UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

Thursday 7 June 2007 1.30 to 3.30

NONPARAMETRIC STATISTICAL THEORY

*Attempt **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Let $n \geqslant 3$ and $a \leqslant x_1 < x_2 < \ldots < x_n \leqslant b$. What is meant by a *cubic spline* with knots at $x_1, \ldots, x_n$? What is meant by a *natural cubic spline*?

Write $S_2[a,b]$ for the class of all real-valued functions on $[a,b]$ having two continuous derivatives, and let $\mathbf{g} = (g_1, \ldots, g_n)^T$. Find the function $\tilde{g} \in S_2[a,b]$ that minimises $R(\tilde{g}'') = \int_a^b \tilde{g}''(x)^2\, dx$ subject to $\tilde{g}(x_i) = g_i$ for $i = 1, \ldots, n$.

[*You may assume that there is a unique natural cubic spline g with knots at $x_1, \ldots, x_n$ such that $g(x_i) = g_i$ for $i = 1, \ldots, n$.*]

Consider the fixed design nonparametric regression model

$$Y_i = g(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent with mean 0 and variance σ^2. Define the *penalised sum of squares* $S_\lambda(\tilde{g})$, for $\lambda \in (0, \infty)$, and show that there is a unique minimiser $\hat{g}_\lambda$ of $S_\lambda(\tilde{g})$ over $\tilde{g} \in S_2[a,b]$.

[*You may assume that if g is a natural cubic spline with knots at $x_1, \ldots, x_n$, then there is a symmetric matrix K such that $z^T K z \geqslant 0$ for all $z \in \mathbb{R}^n$ and such that $\int_a^b g''(x)^2\, dx = \mathbf{g}^T K \mathbf{g}$, where $\mathbf{g} = (g_1, \ldots, g_n)^T$ and $g_i = g(x_i)$ for $i = 1, \ldots, n$.*]

2 Consider the nonparametric regression model

$$Y_i = m(x_i) + \epsilon_i,$$

where $x_i = i/n$ for $i = 1, \ldots, n$, and $\epsilon_1, \ldots, \epsilon_n$ are independent with mean 0 and variance σ^2. We are interested in estimating the regression function $m(x)$. State the weighted least squares problem that is solved by the local polynomial kernel estimator $\hat{m}_h(x; p)$ of degree p and bandwidth h. Give an expression, in terms of matrices W and X and a vector Y, all of which you should define, for a vector whose first component is $\hat{m}_h(x; p)$.

Assume that m is twice continuously differentiable on $[0, 1]$, that the kernel K is non-negative, symmetric, continuously differentiable on $[-1, 1]$ and zero outside $[-1, 1]$, and that $h \to 0$ as $n \to \infty$ but $nh^2 \to \infty$ as $n \to \infty$. Show that, for $x \in (0, 1)$,

$$\mathbb{E}\{\hat{m}_h(x; 0)\} - m(x) = \frac{1}{2} h^2 \mu_2(K) m''(x) + o(h^2)$$

as $n \to \infty$, where $\mu_r(K) = \int_{-1}^{1} y^r K(y)\, dy$.

[*You may assume that under the stated conditions,*

$$\frac{1}{nh} \sum_{i=1}^{n} (x_i - x)^r K\Big(\frac{x_i - x}{h}\Big) = \begin{cases} h^r \mu_r(K) + o(h^r) & \text{if } r \text{ is even} \\ O(h^{r-1}/n) & \text{if } r \text{ is odd.} \end{cases} \quad]$$

Now suppose that $z_n = \alpha h$, for some $\alpha \in [0, 1)$. Derive the corresponding expansion for the bias of $\hat{m}_h(z_n; 0)$ in this case.

[*You may assume that if* $\mu_{r,\alpha}(K) = \int_{-\alpha}^{1} y^r K(y)\, dy$, *then*

$$\frac{1}{nh} \sum_{i=1}^{n} (x_i - \alpha h)^r K\Big(\frac{x_i - \alpha h}{h}\Big) = h^r \mu_{r,\alpha}(K) + o(h^r),$$
$$\frac{1}{nh} \sum_{i=1}^{n} |x_i - \alpha h|^r K\Big(\frac{x_i - \alpha h}{h}\Big) = O(h^r). \quad]$$

State the order of the bias of the local linear kernel estimator $\hat{m}_h(\cdot; 1)$ both at an interior point $x \in (0, 1)$ and at a sequence of boundary points $z_n = \alpha h$.

3 What does it mean to say that a non-degenerate distribution function is *max-stable*? Explain without proof the relevance of this concept for determining possible limiting distributions of appropriately normalised sample maxima of independent and identically distributed random variables. Give expressions for the three extreme value types, and state a necessary and sufficient condition (in terms of these types) for a distribution function to be max-stable.

Let (X_n) be a sequence of independent and identically distributed random variables with distribution function F and let $X_{(n)} = \max_{1 \leqslant i \leqslant n} X_i$. Let $\tau \in [0, \infty)$ and let (u_n) be a real-valued sequence. Prove that $\mathbb{P}(X_{(n)} \leqslant u_n) \to e^{-\tau}$ as $n \to \infty$ if and only if $n\{1 - F(u_n)\} \to \tau$ as $n \to \infty$.

Let $S_n = \sum_{i=1}^{n} \mathbb{1}_{\{X_i > u_n\}}$ and suppose that $n\{1-F(u_n)\} \to \tau \in [0, \infty)$. Prove that for each $k \in \{0, 1, 2, \ldots\}$,

$$\mathbb{P}(S_n \leqslant k) \to e^{-\tau} \sum_{s=0}^{k} \frac{\tau^s}{s!}$$

as $n \to \infty$.

[*The standard result about convergence of binomial distributions to a Poisson distribution may be used without proof.*]

Now suppose that there exist constants $a_n > 0$, b_n and a non-degenerate distribution function G such that for all $x \in \mathbb{R}$,

$$\mathbb{P}\Big(\frac{X_{(n)} - b_n}{a_n} \leqslant x\Big) \to G(x)$$

as $n \to \infty$. Let W_n denote the second largest of $X_1, \ldots, X_n$. By considering the events $\{W_n \leqslant u_n\}$ and $\{S_n \leqslant 1\}$, show that whenever $G(x) > 0$,

$$\mathbb{P}\Big(\frac{W_n - b_n}{a_n} \leqslant x\Big) \to G(x)\{1 - \log G(x)\}$$

as $n \to \infty$.

4 Write an essay on the theory of canonical kernels and optimal kernel choice in kernel density estimation.

[*You may find the following formula helpful:*

$$AMISE(\hat{f}_h) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 \mu_2^2(K) R(f'').$$]

END OF PAPER