IMPERIAL

**MATH70049**

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2024

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

## Introduction to Statistical Learning

Date: Monday, May 20, 2024

Time: 10:00 – 12:30 (BST)

Time Allowed: 2.5 hours

**This paper has 5 Questions.**

**Please Answer All Questions in 1 Answer Booklet**

Candidates should start their solutions 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**

Table 1: Correlation matrix for all `BestWidge` variables.

| | cost.man | time.man | size | price |
|---|---|---|---|---|
| cost.man | 1.00 | 0.88 | -0.08 | 0.44 |
| time.man | 0.88 | 1.00 | -0.12 | 0.37 |
| size | -0.08 | -0.12 | 1.00 | 0.29 |
| price | 0.44 | 0.37 | 0.29 | 1.00 |

1. (a) Let $n \times 1$ vector $Y$ be a response variable and let $X$ be an associated design matrix of order $n \times p$, $n > p$ and $X$ is of full rank . Suppose we formulate the linear model:

$$Y = X\beta + \epsilon, \tag{1}$$

where $\beta$ is a $p \times 1$ vector of parameters and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ is a set of independent normal random variables with mean zero and variance of $\sigma^2 > 0$.

Using matrix-vector notation state the least-squares criterion for linear modelling for estimation of $\beta$ and derive the least-squares estimator $\hat{\beta}$ of $\beta$. State the formula for the ridge regression estimator, $\hat{\beta}^{\text{ridge}}(\lambda)$, for $\beta$ with penalty parameter $\lambda > 0$. (3 marks)

(b) Prove that the ridge regression estimator is identical to the posterior mean for the linear model given in (1) when the parameters, $\beta$, have independent and identically distributed prior distributions $\beta_j \sim N(0, \tau^2)$ for $j = 1, \ldots, p$ and assuming that $\sigma^2$ is fixed and known. (6 marks)

(c) `BestWidge` is a manufacturer that makes 43 different kinds of widgets and collects data for each kind on the following variables: *Price:* the selling price in pence, *cost.man:* the cost of manufacture in pence, *time.man:* the time taken to manufacture in minutes and the *size* of the widget in centimetres. `BestWidge` wants to model the dependent variable *Price* in terms of the explanatory variables *cost.man, time.man* and *size*.

The explanatory variables are plotted against each other in Figure 1. The correlation matrix between all variables is shown in Table 1.

(i) A least-squares model for *Price* in terms of all the explanatory variables was fitted in R with the following output:

```
 Call:
lm(formula = price ~ cost.man + time.man + size, data = the.df)

Residuals:
    Min      1Q  Median      3Q     Max
-680.34 -237.24   20.08  216.12  776.61

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.698    395.000   0.247   0.8059
```

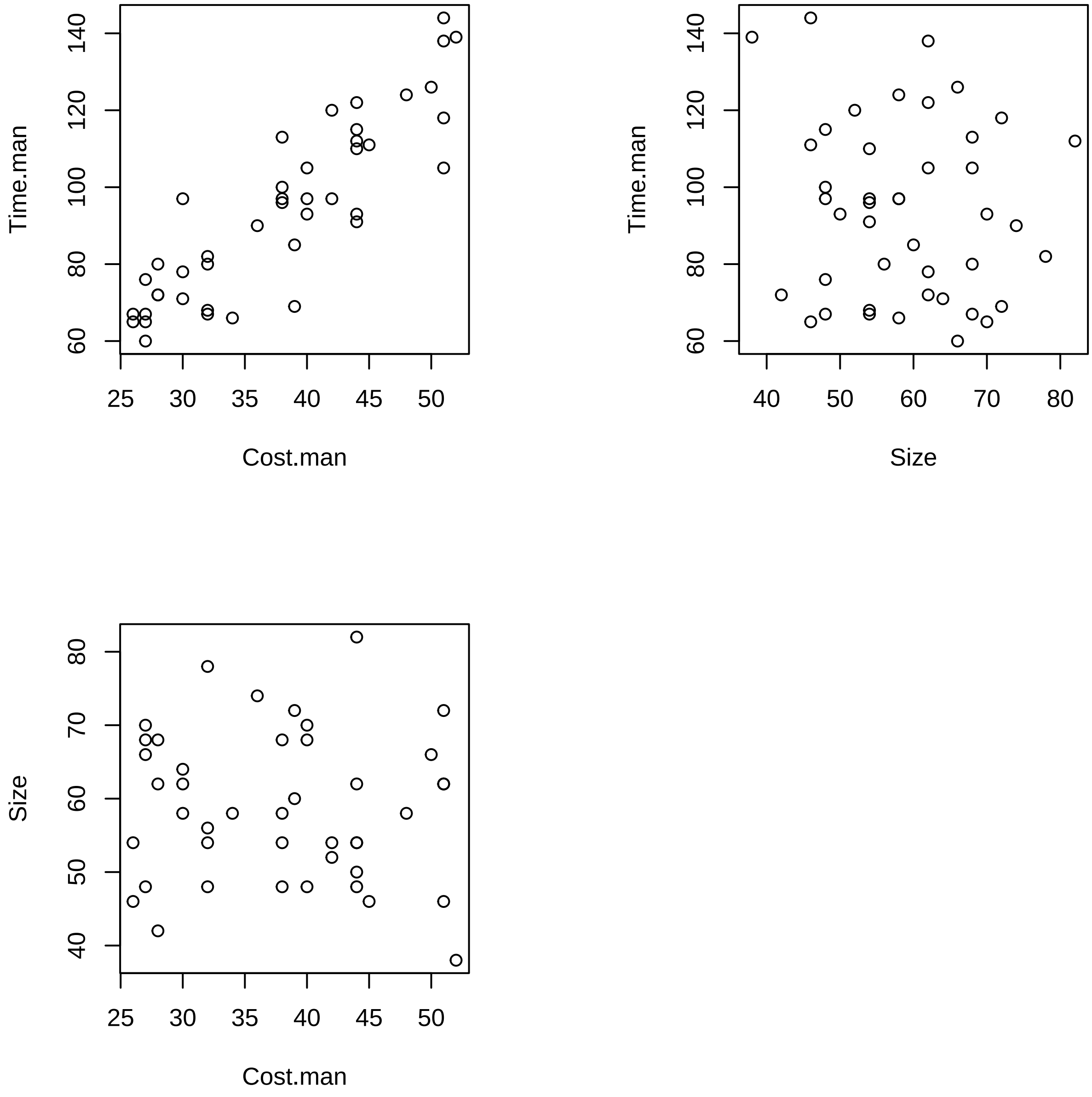

Figure 1: `BestWidge` explanatory variables *cost.man, time.man* and *size* plotted against each other.

```
cost.man      20.447     12.992   1.574   0.1236
time.man       0.338      4.675   0.072   0.9427
size          12.267      5.046   2.431   0.0198 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 328.6 on 39 degrees of freedom
Multiple R-squared:  0.2994,Adjusted R-squared:  0.2455
F-statistic: 5.556 on 3 and 39 DF,  p-value: 0.002839
```

Then, the same model was fitted without the *time.man* variable and the results of this fit are shown next:

```
 all:
lm(formula = price ~ cost.man + size, data = the.df)

Residuals:
    Min      1Q  Median      3Q     Max
-672.85 -235.76   20.22  214.60  776.26

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  100.607    388.029   0.259  0.79675
cost.man      21.273      6.091   3.493  0.00118 **
size          12.225      4.950   2.470  0.01789 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 324.5 on 40 degrees of freedom
Multiple R-squared:  0.2993,Adjusted R-squared:  0.2643
F-statistic: 8.544 on 2 and 40 DF,  p-value: 0.0008132
```

