Imperial College London

MATH97287

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May-June 2020

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

Introduction to Statistical Learning

Date: 7th May 2020

Time: 13.00pm - 15.30pm (BST)

Time Allowed: 2 Hours 30 Minutes

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) This question concerns regression models with two or more regressor variables. Using a suitable notation, write down the optimisation problems that define the *least squares*, the *ridge regression* and the *lasso* estimators. (3 marks)

(b) Working from the optimisation problems, derive explicit formulae for the least squares and ridge regression estimators, stating any assumptions that you need. [Hint: Let $u = \beta^T v$, where v is a p-vector, then $\frac{\partial u}{\partial \beta} = v$. If $w = \beta^T A \beta$ for some symmetric matrix A then $\frac{\partial w}{\partial \beta} = 2A\beta$.] (4 marks)

(c) Now assume that the design matrix is orthogonal and the components of the least squares estimator $\hat{\beta}_{LS,j} > 0$ for all $j = 1, \ldots, p$. Derive an explicit formula for the lasso estimator in terms of the least squares estimator and its smoothing parameter λ. Explain why lasso is called a shrinkage estimator. (5 marks)

(d) Compute the variance matrix of the least squares estimator $\hat{\beta}$ and (i) derive a 95% confidence interval for β_1 [HINT: define $x^{\dagger}_{1,1} = \{(X^T X)^{-1}\}_{1,1}$ and assume $\hat{\sigma}$ is the usual least squares estimator of σ.] (ii) explain the motivation for the ridge regression estimator. (6 marks)

(e) A road safety study carried out in the USA collected data on the annual number of road accident deaths in 26 US states. The response variable was the number of deaths. There were five explanatory variables: `drivers` the number of drivers (in 10000s); `popden` the population density in people per square mile; `rural` the length of rural roads, in 1000s of miles; `temp` the average daily temperature in January and `fuel` fuel consumption in 10×10^6 gallons per year. All the variables were put into the data frame called `road` of dimension 26×6.

A least squares linear regression was fitted and the following table extracted from `summary(.)` applied to the results.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -155.94105  238.79327  -0.653    0.521
drivers        4.44399    0.39618  11.217 4.44e-10 ***
popden        -0.01318    0.02458  -0.536    0.598
rural          2.55112    1.89771   1.344    0.194
temp           6.12376    4.55712   1.344    0.194
fuel          -0.93411    0.87527  -1.067    0.299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A set of ridge regression estimates was computed and the ridge variable plot is shown in Figure 1 (next page). Unfortunately, the plot was accidentally only produced in black and white. Consequently, correctly identify which line is associated with which variable. Briefly intepret the results from the above analyses. (2 marks)

(Total: 20 marks)

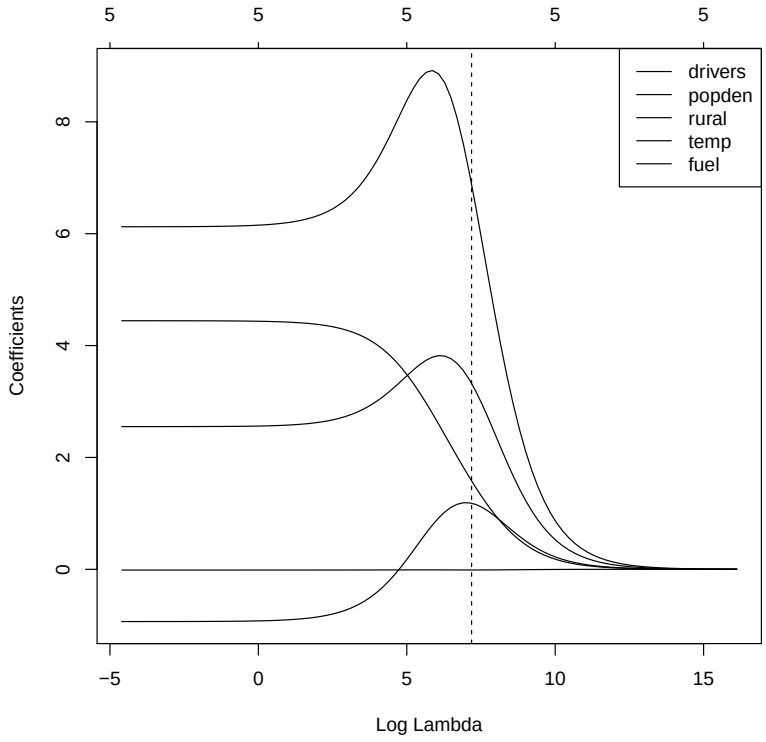

Figure 1: Ridge regression parameters against log of ridge parameter λ for the `road` data. Vertical dotted line corresponds to 'best' λ judged by cross-validation.

2. (a) Explain the purpose of classical multidimensional scaling. (2 marks)

(b) Given a configuration X of n individuals on p variables show how to compute the inner product matrix B from X, and then how to compute the (squared) Euclidean distance matrix E from B, if the entries of E are $e_{m,\ell} = \sum_{i=1}^{p}(X_{m,i} - X_{\ell,i})^2$ for all $m, \ell = 1, \ldots n$.
(3 marks)

(c) Explain what happens to the position and orientation information in X, when computing B, then E. (5 marks)

(d) Suppose we now obtain a distance matrix E from some applied problem. Paying careful attention to the reinstatement of position and orientation information, show how to obtain

(i) a suitable inner product matrix, $\hat{B}$ from E, and (5 marks)

(ii) a recovered configuration, $Y = \hat{X}$, from $\hat{B}$ (3 marks)

[HINT: You can assume that the condition $B^T\mathbf{1} = 0$ and $\hat{B}^T\mathbf{1} = 0$ are imposed.]

(e) An ecologist collects the number of occurrences of each of 30 species of vegetation at 20 sites. Classical scaling is carried out on Euclidean distances and the first two dimensions of the resulting solution are shown in Figure 2 (next page). The eigenvalues from the scaling solution (rounded to the nearest integer) were extracted from R and were

```
 [1] 681 435 235 141 126  95  66  53  49  41  33  26  18
[14]  17  10   8   4   4   2   2   0   0   0   0   0   0
[27]   0   0   0   0
```

The sum of the eigenvalues is 2046. From the plot and the list of eigenvalues comment on the appropriateness of presenting the scaling results as a two-dimensional plot. Is there any evidence of non-Euclidean error?
(2 marks)

(Total: 20 marks)

