Imperial College
London

MATH70081

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)
May 2023

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

Nonparametric Statistics

Date: 12 May 2023

Time: 10:00 – 11:30 (BST)

Time Allowed: 1.5 hrs

This paper has 3 Questions.

Please Answer All Questions in 1 Answer Booklet

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

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

Any required additional material(s) will be provided.

Allow margins for marking.

Credit will be given for all questions attempted.

Each question carries equal weight.

DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO

The open-book material allowed during the examinations consists of any material provided by the lecturers and annotated by the students, i.e. annotated lecture notes, annotated slides, and annotated problem class sheets. Books and electronic devices are not allowed.

1. Let $X_1, \dots, X_n$ be i.i.d. random variables having some probability density function $f : \mathbb{R} \to [0, \infty)$, and let E denote the corresponding expectation under the joint distribution of $X_1, \dots, X_n$.

(i) Define the kernel density estimator $\hat{f}_n = \hat{f}_{n,h}$ of f based on a kernel K and bandwidth $h > 0$. Define the mean squared error of $\hat{f}_n$ at a point $x \in \mathbb{R}$, denoted by $\mathrm{MSE}(\hat{f}_n(x))$, and state the bias-variance decomposition.

Suppose now that $X_1, \dots, X_n \sim^{iid} N(0,1)$, i.e. $f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density function. Let

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

denote the corresponding cumulative distribution function, so that $\Phi'(x) = \varphi(x)$. Consider a kernel density estimator $\hat{f}_n$ with rectangular kernel $K(x) = 1_{[-1/2,1/2]}(x)$.

(ii) Show that

$$E\hat{f}_n(x) = \frac{1}{h} [\Phi(x + h/2) - \Phi(x - h/2)] .$$

Hence or otherwise, deduce that the bias satisfies

$$\mathrm{Bias}(\hat{f}_n(x)) = \alpha(x^2 - 1)\varphi(x)h^2 + O(h^3) \qquad \text{as } h \to 0,$$

where $\alpha \in \mathbb{R}$ is a constant you should specify and the $O(h^3)$ term can depend on x.

[It may simplify your computations to note that $\varphi'(x) = -x\varphi(x)$.]

(iii) You are given that

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh^2} [\Phi(x + h/2) - \Phi(x - h/2)] - \frac{1}{nh^2} [\Phi(x + h/2) - \Phi(x - h/2)]^2 .$$

(you do not need to show the above). Using this, show that

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh}\varphi(x) + o(1/(nh)) \qquad \text{as } h \to 0,$$

where the $o(1/(nh))$ term can depend on x.

[Recall that $f(n) = o(g(n))$ means that $\lim_{n\to\infty} |f(n)/g(n)| = 0$, i.e. f is of strictly smaller order than g as $n \to \infty$.]

(iv) Fix $x \neq \pm 1$. Show that as $n \to \infty$, the mean-squared error $\mathrm{MSE}(\hat{f}_n(x))$ is minimized by the bandwidth choice $h = h_n = c(x, \alpha)n^{-1/5}$, where $c(x, \alpha)$ is a function of x and α that you should evaluate.

Would a bandwidth choice of the form $h \propto n^{-1/5}$ also be optimal at $x = \pm 1$? Briefly justify your answer.

[You do **not** need to derive the optimal choice of bandwidth for $x = \pm 1$.]

[This question continues on the next page . . .]

(v) Fix again $x \neq \pm 1$ and let $h = c(x, \alpha)n^{-1/5}$ be the optimal choice of bandwidth in (iv). Is the resulting mean-squared error $\mathrm{MSE}(\hat{f}_n(x))$ best as a function of n among all kernel density estimators when estimating $\varphi(x)$? If not, explain how you could modify your kernel density estimator to achieve a faster rate of convergence.