Explain why the value of the *cost.man* variable is different in the two model fits and why it is not statistically significant in the first, but it is in the second. (2 marks)

(ii) A lasso model is fitted to all variables and the cross-validated parameter estimates for *cost.man* is equal to $19.14$, *size* is equal to $10.50$ and *time.man* is set to exactly zero. A plot of the lasso estimators for various values of the tuning parameter, $\lambda$ is shown in Figure 2. What are the values of the parameters at the far left-hand and the far right-hand of the plot and, with reference to the lasso objective function, explain why they are these values? How do the cross-validated lasso parameter estimates compare to the least squares ones? (3 marks)

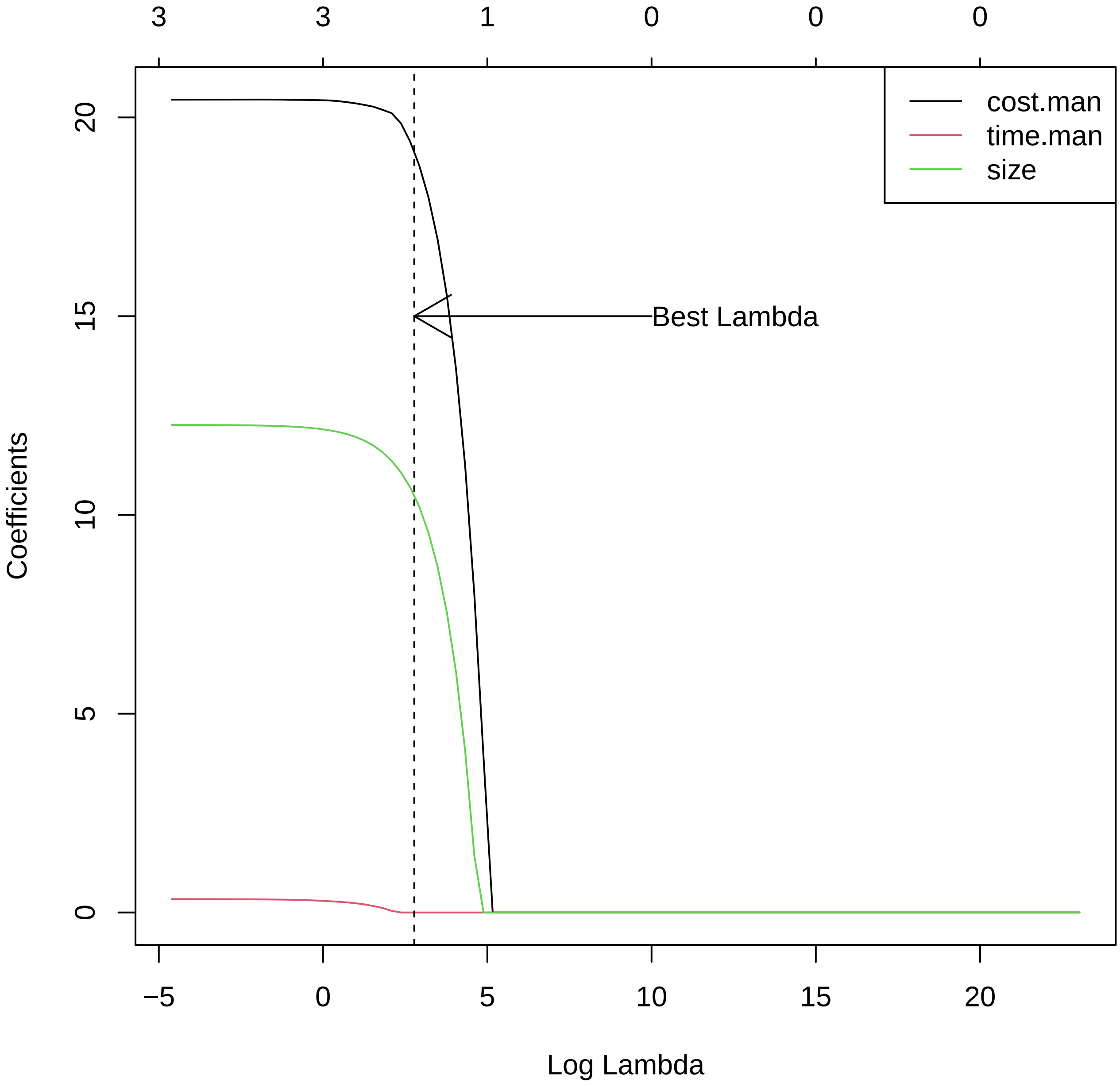

Figure 2: Lasso estimators for various values of tuning parameter $\lambda$ for the `BestWidge` variables..

(d) Now consider the ridge regression estimator $\hat{\beta}^{\text{ridge}}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y$, where $\lambda > 0$ is the ridge parameter. Define the squared two-norm of a $p \times 1$ vector $w$ to be $||w||_2^2 = w^T w$. From the formula or the defining ridge optimisation problem or otherwise, show that

(i) $\lim_{\lambda \to \infty} ||\hat{\beta}^{\text{ridge}}(\lambda)||_2^2 = 0$. (1 mark)

(ii) $\dfrac{\mathrm{d}||\hat{\beta}^{\text{ridge}}(\lambda)||_2^2}{\mathrm{d}\lambda} < 0$, for $\lambda > 0$. [Hint: the singular value decomposition $X = UDV^T$ might be useful]. (5 marks)

(Total: 20 marks)

2. (a) Suppose $X$ is an $n \times p$ data matrix with $n > p$. Explain how to form the centred data matrix $X_C$. (1 mark)

(b) Define the inner product matrix $B$ associated with $X$ from part (a) and write down its dimension. Explain why any orthogonal rotation of $X$ has the same $B$ matrix.

(2 marks)

(c) Let $E = (e_{i,j})$ be the squared Euclidean distance matrix associated with $X$ from part (a). Derive a formula for $E$ in terms of $B$ or entries of $B$. For any Euclidean distance matrix defined in this way, what are its diagonal entries? Compute both $B$ and $E$ when $X^T = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$.

(7 marks)

(d) Define $e_{\bullet,\ell} = \sum_{m=1}^{n} e_{m,\ell}$ and similarly for $e_{m,\bullet}$, $b^*_{\bullet,\bullet} = \sum_{m=1}^{n} b_{m,m}$ and $e_{\bullet,\bullet} = \sum_{m=1}^{n} \sum_{\ell=1}^{n} e_{m,\ell}$. You are given

$$e_{m,\bullet} = b^*_{\bullet,\bullet} + n b_{m,m} \tag{2}$$

and

$$e_{\bullet,\bullet} = 2n b^*_{\bullet,\bullet}. \tag{3}$$

Derive a formula to obtain a suitable inner product matrix $B$ from a general Euclidean matrix $E$. [Hint: you can assume that row and column sums of $B$ will be zero]. Use your formula to calculate a suitable $B'$ from the Euclidean matrix $E$ you obtained in part (c) that was obtained from $X^T = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}$. Why is $B'$ different to the $B$ you obtained in part (c)?

