

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2022

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

Introduction to Statistical Learning

Date: 30 May 2022

Time: 09:00 – 11:30 (BST)

Time Allowed: 2:30 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 AS SEPARATE PDFs TO THE RELEVANT DROPBOXES ON BLACKBOARD (ONE FOR EACH QUESTION) WITH COMPLETED COVERSHEETS WITH YOUR CID NUMBER, QUESTION NUMBERS ANSWERED AND PAGE NUMBERS PER QUESTION.

1. (a) Briefly explain the geometric representation of multivariate linear regression in the context of least squares (LS) estimation (p estimators and n observations). You should include the description of the corresponding fitted line, the LS estimates, and the normal equations. Introduce the necessary notation and the definitions. (4 marks)

- (b) Consider a multivariate linear regression model

$$Y = \beta_0 + X\beta + \epsilon, \quad (1)$$

where $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ and $\epsilon \sim N(0, \sigma^2)$, for some constant $\sigma^2 > 0$. Assume the availability of training data $\{(\underline{x}_i, y_i)\}_{i=1}^n$, with $\underline{x}_i \in \mathbb{R}^p$ for each i . Under the standard assumptions on the error terms and the assumption that the associated design matrix $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ is of full-rank, provide the Maximum Likelihood (ML) estimates, $\hat{\beta}^{ML}$, for the regression coefficients $\beta = (\beta_0, \dots, \beta_p)^T$. Make sure that you indicate clearly how the different assumptions are used. (3 marks)

- (c) Under the same multivariate linear regression setting in (b), briefly explain the main advantage of the ridge regression estimates over LS estimates. (2 marks)

- (d) Consider the following design matrix

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} = \begin{pmatrix} z_{11} & z_{11} \\ z_{22} & z_{22} \end{pmatrix}, \quad (2)$$

for some $z_{11}, z_{22} \in \mathbb{R}$. Assume that your training data is given by \mathbf{X} and $y = (y_1, y_2)^T$. Also suppose that $y_1 + y_2 = 0$, and that $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$. In the following (and when appropriate) indicate clearly how the different assumptions are used.

- (i) Briefly explain whether LS estimation or ridge regression estimation would be more appropriate to fit a linear model with two predictors and an intercept using the training data above. (2 marks)
 - (ii) Write down the generic model to estimate the response Y using linear regression, and *define* the corresponding ridge coefficient estimates in terms of the training data provided above. (2 marks)
 - (ii) Compute explicitly the ridge regression estimates for $\lambda = 2$. (4 marks)
- (e) Suppose that in a given set with $p = 3$ predictors, say X_1, X_2 and X_3 , the best possible one-variable model contains the predictor X_2 , and the best possible two-variable model contains predictors X_1 and X_3 . Giving details, explain why forward-stepwise selection will fail to select the best possible two-variable model. (3 marks)

(Total: 20 marks)

End of Question 1

2. (a) Suppose that you have been hired to evaluate the relationship between annual salaries Y of statisticians (in thousand dollars) and the following continuous variables: X_1 = index of work quality, X_2 = number of years of experience, and X_3 = index of publication success. The following is a summary output from a multiple linear regression model:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	17.84693	2.00188	8.915	2.10e-08
X1	1.10313	0.32957	3.347	0.003209
X2	0.32152	0.03711	8.664	3.33e-08
X3	1.28894	0.29848	4.318	0.000334

Residual standard error: 1.753 on 20 degrees of freedom

F-statistic: 68.12 on 3 and 20 DF, p-value: 1.124e-10

- (i) Comment on the significance of each predictor variable on the linear regression at the significance level $\alpha = 0.05$. Write down the corresponding hypothesis tests and briefly explain your answers. (2 marks)
- (ii) Test a suitable hypothesis to check the overall fit of the regression equation at a significance level $\alpha = 0.05$. (2 marks)
- (b) Determine whether the following statements are TRUE or FALSE. In each case, briefly justify your answers.
- (i) The lasso regression coefficient estimates, $\hat{\beta}_j^{lasso}(\lambda)$, for $\lambda \geq 0$, are always given by
- $$\hat{\beta}_j^{lasso}(\lambda) = \text{sign}(\hat{\beta}_j^{LS})(|\hat{\beta}_j^{LS}| - \lambda)_+, \quad (3)$$
- where $\hat{\beta}^{LS}$ denotes the LS estimates, $\text{sign}(\cdot)$ denotes the sign function, and x_+ denotes the positive part of x . (1 mark)
- (ii) The ridge regression method can also be thought of as a subset selection technique because we can always choose a sufficiently large tuning parameter λ . (2 marks)
- (c) Briefly explain the main characteristics of the K -means approach for clustering (at most 6 lines). (2 marks)

Question 2 continued next page ...

