UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Tuesday, 15 June, 2021 12:00 pm to 2:00 pm

PAPER 223

ROBUST STATISTICS

Before you begin please read these instructions carefully

Candidates have TWO HOURS to complete the written examination.

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
Rough 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 Changing neighborhoods.
Consider the minimax bias problem

$$\min_{\{T_n\}\subseteq\mathcal{T}} \max_{F\in\mathcal{P}_\epsilon^K(\Phi)\cap\mathcal{M}} b(\{T_n\}, F).$$

Here, we write

$$\mathcal{P}_\epsilon^K(\Phi) := \left\{F : \sup_{t\in\mathbb{R}} |F(t)-\Phi(t)| \leqslant \epsilon\right\}$$

to denote the Kolmogorov ϵ-neighborhood of the standard normal distribution, and let $\mathcal{M}$ denote the class of distributions with a finite variance and a probability density function that is nonzero on all of $\mathbb{R}$. We write $\mathcal{T}$ to denote the class of translation-invariant estimators for which the asymptotic bias $b(\{T_n\}, F) = |\lim_{n\to\infty} \mathbb{E}_F(T_n)|$ is well-defined for all $F \in \mathcal{M}$. Suppose $\epsilon \in \left(0, \frac{1}{2}\right)$.

(a) Show that when $\{T_n\}$ corresponds to the sample median, we have the upper bound

$$\max_{F\in\mathcal{P}_\epsilon^K(\Phi)\cap\mathcal{M}} b(\{T_n\}, F) \leqslant \Phi^{-1}\left(\frac{1}{2}+\epsilon\right) := b_1.$$

[Hint: You may use, without proof, the fact that $b(\{T_n\}, F) = F^{-1}\left(\frac{1}{2}\right)$ for $F \in \mathcal{M}$.]

(b) Suppose we can construct symmetric distributions $F_+, F_- \in \mathcal{P}_\epsilon^K(\Phi)\cap\mathcal{M}$, centered at $\pm b_1$, such that $F_-(t) = F_+(t+2b_1)$. Show that this implies the lower bound

$$\min_{\{T_n\}\subseteq\mathcal{T}} \max_{F\in\mathcal{P}_\epsilon^K(\Phi)\cap\mathcal{M}} b(\{T_n\}, F) \geqslant b_1.$$

Thus, the median is minimax optimal.

(c) Now find a construction of F_+ and F_- according to the prescription in part (b).

2 Influence vs. breakdown.
For $0 \leqslant \alpha < \frac{1}{2}$, consider the α-trimmed mean, defined as

$$T_n(x_1, \ldots, x_n) := \frac{1}{n-2m} \sum_{i=m+1}^{n-m} x_{(i)},$$

where $m = \lfloor \alpha n \rfloor$ and $x_{(i)}$ denotes the i^{th} order statistic. If we define the functional

$$T(F) = \frac{1}{1-2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(s) ds,$$

it can be shown (under appropriate regularity conditions) that $T_n \xrightarrow{P} T(F)$ when $x_i \overset{i.i.d.}{\sim} F$.

(a) Compute the influence function $IF(x; T, F)$ when F is a differentiable, strictly monotonic cumulative distribution function of a distribution with a probability density function which is symmetric around 0.

[You may assume that interchanging derivatives and integrals is allowed, and also use the fact that

$$\frac{d}{dt} F_t^{-1}(s) \Big|_{t=0} = \frac{s - \Delta_x(F^{-1}(s))}{F'(F^{-1}(s))},$$

when $F_t = (1-t)F + t\Delta_x$, without proof.]

[Hint: Your answer should be the influence function of the Huber M-estimator with parameter $k = -F^{-1}(\alpha)$. You may find it useful to note that $F^{-1}(1-t) = -F^{-1}(t)$ and F' is even. Also recall the formula $\frac{dF^{-1}(t)}{dt} = \frac{1}{F'(F^{-1}(t))}$.]