[Total 25 marks]

2. Let $(V_j)_{j\geq 0}$ be the Haar wavelet multiresolution analysis in $L^2(\mathbb{R})$. Define the Haar scaling function

$$\phi(x) = 1_{(0,1]}(x) = \begin{cases} 1 & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and Haar wavelet

$$\psi(x) = 1_{[0,1/2]}(x) - 1_{(1/2,1]}(x) = \begin{cases} 1 & 0 \leq x \leq 1/2 \\ -1 & 1/2 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

such that $\{\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k) : k \in \mathbb{Z}\}$ is an orthonormal basis for V_j and $\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k) : k \in \mathbb{Z}\}$ is an orthonormal basis for the detail space $W_j = V_{j+1} \ominus V_j$.

(i) Describe the space of functions V_J.
Let $P_J f = P_{V_J} f$ denote the projection of $f \in L^2(\mathbb{R})$ onto V_J. Define the Haar approximation $P_J f$.

(ii) Let

$$f(x) = \begin{cases} 1 + x & -1 < x \leq 0 \\ 1 - x & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Give the L^2-orthogonal projection of f onto the space V_1, leaving your answer in terms of $\phi(x)$ and $\psi(x)$.

(iii) State whether the following statements are True or False, justifying your answers.

(a) All functions in V_j with $j \geq 0$ are continuous.

(b) If $f \in W_j$ with $j \geq 0$, then $\int_{-\infty}^{\infty} f(x)dx = 0$.

(c) If $f \in V_j$ and $g \in V_{j'}$ for $j \neq j'$, then $\langle f, g\rangle_2 = \int_{-\infty}^{\infty} f(x)g(x)dx = 0$.

(iv) Suppose we observe $X_1, \dots, X_n \sim^{iid} f$ with f a probability density on $[0, 1]$. Consider the *wavelet thresholding density estimator*:

$$\hat{f}_n(x) = \hat{\alpha}\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{jk} 1\{|\hat{\beta}_{jk}| > \tau\}\psi_{jk}(x),$$

where $\tau > 0$ is a threshold, J is an integer such that $n/\log n \leq 2^J \leq 2n/\log n$ and

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i), \qquad \hat{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(X_i)$$

are the empirical wavelet coefficients. Explain why $\hat{\alpha} = 1$ and find the limiting distribution of each $\sqrt{n}(\hat{\beta}_{jk} - c_{jk}(f))$ as $n \to \infty$, where you should specify $c_{jk}(f)$.

Suppose that $\sup_{x\in[0,1]} |f(x)| \leq M$ for some known constant $M > 0$. Motivate a reasonable choice of threshold $\tau = \tau_n$ as a function of n [i.e. you do not need to worry about exact constants].

[This question continues on the next page . . .]

You may use the fact that if $Z_i \sim N(0, \sigma_i^2)$ for $i = 1, \ldots, m$ not necessarily independent, then

$$E\left[\max_{i=1,\ldots,m} |Z_i|\right] \leq \sqrt{2\log(2m)} \max_{i=1,\ldots,m} \sigma_i.$$

[Total 25 marks]

3. Let $a < t_1 < ... < t_n < b$ be known and consider the standard nonparametric regression model

$$Y_i = m(t_i) + \varepsilon_i, \qquad i = 1, \dots, n,$$

where $\varepsilon_i \sim^{iid} N(0, 1)$ and $m : [a, b] \to \mathbb{R}$ is an unknown regression function.

(i) Briefly outline how to use *natural cubic splines* to estimate the regression function m, carefully stating the minimization problem being solved.

Let $a < x_1 < ... < x_n < b$ be known with $n \geq 3$, and consider instead the following regression model

$$Y_i = \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} g(u)du + \varepsilon_i, \qquad i = 1, \dots, n-1,$$

where $\varepsilon_i \sim^{iid} N(0, 1)$. Fix $\lambda > 0$ and consider the penalized least squares objective function

$$S_\lambda(h) = \sum_{i=1}^{n} \left(Y_i - \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} h(u)du \right)^2 + \lambda \int_a^b h'(u)^2 du.$$

We wish to estimate the regression function g by minimizing S_λ over all functions $h \in C^1[a, b]$, where $C^1[a, b]$ is the set of functions $h : [a, b] \to \mathbb{R}$ having one continuous derivative (set $S_\lambda(h) = \infty$ for $h \notin C^1[a, b]$).