- (d) Consider the following observations consisting of $n = 5$ data points with two inputs, x_1 and x_2 :

i	x_1	x_2
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5

Implement the K -means algorithm to cluster the 5 data points above into two clusters (suppose that the algorithm is initialised to two initial clusters $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5\}$). In case of ties in the distance, classify to the centroid with the smallest x_1 value. Show your computations for each iteration until convergence. (5 marks)

- (e) Briefly explain the purpose of classical multidimensional scaling. Briefly explain also the importance of the eigen-decomposition of the inner product matrix in this context. Introduce all the necessary notation and/or definitions. (3 marks)
- (f) In the context of classical multidimensional scaling, suppose that the Euclidean distance matrix E corresponds to an $n \times q$ matrix of full rank. How many eigenvalues in the eigen-decomposition of the corresponding inner product matrix would be zero? Briefly explain why. (1 mark)
- (g) Suppose that you want to fit the model $Y = f(X) + \epsilon$ for a more flexible function f (not necessarily linear). Briefly explain how linear modelling is still useful in the context of a *basis expansion* approach. (2 marks)

(Total: 20 marks)

End of Question 2

3. (a) How many degrees of freedom has a cubic spline with K knots. Briefly justify your answer. (3 marks)
- (b) Consider the model $Y = f(X) + \epsilon$, where Y is the response variable and X is a single explanatory variable. Suppose you wish to fit a cubic spline f with three knots.
- (i) Write down the corresponding basis functions $\{h_m\}$ for this model. (2 marks)
- (ii) Assume that you have training data $\{(x_i, y_i)\}_{i=1}^{10}$, write down the corresponding fitted cubic spline using least squares estimation. You should briefly explain the steps to obtain the fitted cubic spline and define all the necessary notation. You may assume that any necessary operation between vectors and matrices can be obtained. (5 marks)
- (c) Define the concept of smoothing spline and give three of its main characteristics. (3 marks)
- (d) Write down the steps to obtain the fitted spline and the smoother matrix for a fixed given $\lambda \geq 0$. Explicit calculations are not required, but a justification in each step is required (introduce all the necessary notation). (4 marks)
- (e) Let $\phi(x)$ and $\psi(x)$ be the Haar father and mother wavelet respectively. The Haar (mother) wavelet coefficients of the function $f : \mathbb{R} \rightarrow \mathbb{R}$ are given by

$$d_{j,k} = \int f(x) \psi_{j,k}(x) dx, \quad (4)$$

where integrals are over \mathbb{R} and the Haar father wavelet coefficients, $\{c_{j,k}\}_{j,k}$ are constructed similarly replacing ψ by ϕ in the right-hand side of (4).

Now let $f(x)$ additionally be a *unknown* probability density function and let X_1, \dots, X_n , $n \in \mathbb{N}$, be an independent random sample drawn from f . Suppose we wish to estimate $f(x)$ from X_1, \dots, X_n . Since f is a probability density we can rewrite (4) as the mean of $\psi_{j,k}(X)$ as $d_{j,k} = \mathbb{E}\{\psi_{j,k}(X)\}$, where X is a sample from f . Hence, we can estimate $d_{j,k}$ by the following empirical mean

$$\hat{d}_{j,k} = n^{-1} \sum_{i=1}^n \psi_{j,k}(X_i). \quad (5)$$

A hypothesis test for whether $d_{j,k} = 0$ can be performed using the variance of $\hat{d}_{j,k}$. Show that $\text{var}\{\hat{d}_{j,k}\} = n^{-1}(2^{j/2}c_{j,k} - d_{j,k}^2)$.

(3 marks)

(Total: 20 marks)

End of Question 3

4. (a) Define the principal components for a given (standardised and centred) data matrix $X \in \mathbb{R}^{n \times p}$. (2 marks)
- (b) Briefly explain how the principal components regression works, and explain how the corresponding fitted line can be obtained. (2 marks)
- (c) Briefly explain two advantages of principal components regression compared to standard multivariate regression. (2 marks)
- (d) Briefly explain the importance of the negative entropy of a function f (denoted in lectures by $H(f)$) in the context of exploratory projection pursuit. (3 marks)
- (e) In kernel regression, the bandwidth h needs to be chosen small to guarantee that the mean-squared error is minimised. TRUE or FALSE? Briefly explain why. (1 mark)
- (f) Consider the following partitioned feature space:

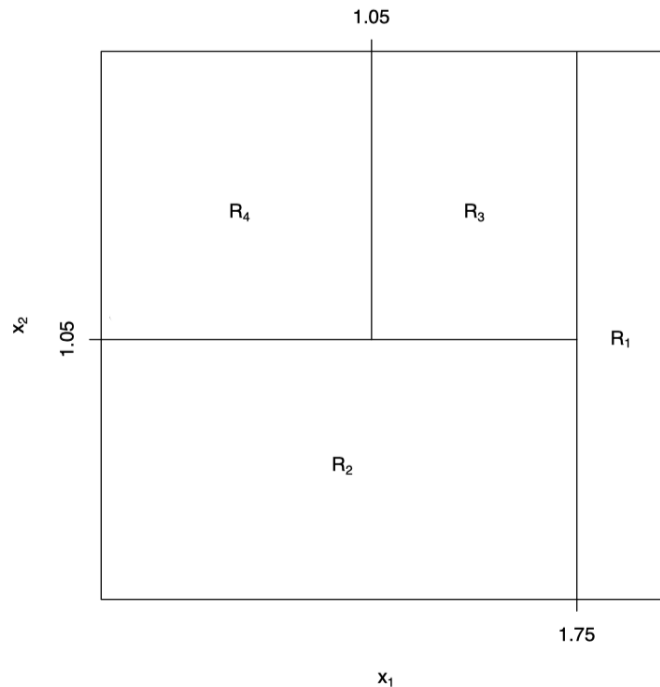


Figure 1

- (i) Draw the classification tree corresponding to Figure 1. (2 marks)

Question 4 continued next page ...