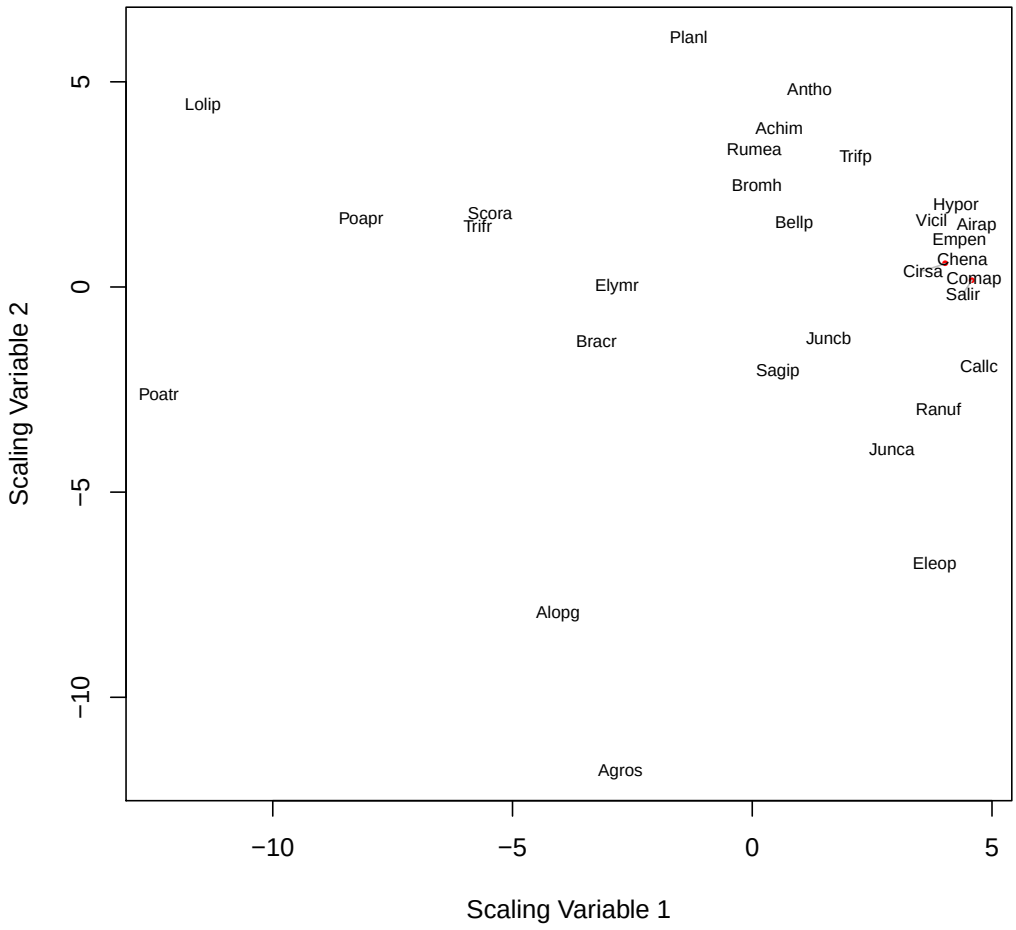

Figure 2: First two dimensions of the scaling solution of the dune data.

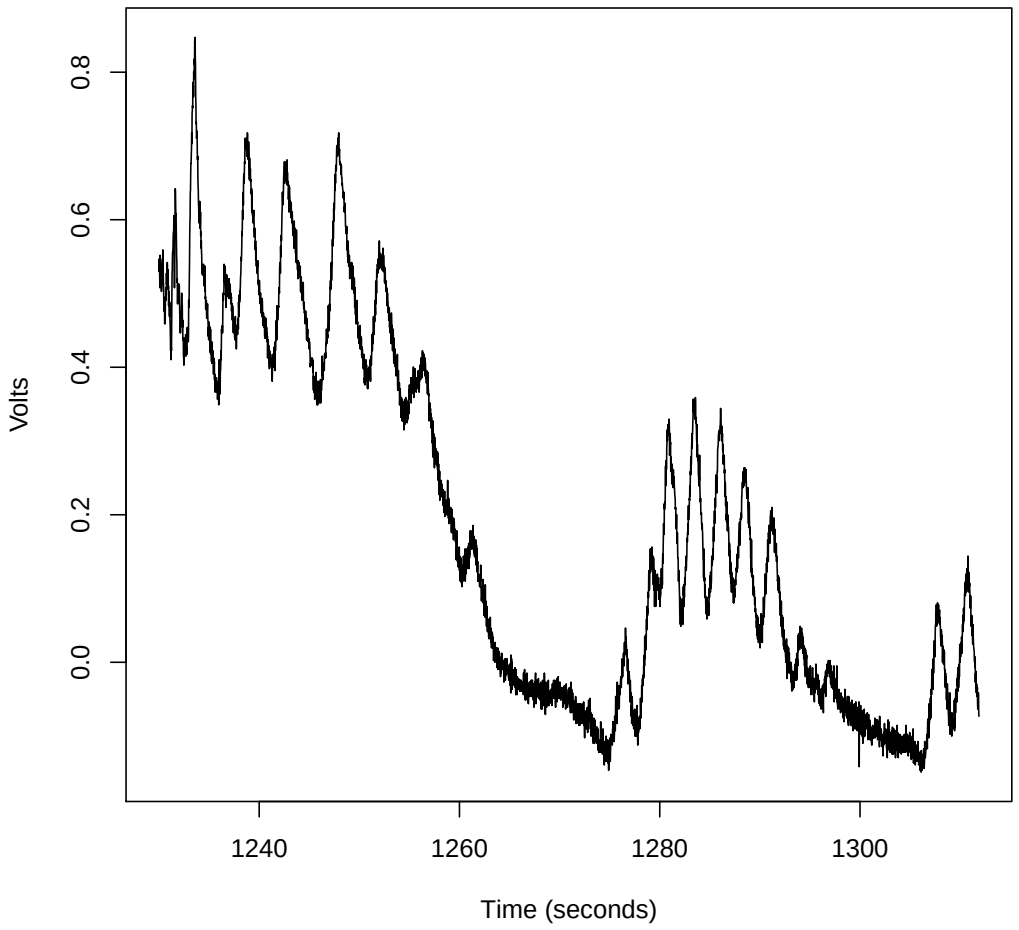

Figure 3: Time series plot of inductance plethysmograph data.

3. During a study into how people breathe in a hospital environment a team of scientists measure the breathing patterns of a set of patients under different conditions. The breathing patterns are measured by a device called an inductance plethysmograph, which is a band worn by the patient around their chest, which measures the expansion and contraction of the chest as the patient breathes. Let Y measure the voltage changes in the plethysmograph as the patient breathes and let X be time. Figure 3 (above) shows a trace from a single patient.

   (a) The scientists want to infer the underlying trend of the traces and decide to use the smoothing spline to estimate $\mathbb{E}(Y|X)$. They assume that the underlying function, $f \in L_2(\mathbb{R})$ for a trace can be estimated by minimising

$$\mathrm{RSS}(f, \lambda) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_{\mathbb{R}} \{f''(x)\}^2 \, dx, \tag{1}$$

   where (x_i, y_i) are the coordinates of the observed inductance plethysmography trace and λ is fixed and chosen by the scientists, and f is at least twice differentiable.

