UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 13 June, 2022 1:30 pm to 3:30 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

Suppose $T = \{T_n\}$ is the median estimator, and let $A(T, F)$ denote the asymptotic variance, computed with respect to i.i.d. samples from a distribution with cdf F.

(a) Suppose F has a valid pdf f which is positive on all of $\mathbb{R}$. Derive an expression for $A(T, F)$. [Your answer should be in terms of F and f, and should be valid even when F does not correspond to a symmetric distribution.]

(b) What is $\max_{F \in \mathcal{P}_\epsilon(\Phi)} A(T, F)$, where Φ denotes the standard normal cdf and $\mathcal{P}_\epsilon(\Phi)$ consists of (possibly asymmetric) distributions in the Huber ϵ-contamination neighborhood around Φ with a valid pdf?

(c) Now solve the problem in part (b), where we optimize only over distributions in $\mathcal{P}_\epsilon(\Phi)$ which are symmetric. How does the answer compare to the one obtained in part (b)?

[You may quote any result from the lectures that you need, without proof.]

2

Consider the normal location family, where F_θ is the cdf of a $N(\theta, 1)$ distribution.

(a) Derive the form of an optimal B-robust M-estimator for θ. [Provide the univariate function defining the optimal M-estimator, as well as the explicit form of the estimator computed from a set of finite samples $\{x_1, \ldots, x_n\}$. Your answer should involve a truncation parameter $b > 0$.]

(b) As b ranges over the interval $(0, \infty)$, what are the corresponding values of the upper bound c on the gross error sensitivity of the estimator in the optimization problem? [*Hint: It may be helpful to show that the map $b \mapsto c$ is monotonic. You may use the fact that $\mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}}$ when $X \sim N(0, 1)$, without proof.*]

(c) What is the most B-robust location M-estimator for θ, and what is its gross error sensitivity?

[You may quote any result from the lectures that you need, without proof.]

3

Suppose $\{x_i\}_{i=1}^n$ are i.i.d. samples from a normal location family $N(\theta, 1)$, and consider a simple hypothesis test of

$$H_0 : \theta = \theta_0 \text{ vs. } H_1 : \theta = \theta_1,$$

where $\theta_1 > \theta_0$. Let F_θ denote the cdf of $N(\theta, 1)$.

(a) Let $T_n = \frac{1}{n}\sum_{i=1}^n (\theta_1 - \theta_0)\left(x_i - \frac{\theta_0+\theta_1}{2}\right)$. Show that a likelihood ratio test should reject H_0 when $T_n > C_\alpha$, where C_α is a critical value depending on the level α of the test.

(b) Let $T(F) = \mathbb{E}_F[T_n]$. Are the influence functions $IF_{test}(x; T, F_{\theta_0})$ and $IF_{test}(x; T, F_{\theta_1})$ bounded in x?

(c) Now consider the truncated statistic

$$S_n = \frac{1}{n}\sum_{i=1}^n \left[(\theta_1 - \theta_0)\left(x_i - \frac{\theta_0+\theta_1}{2}\right)\right]_a^b,$$

for some truncation parameters $a < b$. Let $S(F) = \mathbb{E}_F[S_n]$. Are the influence functions $IF_{test}(x; S, F_{\theta_0})$ and $IF_{test}(x; S, F_{\theta_1})$ bounded in x? [*Hint: It may be helpful to work with the notation* $S_n = \frac{1}{n}\sum_{i=1}^n \psi(x_i)$.]

(d) Suppose $\epsilon \in (0, 1)$.

(i) Show that there exist $c, d \in \mathbb{R}$ such that the functions

$$g_0(x) = \begin{cases} (1-\epsilon) f_{\theta_0}(x) & \text{if } \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} < c, \\ \frac{(1-\epsilon) f_{\theta_1}(x)}{c} & \text{if } \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \geqslant c, \end{cases}$$

$$g_1(x) = \begin{cases} (1-\epsilon) f_{\theta_1}(x) & \text{if } \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > d, \\ d(1-\epsilon) f_{\theta_0}(x) & \text{if } \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \leqslant d, \end{cases}$$

are densities corresponding to distributions G_0 and G_1 which lie in the Huber ϵ-contamination neighborhoods around F_{θ_0} and F_{θ_1}, respectively.

(ii) It can be shown that for sufficiently small ϵ, we have $c > d$. Assuming this fact, show that a hypothesis test based on S_n corresponds to a likelihood ratio test of

$$H_0 : F = G_0 \text{ vs. } H_1 : F = G_1,$$

for appropriate values of (a, b).

[The above results can be used to show that a test based on S_n is a minimax optimal test between distributions in the ϵ-contamination neighborhoods, but you do not have to prove this.]

4

Suppose $x_1, \ldots, x_n$ are i.i.d. samples from the d-dimensional standard normal distribution $N(0, I_d)$, and let $\widehat{\mu}$ denote the coordinatewise median. [In other words, the j^{th} coordinate of $\widehat{\mu}$ is defined to be the sample median of the j^{th} coordinates $\{x_{ij}\}_{j=1}^n$ of the data points.]

(a) Let $\epsilon \in (0, 1)$. Taking the case $d = 1$, show that for sufficiently large n, we have

$$\mathbb{P}\left(\sup_{\{\widetilde{x}_i\}_{i=1}^n} |\widehat{\mu}(\widetilde{x}_1, \ldots, \widetilde{x}_n)| > \epsilon \right) \geqslant 1 - \exp(-cn),$$

where $\{\widetilde{x}_i\}_{i=1}^n$ denotes an adversarial ϵ-perturbation of $\{x_i\}_{i=1}^n$, and c is a constant which does not depend on n (but may depend on ϵ).

[*Hint: Let $\{\widetilde{x}_i\}_{i=1}^n$ be constructed by moving the last $\lfloor \epsilon n \rfloor$ data points to a point mass at 100. Note that the probability in question is lower-bounded by the probability that at least $\frac{n+1}{2} - \lfloor \epsilon n \rfloor$ of the unperturbed points are greater than ϵ. Hoeffding's inequality, which states that for i.i.d. Bernoulli random variables $Y_1, \ldots, Y_m$ with mean p, we have*

$$\mathbb{P}\left(\frac{1}{m} \sum_{i=1}^m (Y_i - p) \leqslant -t \right) \leqslant \exp(-2mt^2),$$

for any $t > 0$, may be used without proof.]

(b) Now show how to adapt the argument in part (a) to conclude that for arbitrary $d \geqslant 1$, and for sufficiently large n (as a function of d and ϵ), we have

$$\mathbb{P}\left(\sup_{\{\widetilde{x}_i\}_{i=1}^n} \|\widehat{\mu}(\widetilde{x}_1, \ldots, \widetilde{x}_n)\|_2 > \epsilon\sqrt{d} \right) \geqslant \frac{1}{2}.$$

[*Hint: Bernoulli's inequality, which states that $(1 - x)^d \geqslant 1 - dx$ when $d \geqslant 1$ and $x \in [0, 1]$, may be used without proof.*]

(c) How does the bound in part (b) compare with the high-probability bound satisfied by the Tukey median for adversarially contaminated data?

[You may quote any result from the lectures that you need, without proof.]

END OF PAPER