UNIVERSITY OF CAMBRIDGE

MAMA/205, NST3AS/205, MAAS/205

MAT3

MATHEMATICAL TRIPOS Part III

Thursday, 1 June, 2023 1:30 pm to 4:30 pm

PAPER 205

MODERN STATISTICAL METHODS

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

Attempt no more than **FOUR** questions.
There are **SIX** 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 Let $\mathcal{X}$ be a non-empty set. What is a *positive definite kernel*? In the following, we refer to a positive definite kernel simply as a kernel.

(a) (i) Write down the Gaussian kernel with bandwidth parameter $\sigma^2 > 0$. [You need not show it is a kernel.]

(ii) Suppose $k_\tau : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel for each $\tau \in \mathbb{R}$ and that

$$k(x, y) := \int_{-\infty}^{\infty} k_\tau(x, y)\, d\tau$$

is finite whenever $x = y \in \mathcal{X}$. Show that for all $x, y \in \mathcal{X}$,

$$\int_{-\infty}^{\infty} |k_\tau(x, y)|\, d\tau < \infty,$$

and that k is a kernel.

(iii) Show that $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $k(x, y) := (\alpha + \|x - y\|_2^2)^{-1/2}$ is a kernel for each $\alpha > 0$.

(b) (i) Suppose $\hat{\phi} : \mathbb{R}^d \to [-M, M]$ for $M > 0$ is a random feature map and define $k(x, y) := \mathbb{E}[\hat{\phi}(x)\hat{\phi}(y)]$, for $x, y \in \mathbb{R}^d$. Show that k is a kernel.

(ii) Show that $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by $k(x, y) := \exp(-\lambda\|x - y\|_1)$ is a kernel for each $\lambda > 0$. [*Hint: Use the fact that if V is a standard Cauchy random variable, then* $\mathbb{E}\exp(itV) = \mathbb{E}\cos(tV) = e^{-|t|}$].

[*Throughout this question you may use any results or derivations from the course without proof.*]

2 Suppose we have m null hypotheses $H_1, \ldots, H_m$ with associated p-values $p_1, \ldots, p_m$. Let $I_0 \subseteq \{1, \ldots, m\}$ be the set of true nulls. What is the *family-wise error rate* FWER? Describe the *Bonferroni correction* and prove that it can be used to control the FWER.

Describe the *closed testing procedure*, introducing any other tests that are needed in order for it to work. Prove that the closed testing procedure controls the FWER.

Now let $w_1, \ldots, w_m$ be positive deterministic weights. Show that the procedure (A) that rejects H_i if and only if

$$\frac{p_i}{w_i} \leqslant \alpha \left(\sum_{i=1}^{m} w_i \right)^{-1}$$

controls the FWER at level α

Define $q_i := p_i / w_i$ and assume for simplicity that the q_i for $i = 1, \ldots, m$ are all distinct. Let $q_{(1)} < \cdots < q_{(m)}$ so (i) is the index of the ith smallest value among $q_1, \ldots, q_m$ (note for instance in the description below, $w_{(1)}$ refers to the weight corresponding to the smallest q_i). Prove that the multiple testing procedure (B) consisting of the following steps (starting with Step 1) controls the FWER.

Step i (for $i < m$): If $q_{(i)} \leqslant \alpha / \sum_{j=i}^{m} w_{(j)}$, reject $H_{(i)}$ and go to step $i+1$; otherwise accept $H_{(i)}, \ldots, H_{(m)}$ and stop.

Step m: If $q_{(m)} \leqslant \alpha / w_{(m)}$, reject $H_{(m)}$; otherwise accept $H_{(m)}$.

Explain carefully why procedure (B) is preferable to procedure (A).

3 Suppose data $(X, Y, Z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times p}$ are formed of i.i.d. observations $(x_i, y_i, z_i) \in \mathbb{R}\times\mathbb{R}\times\mathbb{R}^p$ for $i = 1, \ldots, n$. We wish to test the null hypothesis H_0: $x_1 \perp\!\!\!\perp y_1 \,|\, z_1$ using the test statistic

$$T := \sqrt{n}\frac{\tau_N}{\tau_D},$$

where

$$\tau_N := \frac{1}{n}\sum_{i=1}^{n}\{x_i - \hat{f}(z_i)\}\{y_i - \hat{g}(z_i)\}, \qquad \tau_D^2 := \frac{1}{n}\sum_{i=1}^{n}[\{x_i - \hat{f}(z_i)\}\{y_i - \hat{g}(z_i)\}]^2$$

and estimated regression functions $\hat{f}$ and $\hat{g}$ are formed through regressing each of X and Y on Z respectively. Let $\varepsilon_i := x_i - f(z_i)$ and $\xi_i := y_i - g(z_i)$ where $f(\cdot) = \mathbb{E}(x_1 \,|\, z_1 = \cdot)$ and $g(\cdot) = \mathbb{E}(y_1 \,|\, z_1 = \cdot)$. In all that follows, we assume that H_0 is true.

