$  from probability density function  $f : \mathbb{R} \rightarrow [0, \infty)$ . Suppose we estimate  $f(x)$  using the kernel density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right), \quad (3)$$

where  $K(y)$  is the *triangular* kernel function defined by

$$K(y) = \begin{cases} 1 - |y|, & \text{for } y \in (-1, 1), \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Suppose that  $f(x)$  is twice-differentiable with continuous second derivative and that all kernel functions below are symmetric about zero. You are given that the expectation of  $\hat{f}(x)$  is

$$\mathbb{E}\{\hat{f}(x)\} = f(x) + \frac{1}{2}C_3 h^2 f''(x) + \mathcal{O}(h^3), \quad (5)$$

where  $C_3 = \int_{-\infty}^{\infty} v^2 K(v) dv$ .

This question is about deriving an expression for the variance of  $\hat{f}(x)$ .

- (i) Show that

$$\mathbb{E}\{\hat{f}(x)\} = \frac{1}{h} \mathbb{E}\left\{K\left(\frac{X_i - x}{h}\right)\right\}, \quad (6)$$

and

$$\text{var}\{\hat{f}(x)\} = \frac{1}{nh^2} \text{var}\left\{K\left(\frac{X_i - x}{h}\right)\right\}. \quad (7)$$

(1 mark)

- (ii) Compute the value of  $C_3$  for the triangular kernel function. (2 marks)

- (iii) Show that, for the triangular kernel function,

$$\mathbb{E}\left\{K^2\left(\frac{X_i - x}{h}\right)\right\} = \frac{2h}{3}f(x) + \frac{h^3}{30}f''(x) + \mathcal{O}(h^4), \quad (8)$$

(8 marks)

and thence

$$\text{var}\{\hat{f}(x)\} = \frac{2}{3nh}f(x) - \frac{1}{n}f^2(x) + \mathcal{O}(h/n). \quad (9)$$

(3 marks)

*Question 3 continued next page ...*

- (b) By properly balancing the squared bias and variance of the kernel density estimator above the optimal bandwidth *can be* shown to be

$$h_{\text{opt}} = \left[ \frac{8f(x)}{3n \{f''(x)\}^2} \right]^{1/5}. \quad (10)$$

[You *do not* have to show this].

Explain why the optimal bandwidth given in (10) is not of direct practical use. (1 mark)

- (c) Let  $f : \mathbb{R} \rightarrow [0, \infty)$  be a probability density function. Suppose that  $\{\psi_{j,k}(x)\}_{j,k \in \mathbb{Z}}$  is an orthonormal wavelet basis of  $L^2(\mathbb{R})$ , with associated father wavelets,  $\{\phi_{j,k}(x)\}_{j,k \in \mathbb{Z}}$ , and that we can write, for some integer  $L > 0$ ,

$$f(x) = \sum_{k \in \mathbb{Z}} c_{L,k} \phi_{L,k}(x) + \sum_{j \in \mathbb{J}_L} \sum_{k \in \mathbb{Z}} d_{j,k} \psi_{j,k}(x), \quad (11)$$

where  $\mathbb{J}_L = \{m \in \mathbb{Z} : m \geq L\}$ ,  $c_{j,k} = \int_{\mathbb{R}} f(x) \phi_{j,k}(x) dx$  and  $d_{j,k} = \int_{\mathbb{R}} f(x) \psi_{j,k}(x) dx$ , where

$$\phi_{j,k}(x) = 2^{j/2} \phi(2^j x - k) \text{ and } \psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k), \quad (12)$$

for  $j \in \mathbb{J}_L, k \in \mathbb{Z}$ .

- (i) Suppose that we obtain an independent and identically-distributed sample  $X_1, \dots, X_n$  from  $f(x)$ . Explain why a reasonable estimator of  $d_{j,k}$  is  $\hat{d}_{j,k} = n^{-1} \sum_{i=1}^n \psi_{j,k}(X_i)$ . Let  $\hat{c}_{j,k}$  be the equivalent estimator of  $c_{j,k}$  using  $\phi_{j,k}$  rather than  $\psi_{j,k}$ . (3 marks)
- (ii) Suppose we want to compute  $\hat{d}_{j,k}$  and  $\hat{c}_{j,k}$  up to some high-resolution (integer) level  $J > L$ . Explain why it is more computationally efficient to use the discrete wavelet transform than to calculate all of the  $\hat{d}_{j,k}$  and  $\hat{c}_{j,k}$  directly using the formulae in part (i)? (2 marks)

(Total: 20 marks)

End of Question 3

4. (a) State three examples of additive basis expansions that are used in statistical learning. (1 mark)
- (b) As part of a statistical learning project a researcher decides to use the `circle_data()` function to generate synthetic data. The circle data function depends on two arguments: an inner limit  $i_r$  and an outer limit  $o_r$ . A sample from the `circle_data()` function involves drawing two random variables,  $X_1, X_2$ , each from a uniform distribution on  $[-o_r, o_r]$ . Then, a Bernoulli response  $Y$  is drawn according to a conditional probability mass function based on the realization  $X_1 = x_1, X_2 = x_2$  as follows:

$$\mathbb{P}(Y = 1|x) = \begin{cases} 1 & \text{if } \|x\| \leq i_r, \\ 1 & \text{if } \|x\| \geq o_r, \\ (o_r - \|x\|)/(o_r - i_r) & \text{otherwise,} \end{cases} \quad (13)$$

where  $\|x\| = (x_1^2 + x_2^2)^{1/2}$  is the usual two-norm.

- (i) Draw a sketch diagram of the conditional probability function (13). (2 marks)
- (ii) Linear discriminant analysis is a technique that draws a line in the  $(x_1, x_2)$  space and allocates  $y = 1$  to points  $(x_1, x_2)$  that are one side of the line and  $y = -1$  to points on the other side. Why would linear discriminant analysis not be a useful method for constructing a classifier for  $X$  in this example? (1 mark)
- (iii) Suppose it is decided to use a classification tree to classify circle data observations in a supervised learning experiment. Explain what is meant by training and test sets and the reason for their use in evaluating classifiers. (3 marks)
- (iv) A *regression* tree is defined by partitioning the space (two-dimensional in the case of the circle data) into  $M$  disjoint regions  $R_1, \dots, R_M$  and then forming the model

$$f(x) = \sum_{m=1}^M c_m \mathbb{I}(x \in R_m), \quad (14)$$

where  $\mathbb{I}(A) = 1$  if  $A$  is true, or 0 otherwise.