(6 marks)

(e) Suppose an inner product matrix $D$ was obtained from a new Euclidean distance matrix $F$ using the method you explained in part (d). Briefly explain how you might recover a suitable configuration $Y$ from $D$ that has Euclidean distances equal to $F$. You do not need to show that $YY^T = D$ nor explain how many of the non-zero eigenvalues to retain.

(2 marks)

(f) Explain why one might use ordinal scaling instead of classical scaling? Name one advantage of classical scaling over ordinal scaling. (2 marks)

(Total: 20 marks)

3. The Haar father wavelet is defined to be

$$\phi(x) = \begin{cases} 1 & x \in (0,1), \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The Haar mother wavelet is defined to be

$$\psi(x) = \begin{cases} 1 & x \in (0,1/2), \\ -1 & x \in [1/2,1), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

For the remainder of this question ALL mother and father wavelets are Haar wavelets.

(a) What is the difference between $\psi^2(x)$ and $\phi(x)$? Define the squared norm of a function $f \in L_2(\mathbb{R})$ to be $||f||^2 = \int_{-\infty}^{\infty} f^2(x)\,dx$. Calculate $||\phi||$ and $||\psi||$. (3 marks)

(b) Define the set of Haar wavelets $\mathcal{W} = \{\psi_{j,k}(x)\}_{j,k\in\mathbb{Z}}$, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, for all $x \in \mathbb{R}$ and $j,k \in \mathbb{Z}$ and $\psi$ is the Haar wavelet. Show that $\mathcal{W}$ is an orthonormal set.

(6 marks)

(c) For a function $f \in L_2(\mathbb{R})$ the Haar wavelet expansion of $f$ into $\mathcal{W}$ is given by

$$f(x) = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x), \tag{6}$$

with coefficients given by

$$d_{j,k} = \int_{-\infty}^{\infty} f(x)\psi_{j,k}(x)\,dx. \tag{7}$$

(i) Suppose now that $f(x)$ is an unknown probability density function on $\mathbb{R}$. Show that $d_{j,k} = \mathbb{E}\{\psi_{j,k}(X)\}$ and $c_{j,k} = \mathbb{E}\{\phi_{j,k}(X)\}$ for all $j,k \in \mathbb{Z}$. (1 mark)

(ii) Show that

$$\mathbb{E}\left\{\psi^2_{j,k}(X_i)\right\} = 2^{j/2}c_{j,k}, \tag{8}$$

for all $j,k \in \mathbb{Z}$. (1 mark)

(iii) Suppose further that we have access to an independent and identically distributed sample of values $\{X_i\}_{i=1}^n$ from $f(x)$. Let $\hat{d}_{j,k}$ and $\hat{c}_{j,k}$ be the empirical estimators of $d_{j,k}$ and $c_{j,k}$ respectively, given by

$$\hat{d}_{j,k} = n^{-1}\sum_{i=1}^n \psi_{j,k}(X_i) \text{ and } \hat{c}_{j,k} = n^{-1}\sum_{i=1}^n \phi_{j,k}(X_i), \tag{9}$$

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

Use the result (8) from the previous part to show that

$$\text{var}(\hat{d}_{j,k}) = n^{-1}(2^{j/2}c_{j,k} - d^2_{j,k}), \tag{10}$$

for all $j,k \in \mathbb{Z}$. Show that the estimator $\hat{d}_{j,k}$ is mean-squared consistent for $d_{j,k}$. Explain why $\text{var}(\hat{d}_{j,k})$ might easily be estimated and why is it useful to know this quantity?

(7 marks)

(d) Suppose we wish to regress $Y_i$ on two variables $X_{1,i}, X_{2,i}$ and we both construct (i) a regression tree and (ii) a two-dimensional wavelet shrinkage (regression) using Haar wavelets. Both methods give a piecewise-constant estimates. Briefly compare and contrast the two methods.

(2 marks)

(Total: 20 marks)

4. (a) Suppose we have a set of data which consists of a $p$-dimensional explanatory variable vector $X_i$ and a set of target values $Y_i \in \mathbb{R}$ for $i = 1, \ldots, n > 0$.

We wish to fit a regression tree to $Y_i$ using explanatory variables $X_i$. The regression tree is defined to be

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{I}(x \in R_m), \tag{11}$$

where $\{R_m\}_{m=1}^{M}$ is a partition of the $x$-variable space. We wish to minimize the criterion $SSQ = \sum_{i=1}^{n} \{Y_i - f(X_i)\}^2$.

For a given fixed variable $j = 1, \ldots, p$ and a split point $s$ (somewhere on $X_j$) define the pair of half-planes:

$$Q_1(j, s) = \{x | X_j < x\} \text{ and } Q_2(j, s) = \{x | X_j \geq s\}. \tag{12}$$

As part of a greedy algorithm to grow a regression tree we solve the following optimisation:

$$\min_{j,s} \left\{ \min_{c_1^*} \sum_{X_i \in Q_1(j,s)} (Y_i - c_1^*)^2 + \min_{c_2^*} \sum_{X_i \in Q_2(j,s)} (Y_i - c_2^*)^2 \right\}. \tag{13}$$

Explain how to find the minimising estimates of $\hat{c}_1^*, \hat{c}_2^*$. Further, briefly describe an efficient method to select the optimal $(j, s)$ and explain what the worst case order of computation would be for your method. (4 marks)

(b) Suppose we altered the objective function in $(13)$ by changing the terms $(Y_i - c_q^*)^2$ to $|Y_i - c_q^*|$ for $q = 1, 2; i = 1, \ldots, n$. How would this change the estimation? (1 mark)

(c) Briefly explain what a random forest is. Why can a random forest often be better than using a single decision tree? (4 marks)

(d) Suppose we have two regressions trees (A and B) that map input $x \in \mathbb{R}^p$ to output $\hat{y}$. Let $\hat{y}_A = f_A(x)$ and $\hat{y}_B = f_B(x)$. Define now the weighted sum of $f_A, f_B$ by:

$$f_C(x) = w f_A(x) + (1 - w) f_B(x), \tag{14}$$

for $w \in (0, 1)$. Is the weighted sum, $f_C$, always equivalent to a (deeper) regression tree? If so, show how it might be constructed. Give an example of such a construction based on two very simple trees and a weight $w = 0.4$. (8 marks)

(e) Given an example of an ethical failure arising from either a statistical, machine learning or artificial intelligence procedure. Suggest a way, or one or two principles, that might have prevented or mitigated such failure. (3 marks)

(Total: 20 marks)

5. (a) Suppose we have $R > 0$ data sets $\{G^{(r)}\}_{r=1}^{R}$ each of length $n > 0$ so that the $r$th data set is $G^{(r)} = (G_1^{(r)}, \ldots, G_n^{(r)})$ for $r = 1, \ldots, R$. Also, the values $G_i^{(r)} \in \mathbb{R}$, $\mathbb{E}G_i^{(r)} = 0$, $\operatorname{var} G_i^{(r)} = \sigma^2 > 0$ and the $G_i^{(r)}$ are all identically and independently distributed with density $f(x)$ for all $i = 1, \ldots n$ and $r = 1, \ldots, R$.

Let $\hat{f}_{r,h}(x)$ be a the kernel density estimator of $f(x)$ for set $G^{(r)}$, for $r = 1, \ldots, R$ and $h > 0$. In other words

$$\hat{f}_{r,h}(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{G_i^{(r)} - x}{h}\right), \tag{15}$$

where $K : \mathbb{R} \to \mathbb{R}$ is a kernel function, $K(u) > 0$ for all $u$, symmetric about zero, $K(u) = K(-u)$, and $\int_{-\infty}^{\infty} K(u)\, du = 1$.