(i) Explain the role of the two terms on the right-hand side of (1) and the role of λ; (3 marks)

(ii) explain what happens if $\lambda = 0$ or λ becomes extremely large. (2 marks)

(b) You are told that the solution to the optimisation problem in (1) is

$$f(x) = \sum_{j=1}^{n} \theta_j N_j(x), \tag{2}$$

where $\{N_j(x)\}_{j=1}^n$ is a natural cubic spline basis.
Show that $\mathrm{RSS}(f, \lambda)$ can be written in matrix terms as

$$\mathrm{RSS}(f, \lambda) = (Y - N\theta)^T (Y - N\theta) + \lambda \theta^T \Omega \theta, \tag{3}$$

and explain how Ω is obtained, where the matrix N is given by $(N)_{i,j} = N_j(x_i)$. (3 marks)

(c) Show how the smoothing spline problem can be recast as a ridge regression problem and derive the equivalent ridge regression estimator. Carefully note any assumptions you use. (3 marks)

(d) Separate spline and wavelet smooths (with Daubechies' wavelets with ten vanishing moments) are fitted to the inductance plethysmography time series given in Figure 3.
Figure 4 (next page) shows the residuals after the spline and wavelet fits. The estimated mean squared errors for the spline and wavelet fits were $14.0 \times 10^{-5}\mathrm{V}^2$ and $8.5 \times 10^{-5}\mathrm{V}^2$, respectively. Briefly interpret these results. (2 marks)

(e) The Fourier Transform of an integrable function $f : \mathbb{R} \to \mathbb{C}$ is

$$\hat{f}(\omega) = F(\omega) := (2\pi)^{-1/2} \int_{\mathbb{R}} f(t) e^{i\omega t}\, dt = \mathcal{F}\{f\}(\omega). \tag{4}$$

Define $f_{j,k}(t) := 2^{j/2} f(2^j t - k)$, for all $t \in \mathbb{R}; j, k \in \mathbb{Z}$ and functions f. Prove that

$$\mathcal{F}\{f_{j,k}\}(\omega) = 2^{-j/2} e^{2^{-j} ik\omega} F(2^{-j}\omega). \tag{5}$$

The Littlewood-Paley wavelet is defined by

$$(t) := (\pi t)^{-1}\{\sin(2\pi t) - \sin(\pi t)\}, \tag{6}$$

and has Fourier Transform given by

$$\hat{\psi}(\omega) = \mathcal{F}\{\psi\}(\omega) = \begin{cases} (2\pi)^{-1/2} & \pi \le |\omega| \le 2\pi, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Using Parseval's relation, or otherwise, show that $< \psi_{0,0}, \psi_{1,0} >= 0$. (7 marks)

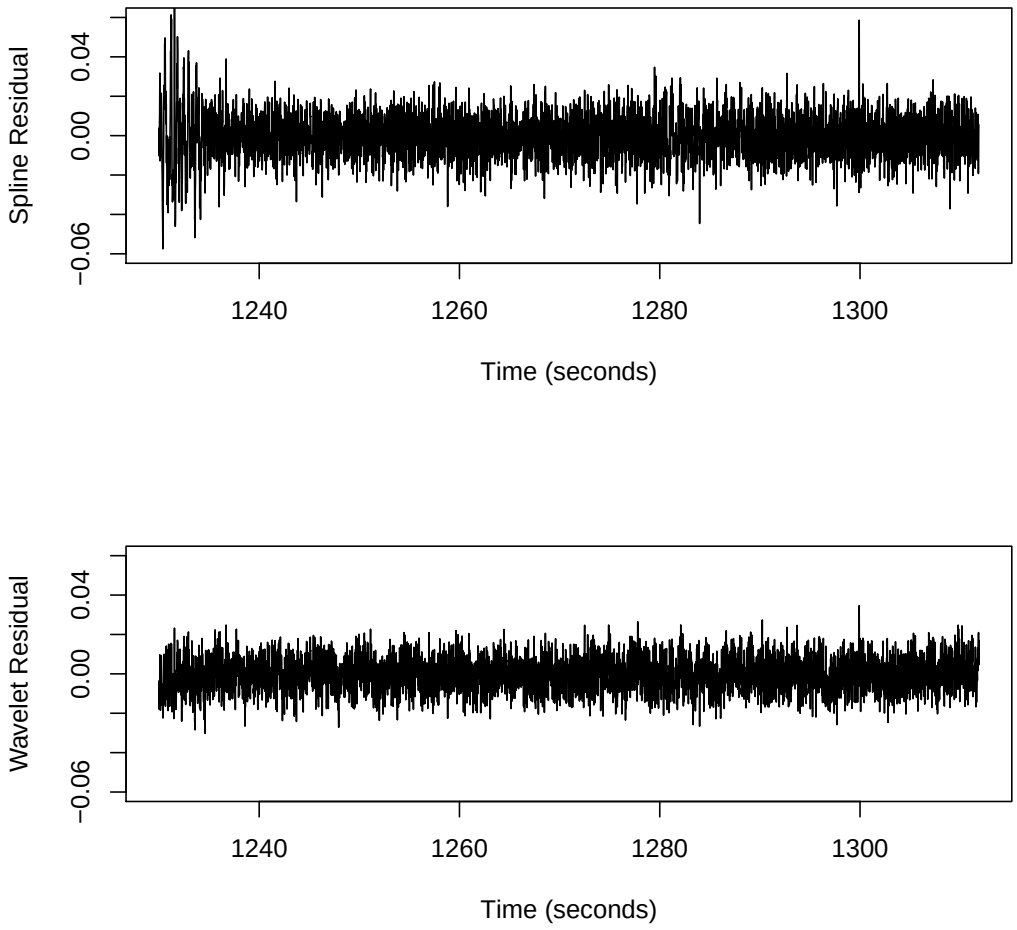

Figure 4: Residuals of spline smooth (top) and wavelet fits (bottom).

(Total: 20 marks)

4. (a) Define both the negative *entropy* $H(f)$ of $f(x)$ and the *Kullback-Leibler divergence*, $D_{KL}(g||f)$, from g to f, where f, g are probability density functions. (2 marks)

   (b) Prove that $D_{KL}(g||f) \geq 0$ for all f, g with equality if $f = g$ almost everywhere. (4 marks)

   (c) Let $\mathcal{F}_{0,\sigma^2}$ be the set of all probability densities with mean zero and variance σ^2. Suppose that $f(x) \in \mathcal{F}_{0,\sigma^2}$ is the density of a $N(0, \sigma^2)$ random variable. Show that $f = \text{argmax}_{g \in \mathcal{F}_{0,\sigma^2}} H(g)$. (5 marks)

   (d) Compare and contrast exploratory projection pursuit and independent components analysis and explain the relevance of the result in (c) to both methods. (4 marks)

   (e) Suppose X is a vector of normally-distributed observations. Are the observations of the (i) centred and (ii) centred and sphered data also normally distributed? Now supposing that X is a data matrix on n observations in p dimensions, comment on the distribution of the centred and sphered versions of multivariate X. (5 marks)

(Total: 20 marks)

5. (a) Let X be a positive random variable with probability density function given by $f_X(x)$. Suppose that it is probable that some realizations of X will be close to zero. What problem might occur if we choose to estimate $f_X(x)$ using the ordinary kernel density estimator with kernel $K : \mathbb{R} \to \mathbb{R}^+$ satisfying $\int_{\mathbb{R}} K(x)\,dx = 1$ and $K(x) = K(-x)$?