(b) Show that the breakdown point of the trimmed mean is $\frac{1}{n}\lfloor \alpha n \rfloor$.

3 Median as a scale M-estimator.

Consider the normal-scale family, where F_θ is the cdf of a $N(0, \theta^2)$ distribution. Let T_n denote the sample median of $\{|x_1|, \ldots, |x_n|\}$ (defined in the usual way as the average of the two middle order statistics when n is even).

(a) Suppose the $|x_i|$'s are unique (which happens with probability 1 when $x_i \overset{i.i.d.}{\sim} F_\theta$). Show that T_n is a solution to the estimating equation $\frac{1}{n}\sum_{i=1}^{n} \psi\left(\frac{x_i}{t}\right) = 0$, where $\psi(u) = \text{sign}(|u| - 1)$, and $\text{sign}(u)$ is defined in the usual way to be ± 1 for $\pm u > 0$ and 0 for $u = 0$. Thus, T_n is a scale M-estimator.

(b) What is the asymptotic distribution of T_n when $x_i \overset{i.i.d.}{\sim} F_\theta$? Use the result to derive an asymptotically valid level-α hypothesis test for

$$H_0 : \theta^2 = 1 \quad \text{vs.} \quad H_1 : \theta^2 > 1$$

based on T_n.

[Hint: One way to approach this problem is to use the general theorem about location M-estimators, which states that

$$\sqrt{n}(T_n - t_0) \xrightarrow{d} N\left(0, \frac{\sigma^2(t_0)}{(\lambda'(t_0))^2}\right),$$

where t_0 is a root of $\lambda(t) = \mathbb{E}_F[\psi(x_i - t)]$ and $\sigma^2(t) = \mathbb{E}_F[\psi^2(x_i - t)] - \lambda^2(t)$.]

4 Not quite an M-estimator.

Suppose $\{x_i\}_{i=1}^n$ are i.i.d. with $\mu := \mathbb{E}(x_i)$ and $\mathbb{E}(x_i^2) \leqslant \sigma^2 < \infty$. Let ψ be a non-decreasing function satisfying

$$-\log\left(1 - t + \frac{t^2}{2}\right) \leqslant \psi(t) \leqslant \log\left(1 + t + \frac{t^2}{2}\right).$$

(a) Show that for any $\theta > 0$, we have

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^n (\psi(\theta x_i) - \theta\mathbb{E}(x_i))\right)\right] \leqslant \exp\left(\frac{\theta^2}{2}\sum_{i=1}^n \mathbb{E}(x_i^2)\right),$$
$$\mathbb{E}\left[\exp\left(\sum_{i=1}^n (\theta\mathbb{E}(x_i) - \psi(\theta x_i))\right)\right] \leqslant \exp\left(\frac{\theta^2}{2}\sum_{i=1}^n \mathbb{E}(x_i^2)\right).$$

[Hint: The inequality $1 + x \leqslant \exp(x)$ for all $x \in \mathbb{R}$ may be helpful.]

(b) Let $\widehat{\mu}_\theta = \frac{1}{n\theta}\sum_{i=1}^n \psi(\theta x_i)$. Use the inequalities in part (a) to show that

$$\mathbb{P}\left(\left|\frac{1}{\theta}\sum_{i=1}^n (\psi(\theta x_i) - \theta\mathbb{E}(x_i))\right| \geqslant t\right) \leqslant 2\exp\left(-\theta t + \frac{\theta^2\sigma^2 n}{2}\right),$$

for any $\theta, t > 0$. Conclude that for $\delta > 0$, taking $\theta = \frac{\sqrt{2\log(2/\delta)}}{\sigma\sqrt{n}}$ and $t = \sigma\sqrt{2n\log(2/\delta)}$ gives

$$\mathbb{P}\left(|\widehat{\mu}_\theta - \mu| \geqslant \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\right) \leqslant \delta.$$

[Hint: Use Markov's inequality after exponentiating the appropriate quantities.]

END OF PAPER