MAT3

MATHEMATICAL TRIPOS Part III

Monday 15 June 2026 2:00pm to 5:00pm

PAPER 223

ROBUST STATISTICS

Before you begin please read these instructions carefully

Candidates have TWO HOURS to complete the written examination.

Attempt **ALL** questions.
There are **THREE** 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 Consider the following location estimators for the normal location model, i.e., $x_i \overset{i.i.d.}{\sim} N(\theta, 1)$: the median $\widehat{\theta}_{med}$, the Huber estimator $\widehat{\theta}_{Hub,k}$, and the trimmed mean $\widehat{\theta}_{trim,\gamma} = \frac{1}{n-2\lfloor \gamma n \rfloor} \sum_{i=\lfloor \gamma n \rfloor+1}^{n-\lfloor \gamma n \rfloor} x_{(i)}$.

(a) Calculate the gross error sensitivity of the median $\widehat{\theta}_{med}$, Huber estimator $\widehat{\theta}_{Hub,k}$ (as a function of the Huber parameter k), and trimmed mean $\widehat{\theta}_{trim,\gamma}$ (as a function of the trimming parameter γ). [You may use, without proof, the following formula for the influence function of the trimmed mean:

$$IF(x; T_\gamma, F) = \begin{cases} \frac{F^{-1}(\gamma)-\mu(F)}{1-2\gamma}, & \text{if } x < F^{-1}(\gamma), \\ \frac{x-\mu(F)}{1-2\gamma}, & \text{if } x \in [F^{-1}(\gamma), F^{-1}(1-\gamma)], \\ \frac{F^{-1}(1-\gamma)-\mu(F)}{1-2\gamma}, & \text{if } x > F^{-1}(1-\gamma), \end{cases}$$

where $T_\gamma(F) = \frac{1}{1-2\gamma} \int_\gamma^{1-\gamma} F^{-1}(u)du$ and $\mu(F)$ is the distribution mean.]

(b) Prove that the breakdown point of any translation-invariant estimator is at most $\frac{1}{n}\lfloor \frac{n-1}{2} \rfloor$.

(c) What are the breakdown points of the median $\widehat{\theta}_{med}$, Huber estimator $\widehat{\theta}_{Hub,k}$, and trimmed mean $\widehat{\theta}_{trim,\gamma}$? *[Hint: For the Huber estimator, consider the estimating equation evaluated at t and argue about the two cases $t < x_{(1)} - k$ and $t > x_{(n)} + k$.]*

(d) Suppose we wish to perform a hypothesis test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$. Provide an asymptotically level-α test based on each of the three test statistics above. In what sense are these tests robust? [You should derive an explicit expression for the rejection region of each test, but you do not need to simplify it.]

[You may quote any result from the lectures that you need, without proof. In particular, you may assume that all three estimators are asymptotically normal, with the usual formula for the asymptotic variance in terms of the influence function, without checking any regularity conditions.]

2 Suppose we observe data points $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.

(a) Show that, for a fixed value of $\sigma > 0$, the regression M-estimator

$$\min_\theta \sum_{i=1}^n \rho_k \left(\frac{y_i - x_i^T \theta}{\sigma} \right),$$

where ρ_k is the Huber loss with parameter k, is equivalent to the penalized least squares problem

$$\min_{(\theta,\gamma)} \left\{ \frac{1}{2\sigma} \|y - X\theta - \gamma\|_2^2 + k\|\gamma\|_1 \right\}.$$

(b) Consider the penalty function $g_\ell(\gamma) := \sum_{i=1}^n g_\ell(\gamma_i)$, defined for a parameter $\ell > 0$, where for $u \in \mathbb{R}$, we define

$$g_\ell(u) := \ell \cdot 1\{u \neq 0\}.$$

Show that the penalized least squares problem

$$\min_{(\theta,\gamma)} \left\{ \|y - X\theta - \gamma\|_2^2 + g_\ell(\gamma) \right\}$$

is equivalent to the *skipped-mean regression* M*-estimator*

$$\min_\theta \sum_{i=1}^n \rho_L(y_i - x_i^T \theta),$$

where

$$\rho_L(u) := \begin{cases} u^2, & \text{for } |u| \leqslant L, \\ L^2, & \text{for } |u| > L, \end{cases}$$

is the "skipped-mean" loss, for a value of L you should determine as a function of ℓ. *[Hint: Consider the penalized least squares objective where we optimize over each component of γ.]*

(c) Write down the estimating equation corresponding to the skipped-mean regression M-estimator in (b).

(d) Suppose the breakdown point of the skipped mean regression M-estimator is defined to be the threshold at which there exist solutions to the estimating equation in part (c) that grow arbitrarily large (where we can contaminate both x_i's and y_i's). Find the breakdown point of the estimator, assuming for simplicity that $X = \{x_i\}_{i=1}^n$ has $x_i \neq (0, 0, \dots, 0)^T$ for all i. *[Hint: Let $v \in \mathbb{R}^p$ be such that $x_i^T v \neq 0$ for all i. Define $\theta_t = tv$, and argue about the behavior of the estimating equations as $t \to \infty$.]*

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

3 Suppose $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a distribution with mean μ and variance σ^2. Suppose $\delta \in (0,1)$ is fixed, and let $k = \lceil 8\log(1/\delta) \rceil$. Suppose $\frac{n}{k}$ is an integer.

(a) Let $\widehat{\mu}_{MOM}$ denote the median-of-means estimator based on k groups. Define $\widehat{\mu}_{MOM}$ in terms of $\{x_i\}_{i=1}^n$, and state a deviation inequality that holds for $|\widehat{\mu}_{MOM} - \mu|$ with probability at least $1 - \delta$.

(b) Suppose we define $\widehat{\mu}_{MOM'}$ to be the *mean-of-medians estimator*, i.e., randomly split the data into k groups and compute the mean of the k group medians. Prove or disprove the statement that $|\widehat{\mu}_{MOM'} - \mu|$ satisfies a similar deviation inequality to the one in part (a).

(c) For a fixed value of δ (and therefore also k), derive the asymptotic distribution of $\widehat{\mu}_{MOM}$, as $n \to \infty$. Assume for simplicity that k is an odd integer. [Your answer should be a function of k, and it is enough to derive an expression for the limiting cumulative distribution function.]

Now suppose n is a perfect square and we calculate $\widehat{\mu}_{MOM}$ based on $k = \sqrt{n}$ groups.

(d) State a conjecture about the asymptotic distribution of $\sqrt{n}(\widehat{\mu}_{MOM} - \mu)$ as $n \to \infty$. (You do not need to prove your conjecture, but you should explain your logic.)

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

END OF PAPER