A *natural quadratic spline (NQS)* h with knots $a = x_0 < x_1 < ... < x_n < x_{n+1} = b$ is any function that is (1) piecewise *quadratic* on each interval $[x_i, x_{i+1}]$, (ii) h and h' are continuous everywhere and (iii) h is constant on $[a, x_1]$ and $[x_n, b]$.

You may **assume** that for any $(y_1, \dots, y_{n-1}) \in \mathbb{R}^{n-1}$, there exists a unique natural quadratic spline g with knots at $x_1, \dots, x_n$ such that $\frac{1}{x_{i+1}-x_i} \int_{x_i}^{x_{i+1}} g(u)du = y_i$ for $i = 1, \dots, n-1$ (you do **not** need to prove this).

(ii) Let $y_1, \dots, y_{n-1} \in \mathbb{R}$ and let g be the unique NQS such that $\frac{1}{x_{i+1}-x_i} \int_{x_i}^{x_{i+1}} g(u)du = y_i$ for $i = 1, \dots, n-1$. Show that for any other function $h \in C^1[a, b]$ such that $\frac{1}{x_{i+1}-x_i} \int_{x_i}^{x_{i+1}} h(u)du = y_i$ for $i = 1, \dots, n-1$, then

$$\int_a^b h'(u)^2 du \geq \int_a^b g'(u)^2 du$$

with equality if and only if $h = g$.

(iii) Prove that any minimizer of $S_\lambda(h)$ over $h \in C^1[a, b]$ must be a natural quadratic spline with knots $x_1, \dots, x_n$.

(iv) Discuss the effect that the choice of $\lambda > 0$ has on the minimizer of S_λ, including the cases $\lambda \to 0$ and $\lambda \to \infty$.

(v) *State* two strategies on how you might pick λ in practice, briefly comparing their advantages and disadvantages.

[Total 25 marks]

Module:	MATH70081
Setter:	K. Ray
Checker:	Akyildiz
Editor:	Varty
External:	Woods
Date:	March 28, 2023

MSc EXAMINATIONS (STATISTICS)

MATH70081 Nonparametric Statistics
Time: 1 hour 30 minutes

Setter's signature

. . . .

Checker's signature

. . . .

Editor's signature

. . . .

1. Let $X_1, \dots, X_n$ be i.i.d. random variables having some probability density function $f : \mathbb{R} \to [0, \infty)$, and let E denote the corresponding expectation under the joint distribution of $X_1, \dots, X_n$.

(i) Define the kernel density estimator $\hat{f}_n = \hat{f}_{n,h}$ of f based on a kernel K and bandwidth $h > 0$. Define the mean squared error of $\hat{f}_n$ at a point $x \in \mathbb{R}$, denoted by $\mathrm{MSE}(\hat{f}_n(x))$, and state the bias-variance decomposition.

ANSWER: (SEEN) A kernel is an integrable function $K : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{\mathbb{R}} K(x)dx = 1$. For a bandwidth $h > 0$, the kernel density estimator is

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

The MSE is

$$\mathrm{MSE}(\hat{f}_h(x)) = E(\hat{f}_n(x) - f(x))^2 = \mathrm{Bias}_f(\hat{f}_n(x))^2 + \mathrm{Var}(\hat{f}_n(x)).$$

[5 marks]

Suppose now that $X_1, \dots, X_n \sim^{iid} N(0, 1)$, i.e. $f(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density function. Let

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

denote the corresponding cumulative distribution function, so that $\Phi'(x) = \varphi(x)$. Consider a kernel density estimator $\hat{f}_n$ with rectangular kernel $K(x) = 1_{[-1/2,1/2]}(x)$.

(ii) Show that

$$E\hat{f}_n(x) = \frac{1}{h} [\Phi(x + h/2) - \Phi(x - h/2)].$$

Hence or otherwise, deduce that the bias satisfies

$$\mathrm{Bias}(\hat{f}_n(x)) = \alpha(x^2 - 1)\varphi(x)h^2 + O(h^3) \qquad \text{as } h \to 0,$$

where $\alpha \in \mathbb{R}$ is a constant you should specify and the $O(h^3)$ term can depend on x.