(2 marks)

(b) Suppose we choose to transform the data by creating a new independent and identically distributed sample $Y_i = \log(X_i)$, $i = 1, \ldots, n$, which have density $f_Y(y)$. Using distribution function arguments, or otherwise, derive a formula for f_X in terms of f_Y and vice versa.

(3 marks)

(c) Suppose we choose to estimate $f_Y(y)$ using an ordinary kernel density estimator, $\hat{f}_Y$, with the normal kernel, $K = \phi$ and bandwidth h. Show that the induced density estimator for X_i is given by

$$\hat{f}_X(x) = (nhx)^{-1} \sum_{i=1}^{n} \phi \left\{ \frac{\log(x) - \log(X_i)}{h} \right\}. \tag{8}$$

Show further that $\int_0^\infty \hat{f}_X(x)\,dx = 1$. (2 marks)

(d) The probability density function of a *log-normal* distribution with parameters μ, σ^2, denoted $\mathrm{Lognormal}(\mu, \sigma^2)$ is

$$\mathrm{Ln}(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left[\frac{\{\log(x) - \mu\}^2}{2\sigma^2} \right] \right). \tag{9}$$

Let $L \sim \mathrm{Ln}(x; \mu, \sigma)$.

Compute $\mathbb{E}(L)$.

You are given that $\mathrm{var}(L) = \{\exp(\sigma^2) - 1\} \exp(2\mu + \sigma^2)$. Show that the kernel estimator $\hat{f}_X(x)$ given in (8) is adaptive in that the kernel functions depend both on the bandwidth, h, and their location. (7 marks)

(e) It can be shown that the highest order term of the approximate bias of the kernel density estimate $\hat{f}_Y$ is equal to $Ch^2 f''(x)/2$, where C is constant. Use this approximate bias and the transformations between the densities of Y and X you derived in part (b) to compute the approximate bias for $\hat{f}_X(\epsilon)$ when ϵ is small. Show that the bias is zero at $\epsilon = 0$ if $f_X(0) = 0$.

(6 marks)

(Total: 20 marks)

MS320 Introduction to Statistical Learning
Exam 2019/20
SOLUTIONS

1. (a) We assume that the statistical model is $Y = X\beta + \epsilon$, where X is a design matrix, β is a p-dimensional parameter vector. Y is an n-dimensional vector of observations and ϵ is an n-dimensional vector of errors, which are IID with mean zero and variance of σ^2. The problem is to estimate β given a set of observations $(y_1, \ldots, y_n)$, a realization of Y.

   In the following, credit will be given to solutions that explain why setting the derivative to zero results in a minimum in any of the solutions. The least-squares estimator is the solution to

$$\min_{\beta \in \mathbb{R}^p} (Y - X\beta)^T (Y - X\beta). \tag{1}$$

   The ridge regression estimator is the solution to

$$\min_{\beta \in \mathbb{R}^p} (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta. \tag{2}$$

   The lasso estimator is the solution to

$$\min_{\beta \in \mathbb{R}^p} (Y - X\beta)^T (Y - X\beta)/2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3}$$

   CatA ([3] Marks) [seen]

   (b) For least squares, we proceed by differentiation. Rewrite the objective function as

$$\mathrm{RSS}_{LS}(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta. \tag{4}$$

   Then, using the hint

$$\frac{\partial \, \mathrm{RSS}_{LS}(\beta)}{\partial \beta} = -2X^T Y + 2X^T X \beta. \tag{5}$$

   Solving $\frac{\partial \, \mathrm{RSS}_{LS}(\beta)}{\partial \beta} = 0$ gives

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T Y, \tag{6}$$

   assuming that $X^T X$ is invertible. For ridge regression the objective is

$$\mathrm{RSS}_R(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta + \lambda \beta^T \beta. \tag{7}$$

   Hence

$$\frac{\partial \, \mathrm{RSS}_R(\beta)}{\partial \beta} = -2X^T Y + 2X^T X \beta + 2\beta = -2X^T Y + 2(X^T X + \lambda I_p), \tag{8}$$

where I_p is the $p \times p$ identity matrix. Hence

$$\hat{\beta}_R = (X^T X + \lambda I_p)^{-1} X^T Y, \tag{9}$$

assuming that $X^T X + \lambda I_p$ is invertible.
CatA ([4] Marks) [seen]

(c) For the lasso, we can write the objective function as

$$\mathrm{RSS}_{LA}(\beta) = \frac{1}{2} Y^T Y - Y^T X \beta + \frac{1}{2} \beta^T X^T X \beta + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{10}$$

The lasso estimator

$$\hat{\beta}_{LA} = \mathrm{argmin}_\beta \, \mathrm{RSS}_{LA}(\beta) \tag{11}$$

$$= \mathrm{argmin}_\beta \, -Y^T X \beta + \frac{1}{2} \beta^T \beta + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{12}$$

since X is orthogonal and $Y^T Y$ does not depend on β. Now, if X is orthogonal, the least squares estimator is just $\hat{\beta}_{LS} = X^T Y$ and so the lasso estimator can be written as

$$\hat{\beta}_{LA} = \mathrm{argmin}_\beta \, -(\hat{\beta}_{LS})^T \beta + \frac{1}{2} \beta^T \beta + \lambda \sum_{j=1}^{p} |\beta_j| \tag{13}$$

$$= \mathrm{argmin}_\beta \sum_{j=1}^{p} \left(-\hat{\beta}_{LS,j} \beta_j + \frac{1}{2} \beta_j^2 + \lambda |\beta_j| \right). \tag{14}$$

Each term in the sum involves one, and only one, β_j, so we can minimise the quantity component by component. So, minimise

$$M_j = \mathrm{argmin}_{\beta_j} \, -\hat{\beta}_{LS,j} \beta_j + \frac{1}{2} \beta_j^2 + \lambda |\beta_j| \tag{15}$$