Suppose we consider splitting on variable  $j$  at split point  $s$  (somewhere on  $X_j$ ) and define the pair of half-planes

$$R_1(j, s) = \{x | X_j < s\} \quad \text{and} \quad R_2(j, s) = \{x | X_j > s\}. \quad (15)$$

Then, we seek the splitting variable  $j$  and split point  $s$  that solves:

$$\min_{j,s} \left\{ \min_{c_1} \sum_{X_i \in R_1(j,s)} (Y_i - c_1)^2 + \min_{c_2} \sum_{X_i \in R_2(j,s)} (Y_i - c_2)^2 \right\}. \quad (16)$$

For a given  $j, s$  derive the inner minimisers in (16) over  $c_1$  and  $c_2$  and label them  $\hat{c}_1, \hat{c}_2$  respectively. (2 marks)

*Question 4 continued next page ...*

- (v) For a given variable  $j$  explain how to rapidly compute the best split point  $s$  for the following objective function:

$$\sum_{X_i \in R_1(j,s)} (Y_i - \hat{c}_1)^2 + \sum_{X_i \in R_2(j,s)} (Y_i - \hat{c}_2)^2. \quad (17)$$

(2 marks)

- (vi) Suppose that we now choose to split variables into three regions with two split points  $s_1, s_2$ . Define a suitable new objective function and how to minimize it given a variable  $j$ . Give a reason why the option of splitting into more than two regions for each variable is not usually done. (3 marks)

- (c) (i) Briefly explain the rationale behind the AdaBoost.M1 classifier. (2 marks)  
(ii) The AdaBoost.M1 binary classifier can be written as

$$f(x) = \sum_{m=1}^M \beta_m G_m(x; \gamma_m), \quad (18)$$

where  $G_m(x, \gamma_m)$  is one of the classifiers and  $G_1(x; \gamma_m)$  is the initial ‘weak’ classifier, where  $G_i(x) \in \{-1, 1\}$  and  $\gamma_m$  are parameters of the model. The fitted model can be obtained by

$$\min_{\{\beta_m, \gamma_m\}_{m=1}^M} \sum_{i=1}^N L \left\{ y_i, \sum_{m=1}^M \beta_m G(x_i, \gamma_m) \right\}, \quad (19)$$

where  $L\{y, f(x)\}$  is the loss function between observation  $y$  and model  $f(x)$ . We assume that the costs of misclassification are equal.

AdaBoost.M1 can be shown to use the loss function  $L\{y, f(x)\} = \exp\{-yf(x)\}$ . Show that the population minimiser is given by

$$f^*(x) = \arg \min_{f(x)} \mathbb{E}_{Y|x} \{e^{-Yf(x)}\} = \frac{1}{2} \log \left\{ \frac{\mathbb{P}(Y = 1|x)}{\mathbb{P}(Y = -1|x)} \right\}, \quad (20)$$

one half of the log-odd ratio  $\mathbb{P}(Y = 1|x)$ . Explain why this means that the sign operator applied to  $\{f(x)\}$  from (18) is the appropriate function to use to build the operational classifier. (4 marks)

(Total: 20 marks)

End of Question 4

5. (a) This part considers the case of multiple linear regression model

$$Y = X\beta + \epsilon, \quad (21)$$

where  $\beta$  is a  $p$ -vector of unknown parameters,  $X$  is an  $n \times p$  matrix of explanatory variables,  $Y$  is an  $n$ -vector dependent variable and  $\epsilon$  is an  $n$ -vector of errors where  $\mathbb{E}(\epsilon_i) = 0$ ,  $\text{var}(\epsilon_i) = \sigma^2$  and the errors are all assumed independent and identically distributed.

Suppose we decide to estimate the parameters  $\beta$  using ridge regression with ridge parameter  $\lambda$ . In this context, explain the concept of cross-validation as a technique for choosing a good value of  $\lambda$  and briefly discuss its pros and cons. Explain what  $k$ -fold cross-validation is. What effect might correlation between errors in the model have on the outcome of cross-validation? What effect might outliers have on the outcome?

(10 marks)

- (b) Now suppose we have the model

$$y_i = f(t_i) + \epsilon_i, \quad (22)$$

for  $i = 1, \dots, n$ , where  $f$  is a smooth function  $f : \mathbb{R} \rightarrow \mathbb{R}$ ,  $\{y_i\}$  are observations taken at the points  $\{t_i\}$  and  $\{\epsilon_i\}$  is a set of independent and identically distributed random variables with mean zero and variance of  $\sigma^2$ . Define the full spline smoothing estimator of  $f(t)$  for the whole dataset to be  $\hat{f}_\lambda(t)$ .

In spline smoothing the *ordinary cross-validated* estimate of error is defined by

$$\text{OCV}(\lambda) = n^{-1} \sum_{i=1}^n \{y_i - \hat{f}_\lambda^{-i}(t_i)\}^2. \quad (23)$$

- (i) Explain what  $\hat{f}_\lambda^{-i}$  is. (2 marks)
- (ii) It can be shown that the full spline smoothing estimator can be written as

$$\hat{f}_\lambda = H(\lambda)y, \quad (24)$$

where  $\hat{f}_\lambda = \{\hat{f}_\lambda(t_1), \dots, \hat{f}_\lambda(t_n)\}^T$  and  $y = (y_1, \dots, y_n)^T$  are both  $n$ -vectors and  $H(\lambda)$  is the  $n \times n$  (hat) matrix depending on  $\lambda$ .