[It may simplify your computations to note that $\varphi'(x) = -x\varphi(x)$.]

ANSWER: (SEEN SIMILAR)

$$\begin{aligned} E\hat{f}_n(x) = K_h * \varphi(x) &= \int_{\mathbb{R}} \frac{1}{h} K\left(\frac{x - y}{h}\right) \varphi(y) dy \\ &= \int_{\mathbb{R}} \frac{1}{h} 1\{-1/2 \leq (x - y)/h \leq 1/2\} \varphi(y) dy \\ &= \int_{\mathbb{R}} \frac{1}{h} 1\{x - h/2 \leq y \leq x + h/2\} \varphi(y) dy \\ &= \frac{1}{h} [\Phi(x + h/2) - \Phi(x - h/2)]. \end{aligned}$$

[This question continues on the next page . . .]

Using the Taylor expansion $\Phi(x+\delta) = \Phi(x)+\varphi(x)\delta+\varphi'(x)\delta^2/2+\varphi''(x)\delta^3/6+O(\delta^4)$,

$$\begin{aligned} \mathrm{Bias}(\hat{f}_n(x)) &= E\hat{f}_n(x) - \varphi(x) \\ &= \frac{1}{h}\left[\varphi(x)h + \varphi''(x)\frac{h^3}{24} + O(h^4)\right] - \varphi(x) \\ &= \varphi''(x)h^2/24 + O(h^3), \end{aligned}$$

where we used that the even powers of h in the brackets cancel out. Directly differentiating then gives that $\varphi''(x) = (x^2-1)\varphi(x)$, which gives the result with $\alpha = 1/24$. [8 marks]

(iii) You are given that

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh^2}\left[\Phi(x+h/2) - \Phi(x-h/2)\right] - \frac{1}{nh^2}\left[\Phi(x+h/2) - \Phi(x-h/2)\right]^2.$$

(you do not need to show the above). Using this, show that

$$\mathrm{Var}(\hat{f}_n(x)) = \frac{1}{nh}\varphi(x) + o(1/(nh)) \qquad \text{as } h \to 0,$$

where the $o(1/(nh))$ term can depend on x.

[Recall that $f(n) = o(g(n))$ means that $\lim_{n\to\infty} |f(n)/g(n)| = 0$, i.e. f is of strictly smaller order than g as $n \to \infty$.]

ANSWER: (SEEN SIMILAR) Using again the Taylor expansion for $\Phi(x+\delta)$, this becomes

$$\begin{aligned} \mathrm{Var}(\hat{f}_n(x)) &= \frac{1}{nh^2}\left[\varphi(x)h + O(h^3)\right] - \frac{1}{nh^2}\left[\varphi(x)h + O(h^3)\right]^2 \\ &= \frac{1}{nh}\varphi(x) + o(1/(nh)). \end{aligned}$$

[3 marks]

(iv) Fix $x \neq \pm 1$. Show that as $n \to \infty$, the mean-squared error $\mathrm{MSE}(\hat{f}_n(x))$ is minimized by the bandwidth choice $h = h_n = c(x,\alpha)n^{-1/5}$, where $c(x,\alpha)$ is a function of x and α that you should evaluate.

Would a bandwidth choice of the form $h \propto n^{-1/5}$ also be optimal at $x = \pm 1$? Briefly justify your answer.

[You do **not** need to derive the optimal choice of bandwidth for $x = \pm 1$.]

ANSWER: (UNSEEN) Using the bias-variance decomposition,

$$\mathrm{MSE}(\hat{f}_n(x)) \approx \frac{1}{24^2}(x^2-1)^2\varphi(x)^2h^4 + \frac{1}{nh}\varphi(x)$$

[This question continues on the next page . . .]

as $n \to \infty$. We want to minimize this with respect to h. Differentiating with respect to h gives

$$\frac{4}{24^2}(x^2-1)^2\varphi(x)^2h^3 - \frac{1}{nh^2}\varphi(x) = 0,$$

so that

$$h^5 = \frac{1}{4\alpha^2(x^2-1)\varphi(x)}\frac{1}{n}.$$

Taking the 5^{th} root gives the answer.