The question states that $\hat{\beta}_{LS,j} > 0$. Suppose $\beta_j < 0$. However, we can obtain a smaller M_j by swapping β_j with $-\beta_j$. This swap leaves β_j^2 and $|\beta_j|$ unchanged. Hence, it must be the case that $\beta_j > 0$ and we optimise over this region. Now let's differentiate M_j wrt β_j, and set to zero, to obtain:

$$\frac{\partial M_j}{\partial \beta_j} = -\hat{\beta}_{LS,j} + \hat{\beta}_{LA,j} + \lambda = 0, \tag{16}$$

which implies $\hat{\beta}_{LA,j} = \hat{\beta}_{LS,j} - \lambda$, but this always has to be positive so $\hat{\beta}_{LA,j} = (\hat{\beta}_{LS,j} - \lambda)_+$, where $x_+ = x\mathbf{I}(x > 0)$, where $\mathbf{I}(\cdot)$ is the usual indicator function.
Lasso is shrinkage as it takes the value of the least-squares estimators and shrinks them by λ or sets them to zero if the least-squares estimator is smaller than λ in modulus.
CatB ([5] Marks) [seen, but tricky]

(d) We need to compute $\text{var}(\hat{\beta}_{LS})$. We will need $\mathbb{E}(\hat{\beta}_{LS})$ first. Using that the model tells us that $\mathbb{E}(Y) = X\beta$ we obtain

$$\mathbb{E}(\hat{\beta}_{LS}) = \mathbb{E}\{(X^TX)^{-1}X^TY\} = (X^TX)^{-1}X^T\mathbb{E}(Y) = (X^TX)^{-1}X^TX\beta = \beta. \tag{17}$$

Then

$$\begin{aligned}
\text{var}(\hat{\beta}_{LS}) &= \mathbb{E}\{(\hat{\beta}_{LS} - \mathbb{E}\hat{\beta}_{LS})(\hat{\beta}_{LS} - \mathbb{E}\hat{\beta}_{LS})^T\} && (18)\\
&= \mathbb{E}(\hat{\beta}_{LS}\hat{\beta}_{LS}^T) - \beta\beta^T && (19)\\
&= \mathbb{E}\{(X^TX)^{-1}X^TYY^TX(X^TX)^{-1}\} - \beta\beta^T && (20)\\
&= (X^TX)^{-1}X^T\mathbb{E}(YY^T)X(X^TX)^{-1} - \beta\beta^T && (21)
\end{aligned}$$

Now

$$\sigma^2 I = \text{var}(Y) = \mathbb{E}\{(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T\} = \mathbb{E}(YY^T) - X\beta\beta^TX^T. \tag{22}$$

CatA ([2] Marks) [seen]

So, inserting (22) into (21) gives

$$\begin{aligned}
\text{var}(\hat{\beta}_{LS}) &= (X^TX)^{-1}X^T(\sigma^2 I + X\beta\beta^TX^T)X(X^TX)^{-1} - \beta\beta^T && (23)\\
&= \sigma^2(X^TX)^{-1} + (X^TX)^{-1}X^TX\beta\beta^TX^TX(X^TX)^{-1} - \beta\beta^T && (24)\\
&= \sigma^2(X^TX)^{-1}. && (25)
\end{aligned}$$

To answer (i) 95% Confidence interval for β_1 is $\hat{\beta}_{LS,1} \pm C^*\hat{\sigma}x^{\dagger}_{1,1}$, where C^* is the 2.5% point of the $N(0,1)$ distribution (or t_{n-1}-distribution). (ii) If X^TX is ill-conditioned then it can contain very small eigenvalues and the matrix will be near singular. This means that entries of $(X^TX)^{-1}$ can be very large meaning that the variance of $\hat{\beta}_{LS}$ corresponding to those entries have high variance and the values can be exceedingly large. The penalty on large β values ($\beta^T\beta$) mitigates against those values becoming too large. [(i) not seen/ (ii) seen but subtle]

CatD ([4] Marks)

(e) The coefficients can be worked out from the LHS of the plot and the LS coefficient table. From top to bottom they are `temp`, `drivers`, `rural`, `popden` and `fuel`.

From the table of least squares coefficients it seems that only the `drivers` variable is statistically significant, with a minute p-value. The ridge regression plot shows how the coefficients are shrunk and some are significantly smaller than their LS values.

CatC ([2] Marks) [hardly seen]

2. (a) Classical multidimensional scaling takes a set of Euclidean distances between n objects and creates a configuration of n points in p dimensions.

CatA ([2] Marks) [bookwork/seen]

(b) The Euclidean distances can be written

$$e_{m,\ell} = \sum_{i=1}^{p} (X_{m,i} - X_{\ell,i})^2 \tag{26}$$

$$= \sum_{i=1}^{p} X_{m,i}^2 - 2X_{m,i}X_{\ell,i} + X_{\ell,i}^2. \tag{27}$$

Each of the terms in this expression can be written as inner products between two rows of X, e.g.

$$B_{m,\ell} = (XX^T)_{m,\ell} = \sum_{i=1}^{p} X_{m,i}X_{\ell,i}. \tag{28}$$

So

$$e_{m,\ell} = b_{m,m} - 2b_{m,\ell} + b_{\ell,\ell}, \tag{29}$$

and B is the inner product matrix.

CatA ([3] Marks) [seen]

(c) If P is a p-dimensional rotation matrix, write $Y = XP$. Then $B_Y = YY^T = XPP^TX^T = XI_pX^T = B_X$. So, we lose orientation information when we form B from X in that any rotated configuration has the same B matrix.

CatA ([2] Marks) [seen]

For position, write $X_{(m)}$ for the p-vector corresponding to individual m. Then, we can shift all of the individuals by relocating according to a shift vector μ by $Y_{(m)} = X_{(m)} - \mu$ for $m = 1, \ldots, n$. Then

$$e^Y_{m,\ell} = Y^T_{(m)}Y_{(m)} - 2Y^T_{(m)}Y_{(\ell)} + Y^T_{(\ell)}Y_{(\ell)} \tag{30}$$

$$= X^T_{(m)}X_{(m)} - X^T_{(m)}\mu - \mu^T X_{(m)} + \mu^T\mu \tag{31}$$

$$+X^T_{(\ell)}X_{(\ell)} - X^T_{(\ell)}\mu - \mu^T X_{(\ell)} + \mu^T\mu \tag{32}$$

$$-2(X^T_{(m)}X_{(\ell)} - X^T_{(m)}\mu - \mu^T X_{(\ell)} + \mu^T\mu) \tag{33}$$

$$= e^X_{m,\ell}. \tag{34}$$

Hence, position information is lost on going from B to E.