All integrals below are over $\mathbb{R}$. The following are constants: $C_1 = \int v^2 K(v)\, dv$, $C_2 = \int K^2(v)\, dv$, $C_3 = \int v K^2(v)\, dv$ and $C_4 = \int v^2 K^2(v)\, dv$. As is customary in kernel density estimation we assume $h$ is such that both $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

From the kernel density estimates we compute the following distance between them

$$D_{r,q;h}(x) = \left\{\hat{f}_{r,h}(x) - \hat{f}_{q,h}(x)\right\}^2. \tag{16}$$

Derive the expectation $\mathbb{E}\left\{\hat{f}_{r,h}(x)\right\}$. Write any terms involving polynomials of $h$ of higher order than or equal to three as $\mathcal{O}(h^3)$. (5 marks)

(b) Derive the expectation $\mathbb{E}\left\{\hat{f}_{r,h}^2(x)\right\}$ and state its limit as $n \to \infty$. (8 marks)

(c) Derive the expectation $\mathbb{E}(D_{r,q;h})$ and show that, for fixed $x$, its asymptotic expectation is zero and the leading term of the bias does not depend on $f''(x)$. Show also that the expectation is mathematically simpler when $C_1 = 1$. (5 marks)

(d) Suppose an analyst collects the data $\{G^{(r)}\}_{r=1}^{R}$ as above and then wants to carry out scaling on $E_{r,q;h} = \int D_{r,q,;h}(x)\, dx$ as a dissimilarity. Would classical multidimensional or ordinal scaling be more appropriate to use here? How would you choose a suitable $h$ for this procedure?

(2 marks)

(Total: 20 marks)

BSc and MSci EXAMINATIONS (MATHEMATICS)

May 2024

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

# MATH60049/70049

# Introduction to Statistical Learning (Solutions)

| Setter's signature | Checker's signature | Editor's signature |
|---|---|---|
| ................ | ................ | ................ |

1. (a) The least-squares criterion is matrix-vector notation is $\mathrm{RSS}(\beta) = (Y - X\beta)^T(Y - X\beta)$. We can derive the least-squares estimator, which is the minimiser of $\mathrm{RSS}(\beta)$ by first rewriting the RSS as

seen ⇓

$$\begin{aligned}\mathrm{RSS}(\beta) &= (Y - X\beta)^T(Y - X\beta)\\ &= Y^TY - (X\beta)^TY - Y^TX\beta + (X\beta)^TX\beta\\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta.\end{aligned}$$

Differentiating with respect to $\beta$

$$\frac{\partial\,\mathrm{RSS}(\beta)}{\partial\beta} = -2X^TY + 2X^TX\beta. \tag{1}$$

Setting equal to zero and solving for $\beta$ gives

$$X^TY = X^TX\hat{\beta} \implies \hat{\beta} = (X^TX)^{-1}X^TY,$$

since $X$ is of full rank.

The formula for the ridge regression estimator is

$$\hat{\beta}^{\mathsf{ridge}}(\lambda) = (X^TX + \lambda I_p)^{-1}X^TY, \tag{2}$$

where $I_p$ is the $p$-dimensional identity matrix.

3, A

(b) Bayes theorem states:

seen ⇓

$$p(\beta|Y) = \frac{p(Y|\beta)p(\beta)}{p(Y)}. \tag{3}$$

However, we only need to consider functions of $\beta$ so we focus on

1, A

$$p(\beta|Y) \propto p(Y|\beta)p(\beta). \tag{4}$$

So, let's figure out $p(\beta)$ and $p(Y|\beta)$. The prior density for $\beta_i$ is proportional to

$$p(\beta_i) \propto \exp\left(-\frac{\beta_i^2}{2\tau^2}\right). \tag{5}$$

and for the vector $\beta$ with the component $\beta_i$ being independent gives

$$p(\beta) = \prod_{i=1}^{p} p(\beta_i) \propto \prod_{i=1}^{p}\exp\left(-\frac{\beta_i^2}{2\tau^2}\right) = \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{p}\beta_i^2\right) = \exp\left(-\frac{||\beta||_2^2}{2\tau^2}\right). \tag{6}$$

1, A

Also $Y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2)$ (the regression model), so similar to above

$$p(Y_i|\beta) \propto \exp\left\{-\frac{(Y_i - \beta_0 - x_i^T\beta)^2}{2\sigma^2}\right\} = \exp\left(-\frac{||Y - X\beta||_2^2}{2\sigma^2}\right), \tag{7}$$

incorporating the $\beta_0$ into the $\beta$ vector and the associated vector of 1s into the design matrix $X$. They should get credit for getting it basically right.

1, B

For the posterior density $p(\beta|Y)$ we use (4) to obtain:

$$\begin{aligned}p(\beta|Y) &\propto p(Y|\beta)p(\beta) && (8)\\ &= \exp\left(-\frac{||Y - X\beta||_2^2}{2\sigma^2}\right)\exp\left(-\frac{||\beta||_2^2}{2\tau^2}\right) && (9)\\ &= \exp\left\{-\frac{||Y - X\beta||_2^2 + (\sigma^2/\tau^2)||\beta||_2^2}{2\sigma^2}\right\} && (10)\\ &= \exp\left\{-\frac{||Y - X\beta||_2^2 + \lambda||\beta||_2^2}{2\sigma^2}\right\}, && (11)\end{aligned}$$

where $\lambda = \sigma^2/\tau^2$. 2, B

Therefore, $p(\beta|Y) \propto \exp\left(-\frac{||Y-X\beta||_2^2+(\sigma^2/\tau^2)||\beta||_2^2}{2\sigma^2}\right)$ is also normally distributed (which can be immediately recognised by the form, especially the quadratic and linear bits in $\beta$). Since the distribution is Gaussian, the mean is obtained by maximising the density function. Observe that, inside the exponential, it has the same form (up to a constant) as a ridge regression with $\lambda = \sigma^2/\tau^2$. The mode and the mean of this distribution is

$$\arg\min_{\beta} ||Y - X\beta||_2^2 + (\sigma^2/\tau^2)||\beta||_2^2, \tag{12}$$

which is indeed the ridge regression estimate. 1, B

(c) (i) The `cost.man` and `time.man` variables are highly correlated and this means that their associated parameter estimates are also highly correlated resulting in a high variance in the estimates (and `time.man`'s estimate influences `cost.man`'s and vice versa). This also leads to poor conditioning of $X^TX$, which results in an inflated variance for each parameter estimate. For example, the standard error for `cost.man` where `time.man` is present is 12.992, but it is 6.091 when `time.man` is omitted. This leads to `cost.man` not being significant in the first model, but it is the second model (even though the actual estimate is not much different from one to the other).

seen/sim.seen ⇓

2, B

(ii) The values of the parameter estimates at the left-hand end are just the least squares estimates that were presented in the first table of coefficients in the question's R output (20.447, 0.338,12.267 for `cost.man`, `time.man` and `size` respectively). All coefficients are exactly zero on the right-hand end of the plot. The lasso objective function is

seen/sim.seen ⇓

$$R(\lambda) = ||Y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{13}$$

or it can be written in other ways. For the left-hand side of the plot, this is when $\lambda = 0$, the penalty term is zero and then the situation collapses to the usual least-squares problem and hence this is why the parameters are the least-squares ones. For $\lambda$ very large the penalty becomes dominant and an optimiser shrinks all of the coefficients to exactly zero, this is depicted on the right hand side of the plot. 3, A