- (ii) Suppose that the tree obtained in (i) is a regression tree corresponding to the following training data:

i	1	2	3	4	5	6	7	8
x_1	1.5	0.6	0.9	0.4	1.9	0.2	1.2	1.6
x_2	1.8	0.3	1.4	1.8	1.8	1.2	1.3	0.9
y	0.4	0.6	0.7	0.5	0.2	0.6	0.6	0.5

If the prediction of a new input falling into a region R_j is given by the value of the mean of the response values from the training observations that fall into region R_j , predict the output of a new input $(x_1, x_2) = (0.8, 1.3)$. (4 marks)

- (g) The random forest method is a special version of boosting with trees. TRUE or FALSE? Briefly explain why. (1 mark)
- (h) Briefly explain the importance of bagging in the context of regression trees and how it works for prediction. (2 marks)
- (i) As the bagging approach, boosting involves the use of bootstrap sampling to improve prediction in the context of decision trees. TRUE or FALSE? Briefly explain why. (1 mark)

(Total: 20 marks)

End of Question 4

5. (a) State two main differences between standard logistic and standard linear regression. (2 marks)
- (b) Briefly explain why it would not be feasible to model the conditional probability $P(Y|X)$ of the response Y , given the inputs X , as a linear function of X . (1 mark)
- (c) In a modelling problem, let X be the input variable. Let Y be the response variable which takes one of only two possible values $+1$ and -1 . Let $\{(x_i, y_i)\}_{i=1}^n$ be a set of training data where the data points are mutually independent. Define

$$p(y|x; \beta) = P(Y_1 = y_1, \dots, Y_n = y_n | X_1 = x_1, \dots, X_n = x_n; \beta). \quad (6)$$

Derive an expression for the log-likelihood $\log p(y|x; \beta)$ for the n observations for the logistic regression model in the binary classification case when the output classes are $\{-1, 1\}$. (5 marks)

- (d) Consider a qualitative predictor variable that takes values in $\{\text{green, blue, pink}\}$. How could we encode this qualitative predictor variable so that it can be used for a binary classification problem? (2 marks)
- (e) Consider binary classification where the response variable can either take the value $+1$ or -1 . Show that fitting a linear regression model to the log odds of $p(y = 1|x)$ gives the same model as the one used in logistic regression. (3 marks)
- (f) Consider a logistic regression model for a binary classification problem with response Y taking values in $\{0, 1\}$. Suppose that

$$\beta^T = (\beta_0, \beta_1, \beta_2, \beta_3), \quad \text{and} \quad \hat{\beta} = (3, -1, 3, 2), \quad (7)$$

where β_0 is the intercept, and

$$\log \frac{P(Y = 1|X; \beta)}{P(Y = 0|X; \beta)} = \beta^T X. \quad (8)$$

According to this model, what is the probability that a new input $x = (2, 1, -1)$ belongs to class 0? (2 marks)

- (g) To simplify notation let $P(x) = P(Y = 1|X = x)$. Define the logit function by

$$\text{logit}\{P(x)\} = \log[P(x)/\{1 - P(x)\}]. \quad (9)$$

The nonparametric logistic regression model is defined as

$$\text{logit}\{P(x)\} = \alpha + \sum_{j=1}^p f_j(x_j), \quad \mathbb{E}\{f_j(X_j)\} = 0, \forall j, \quad (10)$$

where x_1, \dots, x_p are a set of covariate variables.

Question 5 continued next page ...

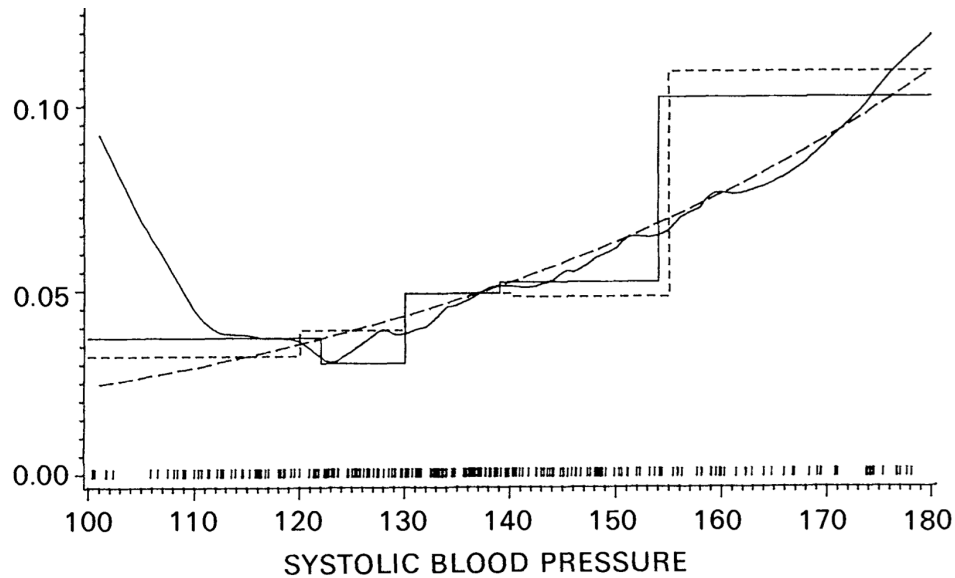


Figure 2: Four model fits to coronary heart disease data as explained in the question. Left-hand axis is $P(Y = 1|X = \text{systolic blood pressure})$. Horizontal axis is systolic blood pressure in mmHg.