CatB ([3] Marks) [seen, but intricate]

(d) i. From (29) we can sum over m (e.g. $e_{\bullet,\ell} = \sum_{m=1}^{n} e_{m,\ell}$, etc) to obtain

$$e_{\bullet,\ell} = b_{\bullet,\bullet} + nb_{\ell,\ell} - 2b_{\bullet,\ell} = b_{\bullet,\bullet} + nb_{\ell,\ell}, \tag{35}$$

as $\mathbf{1}$ is an eigenvector of B^T, and also of B (due to construction of $B = XX^T$), hence row and column sums of B are zero and therefore $b_{\bullet,\ell} = b_{m,\bullet} = 0$. Similarly, summing over the other index, gives

$$e_{m,\bullet} = b_{\bullet,\bullet} + nb_{m,m}. \tag{36}$$

by symmetry. Summing over m and n gives

$$e_{\bullet,\bullet} = nb_{\bullet,\bullet} + nb_{\bullet,\bullet} = 2nb_{\bullet,\bullet}. \tag{37}$$

Now rearrange (29) to give

$$b_{m,\ell} = -\tfrac{1}{2}(e_{m,\ell} - b_{m,m} - b_{\ell,\ell}), \tag{38}$$

and using (35) and (36) we have

$$\begin{aligned} \hat{b}_{m,\ell} &= -\tfrac{1}{2}\left\{e_{m,\ell} - (e_{m,\bullet} - b_{\bullet,\bullet})/n - (e_{\bullet,\ell} - b_{\bullet,\bullet})/n\right\} && (39)\\ &= -\tfrac{1}{2}(e_{m,\ell} - \tfrac{e_{m,\bullet}}{n} - \tfrac{e_{\bullet,\ell}}{n} + 2\tfrac{b_{\bullet,\bullet}}{n}) && (40)\\ &= -\tfrac{1}{2}(e_{m,\ell} - \tfrac{e_{m,\bullet}}{n} - \tfrac{e_{\bullet,\ell}}{n} + \tfrac{e_{\bullet,\bullet}}{n^2}). && (41) \end{aligned}$$

The students might also write this as the following two equivalent ways:

$$\hat{b}_{m,\ell} = -\tfrac{1}{2}(\text{entry} - \text{row av.} - \text{col av.} + \text{grand av.}), \tag{42}$$

or in matrix terms

$$\hat{B} = -\frac{1}{2}(I_n - \mathbf{1}\mathbf{1}^T/n)\, E\, (I_n - \mathbf{1}\mathbf{1}^T/n). \tag{43}$$

CatB ([5] Marks) [seen, but involved]

ii. We now have a $\hat{B}$, which we assume is real symmetric positive semidefinite (as if it were XX^T). So, perform eigendecomposition $\hat{B} = U^T D U$, where D is diagonal matrix containing the eigenvalues $d_j \geq 0$ (assuming symmetric psd assumption, which might be [slightly] invalidated for a real data set). Create new configuration $Y = U^T D^{1/2}$ so that $YY^T = U^T D^{1/2}(D^{1/2})^T U = U^T D U = \hat{B}$, so Y has the inner product matrix that we require (note $D^{1/2}$ is diagonal so automatically symmetric).
CatC ([3] Marks) [seen]

(e) Roughly speaking, the first two eigenvalues account for $(681 + 435)/2046 = 54\%$ of the variation, which is reasonable, but is missing 46%, so a few more dimensions would be useful. The plot does not contribute much to this decision (although it could have done, if there were outlying observations, for example). There is no evidence of non-Euclidean error (negative eigenvalues). [unseen, but seen similar examples]
CatA ([2] Marks)

3. (a) (i) The first of the two terms measures the fidelity of the smooth f to the data $\{y_i\}$. The second measures the 'wiggliness' or degree of average curvature within f. The λ controls the balance between the two terms as presented to the overall objective function. [seen]
CatA ([3] Marks)

(ii) If $\lambda = 0$, then there is no penalty term and the objective function just consists of the fidelity term. The solution to the optimisation is a curve that interpolates the data points. If $\lambda = \infty$, then the penalty term dominates, pushing the optimisation to produce a curve with zero second derivative, and the solution is a straight line.

CatA ([2] Marks) [seen]

(b) For the $N\theta$ we have $f(x_i) = \sum_{j=1}^{n} \theta_j N_j(x_i) = \sum_{j=1}^{n} \theta_j N_{i,j} = N_i^T \theta$, where N_i is the ith row of N. Thence vectorising this gives $Y - N\theta$ and $(Y - N\theta)^T(Y - N\theta) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2$. For the penalty term we can write

$$\int_{\mathbb{R}} \{f''(x)\}^2 \, dx = \int_{\mathbb{R}} \sum_{j=1}^{n} \theta_j N_j''(x) \sum_{k=1}^{n} \theta_k N_k''(x) \, dx \tag{44}$$

$$= \sum_{j=1}^{n} \sum_{k=1}^{n} \theta_j \theta_k \int_{\mathbb{R}} N_j''(x) N_k''(x) \, dx \tag{45}$$

$$= \theta^T \Omega \theta, \tag{46}$$

where Ω is precisely the integral shown in (45). Hence result.

CatA ([3] Marks) [seen]

(c) Consider the parametrisation $\beta = \Omega^{1/2}\theta$.Clearly, Ω is real and symmetric and therefore so is $\Omega^{1/2}$ (according to the recipe given in lectures of eigendecomposition and replacing the diagonal by its square root). Then $\beta^T\beta = \theta^T(\Omega^{1/2})^T\Omega^{1/2}\theta = \theta^T\Omega\theta$. For the fidelity term we replace $Y - N\theta$ by $Y - N\Omega^{-1/2}\beta$, where we assume $\Omega^{1/2}$ (which I think it is, always actually). Now set $X = N\Omega^{-1/2}$ and the objective becomes

$$(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta, \tag{47}$$

which is the ridge regression objective.

CatC ([3] Marks) [was set as optional homework]

(d) Arguably, the wavelet regression gives a better fit as the residuals for the wavelet estimator look more uniform, generally smaller and the spline residuals are much larger near the start. The estimated mean squared error results are perhaps a bit less helpful, as they are calculated on quite a different basis for the two models (spline is 'more' linear and wavelets are nonlinear). Give marks for anything sensible and relevant.

CatB ([2] Marks)[unseen]

(e) First substitute $u = 2^j t - k$, $du = 2^j dt$ into

$$\begin{aligned}
\mathcal{F}\{f_{j,k}\}(\omega) &= (2\pi)^{-1/2} \int_{\mathbb{R}} 2^{j/2} f(2^j t - k) e^{i\omega t}\, dt && (48)\\
&= 2^{-j/2}(2\pi)^{-1/2} \int_{\mathbb{R}} f(u) \exp\{i\omega 2^{-j}(u+k)\}\, du && (49)\\
&= 2^{-j/2}(2\pi)^{-1/2} \exp(i\omega 2^{-j}k) \int_{\mathbb{R}} f(u) \exp\{i\omega 2^{-j}u\}\, du && (50)\\
&= 2^{-j/2}(2\pi)^{-1/2} \exp(i\omega 2^{-j}k) F(2^{-j}\omega), && (51)
\end{aligned}$$

as required. For the second part Parseval's relation gives us

$$< \psi_{0,0}, \psi_{1,0} >=< \hat{\psi}_{0,0}, \hat{\psi}_{1,0} > . \quad (52)$$

Now $\hat{\psi}_{0,0}(\omega) = \hat{\psi}(\omega)$ and, using the result above,

$$\hat{\psi}_{1,0}(\omega) = 2^{-1/2}(2\pi)^{-1/2}\hat{\psi}(2^{-1}\omega). \quad (53)$$

The crucial point is that the supports of $\hat{\psi}_{0,0}$ and $\hat{\psi}_{1,0}$ do not overlap (the former is as in the question, for the latter, its support is $\pi/2 \leq |\omega| \leq \pi$. Hence, their inner product is trivially zero.