The cross-validated coefficients are all shrunk values of the least squares ones.

seen/sim.seen ⇓

(d) (i) We showed in lectures that when $\lambda \to \infty$ the penalty becomes dominant, which forces all of the elements of the $\hat{\beta}$ vector to become zero. The two norm of a zero vector is also zero (we did this for lasso too). 1, A

(ii) The ridge regression estimator is given in the question, so

unseen ⇓

$$||\hat{\beta}^{\text{ridge}}(\lambda)||_2^2 = Y^TX(X^T + \lambda I_p)^{-2}X^TY, \tag{14}$$

as $X^TX + \lambda I_p$ is symmetric. Note $X^T = VD^2V^T$, this was derived in lectures, or it is easily derived here using the SVD. 1, A

Now substitute the SVD given in the hint, $X = UDV^T$, into $(14)$ to derive:

$$\begin{aligned}
||\hat{\beta}^{\text{ridge}}(\lambda)||_2^2 &= Y^TUDV^T(VD^2V^T + \lambda I_p)^{-2}VDU^TY && (15)\\
&= Y^TUDV^T(V[D^2 + \lambda I_p]V^T)^{-2}VDU^TY && (16)\\
&= Y^TUDV^T(V^T)^{-2}(D^2 + \lambda I_p)^{-2}V^{-2}VDU^TY && (17)\\
&= Y^TUD(V^T)^{-1}(D^2 + \lambda I_p)^{-2}V^{-1}DU^TY && (18)\\
&= Y^TUDV(D^2 + \lambda I_p)^{-2}V^TDU^TY && (19)\\
&= a^TD^*(\lambda)a, && (20)
\end{aligned}$$

where $a = V^TDU^TY$ a $p \times 1$ vector, not depending on $\lambda$ and $D^*(\lambda) = (D^2 + \lambda I_p)^{-2}$ is a diagonal matrix. The previous lines use $VV^T = I_p$ to obtain line $(16)$, $(AB)^{-1} = B^{-1}A^{-1}$ to obtain line $(17)$ and $V^T = V^{-1}$ for orthogonal matrices to obtain line $(19)$. $\boxed{2, \mathrm{D}}$

For the derivative we can write

$$a^TD^*(\lambda)a = \sum_{i=1}^{p} a_i^2 D^*_{i,i}(\lambda), \qquad (21)$$

where $D^*_{i,i} = (d^2_{i,i} + \lambda)^{-2}$ and because $D^*(\lambda)$ is diagonal.

For the derivative note that

$$\frac{\mathrm{d}D^*_{i,i}}{\mathrm{d}\lambda} = \frac{\mathrm{d}(d^2_{i,i} + \lambda)^{-2}}{\mathrm{d}\lambda} = -2(d^2_{i,i} + \lambda)^{-3} = E^*_{i,i} < 0, \qquad (22)$$

Hence

$$\frac{\mathrm{d}||\hat{\beta}^{\text{ridge}}(\lambda)||_2^2}{\mathrm{d}\lambda} = a^TE^*(\lambda)a < 0, \qquad (23)$$

where $E^* = \mathrm{diag}(E^*_{i,i})$ since $E^*_{i,i} < 0$ for all $i$. $\boxed{2, \mathrm{B}}$

seen ⇓

2. (a) To centre a matrix merely create a new matrix of the same order with each column of the original replaced by the old column after subtracting the old column's mean. Alternatively, form the centring matrix $C = I_n - 11^T/n$, where 1 is an $n \times 1$ vector containing solely of 1s and then $X_C = CX$.

1, A

(b) The inner product matrix $B = XX^T$ and is of dimension $n \times n$. Suppose $Q$ is an $p \times p$ orthogonal matrix and $Z = XQ$ is a rotated version of $X$. Then $B_Z = ZZ^T = XQQ^TX^T = XI_pX^T = XX^T = B_X$.

2, A

(c) The (squared) Euclidean distance between cases $i, j$ in $X$ is given by

$$\begin{aligned} e_{i,j} &= \sum_{v=1}^{p}(X_{i,v} - X_{j,v})^2 && (24) \\ &= \sum_{v=1}^{p} X_{i,v}X_{i,v} + \sum_{v=1}^{p} X_{j,v}X_{j,v} - 2\sum_{v=1}^{p} X_{i,v}X_{j,v} && (25) \\ &= X_{(i)}^T X_{(i)} + X_{(j)}^T X_{(j)} - 2X_{(i)}^T X_{(j)}, && (26) \end{aligned}$$

where $X_{(i)}^T$ is the $i$th row of $X$, a $p$-dimensional vector, $i, j = 1, \ldots, n$ and these are merely entries in $B$, i.e. $b_{i,j} = X_{(i)}^T X_{(j)}$. Hence

3, A

$$e_{i,j} = b_{i,i} + b_{j,j} - 2b_{i,j}. \tag{27}$$

seen/sim.seen ⇓

The required matrices are $B = \begin{pmatrix} 5 & 11 & 17 \\ 11 & 25 & 39 \\ 17 & 39 & 61 \end{pmatrix}$ and $E = \begin{pmatrix} 0 & 8 & 32 \\ 8 & 0 & 8 \\ 32 & 8 & 0 \end{pmatrix}$.

4, B

seen ⇓

(d) Define $e_{\bullet,\ell} = \sum_{m=1}^{n} e_{m,\ell}$ and similarly for $e_{m,\bullet}$, $b^*_{\bullet,\bullet} = \sum_{m=1}^{n} b_{m,m}$ and $e_{\bullet,\bullet} = \sum_{m=1}^{n}\sum_{\ell=1}^{n} e_{m,\ell}$.

You are given

$$e_{m,\bullet} = b^*_{\bullet,\bullet} + nb_{m,m} \tag{28}$$

and

$$e_{\bullet,\bullet} = 2nb^*_{\bullet,\bullet}. \tag{29}$$

To obtain $b_{m,\ell}$ in terms of the $e_{\cdot,\cdot}$ we first rearrange (27) to give

$$b_{m,\ell} = -\tfrac{1}{2}(e_{m,\ell} - b_{m,m} - b_{\ell,\ell}), \tag{30}$$

and using (28) and (29) we have

$$\begin{aligned} b_{m,\ell} &= -\tfrac{1}{2}\left\{e_{m,\ell} - (e_{m,\bullet} - b^*_{\bullet,\bullet})/n - (e_{\bullet,\ell} - b^*_{\bullet,\bullet})/n\right\} && (31) \\ &= -\tfrac{1}{2}(e_{m,\ell} - \tfrac{e_{m,\bullet}}{n} - \tfrac{e_{\bullet,\ell}}{n} + 2\tfrac{b^*_{\bullet,\bullet}}{n}) && (32) \\ &= -\tfrac{1}{2}(e_{m,\ell} - \tfrac{e_{m,\bullet}}{n} - \tfrac{e_{\bullet,\ell}}{n} + \tfrac{e_{\bullet,\bullet}}{n^2}) && (33) \\ &= -\tfrac{1}{2}(\text{entry} - \text{row av.} - \text{col av.} + \text{grand av.}), && (34) \end{aligned}$$

or in matrix terms (if they remember this, they should get some credit)

2, A

$$B = -\frac{1}{2}(I_n - \mathbf{1}\mathbf{1}^T/n)\,E\,(I_n - \mathbf{1}\mathbf{1}^T/n). \tag{35}$$

1, A

seen/sim.seen ⇓

The suitable $B'$ they are likely to calculate is $B' = \begin{pmatrix} 8 & 0 & -8 \\ 0 & 0 & 0 \\ -8 & 0 & 8 \end{pmatrix}$. It is different because $B$ was calculated from an uncentred matrix and $B'$ assumes the configuration is centred (hence why the row and column sums were zero).