Figure 2 shows the results of four model fits to a set of patients where variable $Y = 1$, if that patient has coronary heart disease (CHD) and $Y = 0$ if not and X is the patients' systolic blood pressure in units of mmHg. The two piecewise constant ('stepped') lines correspond to dividing the systolic blood pressure into (two different sets of) bins and averaging the patients CHD status over each bin. The dashed curve line corresponds to standard linear logistic regression fit (as in parts (a)–(f) above). The solid line corresponds to the nonparametric logistic regression model fit as defined by (10).

- (i) Why is the dashed line corresponding to the logistic *linear* regression not a straight line on the plot? (1 mark)
- (ii) What advantages might there be in using the nonparametric logistic regression over the other three methods? Can you identify anywhere on Figure 2 where these advantages might be demonstrated? (2 marks)
- (iii) Given a set of data, such as the CHD set, briefly outline how the nonparametric logistic regression model might be fitted? (2 marks)

(Total: 20 marks)

End of Question 5

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2022

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

MATH60049/MATH70049/MATH97287

Introduction to Statistical Learning (Solutions)

Setter's signature

.....

Checker's signature

.....

Editor's signature

.....

1. (a) The fitted regression line using LS estimation is given by $\hat{y} = X\hat{\beta}^{LS}$, where $X = [\underline{x}_1 \cdots \underline{x}_p] \in \mathbb{R}^{p \times p}$ is the design matrix, $y \in \mathbb{R}^n$ is the response, and $\hat{\beta}^{LS}$ denotes the LS estimates of the regression coefficients which are obtained by solving the normal equations

$$X^T(y - X\beta) = 0.$$

The fitted regression line corresponds to the orthogonal projection of y onto the column space of X , $\text{col}(X)$ (the space spanned by the columns of X), and so \hat{y} takes the form of a linear combination of the columns of X with coefficients given by the least squares estimates $\hat{\beta}^{LS}$. The normal equations guarantee that the columns of X are orthogonal to the residuals $r = y - X\hat{\beta}^{LS}$ and so \hat{y} is an element in $\text{col}(X)$ which is closest to y .

- (b) It was seen in lectures that the ML estimates for the regression coefficients coincide with the LS estimates when *the distribution of the error terms are normally distributed*.

Hence, we only need to observe that $\mathbf{X}^T \mathbf{X}$ is invertible (as *the design matrix is of full-rank*), and the errors satisfy the standard assumptions.

We conclude then that $\hat{\beta}^{ML} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$.

- (c) Since the LS estimates are given by

$$\hat{\beta}^{LS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y,$$

where \mathbf{X} is the design matrix, LS estimates depend on the invertibility of the matrix $\mathbf{X}^T \mathbf{X}$ which cannot be always guaranteed. More generally, the *condition number* of the matrix $\mathbf{X}^T \mathbf{X}$ determines the variance of the estimates.

However, since the ridge regression estimates take the form

$$\hat{\beta}^{LS} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbb{I})^{-1} \mathbf{X}^T y,$$

the possible singularity issues of the matrix $\mathbf{X}^T \mathbf{X}$ (and so the variance of the estimates) can be mitigated via the tuning parameter λ in ridge regression.

- (d) (i) In this setting, there are $n = 2$ observations and $p = 2$ predictors. Since the two columns in the design matrix are the same, the observations are perfectly correlated, which then implies that the design matrix is not of full-rank. So the LS estimation is not possible in this case due to the singularity of the matrix $\mathbf{X}^T \mathbf{X}$. Ridge regression would be a better option to control the singularity of the matrix.

seen ↓

1, A

1, B

1, B

1, C

seen ↓

1, B

1, B

1, B

seen ↓

1, B

1, A

1, A

1, A

- (ii) The model to explain the response variable Y using two predictors X_1 and X_2 using linear regression takes the form:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

where ϵ denotes the error term satisfying the standard assumptions. The ridge coefficient estimates $\hat{\beta}^{ridge}$ are the coefficients that minimise a penalised residual sum of squares. More precisely, for a given $\lambda \geq 0$,

1, B

$$\hat{\beta}^{ridge} = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \left\{ \sum_{i=1}^2 \left(y_i - \sum_{j=1}^2 \beta_j x_{ij} \right)^2 + \lambda(\beta_1^2 + \beta_2^2) \right\}.$$

1, B

- (iii) We know that the estimation of the coefficients can be broken down into two parts: estimating the intercept by \bar{y} and then estimating β_1 and β_2 using ridge regression without an intercept, using centred data. By assumption $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, hence $\hat{\beta}_0^{ridge} = 0$ and the matrix \mathbf{X} it is centred.

1, C

Now, we can use the ridge regression formula to estimate β_1 and β_2 :

$$\hat{\beta}^{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbb{I}_n)^{-1} \mathbf{X}^T y,$$

where \mathbb{I}_n stands for the n -dimensional identity matrix.