Now define the augmented data set  $\tilde{y}_i = (\tilde{y}_{i,1}, \tilde{y}_{i,2}, \dots, \tilde{y}_{i,n})^T$  for  $i = 1, \dots, n$  where

$$\tilde{y}_{i,j} = \begin{cases} y_j & j \neq i, \\ \hat{f}_\lambda^{-i}(t_i) & \text{for } j = i, \end{cases} \quad (25)$$

for  $j = 1, \dots, n$ . Let  $\tilde{f}_\lambda^{-i}$  be the spline smooth estimator based on the  $\tilde{y}_i$  augmented data set (of  $n$  points).

*Question 5 continued next page ...*

You are given the fact that

$$\tilde{f}_\lambda^{-i}(t_i) = \hat{f}_\lambda^{-i}(t_i), \quad (26)$$

for all  $i = 1, \dots, n$ .

Show that  $\text{OCV}(\lambda)$  can be written alternatively as

$$\text{OCV}(\lambda) = n^{-1} \sum_{i=1}^n \left\{ \frac{y_i - \hat{f}_\lambda(t_i)}{1 - h_{i,i}(\lambda)} \right\}^2, \quad (27)$$

where  $h_{i,j}(\lambda) = \{H(\lambda)\}_{i,j}$ . (6 marks)

- (iii) What is the advantage of using formula (27) instead of (23)? (2 marks)

(Total: 20 marks)

End of Question 5

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2021

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

MATH96067/MATH97287

Introduction to Statistical Learning (Solutions)

Setter's signature

.....

Checker's signature

.....

Editor's signature

.....

1. (a) An estimator  $\tilde{\beta}$  is unbiased if  $\mathbb{E}(\tilde{\beta}) = \beta$  and it is linear if it can be written as a linear combination of  $\{Y_i\}_{i=1}^n$ .
- (b) First, they should show

seen ↓

1, A

seen ↓

$$\begin{aligned}
 \text{var}(\check{\beta}) &= \text{var}(CY) \\
 &= C \text{var}(Y)C^T \\
 &= \sigma^2 CC^T \\
 &= \sigma^2 \{(X^T X)^{-1} X^T + D\} \{(X^T X)^{-1} X^T + D\}^T \\
 &= \sigma^2 \{(X^T X)^{-1} X^T + D\} \{X(X^T X)^{-1} + D^T\} \\
 &= \sigma^2 \{(X^T X)^{-1} X^T X(X^T X)^{-1} + DX(X^T X)^{-1} \\
 &\quad + (X^T X)^{-1} X^T D^T + DD^T\} \\
 &= \sigma^2 \{(X^T X)^{-1} + DD^T\} \\
 &= \text{var}(\hat{\beta}) + \sigma^2 DD^T,
 \end{aligned}$$

since  $\check{\beta}$  is unbiased and  $DX = 0$ .

Now examine an arbitrary linear combination of parameters:  $\theta = \alpha^T \beta$  for some  $p$ -vector  $\alpha$ .

Let  $\hat{\theta} = \alpha^T \hat{\beta}$  and  $\check{\theta} = \alpha^T \check{\beta}$ .

Clearly, both  $\hat{\theta}$ ,  $\check{\theta}$  are unbiased for  $\theta$ .

Now

$$\begin{aligned}
 \text{var}(\alpha^T \check{\beta}) &= \alpha^T \text{var}(\check{\beta}) \alpha \\
 &= \alpha^T \{\text{var}(\hat{\beta}) + \sigma^2 DD^T\} \alpha \\
 &= \text{var}(\alpha^T \hat{\beta}) + \sigma^2 \alpha^T DD^T \alpha.
 \end{aligned}$$

Now  $\alpha^T DD^T \alpha = v^T v = \sum_{i=1}^n v_i^2 \geq 0$ , or recognise  $DD^T$  is positive semi-definite, where  $v$  is some vector  $v = D^T \alpha$ .

Gauss Markov means  $\text{var}(\alpha^T \hat{\beta}) \leq \text{var}(\alpha^T \check{\beta})$ .

And recall  $\check{\beta}$  was arbitrary linear unbiased estimator. So,  $\hat{\beta}$  is best linear unbiased estimator (BLUE).

- (c) Fundamentally, we're interested in estimators with good mean-squared error performance. Since mean-squared error equals the sum of the squared bias and variance, we might wish to have an estimator that has some bias, but much reduced variance and, hence, overall smaller mean-squared error.

4, A

part seen ↓

1, B

part seen ↓

1, B

part seen ↓

2, B

1, C

- (d) (i) For example, where the explanatory variables are highly correlated, which translates into an  $(X^T X)^{-1}$  is ill-conditioned.

- (ii) The Bayesian formulation begins by putting a prior on the parameters of interest  $p(\beta)$  for  $\beta$ , this is a  $p$ -dimensional distribution. Then one collects data and can form a likelihood expression  $p(Y|\beta)$ , which stems from the model  $Y = X\beta + \epsilon$  in this case. Then, we can use Bayes' theorem to find a posterior distribution by  $p(\beta|Y) = p(Y|\beta)p(\beta)$ . The relationship between  $\lambda, \tau^2, \sigma^2$  is  $\lambda = \sigma^2/\tau^2$ .

The fundamental Bayesian formulation does not explicitly optimise anything as it is obtaining a posterior distribution, which obtains a whole distribution and not a single parameter (vector). Of course, when moving to a point estimate from a distribution, some kind of optimisation is often done.

unseen ↓

2, D

part seen ↓

- (iii) The ridge regression objective function in matrix form is (from lecture notes)

$$\text{RSS}(\beta) = (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta. \quad (1)$$

The first derivative, again from lecture notes is

$$\frac{\partial \text{RSS}(\beta)}{\partial \beta} = -2X^TY + 2(X^TX + \lambda I)\beta. \quad (2)$$

Now taking second derivative gives

1, A

$$2(X^TX + \lambda I), \quad (3)$$

and this matrix can be shown to be positive-definite by

$$2z^TX^TXz + \lambda z^Tz > 0, \quad (4)$$

for  $z \neq 0$  and since  $X^TX$  is positive semi-definite.

2, D

meth seen ↓

- (e) There is high correlation between variables Test1, Test2, Test4, but not with Test3 with the others. So, looks like co-linearity is present.