At $x = \pm 1$, the h^2 term in the expansion of the bias is zero and the bias thus has smaller order. The first non-zero term is now of order h^4 and hence one has a different bias-variance tradeoff (the optimal choice is then $h \simeq n^{-1/7}$).

[5 marks]

(v) Fix again $x \neq \pm 1$ and let $h = c(x,\alpha)n^{-1/5}$ be the optimal choice of bandwidth in (iv). Is the resulting mean-squared error $\mathrm{MSE}(\hat{f}_n(x))$ best as a function of n among all kernel density estimators when estimating $\varphi(x)$? If not, explain how you could modify your kernel density estimator to achieve a faster rate of convergence.

ANSWER: (UNSEEN) No, the rate is constrained by the low order of the rectangular kernel. Since φ is infinitely differentiable, one can use a higher order kernel to reduce the bias and hence get a faster rate of convergence.

[4 marks]

[Total 25 marks]

2. Let $(V_j)_{j\geq 0}$ be the Haar wavelet multiresolution analysis in $L^2(\mathbb{R})$. Define the Haar scaling function

$$\phi(x) = 1_{(0,1]}(x) = \begin{cases} 1 & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and Haar wavelet

$$\psi(x) = 1_{[0,1/2]}(x) - 1_{(1/2,1]}(x) = \begin{cases} 1 & 0 \leq x \leq 1/2 \\ -1 & 1/2 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

such that $\{\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k) : k \in \mathbb{Z}\}$ is an orthonormal basis for V_j and $\{\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k) : k \in \mathbb{Z}\}$ is an orthonormal basis for the detail space $W_j = V_{j+1} \ominus V_j$.

(i) Describe the space of functions V_J.

Let $P_J f = P_{V_J} f$ denote the projection of $f \in L^2(\mathbb{R})$ onto V_J. Define the Haar approximation $P_J f$.

ANSWER: (SEEN) V_J consists of the space of functions that are piecewise constant on intervals of the form $(k2^{-J}, (k+1)2^{-J}]$ for $k \in \mathbb{Z}$.

We have either of two equivalent definitions:

$$P_J f = \sum_{k\in\mathbb{Z}} \langle f, \phi_{Jk}\rangle_2 \phi_{Jk}(x) = \sum_{k\in\mathbb{Z}} \phi_{0k}(x) + \sum_{j=0}^{J-1} \sum_{k\in\mathbb{Z}} \langle f, \psi_{jk}\rangle_2 \psi_{jk}(x).$$

[5 marks]

(ii) Let

$$f(x) = \begin{cases} 1 + x & -1 < x \leq 0 \\ 1 - x & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Give the L^2-orthogonal projection of f onto the space V_1, leaving your answer in terms of $\phi(x)$ and $\psi(x)$.

ANSWER: (SEEM SIMILAR) We can write

$$P_1 f(x) = \sum_k c_k \phi(x - k) + \sum_k d_{0k}\psi_{0k}(x).$$

The coefficients $c_k = \langle f, \phi_k\rangle_2 = \int_k^{k+1} f(x)dx$ are all equal to zero except c_{-1} and c_0, which both equal 1/2. For the wavelet coefficients, we have

$$d_{0k} = \langle f, \psi_{0k}\rangle_2 = \int_k^{k+1/2} f(x)dx - \int_{k+1/2}^{k} f(x)dx.$$

[This question continues on the next page . . .]

These are all zero except d_{0-1} and d_{00} by the support of f. Thus,

$$\begin{aligned} d_{0-1} &= \int_{-1}^{-1/2}(1+x)dx - \int_{-1/2}^{0}(1+x)dx \\ &= [x+x^2/2]_{-1}^{-1/2} - [x+x^2/2]_{-1/2}^{0} = -1/4. \end{aligned}$$

Similarly, $d_{00} = 1/4$, giving

$$P_1 f(x) = \frac{1}{2}[\phi(x)+\phi(x+1)] + \frac{1}{4}[\psi(x)-\psi(x+1)].$$

[5 marks]

(iii) State whether the following statements are True or False, justifying your answers.