CatD ([7] Marks) [unseen, and challenging]

4. (a) The entropy of f is given by

$$H(f) = -\int_{\mathbb{R}} f(x) \log\{f(x)\}\, dx. \quad (54)$$

The Kullback-Leibler divergence is

$$D_{KL}(g||f) = \int_{\mathbb{R}} g(x) \log\left\{\frac{g(x)}{f(x)}\right\} dx. \quad (55)$$

CatA ([2] Marks) [seen]

(b) Note that $\log(r) \leq r - 1$, for all $r \in \mathbb{R}^+$. So

$$\begin{aligned}
-D_{KL}(g||f) &= \int_{\mathbb{R}} g(x) \log\left\{\frac{f(x)}{g(x)}\right\} dx && (56)\\
&\leq \int_{\mathbb{R}} g(x) \left\{\frac{f(x)}{g(x)} - 1\right\} dx && (57)\\
&= \int_{\mathbb{R}} f(x)\, dx - \int_{\mathbb{R}} g(x)\, dx && (58)\\
&= 1 - 1 = 0. && (59)
\end{aligned}$$

Hence, $D_{KL}(g||f) \geq 0$ as required. If $f = g$ a.e., then $g/f = 1$ a.e. and $\log(g/f) = 0$ a.e., hence $D_{KL}(g||f) = 0$.

CatC ([4] Marks) [unseen, but told name of result (Gibbs inequality) in lectures, so could have looked up very similar proof]

(c) Let $g \in \mathcal{F}_{0,\sigma^2}$. Then

$$\begin{aligned} 0 &\leq D_{KL}(g||f) && (60)\\ &= \int_{\mathbb{R}} g(x) \log\left\{\frac{g(x)}{f(x)}\right\} dx && (61)\\ &= \int_{\mathbb{R}} g(x) \log\{g(x)\}\, dx - \int_{\mathbb{R}} g(x) \log\{f(x)\}\, dx. && (62) \end{aligned}$$

First integral is $-H(g)$ and since $f(x) = (2\pi\sigma^2)^{-1/2} \exp(-x^2/2\sigma^2)$ we have

$$\log\{f(x)\} = -\tfrac{1}{2}\log(2\pi\sigma^2) - x^2/2\sigma^2. \quad (63)$$

So, the second integral is

$$-\tfrac{1}{2}\log(2\pi\sigma^2)\cancelto{1}{\int_{\mathbb{R}} g(x)\, dx} - (2\sigma^2)^{-1}\cancelto{\sigma^2}{\int_{\mathbb{R}} x^2 g(x)\, dx} = -\tfrac{1}{2}\log(2\pi\sigma^2) - \tfrac{1}{2}. \quad (64)$$

It is the case that $-H(f) = \int_{\mathbb{R}} f(x)\log\{f(x)\}\, dx = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}$ also, as the integration in (64) is the same if $g(x)$ is replaced by $f(x)$.
Hence, substituting $-H(f)$ and $-H(g)$ into (62) gives

$$0 \leq -H(g) + H(f) \implies H(g) \leq H(f), \quad (65)$$

as required.

CatB ([5] Marks) [seen, but involved]

(d) Both EPP and ICA are methods of data analysis, which involve optimisation. They seek departures of a density from normality. ICA is embedded in a factor model, EPP looks at estimated densities of projected data. Both operate on sphered/whitened data. The relevance to (c) is that the normal distribution maximises $H(g)$, and so EPP and ICA both seek to minimise $H(g)$ to get away from normal densities — the more non-normal the better.

CatA ([4] Marks) [seen, but not explicit]

(e) Let's consider the 1D case. Here the data X matrix is a 1D vector $x_1, \ldots, x_n$. Centering subtracts the sample mean $y_i = x_i - \bar{x}$. We know that $\bar{x}$ is normally distributed and thus so is y_i (as the sum of two normals is normal; although they will now be correlated, even if they weren't previously). For sphering, we can divide through by $\hat{\sigma}$, the usual estimate of the standard deviation (with the n^{-1} denominator, not $(n-1)^{-1}$. It is well known that the resulting quantity is distributed according to the Student't t-distribution and hence is not normal.

CatD ([5] Marks) [unseen, challenging]

5. (a) The ordinary KDE can put significant mass on the negative half of the real line. This is a problem as the data are positive $f(x)$ is zero for $x <= 0$.

CatB ([2] Marks)[requires thought and good knowledge of KDEs]

(b) We have

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) && (66)\\ &= \mathbb{P}\{\log(X) \leq y\} && (67)\\ &= \mathbb{P}\{X \leq e^y\} = F_X(e^y). && (68) \end{aligned}$$

Differentiating gives the density $f_Y(y) = F'_Y(y) = \frac{\partial F_X(e^y)}{\partial y} = F'_X(e^y)e^y = e^y f_X(e^y)$. Now let $x = e^y$, so $y = \log(x)$ (for $x > 0$). Hence, $f_X(x) = x^{-1} f_Y\{\log(x)\}$.

CatA ([3] Marks) [standard UG statistics, 2nd year?]

(c) The kde for $\{Y_i\}$ is

$$\hat{f}_Y(y) = (nh)^{-1} \sum_{i=1}^{n} \phi\left(\frac{y - Y_i}{h}\right). \quad (69)$$

for $y \in \mathbb{R}$. From (b) we find the density of X by $\hat{f}_X(x) = x^{-1} f_Y\{\log(x)\}$, which gives

$$\hat{f}_X(x) = (nhx)^{-1} \sum_{i=1}^{n} \phi\left\{\frac{\log(x) - \log(X_i)}{h}\right\}, \quad (70)$$

as $Y = \log(X_i)$, for $x > 0$. For the integral

$$\int_0^\infty \hat{f}_X(x)\, dx = (nh)^{-1} \sum_{i=1}^{n} \int_0^\infty x^{-1} \phi\left\{\frac{\log(x) - \log(X_i)}{h}\right\} dx. \quad (71)$$

Substitute $y = \{\log(x) - \log(X_i)\}/h$, so $dy = (hx)^{-1}dx$ giving

$$\int_0^\infty \hat{f}_X(x)\, dx = n^{-1} \sum_{i=1}^{n} \int_{\mathbb{R}} \phi(y)\, dy = 1. \quad (72)$$

OR bright students will notice that $\hat{f}_X$ is a transform of $\hat{f}_Y$ and

$$\int_0^\infty \hat{f}_X(x) = \int_0^\infty \hat{f}_Y\{\log(x)\}\frac{dx}{x} = \int_{\mathbb{R}} \hat{f}_Y(y)\, dy = 1, \quad (73)$$

with the transformation $y = \log(x), dy = x^{-1}dx$ and knowing that the ordinary KDE integrates to 1 (they might show this, I suppose, but then it's essentially the same proof as above).

