UNIVERSITY OF CAMBRIDGE

# MATHEMATICAL TRIPOS Part III

Thursday, 31 May, 2018 1:30 pm to 3:30 pm

## PAPER 210

## TOPICS IN STATISTICAL THEORY

*Attempt no more than **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 $X_1, \ldots, X_n$ be independent and identically distributed real-valued random variables with density $f$. Given a point $x \in \mathbb{R}$ define the *k-nearest neighbour distance* of $x$, denoted $\rho_{(k)}(x)$.

For $r \geqslant 0$ and $x \in \mathbb{R}$ write $p_x(r) = \int_{x-r}^{x+r} f(y)\,dy$. Show that the random variable defined by $P = p_x(\rho_{(k)}(x))$ has Beta density

$$B_{k,n+1-k}(s) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n+1-k)} s^{k-1}(1-s)^{n-k}$$

for $s \in (0,1)$. Calculate $\mathbb{E}P$ and $\mathbb{E}P^2$.

Henceforth suppose that $f$ is $L$-Lipschitz and strictly positive on all of $\mathbb{R}$. Prove that $|p_x(r) - 2rf(x)| \leqslant Lr^2$ and hence, writing $p_x^{-1}$ for the inverse of $p_x$, verify that

$$|2f(x)p_x^{-1}(s) - s| \leqslant \frac{Ls^2}{f(x)^2}$$

for any $s \in (0,1)$ and $x$ such that $f(x) \geqslant L^{1/2}$.

Write $\hat{f}_{(k)}(x) = \frac{k}{2(n+1)\rho_{(k)}(x)}$ for the $k$-nearest neighbour density estimator at $x$. Prove that

$$\left| \mathbb{E}\left( \frac{f(x)}{\hat{f}_{(k)}(x)} \right) - 1 \right| \leqslant \frac{k+1}{n+2}$$

for any $x$ such that $f(x) \geqslant L^{1/2}$.

**2**

Consider the fixed-design nonparametric regression model in which we observe $Y_i = m(x_i) + \sigma\epsilon_i$ for $i = 1, \ldots, n$, where $x_i = i/n, \sigma \in (0, \infty)$ and $\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d.}}{\sim} N(0, 1)$. For a kernel $K : \mathbb{R} \to [0, \infty)$ and bandwidth $h \in (0, \infty)$ give the definition of the *local polynomial estimator* $\hat{m}_h(x; p)$ of $m(x)$, and derive the Nadaraya–Watson (local constant) estimator $\hat{m}(x)$.

We henceforth restrict attention to the uniform kernel $K(x) = 2^{-1}\mathbb{1}_{\{|x|\leqslant 1\}}$. For $L > 0$ let

$$\Theta_L = \{m : |m(y) - m(x)| \leqslant L|x - y| \text{ for all } x, y \in [0, 1]\}$$

denote the set of $L$-Lipschitz functions on $[0, 1]$. Writing $\mathbb{E}_m$ for the expectation when the true mean function is $m$, show that

$$\sup_{m\in\Theta_L} \mathbb{E}_m\left[\int_h^{1-h} \{\hat{m}_h(x) - m(x)\}^2 \, dx\right] \leqslant L^2h^2 + \frac{3\sigma^2}{2nh}$$

whenever $nh \geqslant 1$.

Deduce that there exists $n_0$ depending only on $L$ and $\sigma^2$ such that, for $n \geqslant n_0$,

$$\inf_{h\in(0,1/3)} \sup_{m\in\Theta_L} \mathbb{E}_m\left[\int_h^{1-h} \{\hat{m}_h(x) - m(x)\}^2 \, dx\right] \leqslant C\left(\frac{L\sigma^2}{n}\right)^{2/3}$$

where $C > 0$ is a constant that you should specify.

Fix $x_0 \in (0, 1)$. State Le Cam's two point lemma and use it to show that there exists a constant $c > 0$ such that

$$\inf_{\tilde{m}} \sup_{m\in\Theta_L} \mathbb{E}_m[\{\tilde{m}(x_0) - m(x_0)\}^2] \geqslant c\left(\frac{L\sigma^2}{n}\right)^{2/3},$$

where the infimum is taken over all estimators $\tilde{m}$ based on $\{Y_1, \ldots, Y_n\}$.

[*Hint: You may wish to consider the function* $m_1(x) = \lambda h K(\frac{x-x_0}{h})$, *where* $K(t) = \exp(-\frac{1}{1-t^2})\mathbb{1}_{\{|t|\leqslant 1\}}$ *and* $\lambda$ *is a scalar to be chosen.*]

**3**

For a non-degenerate distribution function $G$ define the *domain of attraction* of $G$, denoted $D(G)$, in the context of extreme value theory for sample maxima.

Defining the notion of a *regularly varying* function, state necessary and sufficient conditions for a distribution function $F$ to satisfy $F \in D(G)$ for the three cases of $G$ being the Fréchet($\alpha$), the Negative Weibull($\alpha$) and the Gumbel distribution functions. State sufficient conditions in terms of the hazard function.

For a positive integer $m$ let $f_m(x) = \frac{x^{m-1}}{(m-1)!}e^{-x}$ denote the $\Gamma(m,1)$ density. Writing $F_m$ for the corresponding distribution function show that

$$F_m(x+\beta_n)^n \to e^{-e^{-x}}$$

as $n \to \infty$, for all $x \in \mathbb{R}$, where $\beta_n = \log n + (m-1)\log\log n - \log((m-1)!)$. *[Hint: You may first wish to show that*

$$1 - F_m(x) = e^{-x}\sum_{j=0}^{m-1}\frac{x^j}{j!}$$

*for all $x \in [0,\infty)$.]*

**4**

Let $Y_1, \ldots, Y_n$ be independent, mean-zero random variables with $Y_i$ taking values in $[a_i, b_i]$. Prove Hoeffding's inequality,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} Y_i\Big| > \epsilon\Big) \leqslant 2\exp\Big(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\Big),$$

for each $\epsilon > 0$.

Now write $M := \max_{i=1,\ldots,n}\max\{-a_i, b_i\}$ and $\sigma^2 := \max_{i=1,\ldots,n}\mathbb{E}Y_i^2$. Assuming that you may interchange the order of expectation and summation where necessary, prove Bennett's inequality,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n} Y_i\Big| > \epsilon\Big) \leqslant 2\exp\Big(-\frac{n\sigma^2}{M^2}\phi\Big(\frac{M\epsilon}{n\sigma^2}\Big)\Big),$$

for each $\epsilon > 0$, where $\phi(x) = (1+x)\log(1+x) - x$. [*Hint: You may wish to bound a moment generating function using the fact that $\mathbb{E}Y_i^k \leqslant \sigma^2 M^{k-2}$ for all $k \geqslant 2$.*]

Let $X \sim \mathrm{Bin}(n, p_n)$ with $p_n \to 0$ and $np_n \to \infty$ as $n \to \infty$. For a fixed $C > 0$ and large $n$, which of the two inequalities above provides a better bound on $\mathbb{P}\Big(\frac{|X - np_n|}{\sqrt{np_n(1-p_n)}} \geqslant C\Big)$?

**END OF PAPER**