1, A

Using that $y_1 = -y_2$, by setting $a \stackrel{\text{def}}{=} x_{11}^2 + x_{22}^2$ and $b \stackrel{\text{def}}{=} (x_{11} - x_{22})y_1$ we can derive the solution as follows

$$(X^T X + 2I) = \begin{pmatrix} z_{11} & z_{22} \\ z_{11} & z_{22} \end{pmatrix} \begin{pmatrix} z_{11} & z_{11} \\ z_{22} & z_{22} \end{pmatrix} + 2I \quad (1)$$

$$= \begin{pmatrix} z_{11}^2 + z_{22}^2 & z_{11}^2 + z_{22}^2 \\ z_{11}^2 + z_{22}^2 & z_{11}^2 + z_{22}^2 \end{pmatrix} + 2I \quad (2)$$

$$= \begin{pmatrix} a + 2 & a \\ a & a + 2 \end{pmatrix} \quad (3)$$

Then

$$\hat{\beta}^{ridge} = \begin{pmatrix} a+2 & a \\ a & a+2 \end{pmatrix}^{-1} X^T y \quad (4)$$

$$= \frac{1}{(a+2)^2 - a^2} \begin{pmatrix} a+2 & -a \\ -a & a+2 \end{pmatrix} \begin{pmatrix} z_{11} & z_{22} \\ z_{11} & z_{22} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (5)$$

$$= \frac{1}{a^2 + 2a + 4 - a^2} \begin{pmatrix} a+2 & -a \\ -a & a+2 \end{pmatrix} \begin{pmatrix} z_{11}y_1 + z_{22}y_2 \\ z_{11}y_1 + z_{22}y_2 \end{pmatrix} \quad (6)$$

$$= \frac{1}{2a+4} \begin{pmatrix} a+2 & -a \\ -a & a+2 \end{pmatrix} \begin{pmatrix} (z_{11} - z_{22})y_1 \\ (z_{11} - z_{22})y_1 \end{pmatrix} \quad (7)$$

$$= \frac{1}{2(a+2)} \begin{pmatrix} a+2 & -a \\ -a & a+2 \end{pmatrix} \begin{pmatrix} b \\ b \end{pmatrix} \quad (8)$$

$$= \frac{b}{2(a+2)} \begin{pmatrix} a+2 & -a \\ -a & a+2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (9)$$

$$= \frac{b}{2(a+2)} \begin{pmatrix} 2 \\ 2 \end{pmatrix} = \frac{b}{a+2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (10)$$

- (e) Starting from the null model, forward-stepwise selection will choose the best one-variable model which, by assumption, is X_2 (denoted in lectures by \mathcal{M}_1).

Then, the best two-variable model, \mathcal{M}_2 , will need to include variable X_2 (because in forward-stepwise selection the set of active variables at each step is nested).

However, from the assumptions we know that the best two-variable model only includes X_1 and X_3 , and not X_2 . So the forwards-stepwise selection model would fail at finding the best two-variable model.

2, A

sim. seen \Downarrow

1, B

1, C

1, C

2. (a) (i) To check the significance of β_1 , we test

meth seen ↓

$$H_0 : \beta_1 = 0 \quad v \quad H_1 : \beta_1 \neq 0$$

Since the p -value, p_1 , is 0.003209 and $p_1 < \alpha$, we have enough evidence to reject H_0 and conclude that β_1 is significantly different from 0. Similarly, to check the significance of β_i , $i = 2, 3$, we test

1, A

$$H_0 : \beta_i = 0 \quad v \quad H_1 : \beta_i \neq 0$$

Since the corresponding p -value, $p_2 = 3.33E^{-8} < \alpha$, and $p_3 = 0.000334 < \alpha$, we have enough evidence to reject H_0 in both cases and conclude that β_i is significantly different from 0 for each $i = 1, 2$.

1, A

- (ii) To test the overall fit of the model we use the hypothesis test $H_0 = \beta_1 = \beta_2 = \beta_3 = 0$. Now, we use the F -statistics for the overall fit and the corresponding p -value. Since the p -value is $1.124E^{-10} < \alpha$, we reject H_0 and conclude that there is an overall significant fit.

2, A

- (b) (i) FALSE. Lasso regression estimates do not have a closed form in general, the expression given here corresponds only to the case of a design matrix with orthonormal columns.

1, A

1, A

- (ii) FALSE. Ridge regression method is a shrinkage method and a sufficiently large tuning parameter could only shrink the coefficient towards zero but they will not be exactly zero (unlike, e.g., the lasso estimates).

1, A

meth seen ↓

- (c) This method clusters data according to the minimum *Euclidean distance* to the mean of the cluster. It requires the *specification of the desired number of clusters* K and initial starting centroids.

1, A

Then, at each iteration the algorithm updates the cluster by assigning each data to the cluster whose mean has the smallest distance to the given data. It *stops when there is no change* in the cluster configuration.

1, A

- (d) Iteration 1.

Since the initial clusters are $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5\}$, we first compute the means of each cluster. For cluster $C_1 = \{1, 2, 3\}$, the mean is

$$\bar{m}_1 = \left(\frac{1}{3}(1 + 1.5 + 3), \frac{1}{3}(1 + 2 + 4) \right)^T \approx (1.82, 2.33)^T$$

For cluster $C_2 = \{4, 5\}$, the mean is

$$\bar{m}_2 = \left(\frac{1}{2}(5 + 3.5), \frac{1}{2}(7 + 5) \right)^T = (4.25, 6)^T$$

1, A

Computing the Euclidean distances from each point to the means above, it follows that

i	Distance to \bar{m}_1	Distance to \bar{m}_2	Cluster
1	1.57	5.96	1
2	0.47	4.85	1
3	2.03	2.36	1
4	5.64	1.25	2
5	3.14	1.25	2

Since there is no change in the initial clusters, we stop the algorithm. The final clusters are $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5\}$.