(a) All functions in V_j with $j \geq 0$ are continuous.

ANSWER: (UNSEEN) False: e.g. $\phi_{Jk} \in V_J$ is discontinuous. [3 marks]

(b) If $f \in W_j$ with $j \geq 0$, then $\int_{-\infty}^{\infty} f(x)dx = 0$.

ANSWER: (UNSEEN) True: if $f \in W_J$, then it can be expanded in the basis ψ_{jk} as $f = \sum_k d_{jk}\psi_{jk}$ with $d_{jk} = \langle f, \psi_{jk}\rangle_2$. But then $\int_{\mathbb{R}} f(x)dx = \sum_k d_{jk} \int_{\mathbb{R}} \psi_{jk}(x)dx = 0$ using the definition of the Haar wavelet.

[3 marks]

(c) If $f \in V_j$ and $g \in V_{j'}$ for $j \neq j'$, then $\langle f, g\rangle_2 = \int_{-\infty}^{\infty} f(x)g(x)dx = 0$.

ANSWER: (UNSEEN) False: if $j < j'$, then $V_j \subset V_{j'}$. So any $0 \neq f \in V_j$ is also in $V_{j'}$, while $\int ff = \int f^2 \neq 0$. [3 marks]

(iv) Suppose we observe $X_1, \ldots, X_n \sim^{iid} f$ with f a probability density on $[0,1]$. Consider the *wavelet thresholding density estimator*:

$$\hat{f}_n(x) = \hat{\alpha}\phi(x) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \hat{\beta}_{jk} 1\{|\hat{\beta}_{jk}| > \tau\}\psi_{jk}(x),$$

where $\tau > 0$ is a threshold, J is an integer such that $n/\log n \leq 2^J \leq 2n/\log n$ and

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i), \qquad \hat{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(X_i)$$

are the empirical wavelet coefficients. Explain why $\hat{\alpha} = 1$ and find the limiting distribution of each $\sqrt{n}(\hat{\beta}_{jk} - c_{jk}(f))$ as $n \to \infty$, where you should specify $c_{jk}(f)$.

Suppose that $\sup_{x\in[0,1]}|f(x)| \leq M$ for some known constant $M > 0$. Motivate a reasonable choice of threshold $\tau = \tau_n$ as a function of n [i.e. you do not need to worry about exact constants].

[This question continues on the next page . . .]

You may use the fact that if $Z_i \sim N(0, \sigma_i^2)$ for $i = 1, \dots, m$ not necessarily independent, then

$$E\left[\max_{i=1,\dots,m} |Z_i|\right] \leq \sqrt{2\log(2m)} \max_{i=1,\dots,m} \sigma_i.$$

ANSWER: (UNSEEN) Since $X_i \in [0,1]$ with probability one, $\phi(X_i) = 1_{(0,1]}(X_i) = 1$ and hence $\hat{\alpha} = 1$. We have $\hat{\beta}_{jk}$ is the sum of i.i.d. random variables with mean

$$E\psi_{jk}(X_i) = \int \psi_{jk}(x) f(x) dx = \langle f, \psi_{jk} \rangle_2$$

and variance

$$\mathrm{Var}(\psi_{jk}(X_i)) = \int \psi_{jk}^2 f dx - \left(\int f \psi_{jk} dx\right)^2.$$

By the central limit theorem,

$$\sqrt{n}(\hat{\beta}_{jk} - \langle f, \psi_{jk} \rangle_2) \to^d N(0, \mathrm{Var}(\psi_{jk}(X_1))$$

as $n \to \infty$.

We can upper bound the variance of $\hat{\beta}_{jk}$ by $M \int \psi_{jk}^2 dx/n = M/n$ using the orthonormality of ψ_{jk}. We have the number of coefficients $\sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} 1 = \sum_{j=0}^{J-1} 2^j = 2^J - 1$. Thus we have the approximate upper bound

$$\begin{aligned} E \max_{j=0,\dots,J-1} \max_{k=0,\dots,2^j-1} |\hat{\beta}_{jk} - \langle f, \psi_{jk} \rangle_2| &\leq \sqrt{M/n}\sqrt{2\log(2^{J+1})} \\ &\leq C\sqrt{J}/\sqrt{n} \\ &\leq C'\sqrt{(\log n)/n}. \end{aligned}$$

Thus if we $|\hat{\beta}_{jk}| \lesssim \sqrt{(\log n)/n}$, it can be fully explained by the noise, and hence we set the threshold $\tau_n \simeq \sqrt{(\log n)/n}$ exactly as in the regression case.

