UNIVERSITY OF CAMBRIDGE

MAMA/223, NST3AS/223, MAAS/223

MAT3

MATHEMATICAL TRIPOS Part III

Friday 31 May 2024 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 Scale estimation via location estimation

Consider a family of univariate distributions parametrized by scale, where the probability density function of F_σ is defined by

$$f_\sigma(x) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right),$$

for a fixed $f : \mathbb{R} \to \mathbb{R}$.

(a) State the general form of an optimal B-robust scale M-estimator. Simplify the formula in the case where f is the probability density function of a standard normal distribution.

(b) Define G_θ to be the distribution of the random variable $Y = \log X^2$, where $X \sim F_\sigma$ and $\theta = \log \sigma^2$. Show that $\{G_\theta : \theta > 0\}$ forms a location family.

Suppose T_n is a location M-estimator for θ based on observations $y_i = \log x_i^2$, defined in terms of a function $\psi : \mathbb{R} \to \mathbb{R}$ satisfying the Fisher consistency condition

$$\mathbb{E}_{y_i \sim G_\theta}[\psi(y_i - \theta)] = 0.$$

Further suppose ψ is differentiable, ψ' is bounded, and $\mathbb{E}_{y_i \sim G_0}[\psi^2(y_i)] < \infty$.

(c) Derive a result regarding asymptotic normality of (a recentered, rescaled version of) T_n when $x_i \sim F_\sigma$.

(d) Derive a result regarding the asymptotic distribution of $\exp\left(\frac{T_n}{2}\right)$. [Hint: Recall the Delta Method, a fact you may use without proof: If $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} Z$, then $\sqrt{n}(F(\hat{\theta}_n) - F(\theta_0)) \xrightarrow{d} F'(\theta_0)Z$, for a continuously differentiable function F.]

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

2 Breakdown point of trimmed estimator

Let $\rho : \mathbb{R} \to \mathbb{R}$ be a symmetric loss function such that $\rho(|t|)$ is increasing in $|t|$ and $\sup_t \rho(t) = \infty$. For a vector $v \in \mathbb{R}^n$, we write $\rho(v)$ to denote the vector in $\mathbb{R}^n$ with i^{th} component equal to $\rho(v_i)$. For an integer $h \leqslant n$ and a regularization parameter $\lambda > 0$, define the linear regression estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{h} \left(\rho(y - X\theta)\right)_{(i)} + \lambda \|\theta\|_1 \right\}, \tag{1}$$

where $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the vector of responses, $X = \{x_{ij}\} \in \mathbb{R}^{n \times p}$ contains the matrix of predictors $\{x_i\}_{i=1}^n$, and the sum is taken over the smallest h order statistics of the loss applied to the residuals. Define the breakdown point

$$\epsilon^*(y, X, \widehat{\theta}) = \frac{1}{n} \cdot \max_{m \geqslant 0} \left\{ m : \sup_{Z' \in \mathcal{Z}_m} \|\widehat{\theta}(Z') - \widehat{\theta}(Z)\|_2 < \infty \right\},$$

where we write $Z = (X, y)$ and denote by $\mathcal{Z}_m$ the set of data sets where we arbitrarily change at most m points of Z (allowing changes in both x_i and y_i).

(a) Let $\mathcal{L}_{Z'}$ denote the objective function defining the estimator (1), computed on a data set $Z' \in \mathcal{Z}_{n-h}$. Derive the upper bound

$$\mathcal{L}_{Z'}(0) \leqslant h\rho(M_y),$$

where $M_y := \max_{1 \leqslant i \leqslant n} |y_i|$ is the maximum absolute response computed over Z.

(b) Conclude that the breakdown point satisfies $\epsilon^*(y, X, \widehat{\theta}) \geqslant \frac{n-h}{n}$. [Hint: Observe that $\lambda \|\widehat{\theta}(Z')\|_1 \leqslant \mathcal{L}_{Z'}(\widehat{\theta}(Z'))$ and deduce a bound on $\|\widehat{\theta}(Z')\|_2$.]

Now consider the data set $Z_{\gamma,\tau} \in \mathcal{Z}_{n-h+1}$ obtained by moving the last $n-h+1$ observations in Z to $(x_0, y_0) = \big((\tau, 0, \ldots, 0)^T, \gamma\tau\big)$, where $\tau, \gamma > 0$ are parameters to be specified.

(c) Define $\theta_\gamma = (\gamma, 0, \ldots, 0)^T$. Derive the upper bound

$$\mathcal{L}_{Z_{\gamma,\tau}}(\theta_\gamma) \leqslant h\rho\left(M_y + \gamma \max_{1 \leqslant i \leqslant n} |x_{i1}|\right) + \lambda\gamma.$$

(d) If $\theta = (\theta_1, \ldots, \theta_p)^T$ with $\|\theta\|_2 \leqslant \gamma - 1$, derive the lower bound

$$\mathcal{L}_{Z_{\gamma,\tau}}(\theta) \geqslant \rho(\tau).$$

[Hint: Observe that the computation of the objective function in the estimator (1) must include at least one (x_0, y_0).]