Then, when fitting a model with just Test1: it is highly significant. But when adding the other variables the significance of Test1 is smaller (p-value larger!) and Test 4 is not significant at the 5% level, when perhaps it should be

This is reflected in the `glmnet` Lasso analysis, which indicates by cross-validation that only one variable would go into the regression (Test1 and a shrunk version at that (e.g. 0.05 rather than the full LS value of 0.35. This fits with the highly-correlated nature of the variables above.

5, C

2. (a) (i) Given a set of  $\{y_i\}_{i=1}^M$  the Miles algorithm works by

seen  $\downarrow$

1. Write down the  $y_i$ s in singleton blocks.
2. Scanning from left to right, unite any two blocks where there is not a strict increase from left to right (uniting means forming the mean of all elements in both blocks, merging the blocks, keeping the same number of entries, but each entry replaced by the mean).
3. Keep checking until all boundaries are satisfied.

6, A

(ii) Let's do Miles on the following data — with unsatisfied boundaries in red.

meth seen  $\downarrow$

$$|3|9|17|56|8|17|7|22|10| \quad (5)$$

first stage:

$$|3|9|17|\color{red}{32} \color{red}{32} |17|7|22|10| \quad (6)$$

$$|3|9|17|\color{red}{27} \color{red}{27} \color{red}{27}|7|22|10| \quad (7)$$

$$|3|9|17|\color{red}{22} \color{red}{22} \color{red}{22} \color{red}{22}|22|10| \quad (8)$$

$$|3|9|17|\color{red}{22} \color{red}{22} \color{red}{22} \color{red}{22} \color{red}{22}|10| \quad (9)$$

$$|3|9|17|20 \ 20 \ 20 \ 20 \ 20 \ 20| \quad (10)$$

2, B

(b) (i) Suppose that there is an underlying configuration  $X$  which is  $n$  observations on  $p$  variables. Define the inner product matrix  $B = XX^T$  and Euclidean distances can be obtained by the given formula

$$e_{m,\ell} = b_{m,m} + b_{\ell,\ell} - 2b_{m,\ell}. \quad (11)$$

We need to reverse the steps to get  $B$  from  $E$  and then something like  $X$  from  $B$ . First, summing over  $m$  gives us ( $e_{\bullet,\ell} = \sum_{m=1}^n e_{m,\ell}$ )

$$e_{\bullet,\ell} = b_{\bullet,\bullet} + nb_{\ell,\ell} - 2b_{\bullet,\ell}. \quad (12)$$

Since  $\mathbf{1}$  is an eigenvector of  $B^T$ , it is also of  $B$  (due to construction of  $B = XX^T$ , with centred  $X$ ), hence row and column sums of  $B$  are zero. Hence,  $b_{\bullet,\ell} = b_{m,\bullet} = 0$  and

$$e_{\bullet,\ell} = b_{\bullet,\bullet} + nb_{\ell,\ell}. \quad (13)$$

Similarly,

$$e_{m,\bullet} = b_{\bullet,\bullet} + nb_{m,m}. \quad (14)$$

Summing over  $m$  and  $n$  gives

$$e_{\bullet,\bullet} = nb_{\bullet,\bullet} + nb_{\bullet,\bullet} = 2nb_{\bullet,\bullet}. \quad (15)$$

Now rearrange (11) to give

$$b_{m,\ell} = -\frac{1}{2}(e_{m,\ell} - b_{m,m} - b_{\ell,\ell}), \quad (16)$$

and using (13) and (14) we have

$$b_{m,\ell} = -\frac{1}{2} \{e_{m,\ell} - (e_{m,\bullet} - b_{\bullet,\bullet})/n - (e_{\bullet,\ell} - b_{\bullet,\bullet})/n\} \quad (17)$$

$$= -\frac{1}{2}(e_{m,\ell} - \frac{e_{m,\bullet}}{n} - \frac{e_{\bullet,\ell}}{n} + 2\frac{b_{\bullet,\bullet}}{n}) \quad (18)$$

$$= -\frac{1}{2}(e_{m,\ell} - \frac{e_{m,\bullet}}{n} - \frac{e_{\bullet,\ell}}{n} + \frac{e_{\bullet,\bullet}}{n^2}) \quad (19)$$

$$= -\frac{1}{2}(e_{m,\ell} - \frac{e_{m,\bullet}}{n} - \frac{e_{\bullet,\ell}}{n} + \frac{e_{\bullet,\bullet}}{n^2}) \quad (20)$$

$$= -\frac{1}{2}(\text{entry} - \text{row av.} - \text{col av.} + \text{grand av.}), \quad (21)$$

or in matrix terms

$$B = -\frac{1}{2}(I_n - \mathbf{1}\mathbf{1}^T/n) E (I_n - \mathbf{1}\mathbf{1}^T/n). \quad (22)$$

Since  $B = XX^T$  it is positive semi-definite and symmetric and hence would be diagonalizable. The  $B$  we get from  $E$  should be more or less like this. So, we use an eigendecomposition

$$B = \sum_{i=1}^n \lambda_i e^{(i)} e^{(i)T}, \quad (23)$$

where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n'} > 0$  and  $\lambda_{n'+1}, \dots, \lambda_n = 0$ .

- (ii) At least one of the eigenvalues has to be precisely zero as  $\mathbf{1}$  is an eigenvector of  $B$  with eigenvalue of 0.
- (iii) From the plot, it looks like two dimensions is suitable because the first two eigenvalues have values that are considerably larger than the others. [Or, those two are bigger than the smaller ones on the RHS of the plot].

6, A

seen ↓

1, B

sim. seen ↓

1, C

unseen ↓

- (c) We assume that  $m_1 = x_1 = 1$  as directed by the question. So, we need to work out what happens when the other cluster centre is taken to be each of the points  $x_2, x_3, x_4$ . Let's do the first one in a bit more detail.