2, B

1, B

(e) First compute an eigendecomposition of $D$, i.e.

$\boxed{\text{seen} \Downarrow}$

$$D = \sum_{i=1}^{n} \lambda_i d^{(i)} d^{(i)T}, \tag{36}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n'} > 0$ and $\lambda_{n'+1}, \ldots, \lambda_n = 0$ and the $\{d^{(i)}\}$ are the eigenvectors of $D$. Define a new set of vectors $g^{(i)} = \sqrt{\lambda_i} d^{(i)}$ and define

$$Y_{n \times n'} = \begin{pmatrix} \vdots & \vdots & \cdots & \vdots \\ g^{(1)} & g^{(2)} & \cdots & g^{(n')} \\ \vdots & \vdots & \cdots & \vdots \end{pmatrix}. \tag{37}$$

This configuration will do the job. $\boxed{2, \text{A}}$

(f) Ordinal scaling is used when it is known, or suspected, that the dissimilarity measure used is not Euclidean or deviates from Euclidean distances. This can often be when non-Euclidean distance metrics have not been used and others such as Jaccard's distance or Gower's similarity coefficient has been used.

$\boxed{\text{seen/sim.seen} \Downarrow}$

An advantage of classical scaling over ordinal scaling is that the former is not based on an iterative algorithm, which can be computationally expensive and also get stuck in local optima, whereas classical scaling relies on simple finite calculations and an eigendecomposition, which is comparatively fast and one knows that one obtains the actual solution (not a local optima). $\boxed{2, \text{C}}$

3. (a) It is easy to see that $\psi^2(x) = \phi(x)$. The norms are as follows $||\phi||^2 = \int \phi^2(x)\,dx = \int_0^1 1\,dx = 1$. For the wavelet either calculate this directly or notice that $\psi^2(x) = \phi(x)$ and use result for $||\phi||$.

$\boxed{\text{seen} \Downarrow}$

$\boxed{2, \text{A}}$

$\boxed{1, \text{B}}$

$\boxed{\text{meth seen} \Downarrow}$

(b) To show that $\mathcal{W}$ is an orthonormal set we need to do two things (i) show that each member has unit norm and (ii) show that any two different members have zero inner product. Let's do this. The norm of an arbitrary member:

$$\begin{aligned} ||\psi_{j,k}||^2 &= \int_{-\infty}^{\infty} \psi_{j,k}^2(x)\,dx \qquad (38)\\ &= 2^j \int_{-\infty}^{\infty} \psi^2(2^j x - k)\,dx \qquad (39)\\ &= \int_{-\infty}^{\infty} \psi^2(v)\,dv \qquad (40)\\ &= ||\psi||^2 = 1, \qquad (41)\end{aligned}$$

where we substituted $v = 2^j x - k$, $dv = 2^j dx$ in equation (39).

$\boxed{1, \text{A}}$

Let $j, k, \ell, m$ be such that $j \neq \ell$ and $k \neq m$. Then

$\boxed{1, \text{B}}$

$$\begin{aligned} < \psi_{j,k}, \psi_{\ell,m} > &= \int_{-\infty}^{\infty} \psi_{j,k}(x)\psi_{\ell,m}(x)\,dx \qquad (42)\\ &= 2^{(j+\ell)/2} \int_{-\infty}^{\infty} \psi(2^j x - k)\psi(2^\ell x - m)\,dx \qquad (43)\\ &= 2^{(\ell-j)/2} \int_{-\infty}^{\infty} \psi(v)\psi\{2^{\ell-j}v + (2^{\ell-j}k - m)\}\,dv \qquad (44)\\ &= 2^{(\ell-j)/2} \int_{-\infty}^{\infty} \psi(v)\psi(2^r v - q)\,dv, \qquad (45)\end{aligned}$$

where $r = 2^{\ell-j}$ and $q = 2^{\ell-j}k - m$, where we substituted $v = 2^j x - k$ in line (43). Hence

$$< \psi_{j,k}, \psi_{\ell,m} >= 2^{(\ell-j)/2} \left\{ \int_0^{1/2} \psi_{r,q}(v)\,dv - \int_{1/2}^{1} \psi_{r,q}(v)\,dv \right\}. \qquad (46)$$

Since $j \neq \ell$, then $r$ cannot be equal to $1$ so the scale of $\psi_{r,q}(v)$ cannot be the same as the mother wavelet. If $\psi_{r,q}(v)$ is twice as wide as the mother wavelet (or wider), then either (i) the $\psi_{r,q}(v)$ wavelet does not overlap $(0, 1)$ at all, or a constant portion overlaps $(0, 1)$ and then the integral in (46) has to be zero (since the values on $(0, 1)$ are the same, and the integrals subtraction cancels. If $\psi_{r,q}(v)$ is half as wide, or any power of two fraction of the width, then $\psi_{r,q}(v)$ is either outside of $(0, 1)$ in which case the value of each integral is zero, or $\psi_{r,q}(v)$ is inside one, and only one, of the intervals $(0, 1/2)$ or $(1/2, 1)$. In each case, the 'zero-integral' property of the wavelet means that (46) is zero.

$\boxed{4, \text{C}}$

A good picture covering all bases would also be acceptable.

$\boxed{\text{seen} \Downarrow}$

(c) (i) Due to the 'law of the unconscious statistician' we have $d_{j,k} = \int \psi_{j,k}(x) f(x)\,dx = \mathbb{E}\{\psi_{j,k}(X)\}$ and similarly for $\phi$.

$\boxed{1, \text{A}}$

unseen ⇓

(ii) We calculate

$$\begin{aligned}
\mathbb{E}\left\{\psi_{j,k}^2(X_i)\right\} &= \int_{-\infty}^{\infty} \psi_{j,k}^2(x) f(x)\,dx \qquad (47)\\
&= 2^j \int_{-\infty}^{\infty} \psi^2(2^j x - k) f(x)\,dx \qquad (48)\\
&= 2^{j/2} \int_{-\infty}^{\infty} 2^{j/2}\phi(2^j x - k) f(x)\,dx \qquad (49)\\
&= 2^{j/2} \int_{-\infty}^{\infty} \phi_{j,k} f(x)\,dx \qquad (50)\\
&= 2^{j/2} c_{j,k}, \qquad (51)
\end{aligned}$$

as required, using the fact that $\psi^2(x) = \phi(x)$ from part (a), and the definitions of $\psi_{j,k}(x), \phi_{j,k}(x)$. 1, C