(e) Conclude that the breakdown point also satisfies $\epsilon^*(y, X, \widehat{\theta}) \leqslant \frac{n-h}{n}$. [Hint: Suppose $\sup_{\tau,\gamma} \|\widehat{\theta}(Z_{\gamma,\tau})\|_2 \leqslant M$. Choose γ and τ appropriately and use parts (c) and (d) to obtain a contradiction.]

3 Two views of robust hypothesis testing

(a) State a general formula for the influence function $IF(x; T, F)$, when F is a distribution on $\mathbb{R}$ and T corresponds to the location M-estimator associated to $\psi : \mathbb{R} \to \mathbb{R}$. [You do not need to rigorously justify the formula, and may assume that F and ψ satisfy the usual regularity conditions assumed in lecture.]

(b) Using the formula in part (a), derive an expression for $IF(x; T, F)$ when T is the Huber location M-estimator with parameter k and F is the distribution $N(\theta, 1)$.

Suppose we have two distributions $P_0 \neq P_1$ and we wish to test the hypotheses

$$H_0 : P \in \mathcal{P}_\epsilon(P_0) \quad \text{vs.} \quad H_1 : P \in \mathcal{P}_\epsilon(P_1),$$

based on n samples $x_i \overset{i.i.d.}{\sim} P$.

(c) Define what is meant by a *maximin test* at level $\alpha \in (0, 1)$.

(d) Suppose $P_0 = N(-\theta_0, 1)$ and $P_1 = N(\theta_0, 1)$, for some $\theta_0 > 0$. Derive the form of a maximin test, assuming ϵ is sufficiently small. [You should simplify the rejection rule as much as possible, but you need not explicitly compute any threshold or truncation parameters as functions of $(\theta_0, \epsilon, \alpha)$.]

(e) What is the influence function $IF_{test}(x; T, F)$ of the test in part (d) when $F = P_0$? Is IF_{test} bounded as a function of x?

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

4 Another median-of-means estimator

Consider the d-dimensional multivariate location estimator defined by

$$\widehat{\mu}(X) \in \arg\min_{\mu \in \mathbb{R}^d} \sup_{\|v\|_2 \leqslant 1} |MOM_k(Xv) - \mu^T v|, \tag{1}$$

where $X \in \mathbb{R}^{n \times d}$ is the data matrix containing n observations and $MOM_k : \mathbb{R}^n \to \mathbb{R}$ denotes the univariate median-of-means estimator, which computes the median of means taken over k blocks. For simplicity, we assume that n is a multiple of k.

(a) Show that if we replace $MOM_k(Xv)$ in the definition of the estimator (1) by $\mu_0^T v$, for a fixed $\mu_0 \in \mathbb{R}^d$, the unique minimizer is μ_0.

(b) Prove that for any data matrix $X \in \mathbb{R}^{n \times d}$ and any $\mu_0 \in \mathbb{R}^d$, we have

$$\|\widehat{\mu}(X) - \mu_0\|_2 \leqslant 2 \sup_{\|v\|_2 \leqslant 1} |MOM_k(Xv) - \mu_0^T v|.$$

[Hint: Use the triangle inequality and optimality of $\widehat{\mu}$.]

In the remainder of the question, you may use the following fact: Suppose $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^d$ with mean μ_0 and covariance Σ, such that $\mathbb{E}[\|X_i\|_2^2] < \infty$. For any $\alpha \in (0,1)$, there exists a constant c_α such that, for any $k \geqslant \frac{1}{\alpha}$, with probability at least $1 - \exp(-k/c_\alpha)$, there exist at least $(1-\alpha)k$ blocks B_j satisfying

$$\sup_{\|v\|_2 \leqslant 1} \left| \frac{1}{|B_j|} \sum_{i : X_i \in B_j} X_i^T v - \mu_0^T v \right| \leqslant c_\alpha \sqrt{\frac{\max\{\mathrm{tr}(\Sigma), \|\Sigma\|_2 k\}}{n}},$$

where c_α does not depend on (n, d), and where $\|\Sigma\|_2$ is the operator norm of Σ.

(c) Using part (a) and the fact above, prove that for a suitably defined constant c and sufficiently large k, with probability at least $1 - \exp(-k/c)$, we have

$$\|\widehat{\mu}(X) - \mu_0\|_2 \leqslant 2c \sqrt{\frac{\max\{\mathrm{tr}(\Sigma), \|\Sigma\|_2 k\}}{n}}.$$

(d) Now suppose ϵn of the data points are contaminated arbitrarily, i.e., up to ϵn rows of the data matrix X are arbitrarily replaced, to obtain an observation matrix Z. Prove that if $k \geqslant 4 \max\{n\epsilon, 1\}$, with probability at least $1 - \exp(-c'k)$, we have

$$\|\widehat{\mu}(Z) - \mu_0\|_2 \leqslant c'' \sqrt{\frac{\max\{\mathrm{tr}(\Sigma), \|\Sigma\|_2 k\}}{n}},$$

for constants (c', c'') that do not depend on (n, k, d, ϵ). [Hint: Note that the contaminated data can affect the means of at most ϵn blocks.]

(e) Compare the error bound for $\widehat{\mu}$ obtained in part (d) with the error guarantee for the Tukey median, as functions of ϵ and d, in the case when $X_i \overset{i.i.d.}{\sim} N(\mu_0, I_d)$.

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

END OF PAPER