Assignment:  $m_2 = x_2 = 2 + \epsilon$ . So, initially cluster 1 has point  $x_1$  and cluster 2 has point  $x_2 = 2 + \epsilon$ . Points  $x_3, x_4$  are closer to  $m_2$  so, in the first step, they go into cluster 2. Then, we iterate. The mean of the first cluster is 1. The mean of the second cluster is  $m_2 = (2 + \epsilon + 3 + 4)/3 = 3 + \epsilon/3$ . Clearly, point 1 goes into cluster 1 and points 3, 4 go into cluster 2. What about point  $2 + \epsilon$ ? This is nearer to  $m_2 = 3 + \epsilon/3$  (distance  $1 - 2\epsilon/3$ ) than  $m_1 = 1$  (distance  $1 + \epsilon$ ). So, point  $x_2$  remains in cluster 2 and since point allocation has not changed, this is the end for this run.

In summary: cluster 1 contains  $x_1$  with  $m_1 = 1$ . Cluster 2 contains  $x_2, x_3, x_4$  with  $m_2 = 3 + \epsilon_3$ .

Assignment:  $m_2 = x_3 = 3$ . So, initially, cluster 1 has point  $x_1$  and cluster 2 has point  $x_3$ . First allocation for remaining points: point  $x_4 = 4$  clearly to cluster 2. And, since  $x_2 = 2 + \epsilon$  is closer to 3 than 1 it also goes into cluster 2. We then stop, and the answer is the same as the previous run (since the allocations are now the same). So:

In summary: cluster 1 contains  $x_1$  with  $m_1 = 1$ . Cluster 2 contains  $x_2, x_3, x_4$  with  $m_2 = 3 + \epsilon_3$ .

Assignment:  $m_2 = x_4 = 4$ . So, initially, cluster 1 has point  $x_1$  and cluster 2 has point  $x_4$ . Now  $x_3 = 3$  is nearer to 4 than 1, so it goes into cluster 2. However,  $x_2 = 2 + \epsilon$  is nearer to 1 than 4, so it goes into cluster 1. So, cluster 1 contains  $x_1, x_2$  and cluster 2 contains  $x_3, x_4$  with respective cluster means of  $m_1 = 1.5 + \epsilon/2$  and  $m_2 = 3.5$ . It's easy to see that points  $x_1, x_3$  and  $x_4$  stay where they are. For  $x_2$  it is  $2 + \epsilon - 1.5 - \epsilon/2 = 0.5 + \epsilon$  away from  $m_1$  and  $2 + \epsilon - 3.5 = -1.5 + \epsilon$  away from  $m_2$ , so it stays in cluster 1. So:

In summary: cluster 1 contains  $x_1, x_2$  with  $m_1 = 1.5 + \epsilon/2$ . Cluster 2 contains  $x_3, x_4$  with  $m_2 = 3.5$ .

Hence, overall, there are two possible clusterings, depending on how we initialize the cluster means.

1, A

2, B

1, C

3. (a) Suppose we have a sample of  $n$  observations,  $X_1, \dots, X_n$  from probability density function  $f : \mathbb{R} \rightarrow [0, \infty)$  and we estimate  $f(x)$  using the kernel density estimator

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right), \quad (24)$$

where here  $K(y)$  is the triangular kernel defined by

$$K(y) = \begin{cases} 1 - |y|, & \text{for } y \in (-1, 1), \\ 0, & \text{otherwise.} \end{cases} \quad (25)$$

Suppose that  $f(x)$  is twice-differentiable with continuous second derivative. The expectation of  $\hat{f}(x)$  for a general kernel that is symmetric about zero is given by

$$\mathbb{E}\{\hat{f}(x)\} = f(x) + \frac{1}{2}C_3 h^2 f''(x) + \mathcal{O}(h^3), \quad (26)$$

where  $C_3 = \int_{-\infty}^{\infty} v^2 K(v) dv$ .

- (i) The expectation is easy. For the variance

$$\text{var}\left\{\hat{f}(x)\right\} = \text{var}\left[(nh)^{-1} \sum_{i=1}^n \left\{K\left(\frac{X_i - x}{h}\right)\right\}\right] \quad (27)$$

$$= (nh)^{-2} \sum_{i=1}^n \text{var}\left\{K\left(\frac{X_i - x}{h}\right)\right\} \quad (\text{independence of } X_i)$$

$$= n^{-1} h^{-2} \text{var}\left\{K\left(\frac{X_i - x}{h}\right)\right\}, \quad (28)$$

as the variance in the sum does not depend on  $i$ .

sim. seen ↓

- (ii) Let's calculate  $C_3$  for the triangular kernel:

$$C_3 = \int_{-\infty}^{\infty} v^2 K(v) dv \quad (29)$$

$$= \int_{-1}^1 v^2 (1 - |v|) dv \quad (30)$$

$$= 2 \int_0^1 v^2 (1 - v) dv \quad (\text{symmetry}) \quad (31)$$

$$= 2 \int_0^1 v^2 - v^3 dv \quad (32)$$

$$= 2 [v^3/3 - v^4/4]_0^1 \quad (33)$$

$$= 2(1/3 - 1/4) = 1/6. \quad (34)$$

1, B

sim. seen ↓

- (iii) Now let

$$(*) = \mathbb{E}\left\{K^2\left(\frac{X_i - x}{h}\right)\right\} \quad (35)$$

$$= \int_{-\infty}^{\infty} K^2\left(\frac{y - x}{h}\right) f(y) dy \quad (36)$$

$$= h \int K^2(v) f(x + vh) dv. \quad (37)$$

2, B

unseen ↓

Now use the Taylor series

$$f(x + \delta) = f(x) + \delta f'(x) + \delta^2 f''(x)/2 + \mathcal{O}(\delta^3), \quad (38)$$

to give

$$(*) = h \left\{ \int K^2(v) f(x) dv + \int v h K^2(v) f'(x) dv \right. \quad (39)$$

$$\left. + \int \frac{v^2 h^2}{2} K^2(v) f''(x) dv + \mathcal{O}(h^3) \right\} \quad (40)$$

$$= h \left\{ f(x) K_1 + h f'(x) K_2 + \frac{h^2}{2} f''(x) K_3 + \mathcal{O}(h^3) \right\}, \quad (41)$$