CatB ([2] Marks) [unseen]

(d) Expectation of a lognormal is

$$\begin{aligned} \mathbb{E}(L) &= \int_0^\infty x f_L(x)\, dx && (74)\\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left[\frac{\{\log(x) - \mu\}^2}{2\sigma^2}\right]\right) dx. && (75) \end{aligned}$$

Now substitute $u = \{\log(x) - \mu\}/\sigma$, $du = (x\sigma)^{-1}dx$, which implies $dx = x\sigma du = e^{\sigma u+\mu}\sigma du$ as $\log(x) = \sigma u + \mu$ and $x = e^{\sigma u+\mu}$. Thus

$$\begin{aligned}\mathbb{E}(L) &= \int_{\mathbb{R}} e^{\sigma u+\mu}\phi(u)\,du &(76)\\ &= e^{\mu}\int_{\mathbb{R}} e^{\sigma u}\phi(u)\,du &(77)\\ &= e^{\mu}\int_{\mathbb{R}} (2\pi)^{-1/2}\exp(\sigma u - u^2/2)\,du. &(78)\end{aligned}$$

We now complete the square on the exponent of the exponential in the integral.

$$-u^2/2 + \sigma u = (u/\sqrt{2} - \sigma/\sqrt{2})^2 - \sigma^2/2. \quad (79)$$

So

$$\begin{aligned}\mathbb{E}(L) &= e^{\mu+\sigma^2/2}\int_{\mathbb{R}} (2\pi)^{-1/2}\exp\left\{-(u-\sigma)^2/2\right\}du &(80)\\ &= e^{\mu+\sigma^2/2}\int_{\mathbb{R}} \phi(u-\sigma)\,du &(81)\\ &= e^{\mu+\sigma^2/2}, &(82)\end{aligned}$$

as required. CatB ([3] Marks) [unseen, but standard]

The kernel density estimator in formula (9) in the question paper additively combines kernel functions centred at $\mu = \log(X_i)$. The second parameter in the estimate is $\sigma = h$. However, the variance of the ith one is

$$\{\exp(\sigma^2) - 1\}\exp(2\mu + \sigma^2) = \{\exp(h^2) - 1\}\exp\{2\log(X_i) + h^2\}, \quad (83)$$

so the spread of the kernel definitely depends on $\log(X_i)$ and will be wider for X_i further away from 0.

CatC ([4] Marks) [unseen, requires thoughts]

(e) Since $\hat{f}_Y$ is an ordinary KDE its bias near zero is

$$\text{bias}\{\hat{f}_Y(\epsilon)\} = \mathbb{E}\{\hat{f}_Y(\epsilon)\} - f_Y(\epsilon) \sim \frac{h^2}{2}f_Y''(\epsilon). \quad (84)$$

From our variable transformation and (84) we have

$$\mathbb{E}\{\hat{f}_X(\epsilon)\} = \frac{1}{\epsilon}\mathbb{E}\{\hat{f}_Y\{\log(\epsilon)\} \sim \frac{1}{\epsilon}\left[f_Y\{\log(\epsilon)\} + \frac{h^2}{2}f_Y''\{\log(\epsilon)\}\right]. \quad (85)$$

Since $f_Y\{\log(x)\} = xf_X(x)$, it follows that

$$\mathbb{E}\{\hat{f}_X(\epsilon)\} \sim f_X(\epsilon) + \frac{h^2}{2\epsilon}f_Y''\{\log(\epsilon)\}. \quad (86)$$

Now for the f''_Y term — representing it in terms of f_X and its derivatives. We take $f_Y(y) = e^y f_X(e^y)$ and differentiating once gives

$$f'_Y(y) = e^y f_X(e^y) + e^{2y} f'_X(e^y), \tag{87}$$

and again gives

$$\begin{aligned} f''_Y(y) &= e^y f_X(e^y) + e^{2y} f'_X(e^y) + 2e^{2y} f'_X(e^y) + e^{3y} f''_X(e^y) && (88)\\ &= e^y f_X(e^y) + 3e^{2y} f'_X(e^y) + e^{3y} f''_X(e^y). && (89) \end{aligned}$$

So, substituting $y = \log(\epsilon)$ gives

$$f''_Y\{\log(\epsilon)\} = \epsilon f_X(\epsilon) + 3\epsilon^2 f'_X(\epsilon) + \epsilon^3 f''_X(\epsilon). \tag{90}$$

Hence, putting it all together gives

$$\text{bias}\{\hat{f}_X(\epsilon)\} \sim \frac{h^2}{2}\left\{f_X(\epsilon) + 3\epsilon f'_X(\epsilon) + \epsilon^2 f''_X(\epsilon)\right\}. \tag{91}$$

Hence, when $f_X(0) = 0$, if $\epsilon = 0$ then the bias is zero.
CatD ([6] Marks) [unseen, challenging]

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 sperate pdf file with your email.

ExamModuleCode	Question	Comments for Students	
MATH97287	1	Parts a to c which contained seen material where answered well. Marks have been removed when notation where not defined or some justification or assumption missing. In part d, some students attempted to perform hypothesis testing instead of deriving a confidence interval. In part e, some studnets did not correctly identify the order of the variables.	
MATH97287	2	This question was overall very well answered. Parts a to d concerned material directly seen in lectures. In part 2, most of student correctly identified that there was no evidence of non-Euclidean error; the justification regarding appropriateness of presenting the scaling results as a two-dimensional plot was often too vague.	
MATH97287	3	Parts a, b answered well. Part c: people forgot to mention assumptions about invertibility. Part d: mostly answered well: people either mentioned overall MSE comparison or the fact that spline residuals were larger at the start, but often both. Part e: not answered well in the main.Some people managed the first part, but almost nobody the second part involving Parseval.	
MATH97287	4	Part (a) and c answered well by nearly everybody. Part b was answered well by some. Some just thought the result was due to densities >=0, but forgetting that log (r) can be negative if r < 1. Part d was attemped by most who put sensible things, but the expression and language was often poor. Few people got part e correct and, I suspect, there was a lot of guessing. A few people thought about it deeply and came up with good answers	

MATH97287	5	The question was answered poorly. It may be that students ran out of time? This question required a deeper level of thinking, but relied on lower level techniques. For example, part b was a simple one-dimensional change of variable of a pdf, part c required some relatively simple integrations. Part d. contained a typo (an extra -1/2 in equation (9)). However, this exam being open book and the lognormal distribution and quantities displayed on Wikipedia, this was not really the issue. Part (e) was challenging and at the upper end of what we might expect from a Masters student. However, two or three students did get this correct and one student managed to score 20/20 on this question.	