[6 marks]

[Total 25 marks]

3. Let $a < t_1 < ... < t_n < b$ be known and consider the standard nonparametric regression model

$$Y_i = m(t_i) + \varepsilon_i, \qquad i = 1, ..., n,$$

where $\varepsilon_i \sim^{iid} N(0, 1)$ and $m : [a, b] \to \mathbb{R}$ is an unknown regression function.

(i) Briefly outline how to use *natural cubic splines* to estimate the regression function m, carefully stating the minimization problem being solved.

ANSWER: (SEEN) One can consider the estimator obtained by minimizing the objective function

$$Q_\lambda(m) = \sum_{i=1}^{n} (Y_i - m(t_i))^2 + \lambda \int_a^b m''(x)^2 dx$$

over all twice differentiable functions (else the objective function is infinite). It is shown in the module that the minimizer of such an objective function will be a natural cubic spline with knots at the observed covariates/design points $t_1, ..., t_n$.

[5 marks]

Let $a < x_1 < ... < x_n < b$ be known with $n \geq 3$, and consider instead the following regression model

$$Y_i = \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} g(u)du + \varepsilon_i, \qquad i = 1, ..., n - 1,$$

where $\varepsilon_i \sim^{iid} N(0, 1)$. Fix $\lambda > 0$ and consider the penalized least squares objective function

$$S_\lambda(h) = \sum_{i=1}^{n} \left(Y_i - \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} h(u)du \right)^2 + \lambda \int_a^b h'(u)^2 du.$$

We wish to estimate the regression function g by minimizing S_λ over all functions $h \in C^1[a, b]$, where $C^1[a, b]$ is the set of functions $h : [a, b] \to \mathbb{R}$ having one continuous derivative (set $S_\lambda(h) = \infty$ for $h \notin C^1[a, b]$).

A *natural quadratic spline (NQS)* h with knots $a = x_0 < x_1 < ... < x_n < x_{n+1} = b$ is any function that is (1) piecewise *quadratic* on each interval $[x_i, x_{i+1}]$, (ii) h and h' are continuous everywhere and (iii) h is constant on $[a, x_1]$ and $[x_n, b]$.

You may **assume** that for any $(y_1, ..., y_{n-1}) \in \mathbb{R}^{n-1}$, there exists a unique natural quadratic spline g with knots at $x_1, ..., x_n$ such that $\frac{1}{x_{i+1}-x_i} \int_{x_i}^{x_{i+1}} g(u)du = y_i$ for $i = 1, ..., n - 1$ (you do **not** need to prove this).

[This question continues on the next page . . .]

(ii) Let $y_1, \ldots, y_{n-1} \in \mathbb{R}$ and let g be the unique NQS such that $\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} g(u)du = y_i$ for $i = 1, \ldots, n-1$. Show that for any other function $h \in C^1[a, b]$ such that $\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} h(u)du = y_i$ for $i = 1, \ldots, n-1$, then

$$\int_a^b h'(u)^2 du \geq \int_a^b g'(u)^2 du$$

with equality if and only if $h = g$.

ANSWER: (SEEN SIMILAR) Set $\varphi(u) = h(u) - g(u)$, so that $\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} \varphi(u)du = y_i - y_i = 0$. Since g is a NQS, we have g is linear on $[a, x_1]$ and $[x_n, b]$, and hence $g'' = 0$ on these intervals, while $g'' = a_i \in \mathbb{R}$ is constant on all other intervals of the form $[x_i, x_{i+1}]$ since g is piecewise quadratic.