(iii) It's easy to show that $\mathbb{E}(\hat{d}_{j,k}) = n^{-1}\sum_{i=1}^{n} \mathbb{E}\left\{\psi_{j,k}(X_i)\right\} = d_{j,k}$ from part (i) above. So, to find $\mathrm{var}(\hat{d}_{j,k})$ we now compute

$$\begin{aligned}
\mathbb{E}(\hat{d}_{j,k}^2) &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left\{\psi_{j,k}(X_i)\psi_{j,k}(X_j)\right\} \qquad (52)\\
&= n^{-2}\left[\sum_{i=1}^{n} \mathbb{E}\{\psi_{j,k}^2(X_i)\} + \sum_{j=1, j\neq i}^{n} \mathbb{E}\left\{\psi_{j,k}(X_i)\psi_{j,k}(X_j)\right\}\right] \qquad (53)\\
&= n^{-2}\left\{n 2^{j/2} c_{j,k} + n(n-1) d_{j,k}^2\right\} \qquad (54)\\
&= n^{-1} 2^{j/2} c_{j,k} + n^{-1}(n-1) d_{j,k}^2, \qquad (55)
\end{aligned}$$

since $X_i$ and $X_j$ are independent for $i \neq j$. Hence, putting together 2, D

$$\begin{aligned}
\mathrm{var}(\hat{d}_{j,k}) &= \mathbb{E}(\hat{d}_{j,k}^2) - \mathbb{E}(\hat{d}_{j,k})^2 \qquad (56)\\
&= n^{-1}\left\{2^{j/2} c_{j,k} + (n-1) d_{j,k}^2 - n d_{j,k}^2\right\} \qquad (57)\\
&= n^{-1}(2^{j/2} c_{j,k} - d_{j,k}^2) = \sigma_{j,k}^2, \qquad (58)
\end{aligned}$$

as required. 2, D

Since $\hat{d}_{j,k}$ is an unbiased estimator of $d_{j,k}$ and $\mathrm{var}(\hat{d}_{j,k}) \to 0$ as $n \to \infty$, the estimator $\hat{d}_{j,k}$ is mean-squared consistent for $d_{j,k}$. The quantity $\mathrm{var}(\hat{d}_{j,k})$ might be easily estimated as $c_{j,k}$ and $d_{j,k}$ are both automatically computed as part of the Haar wavelet (pyramid) transform. It's useful to know the variance to help form a hypothesis test for $d_{j,k}$, for example (but there might be other reasons). 1, B

2, C

(d) The piecewise constant heights of the wavelet transform are restricted to a dyadic grid — i.e. the lengths of all sides are powers of two. This is not the case for trees, where the regions can be of any size. Trees are very easy to interpret, wavelets less so. Trees can easily handle variables of different types (ordinal, nominal, ratio, interval), whereas standard wavelets really only work for ratio or interval variables. unseen ⇓

2, D

4. (a) Conditional on $j, s$, then we can minimise the $R(c_q) = \sum_{X_i \in Q_q(j,s)} (Y_i - c_q)^2$ by simple differentiation and setting the differential equal to zero, i.e. seen ⇓

$$\frac{\mathrm{d}R(c_q)}{\mathrm{d}c_q} = 2 \sum_{X_i \in Q_q(j,s)} (Y_i - c_q) = 0 \qquad (59)$$

$$\implies \hat{c}_q = n_q^{-1} \sum_{X_i \in Q_q(j,s)} Y_i, \qquad (60)$$

where $n_q$ is the number of $Y_i$ in $Q_q$. I.e. the estimator $\hat{c}_q$ is the mean of the $Y_i$ in half-plane $Q_q$ for $q = 1, 2$. 2, A

Given a $j$, the best split point can be efficiently found as follows. The objective function inside the brackets of (11) only changes when $s$ crosses a $X_i$. Hence, there are only $n+1$ different values of the objective function as $s$ varies. This calculation can be performed once for each $j$, which results in the worst case computation order of $\mathcal{O}\{p(n+1)\}$. 2, A

unseen ⇓

(b) If we change the $\cdot^2$ to $|\cdot|$, then the estimator $\hat{c}_q = \mathrm{median}_{X_i \in Q_q(j,s)} Y_i$. 1, B

(c) One can often improve estimators by bagging. This consists of forming an average over a number of regression models produced on bootstrapped input data. For trees, it can be shown that the bootstrapped trees are very similar and, indeed, highly correlated and the bagged estimate's variance does not tend to zero for increased number of bagged trees. seen ⇓

Random forests are a way of injecting some more randomness into trees and makes them less correlated. Essentially, at each tree split decision a random selection of variables is chosen to search for the best split. If this step is not present, then dominant variables can essentially get 'used' far too many times, resulting in similar trees each time. This results in a less correlated ensemble of trees. 3, A

1, B

(d) The answer is yes, it can always be equivalent. One way to construct such a tree is to paste tree $f_B$ onto every terminal node of $f_A$. Then, the value of the output will be the value of the terminal node of $f_B$ combined in the same weighted combination of the terminal node of $f_A$ that the $f_B$ tree was pasted on to. unseen ⇓ 4, D

An example should be given according to the above prescription. 4, D

(e) This could be almost any example, either from the notes or from something that has recently occurred (for example, current generative AI systems hallucinating in a certain way, or having biases built-in to the system). unseen ⇓ 3, C

5. (a) To find the expectation of $\hat{f}_{r,h}(x)$ we apply expectation to formula (13) in the exam paper. seen ⇓

$$\mathbb{E}\left\{\hat{f}_{r,h}(x)\right\} = (nh)^{-1}\sum_{i=1}^{n}\mathbb{E}\left\{K\left(\frac{G_i^{(r)}-x}{h}\right)\right\}, \tag{61}$$

Let us now work out the expectation of the term on the RHS of (61), let's call this quantity $(*)$: 1, M

$$\begin{aligned}(*) &= \mathbb{E}\left\{K\left(\frac{G_i^{(r)}-x}{h}\right)\right\} &(62)\\ &= \int K\left(\frac{y-x}{h}\right) f(y)\,dy &(63)\\ &= h\int K(v)f(x+vh)\,dv, &(64)\end{aligned}$$

by the substitution $y = x + vh$, $dy = hdv$, as $h > 0$. We now use the Taylor expansion of $f$ around $x$ 1, M

$$f(x+vh) = f(x) + hvf'(x) + h^2v^2f''(x)/2 + \mathcal{O}(h^3). \tag{65}$$

Now substitute (65) into (64) to give 1, M

$$\begin{aligned}(*) &= hf(x)\cancelto{1}{\int K(v)\,dv} + h^2f'(x)\cancelto{0}{\int vK(v)\,dv}\\ &\quad +h^3f''(x)\cancelto{C_1}{\int v^2K(v)\,dv} + \mathcal{O}(h^4), &(66)\end{aligned}$$

as $K$ integrates to one and is symmetric about zero. Then, substituting (66) into (61) gives 1, M

$$\mathbb{E}\left\{\hat{f}_{r,h}(x)\right\} = f(x) + h^2C_1f''(x) + \mathcal{O}(h^3). \tag{67}$$

1, M

(b) We will calculate the expectation of the square in a similar way. meth seen ⇓

$$\begin{aligned}\mathbb{E}\left\{\hat{f}^2_{r,h}(x)\right\} &= (nh)^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left\{K\left(\frac{G_i^{(r)}-x}{h}\right)K\left(\frac{G_j^{(r)}-x}{h}\right)\right\}, &(68)\\ &= (nh)^{-2}\left[\sum_{i=1}^{n}\mathbb{E}\left\{K^2\left(\frac{G_i^{(r)}-x}{h}\right)\right\}\right. &(69)\\ &\quad \left.+\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathbb{E}\left\{K\left(\frac{G_i^{(r)}-x}{h}\right)\right\}\mathbb{E}\left\{K\left(\frac{G_j^{(r)}-x}{h}\right)\right\}\right].\end{aligned}$$