where

$$K_1 = \int K^2(v) dv \quad (42)$$

$$= \int_{-1}^1 (1 - |v|)^2 dv = 2 \int_0^1 (1 - v)^2 dv \quad (43)$$

$$= 2 \int_0^1 (1 - 2v + v^2) dv = 2[v - v^2 + v^3/3]_0^1 \quad (44)$$

$$= 2(1 - 1 + 1/3) = 2/3, \quad (45)$$

and

$$K_2 = \int_{-1}^1 v K^2(v) dv = 0, \quad (46)$$

as the integrand is an odd function, and

$$K_3 = \int v^2 K^2(v) dv = \int_{-1}^1 v^2 K^2(v) dv = 2 \int_0^1 v^2 K^2(v) dv \quad (47)$$

$$= 2 \int_0^1 v^2 (1 - v)^2 dv = 2 \int_0^1 v^2 (1 - 2v + v^2) dv \quad (48)$$

$$= 2 \int_0^1 v^2 - 2v^3 + v^4 dv = 2[v^3/3 - v^4/2 + v^5/5]_0^1 \quad (49)$$

$$= 2(1/3 - 1/2 + 1/5) = 1/15. \quad (50)$$

Hence,

$$\mathbb{E} \left\{ K^2 \left( \frac{X_i - x}{h} \right) \right\} = \frac{2h}{3} f(x) + \frac{h^3}{30} f''(x) + \mathcal{O}(h^4). \quad (51)$$

8, D

Thus

$$\text{var} \left\{ K^2 \left( \frac{X_i - x}{h} \right) \right\} = (51) - (h \times \text{Eqn (5) from exam q})^2 \quad (52)$$

$$= \frac{2h}{3} f(x) + \frac{h^3}{30} f''(x) + \mathcal{O}(h^4) \quad (53)$$

$$- \left\{ h f(x) + \frac{h^3}{12} f''(x) + \mathcal{O}(h^3) \right\}^2 \quad (54)$$

$$= (51) - h^2 f^2(x) \quad (55)$$

$$- 2h f(x) \left\{ \frac{h^3}{12} f''(x) + \mathcal{O}(h^3) \right\} \quad (56)$$

$$+ \left\{ \frac{h^3}{12} f''(x) + \mathcal{O}(h^3) \right\}^2 \quad (57)$$

$$= (51) - h^2 f^2(x) + \mathcal{O}(h^4) + \mathcal{O}(h^4) + \mathcal{O}(h^6) \quad (58)$$

$$= \frac{2h}{3} f(x) - h^2 f^2(x) + \frac{h^3}{30} f''(x) + \mathcal{O}(h^4). \quad (59)$$

3, C

- (b) The bandwidth is not of direct practical use because it depends on the quantity we are trying to estimate.

seen ↓

1, B

part seen ↓

- (c) (i) The exam paper mentions that  $d_{j,k} = \int_{\mathbb{R}} f(x) \psi_{j,k}(x) dx$ , which can be written equivalently as  $\mathbb{E}\{\psi_{j,k}(X)\}$ , where  $X \sim f$ . This expectation can be estimated by the empirical expectation  $\hat{d}_{j,k} = n^{-1} \sum_{i=1}^n \psi_{j,k}(X_i)$  and similarly for  $c_{j,k}$ .

3, A

unseen ↓

- (ii) Each  $d_{j,k}$  using the empirical formula requires  $n$  additions (and multiplications and function evaluations). There are  $n$  coefficients, so the whole computation would take  $\mathcal{O}(n^2)$  operations. If one used the discrete wavelet transform, its recursive pyramid nature only takes  $\mathcal{O}(n)$  operations for all coefficients.

2, B

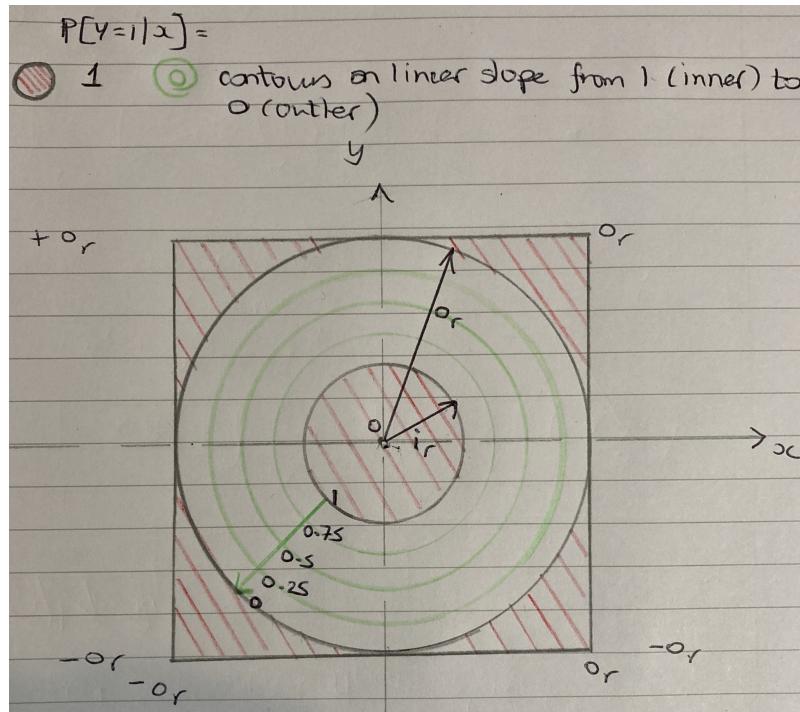
4. (a) Three of wavelets, splines, neural networks, boosting, linear regression, orthogonal polynomials

seen ↓

1, A

unseen ↓

- (b) (i) Sketch is



**NB.** The axes should be  $x_1, x_2$  not  $x, y$ .

- (ii) The  $Y$  probability is rotationally symmetrical about the origin. All straight lines will be not very useful, but LDA ones through the mean will be totally useless and not separate the data into  $+1$  and  $-1$  (not unless there is a failure of the laws of probability/nature).