- (e) Classical multidimensional scaling is a distance-based method that allows us to recover the configuration of a data matrix X given only its associated matrix of Euclidean distances E .

The eigen-decomposition of the inner product matrix B (defined as the product $X^T X$) plays an important role as it gives us a closed form to obtain X from B .

Indeed, once we have obtained B from the distance matrix E , using the eigen-decomposition of B , say $B = U\Lambda U^T$, where U is the corresponding matrix of eigenvectors (properly normalised) and Λ is the diagonal matrix of eigenvalues of B , then we can recover the configuration X by setting $X = U_1\Lambda_1^{1/2}$, where U_1 is the matrix that contains only the eigenvectors corresponding to non-zero eigenvalues and Λ_1 is the diagonal matrix with the non-zero eigenvalues.

- (f) Since E arises from a $n \times q$ matrix of full rank, it follows that the matrix B would be of rank q , and so there would be $n - q$ eigenvalues equal to zero.

- (g) In this context, the basis expansion approach allows more flexibility to model $f(X) = \mathbf{E}[Y|X]$ beyond linearity assumptions. The general idea is to transform the original predictors into new variables, say via a pre-defined transformation $h_m : \mathbb{R}^p \rightarrow \mathbb{R}$ so that $f(X)$ is written as a *linear basis expansion*

$$f(X) = \sum_{m=1}^M \beta_m h_m(X).$$

In this way, we can think of our model as a linear model with new predictors $\tilde{X}_m := h_m(X)$, and so to fit this model we can apply any of the methods known for the linear case.

3. (a) A cubic spline with K knots is a spline of order 4 fitted in $K + 1$ regions, and so there are $4(K + 1)$ parameters (4 parameters per each region), but there are three constraints per knot (as it needs to be C^2 at the knots). Hence, the total number of degrees of freedom is $4(K + 1) - 3K = K + 4$.

meth seen ↓

1, D

1, D

1, D

- (b) (i) The function $f(X)$ can be written as

$$f(X) = \sum_{m=0}^p \beta_m h_m(X),$$

where the $h_m : x \mapsto x^m$, for $m \in \{0, 1, 2, 3\}$,

And $h_{3+i}(x) = (x - \xi_i)_+^3$, where ξ_i denotes the i th knot for $i = 1, 2, 3$.

1, B

1, C

- (ii) Using the transformations h_m , the regression cubic spline gives a linear model in β with an intercept and predictors $\tilde{X}_1 = X^1$, $\tilde{X}_2 = X^2$, $\tilde{X}_3 = X^3$, $\tilde{X}_4 = (X - \psi_1)_+^3$, $\tilde{X}_5 = (X - \psi_2)_+^3$ and $\tilde{X}_6 = (X - \psi_3)_+^3$.

1, C

Hence, defining the design matrix

2, C

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & (x_1 - \xi_1)_+^3 & (x_1 - \xi_2)_+^3 & (x_1 - \xi_3)_+^3 \\ 1 & x_2 & x_2^2 & x_2^3 & (x_2 - \xi_1)_+^3 & (x_2 - \xi_2)_+^3 & (x_2 - \xi_3)_+^3 \\ \vdots & \vdots & \dots & & & \vdots & \vdots \\ 1 & x_{10} & x_{10}^2 & x_{10}^3 & (x_{10} - \xi_1)_+^3 & (x_{10} - \xi_2)_+^3 & (x_{10} - \xi_3)_+^3 \end{pmatrix}$$

and setting $y = (y_1, \dots, y_{10})^T$, we can then re-use the results of standard least squares estimation to obtain that the fitted line is given by

1, C

$$\hat{y} = \tilde{\mathbf{X}} \hat{\beta}^{LS} = \tilde{\mathbf{X}} (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T y.$$

We were allowed to assume that the inverse matrix $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}$ exists.

1, A

- (c) Given the training data $\{(x_i, y_i)\}_{i=1}^n$, a smoothing spline with tuning parameter $\lambda \geq 0$ is a function f which minimises the penalised residual sum of squares:

$$RSS(\lambda, f) = \|y - f(x)\|_2^2 + \lambda \int (f''(x))^2 dz,$$

over a class of functions with continuous first and second derivatives. Here, $y = (y_1, \dots, y_n)$ and $f(x) = (f(x_1), \dots, f(x_n))^T$.

1, A

Three main properties:

* it is a piecewise cubic polynomial with knots at the unique points x_1, \dots, x_n .

1, A

* it is C^2 at the knots, and

1, A

* it satisfies a linear constraint beyond the boundaries knots.

Remark: Other sensible properties are permissible/possible.

(d) Step 1 . Since a smoothing spline is a natural spline with knots at the points $\{x_i\}_{i=1}^n$, we can represent it as

$$f(x) = \sum_{i=1}^n N_i(x)\theta_i,$$

where N_i are the associated basis functions.

1, B

Step 2 . Using the above, for a given $\lambda \geq 0$, rewrite the penalised RSS as

1, B

$$RSS(\lambda, f) = \|y - N\theta\|_2^2 + \lambda\theta^T\Omega\theta,$$

where N is a matrix with entries $N_{ij} = N_j(x_i)$ and Ω is a matrix with entries $\Omega_{jk} = \int N_j''(t)N_k''(t) dt$.