Using these facts,

$$\begin{aligned} \int_a^b \varphi'(u)g'(u) &= [\varphi(u)g'(u)]_a^b - \int_a^b \varphi(u)g''(u)du \\ &= 0 - \sum_{i=1}^{n-1} a_i \int_{x_i}^{x_{i+1}} \varphi(u)du = 0 \end{aligned}$$

using that $g'(a) = g'(b) = 0$ and $\int_{x_i}^{x_{i+1}} \varphi(u)du = 0$ for $i = 1, \ldots, n-1$. Therefore, using that $\varphi = h - g$,

$$\begin{aligned} \int_a^b h'(u)^2 du - \int_a^b g'(u)^2 du &= \int_a^b (\varphi'(u) + g'(u))^2 - g'(u)^2 du \\ &= \int_a^b 2\varphi'(u)g'(u) + \varphi'(u)^2 du = \int_a^b \varphi'(u)^2 du \geq 0 \end{aligned}$$

as desired. We have equality if and only if $\int_a^b \varphi'(u)^2 du = 0$, i.e. $\varphi'(u) = 0$ on $[a, b]$ and thus φ is a constant function. But $\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} \varphi(u)du = 0$ and hence $\varphi \equiv 0$. Thus we have equality if and only if $\varphi(u) = h(u) - g(u) = 0$ for all $u \in [a, b]$, i.e. $g = h$. [8 marks]

(iii) Prove that any minimizer of $S_\lambda(h)$ over $h \in C^1[a, b]$ must be a natural quadratic spline with knots $x_1, \ldots, x_n$.

ANSWER: (SEEN SIMILAR) Suppose that h minimizes S_λ over all functions $h \in C^1[a, b]$. Let g be the unique NQS such that $\frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} g(u)du = \frac{1}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}} h(u)du$ for $i = 1, \ldots, n-1$, which exists by the question assumption. Then

$$\sum_{i=1}^n \left(Y_i - \frac{1}{x_{i+1} - x_i}\int_{x_i}^{x_{i+1}} h(u)du\right)^2 = \sum_{i=1}^n \left(Y_i - \frac{1}{x_{i+1} - x_i}\int_{x_i}^{x_{i+1}} g(u)du\right)^2.$$

[This question continues on the next page . . .]

Since the NQS minimizes $\int_a^b h'(u)^2 du$ over all functions with this property, this implies $\int_a^b h'(u)^2 du \geq \int_a^b g'(u)^2 du$ with equality if and only if $g = h$. Hence $S_\lambda(h) \leq S_\lambda(g)$ with equality if and only if $g = h$. Since h minimizes S_λ by assumption, we must have $h = g$, i.e. h is a natural quadratic spline with knots at $x_1, \dots, x_n$. [5 marks]

(iv) Discuss the effect that the choice of $\lambda > 0$ has on the minimizer of S_λ, including the cases $\lambda \to 0$ and $\lambda \to \infty$.

ANSWER: (SEEN SIMILAR) As λ increases, the penalty term becomes increasingly dominant and hence the derivative of the minimizer will be smaller in magnitude. In the limit $\lambda \to \infty$, we require $h'(u) \equiv 0$ everywhere, and hence we will obtain the best constant approximation to the data in a least-squares sense.

As λ decreases, the fit to the training data becomes more important and so the derivative of the minimizer g can have larger magnitude (i.e. the function can change rapidly). In the limit $\lambda \to 0$, you will recover a function that perfectly minimizes the least squares criterion, i.e. $\frac{1}{x_{i+1}-x_i} \int_{x_i}^{x_{i+1}} g(u)du = y_i$ for $i = 1, \dots, n-1$. [4 marks]

(v) *State* two strategies on how you might pick λ in practice, briefly comparing their advantages and disadvantages.

ANSWER: (SEEN) In practice, two common strategies are cross-validation and using a train-test split. Train-test split is typically faster than cross-validation since we need to refit the model fewer times, and reduces the dependence between parameter tuning and model fitting, which can lead to overfitting. However, since it only uses a fraction of the data for model fitting, it can be inefficient for small or moderate data sizes, whereas cross-validation uses the data more fully. [3 marks]

[Total 25 marks]