2, B

unseen ↓

- (iii) A training set is a subset of the data that you set aside and use for building the model. A test set is a subset of the data that you set aside and only use once the model was built - you can then feed the test data into the estimated model to see how well the model performs on 'new' data. It is important to split like this. If you use one set of data to train and test, then it is possible to overfit the data at the training stage and, then, when the same data is used with that model it tends to predict/fit very accurately, because it is using specific local and not generalizable information present in the training set, that would not be present in a general data set.

1, C

sim. seen ↓

- (iv) For the inner minimiser, we can proceed by calculus. Define  $M(c) = \sum_i(Y_i - c)^2$ . Then  $M'(c) = 2 \sum_i(Y_i - \hat{c}) = 0$ , which implies  $\sum_i Y_i = n\hat{c}$  and hence  $\hat{c} = \bar{Y}$ .

3, A

seen ↓

- (v) The sum in (17) for FIXED  $j$  does not change, except when  $s$  crosses  $X_i$  for some  $i$ . When  $s$  crosses  $X_i$  from left to right, then one point  $(X_i, Y_i)$  moves from the right-hand term, to the left hand-term. So, there will only be  $n+1$  (quickly computable) different values for (17) [for fixed  $j$ ] and we choose the  $s$  that gives us the minimum.

2, A

seen ↓

2, B

- (vi) There are no conceptual difficulties in moving from two to three splits. You need two splitting locations,  $s_1, s_2$  and (17) will have three terms similar to the two we have now. The discovery of the  $\hat{c}$  is the same as before. Now the objective only changes when  $s_1$  or  $s_2$  cross a  $X_i$  so there are now  $(n+1)^2$  possibilities.

unseen ↓

It's not usually done, because one rapidly runs out of data as one descends the tree. E.g. the rate of data 'decline' goes like  $3^{-\ell}$  rather than  $2^{-\ell}$ .

- (c) (i) AdaBoost.M1 builds an effective classifier by constructing a sequence of increasingly stronger classifiers. At each step one (attempts to) improve the classifier by identifying those points that were previously poorly classified and give them more weight in the problem. Then the classifier focus on those and classifies them better next time.

2, A

1, B

seen ↓

- (ii) The quantity to minimise is

$$\mathbb{E}_{Y|x}[e^{-Yf(x)}] = e^{-f(x)}\mathbb{P}[Y=1|x] + e^{f(x)}\mathbb{P}[Y=-1|x] \quad (60)$$

$$= e^{-R}\mathbb{P}[Y=1|x] + e^R\mathbb{P}[Y=-1|x] \quad (61)$$

2, A

unseen ↓

Now differentiate wrt  $R$  gives

$$\frac{\partial}{\partial R} = -e^{-\hat{R}}\mathbb{P}[Y=1|x] + e^{\hat{R}}\mathbb{P}[Y=-1|x] = 0 \quad (62)$$

$$\Rightarrow e^{2\hat{R}} = \frac{\mathbb{P}[Y=1|x]}{\mathbb{P}[Y=-1|x]} \quad (63)$$

$$\Rightarrow \hat{R} = 2^{-1} \log \left\{ \frac{\mathbb{P}[Y=1|x]}{\mathbb{P}[Y=-1|x]} \right\} = \hat{f}(x). \quad (64)$$

So, comparing back to the additive basis expansion if this is positive (sign is one) then this means that the conditional probability  $\mathbb{P}[Y=1|x]$  is greater than  $\mathbb{P}[Y=-1|x]$  and so you want to classify as +1, and similarly for the opposite sign.

4, D

5. (a) Solutions to this question may show some variability. Let's define the ridge-regression estimator of the parameters  $\beta$  by  $\hat{\beta}(\lambda)$ . The fitted values are  $\hat{y} = X\hat{\beta}(\lambda)$ . We can drop the  $i$ th observation from the data set and, using the same ridge regression method, obtain  $\hat{\beta}^{-i}(\lambda)$  to be the estimates of  $\beta$  on this new omitted data set. We can then predict the value of the  $i$ th data point using  $\hat{y}^{-i}(\lambda) = x^{(i)}\hat{\beta}^{-i}(\lambda)$ , where  $x^{(i)}$  is the  $i$ th row of  $X$ . The ordinary cross-validated estimate of error is

$$\text{OCV}(\lambda) = n^{-1} \sum_{i=1}^n \{Y_i - \hat{y}^{-i}(\lambda)\}^2. \quad (65)$$

The rationale is that if  $\lambda$  is a 'good' value, then the fitted values  $\hat{y}^{-i}(\lambda)$  will be a good fit to  $Y_i$ , but  $Y_i$  itself was not involved in the calculation of the fitted values (i.e. the training set and test set are mutually exclusive). This comparison is then averaged over all  $i = 1, \dots, n$ .

Hence, a good bandwidth can be obtained by minimising the  $\text{OCV}(\lambda)$  error as a function of  $\lambda$ .

The  $k$ -fold cross-validation concept is similar. First, randomly permute the data. Then, divide the data into  $k$  equally-sized groups (or as close to as possible). For test  $i = 1, \dots, k$ , hold back group  $i$  as a test set. Then, fit the model with the remaining (training) data and use the trained model to predict the values at the locations of the test set, forming the mean sum of squares between the predicted values and the test values. Repeat this for each  $i = 1, \dots, k$  and average the results.

This can itself be repeated for different values of  $k$ .

Essentially, cross-validation works on the assumption that there is no correlation. Serial correlation can cause problems in several ways. For example, for the random permutation, this might disrupt the correlations in the data, so that where previously there might have been association between neighbouring points (for example), there is now not and this might generate dependences between the training and test sets, which will influence the variance of the OCV score. Another way of looking at this is follows. Suppose the correlation between neighbouring points is highly positive. Then the  $i$ th data point will be strongly similar to the  $i-1$ th and  $i+1$ th data point, then good prediction of the  $i$  point is not a big deal, since the neighbouring points will be close. However, the methodology thinks it is a big deal and rewards it inappropriately.