Step 3 . Find the estimates $\hat{\theta}$ of θ by observing that the minimisation problem is a *generalised ridge regression problem*. Hence,

$$\hat{\theta} = (N^T N + \lambda\Omega)^{-1} N^T y,$$

And so the fitted line is given by $\hat{f} = N\hat{\theta}$.

1, A

Step 4 . By definition, the smoother matrix S_λ is obtained as

1, A

$$S_\lambda = N(N^T N + \lambda\Omega)^{-1} N^T.$$

(e) Let's just directly apply the variance operator:

$$\text{var}(\hat{d}_{j,k}) = n^{-2} \sum_{i=1}^n \text{var} \{ \psi_{j,k}(X_i) \} \quad (11)$$

$$= n^{-2} \sum_{i=1}^n (\mathbb{E} \{ \psi_{j,k}^2(X_i) \} - [\mathbb{E} \{ \psi_{j,k}(X_i) \}]^2) \quad (12)$$

$$= n^{-2} \sum_{i=1}^n \left\{ \int \psi_{j,k}^2(x) f(x) dx - d_{j,k}^2 \right\}. \quad (13)$$

From definition in lecture notes, for Haar, it is easy (but not intuitive) to see that $\psi^2(x) = \phi(x)$. Hence,

$$\{ \psi_{j,k}(x) \}^2 = 2^j \psi(2^j x - k)^2 = 2^j \phi(2^j x - k) = 2^{j/2} \phi_{j,k}(x). \quad (14)$$

Hence, continuing (13) gives

$$\text{var}(\hat{d}_{j,k}) = n^{-1} (2^{j/2} c_{j,k} - d_{j,k}^2), \quad (15)$$

as required.

3, D

4. (a) Given a centred data matrix $X \in \mathbb{R}^{n \times p}$, the principal components z_k are defined as $z_k = Xv_k$, $k = 1, \dots, M$, for some $M \leq p$, (i.e. the projection of the data X onto the vector v_k),

meth seen ↓

where v_k are the eigenvectors corresponding to the eigendecomposition of the matrix $X^T X$.

1, B

- (b) In principal components regression the principal components (as defined above) $z_i = Xv_i$, for $i = 1, \dots, M$, are used to replace the original predictors X_j , $j = 1, \dots, p$.

1, B

meth seen ↓

The corresponding fitted regression line is obtained then by regressing y on z_1, \dots, z_M , for some $M \leq p$. Since the principal components are orthogonal, the fitted line is given by a linear combination of the principal components and takes the form

1, B

$$\hat{y} = \bar{y}\mathbf{1}_n + \sum_{m=1}^M \frac{\langle y, z_m \rangle}{\langle z_m, z_m \rangle} z_m,$$

where $\mathbf{1}_n$ is an n -dimensional vector of ones.

1, D

- (c) Two possible advantages:

1. It can be used as a dimension reduction technique by finding a low-dimensional representation of the data but keeping as much as possible of the variation.
2. It helps to deal with explanatory variables that are highly correlated.

1, B

1, B

- (d) Given a (centred and sphered) data matrix X ,

meth seen ↓

exploratory projection pursuit is a method that searches the unitary directions onto which our *projected data show greatest divergence from a Gaussian distribution*.

1, C

The negative entropy H plays an important role as it is the *criterion used to measure the divergence of the density estimates of the projected data* from the Gaussian distribution.

1, D

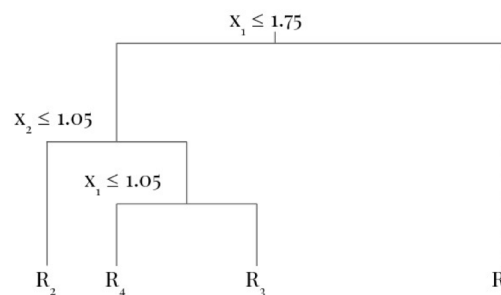
1, C

- (e) FALSE. The bandwidth h needs to be chosen to balance the corresponding bias and variance.

1, A

sim. seen ↓

- (f) (i) The classification tree is given by



2, D

- (ii) First note that $x^* = (x_1, x_2)$ falls into R_4 .

1, A

To predict the output of x^* , we need to compute the mean of all the data that fall into R_4 . Moreover, since

i	1	2	3	4	5	6	7	8
x_1	1.5	0.6	0.9	0.4	1.9	0.2	1.2	1.6
x_2	1.8	0.3	1.4	1.8	1.8	1.2	1.3	0.9
y	0.4	0.6	0.7	0.5	0.2	0.6	0.6	0.5
Region	R_3	R_2	R_4	R_4	R_1	R_4	R_3	R_2

it follows that the mean response for the training observation in region R_4 is $(0.5 + 0.7 + 0.6)/3 = 0.6$. So the prediction of x^* is 0.6.

2, D

1, D

- (g) FALSE. The random forest is a bagging algorithm with trees, not a boosting algorithm.

1, A

- (h) Bagging is a procedure for reducing the variance of a statistical learning method and so it is useful in the context of regression trees which usually suffer from high variance.

1, D

To apply bagging to regression trees one needs to

1, D

- obtain B bootstrapped training sets.
- construct B regression trees using the B bootstrapped training sets.
- average the predictions of the B regression trees.

- (i) FALSE. Boosting does not involve bootstrap sampling of the training data to fit each tree; instead each tree is fit using information from previously grown trees.