Each of the expectations on the far RHS are equal to $(*)$ as calculated in previous answer in (a). Multiplying two $(*)$ together and collecting orders gives 1, M

$$\begin{aligned}(*)^2 &= \{hf(x) + h^3f''(x) + \mathcal{O}(h^4)\}\{hf(x) + h^3f''(x) + \mathcal{O}(h^4)\} &(70)\\ &= h^2f^2(x) + 2h^4f(x)f''(x) + \mathcal{O}(h^5). &(71)\end{aligned}$$

Now for the first term in (69). Let's compute 1, M

$$
\begin{aligned}
K2 &= \int K^2\left(\frac{G_i^{(r)}-x}{h}\right) f(x)\,dx \qquad (72)\\
&= h\int K^2(v) f(x+vh)\,dv \qquad (73)\\
&= h\left\{ f(x)\cancelto{C_2}{\int K^2(v)\,dv} + hf'(x)\cancelto{C_3}{\int vK^2(v)\,dv}\right. \qquad (74)\\
&\qquad \left. +\tfrac{1}{2}h^2\cancelto{C_4}{\int v^2K^2(v)\,dv} + \mathcal{O}(h^3)\right\}. \qquad (75)
\end{aligned}
$$

Hence,

2, M

$$K2 = hC_2f(x) + h^2C_3f'(x) + h^3C_4f''(x) + \mathcal{O}(h^4). \tag{76}$$

Putting these together gives

1, M

$$
\begin{aligned}
\mathbb{E}\left\{\hat{f}^2_{r,h}(x)\right\} &= (nh)^{-2}\left\{nK2 + n(n-1)(*)^2)\right\} \qquad (77)\\
&= (nh)^{-1}\left\{C_2f(x) + hC_3f'(x) + h^2C_4f''(x) + \mathcal{O}(h^3)\right\} \qquad (78)\\
&\quad +\frac{(n-1)}{n}\left\{f^2(x) + 2h^2f(x)f''(x) + \mathcal{O}(h^3)\right\}. \qquad (79)
\end{aligned}
$$

From the question we know that $h \to 0$ and $nh \to \infty$ as $n \to \infty$ so $\mathbb{E}\left\{\hat{f}^2_{r,h}(x)\right\} \to f^2(x)$ as $n \to \infty$.

2, M

1, M

(c) For the expectation of $D_{r,q;h}$ we have

unseen ⇓

$$
\begin{aligned}
\mathbb{E}\{D_{r,q;h}(x)\} &= \mathbb{E}\left\{\hat{f}^2_{r,h}(x)\right\} - 2\mathbb{E}\{\hat{f}_{r,h}(x)\}\mathbb{E}\{\hat{f}_{q,h}(x)\} + \mathbb{E}\left\{\hat{f}^2_{q,h}(x)\right\} \qquad (80)\\
&= 2(nh)^{-1}\left\{C_2f(x) + hC_3f'(x) + h^2C_4f''(x) + \mathcal{O}(h^3)\right\} \qquad (81)\\
&\quad -\frac{2}{n}\left\{f^2(x) + 2h^2f(x)f''(x) + \mathcal{O}(h^3)\right\} \qquad (82)\\
&\quad +\left\{\cancel{2f^2(x)} + 4h^2f(x)f''(x) + \mathcal{O}(h^3)\right\} \qquad (83)\\
&\quad -\cancel{2f^2(x)} - 4h^2C_1f(x)f''(x) + \mathcal{O}(h^3), \qquad (84)
\end{aligned}
$$

where equation (81) copies (78), equations (82) and (83) split (79) and (84) is the square of (67). Clearly, the $f^2(x)$ terms cancel. The expectation is mathematically simplified for $C_1 = 1$ because then the $4h^2f(x)f''(x)$ terms cancel.

3, M

1, M

In this case, the leading term for $n$ large for the expectation of $D_{r,q;h}$ is $\frac{2C_2f(x)}{nh}$ and does not depend on $f''(x)$. Clearly, $\mathbb{E}\{D_{r,q;h}(x)\} \to 0$ as $n \to \infty$.

1, M

(d) Probably classical multidimensional scaling would be more appropriate as (14) and its integral is a Euclidean distance measure. $h$ could be chosen by eye, or by cross-validation for the individual density estimates. Other sensible answers will attract credit.

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

# MATH60049 Introduction to Statistical Learning

| Question | Marker's comment |
|---|---|
| 1 | 1a Was answered well by nearly all students1b Was answered poorly by nearly all students, including 70049. This was somewhat of a surprise because the question was identical to a homework solution.1c(i) Was answered well by nearly all students.1c(ii) Was answered well by most students1d(i) Was answered reasonably well by most students1d(ii) Was attempted by some students. Very few attempts were credible and the answers were mostly vague. A few students did answer the question extremely well. |
| 2 | 2(a) About one quarter of the students answered the question well. With another quarter it was clear that students intuitively knew they answer, but found it difficult to articulate the answer either in words or mathematically. About half the students could not answer.2(b) This part was answered well in the main. About 2/3 of students got this right. About 1/6 of students wrote B as X^T X which resulted in a pxp matrix and not an nxn matrix. Apart from it being a basic definition, as B leads to E, and E is a distance matrix, which contains the pairwise distances between n points, it’s surprising that those students did not realize that a pxp dimension could not be right. This mean that their next calculation was wrong and this persisted into later questions. Those students got partial credit for this.2(c) This was answered well in the main. Those students who in (b) got the matrices the wrong way round often then got the wrong answer for B and E.2(d) Was answered well by most2(e) Was answered well by about half. Another quarter of students again showed some intuition here, but were unable to articulate the answer either in words or mathematically. The remainder did not answer.2(f) This was answered well by most students. |
| 3 | 3(a) Nearly all students got this right. A surprising minority could not write/see or realize that $\psi^2 = \phi$ as this was mentioned in the notes, has appeared in previous exam papers and is at an extremely basic level.3(b) This question was poorly done by nearly everybody. About 1/4 of students could show the normality part. However, the inner product orthogonality part stumped nearly everyone. About 1/4 students partially answered by considering a special case. A few students used a graphical approach, which was totally acceptable and these students got high marks.3(c) Was answered well. Some students did not answer about the consistency, computation nor use at the end of the question.3(d) About 1/8 of students answered this question well. About another 1/2 mentioned ease of interpretation which gained a mark, but very few other acceptable answers were produced. |
| 4 | 4(a) The question asked candidates to explain how to obtain the c, many students just stated this, which lost a few marks. Probably most students got the answer, but about 1/3 explained. That 1/3 also usually got the order of computation right and explained how.(b) Very few nbsp;students got this part correct.(c) This was answered extremely well by most students(d) Many solutions were proposed, but few students got this exactly right.(e) This was answered well by many students, with many innovative and interesting answers.nbsp; |

# MATH70049 Introduction to Statistical Learning

| Question | Marker's comment |
|---|---|
| 1 | See 60049 comments |
| 2 | See 60049 comments |
| 3 | See 60049 comments |
| 4 | See 60049 comments |
| 5 | 5(a) The majority of students answered this well. (b) About 1/3 of students did this well. Another 1/3 got a reasonable way through, but not all the way.(c) If students had got (b) right, then they usually went on to get (c) correct. However, those that got part-way through and followed sensible procedures got partial credit.(d) About 1/3 students answered this and quite a few of those gave a good answer. |