Further, specific correlation structure can interfere in other ways. For example, if correlation is related to a particular scale, then  $k$ -fold cross-validation at the same scale (e.g.  $n/k$  similar) can be disrupted. Outliers can also disrupt cross-validation as there is an implicit assumption that all errors are iid with the same variance and an outlier can be seen to be anomalous. When the outlier is 'left out', then the prediction error for that point will be very large and, potentially, dominate the OCV score unduly.

- (b) (i) The quantity  $\hat{f}_\lambda^{-i}$  is the smoothing spline estimator applied to the data set of length  $n-1$  that omits the  $i$ th point.
- (ii) The full estimator can be written using the hat matrix as

$$\hat{f}_\lambda(t_i) = \sum_{j=1}^n h_{i,j} y_j, \quad (66)$$

seen ↓

10, M

meth seen ↓

2, M

for  $i = 1, \dots, n$  and suppressing the  $\lambda$  from the  $h$  notation.

Now using the given result gives

$$\hat{f}_\lambda^{-i}(t_i) = \tilde{f}_\lambda^{-i}(t_i) = \sum_{j=1}^n h_{i,j} \tilde{y}_j \quad (67)$$

$$= \sum_{j \neq i} h_{i,j} y_j + h_{i,i} \hat{f}_\lambda^{-i}(t_i). \quad (68)$$

Now subtract (68) from (66) to obtain

$$\hat{f}_\lambda(t_i) - \hat{f}_\lambda^{-i}(t_i) = h_{i,i} \{y_i - \hat{f}_\lambda^{-i}(t_i)\}. \quad (69)$$

Now, on the LHS of (69) subtract and add  $y_i$  to obtain

$$\hat{f}_\lambda(t_i) - y_i + y_i - \hat{f}_\lambda^{-i}(t_i) = h_{i,i} \{y_i - \hat{f}_\lambda^{-i}(t_i)\}. \quad (70)$$

Collect terms and rearranging gives

$$\{y_i - \hat{f}_\lambda^{-i}(t_i)\} (1 - h_{i,i}) = y_i - \hat{f}_\lambda(t_i), \quad (71)$$

and hence

$$y_i - \hat{f}_\lambda^{-i}(t_i) = \frac{y_i - \hat{f}_\lambda(t_i)}{1 - h_{i,i}} \quad (72)$$

as required. 6, M

- (iii) The advantage is computational. The original formulation of the OCV requires  $\hat{f}_\lambda^{-i}$  to be computed for each  $i = 1, \dots, n$ , in other words  $n$  times. The new formulation merely requires computation of  $\hat{f}_\lambda(t_i)$ , the full estimator on the whole data set and knowledge of the hat matrix entries — again that only need to be computed once (essentially the first step on the way to the generalized cross-validation estimate). 2, M

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.

Please record below, some brief but non-trivial comments for students about how well (or otherwise) the questions were answered.

For example, you may wish to comment on common errors and misconceptions, or areas where students have done well. These comments should note any errors in and corrections to the paper. These comments will be made available to students via the MathsCentral Blackboard site and should not contain any information which identifies individual candidates. Any comments which should be kept confidential should be included as confidential comments for the Exam Board and Externals. If you would like to add formulas, please include a separate pdf file with your email.

| ExamModuleCode      | QuestionNumber | Comments for Students   |
|---------------------|----------------|---|
| MATH96067/MATH97287 | 1              | Part (a) most students correctly defined the unbiasedness concept, but many forgot to define "linear". Parts (b), (c), d(i) were done well. Part d(ii) was done well by many students, but some candidates had trouble identifying the key differences between the Bayesian and ridge approach. In d(iii) a surprising number did not attempt taking the second derivative (in spite of successfully taking the first one) and/or use another method to prove that the ridge solution obtained a minimum. Part (e) was reasonably well done. Most people identified that Test1, Test2 and Test4 all provided some explanation of the response, but they were highly correlated. Some candidates forgot to address the correlation and/or mention anything about the Lasso plot. |
| MATH96067/MATH97287 | 2              | Most of the question was very well done. In part (a), most of the students correctly described the Miles algorithm and successfully applied it to the provided example; some students made small computation mistakes or did not carefully respect the order. Part (b) was mainly seen material and correctly done. Part (c) was an unseen question; many students correctly identified the two possible clusters, however the proof that only two clusters could be obtained was often too imprecise and the cluster means were not always stated.   |

|                     |   |   |
|---------------------|---|---|
| MATH96067/MATH97287 | 3 | Part(a)(i) Nearly all candidates got this part correct. A surprising number were not able to perform the simple integration C_3 (which was essentially A level) in part 3(ii). Although none of the integrals in this question were hard (again, all A level standard), there were quite a few of them. Candidates were given marks for method, even if the final answer was wrong due to small slips. Most candidates did 3a(iii) and (b) well. 3(c)(i) was essentially about recognising the link between population and empirical means and 3(c)(ii) relied on lecture notes about the efficiency of the discrete wavelet transform. The majority of candidates attempted these, but a minority did not.   |
| MATH96067/MATH97287 | 4 | Part (a) was really about basis expansion methods such as wavelets, splines, orthogonal polynomials, but some candidates interpreted the question to mean types of basis function (such as linear, quadratic, step function etc). Credit was given for either interpretation. Many candidates got the plot for part b(i) correct and identified why linear discriminant analysis would not be able to help, b(ii). Most candidates gave an adequate description about training and test sets in answer to 4b(iii). Part b(iv) and (v) were book work that most candidates did well. Part (vi) was an extension of the ideas in (iv, v) and was done well by those candidates who attempted it. Most candidates answered c(i) adequately although focused on a bald description of the algorithm, rather than the rationale (reason why). Very few candidates attempted part c(ii). A few did manage to get it completely correct. |
| MATH97287           | 5 | Overall, the response to this question was disappointing. Most candidates attempted an informal description of cross-validation in part (a), and these attracted marks, but these answers were often vague and the responses to the concepts of k-fold, correlation and outliers were a bit mixed. Part (b)(i and iii) were generally well done, but part (ii) was not attempted by most. A few people made reasonable attempts and two solutions were particularly well done (although using a more complex route than the sample solution).   |