1, C

5. (a) Here some possible differences:

2, M

1. In linear regression models we are interested in estimating the true relationship f between the response Y and the input variable X as a conditional expectation $f(X) = E[Y|X]$, whereas logistic regression is interested in estimating conditional probabilities $P(Y|X)$.
2. Linear regression models require quantitative response variables, whereas logistic regression deals with qualitative response variables.
3. Unlike logistic regression, linear regression is not a classification method.

(b) A linear model of the form $P(Y|X) = \beta_0 + \beta_1 X$ would not be feasible for conditional probabilities as one would not be able to guarantee that $P(Y|X)$ takes values between 0 and 1 unless additional restrictions are imposed on X and the coefficients.

1, M

(c) By the *independence* assumption and the logarithm properties

$$\log p(y|X; \beta) = \sum_{i=1}^n \log p(y_i|x_i; \beta) \quad (16)$$

where

$$p(y_i|x_i; \beta) = P[Y = y_i|X = x_i; \beta].$$

1, M

The logistic regression model yields

$$p(y_i = 1|x_i; \beta) = \frac{e^{y_i \beta^T x_i}}{1 + e^{y_i \beta^T x_i}},$$

and

$$p(y_i = -1|x_i; \beta) = 1 - \frac{e^{y_i \beta^T x_i}}{1 + e^{y_i \beta^T x_i}} = \frac{1}{1 + e^{y_i \beta^T x_i}}.$$

1, M

Substituting both expressions above into (16) yields

1, M

$$\log p(y|X; \beta) = \sum_{i=1}^n -\log(1 + \exp(-y_i \beta^T x_i)). \quad (17)$$

2, M

(d) One could use dummy variables X_1 , X_2 and X_3 defined as follows

2, M

$$X_1 = \begin{cases} 1 & \text{if green} \\ 0 & \text{o.w.} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if blue} \\ 0 & \text{o.w.} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if pink} \\ 0 & \text{o.w.} \end{cases} \quad (18)$$

- (e) Set $g(x) = p(y = 1|x)$ and note that $p(y = -1|x) = 1 - g(x)$. The log odds of $g(x)$ are given by

$$\text{log odds} \stackrel{\text{def}}{=} \log \frac{p(y = 1|x)}{p(y = -1|x)}.$$

Hence, fitting a linear regression model to the log odds means

$$\log \frac{g(x)}{1 - g(x)} = \beta^T x.$$

By taking exponentials and rearranging terms

1, M

$$g(x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}},$$

and so

1, M

$$p(y = 1|x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}},$$

and

$$p(y = -1|x) = \frac{1}{1 + e^{\beta^T x}}$$

as required for the logistic model with responses in $\{-1, 1\}$.

1, M

- (f) By assumption on the log odds, and doing similar calculations as those in (e), it follows that

$$P(Y = 1|X; \beta) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}$$

and so

$$P(Y = 0|X; \beta) = \frac{1}{1 + e^{\beta^T X}}.$$

Taking into account the intercept, we set $X = (1, 2, 1, -1)^T$ and then compute $\hat{\beta}^T X = 2$. Therefore, the probability that the input $x = (2, 1, -1)^T$ belongs to class 0 is equal to $1/(1 + e^2) \approx 12\%$.

1, M

1, M

- (g) (i) The term 'linear' here refers to the use of a linear predictor (e.g. $\beta^T x$ in the logistic regression, as shown in part (e) above. However, this predictor then gets transformed through the non-linear logit function so that the line is a curve, not straight.
- (ii) The three methods that are not nonparametric are not as flexible (so the linear logistic has to follow a smooth curve and the two stepped lines can only be piecewise constant and also only on the bins that they are defined on) and so interesting little dips (such as around $x = 125$) would not be shown. The fit might be better and hence the model parameters more accurate, leading to more accurate predictions. The area on Figure 2 is the dip. However, students might draw attention to the noticeable increase in risk for pressures less than 115.

1, M

2, M

- (iii) Any solution that mentions something like the alternating algorithm for fitting projection pursuit regression in lectures. This involves a scatter plot fit of the transformed data for each variable separately. Then the algorithm successively does this fit minimising the RSS for each variable consecutively, holding the others constant.

2, M

[The CHD dataset for this part and figures reproduced from Hastie, T. and Tibshirani, R. (1987) Non-Parametric Logistic and Proportional Odds Regression. *Journal of the Royal Statistical Society, Series C. Applied Statistics*, **36**, 260–276. Under academic fair use by 'Exceptions to copyright' from the Intellectual Property Office, U.K.]

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.

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
----------------	----------------	-----------------------

	1	
--	---	--

		N/A
--	--	-----

	2	
--	---	--

		Part (c) Answered reasonably well although many students did not mention that the K-means was an iterative technique
--	--	--

	3	
--	---	--

		Part (c) Answered reasonably well although many students did not mention that the K-means was an iterative technique. Part (d) was answered extremely well in general. A few people made the occasional slip, but got full credit for getting the method right. Part (e) was answered well. Very few people managed to answer part (f) correctly. Most people did ok on part (g), the key was identifying the resulting linearity in the coefficients.
--	--	--

	4	
--	---	--

		Many students got part (a) correct, but a surprising number didn't. Most students answered part (b) and (c) well. Students managed to write down the specification of the problem well in part(d), but fewer managed to identify the smoother matrix. For part (e) most students got some marks here, and made some progress. Few managed to complete the question entirely correctly.
--	--	--

	5	
--	---	--

		N/A
--	--	-----