(a) Assume that for some $C > 0$, $\mathbb{E}(\varepsilon_1^2 \,|\, z_1) \leqslant C$ and $\mathbb{E}(\xi_1^2 \,|\, z_1) \leqslant C$. Show that $\mathbb{E}(\varepsilon_1^2\xi_1^2) \leqslant C^2$.

(b) Writing $F_i := f(z_i) - \hat{f}(z_i)$ and $G_i := g(z_i) - \hat{g}(z_i)$, further assume that $\mathbb{E}(\frac{1}{n}\sum_{i=1}^n F_i^2) \to 0$ and $\mathbb{E}(\frac{1}{n}\sum_{i=1}^n G_i^2) \to 0$ as $n \to \infty$. Show that

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 G_i^2 \xrightarrow{p} 0. \qquad (1)$$

Show further that

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i\varepsilon_i^2 G_i \xrightarrow{p} 0. \qquad (2)$$

(c) Now additionally assume

$$\frac{1}{n}\sum_{i=1}^{n}F_i^2 G_i^2 \xrightarrow{p} 0, \qquad (3)$$

and show that

$$\frac{1}{n}\sum_{i=1}^{n}F_i G_i \varepsilon_i \xi_i \xrightarrow{p} 0. \qquad (4)$$

Show further that

$$\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i F_i| G_i^2 \xrightarrow{p} 0. \qquad (5)$$

(d) Finally, assuming all of the above and additionally that $\sqrt{n}\tau_N \xrightarrow{d} N(0, \mathbb{E}(\varepsilon_1^2\xi_1^2))$, show carefully that under H_0 we have $T \xrightarrow{d} N(0, 1)$.

4 What does it mean for a random variable to be *sub-Gaussian* with parameter $\sigma > 0$?

Let $(U_1, V_1), \ldots, (U_n, V_n)$ be i.i.d. pairs of random variables with mean zero and $\mathrm{Var}(U_1) = \mathrm{Var}(V_1) = 1$. Suppose U_1 and V_1 are both sub-Gaussian with parameter $\sigma/4 > 0$. Stating any results from lectures that you need and writing $U = (U_1, \ldots, U_n)^T$ and similarly for V, show that for all $t \geqslant 0$,

$$\mathbb{P}(|U^T V/n - \mathbb{E}(U_1 V_1)| \geqslant t) \leqslant 2 \exp\left(-\frac{2nt^2}{\sigma^2(\sigma^2 + t)}\right).$$

Define, for an arbitrary symmetric positive semi-definite $\Sigma \in \mathbb{R}^{p \times p}$ and non-empty proper subset $S \subset \{1, \ldots, p\}$ with $s := |S|$,

$$\phi_\Sigma^2 := s \inf_{\substack{\delta \in \mathbb{R}^p : \|\delta_S\|_1 = 1, \\ \|\delta_{S^c}\|_1 \leqslant 3}} \delta^T \Sigma \delta.$$

Prove that if symmetric positive semi-definite $\Theta \in \mathbb{R}^{p \times p}$ has $\max_{jk} |\Sigma_{jk} - \Theta_{jk}| \leqslant \phi_\Sigma^2/(32s)$, then $\phi_\Theta^2 \geqslant \phi_\Sigma^2/2$.

Now let matrix $X \in \mathbb{R}^{n \times p}$ consist of i.i.d. rows each with variance matrix $\Sigma \in \mathbb{R}^{p \times p}$ where $\Sigma_{jj} = 1$ for all $j = 1, \ldots, p$. Further suppose that each entry of X is mean zero and sub-Gaussian with parameter $\sigma/4 > 0$. Let $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ have entries given by

$$\hat{\Sigma}_{jk} := \frac{X_j^T X_k}{\|X_j\|_2 \|X_k\|_2},$$

where $X_j \in \mathbb{R}^n$ is the jth column of X. Let $\Sigma \in \mathbb{R}^{p \times p}$ be the covariance matrix of a row of X. Let $t := \sigma^2 \sqrt{2 \log(p+1)/n}$ and suppose n and p are such that

$$t \leqslant \min\left(\frac{\sigma^2}{3}, \frac{\phi_\Sigma^2}{64s + \phi_\Sigma^2}\right).$$

Prove that

$$\mathbb{P}(\phi_{\hat{\Sigma}}^2 \geqslant \phi_\Sigma^2/2) \geqslant \frac{p}{p+1}.$$

5 Let $Y \in \mathbb{R}^n$ be a vector of responses and let $X \in \mathbb{R}^{n\times p}$ be a matrix of predictors where each column has been centred and has ℓ_2-norm $\sqrt{n}$.

(a) Write down the optimisation problem solved by the *ridge regression estimator* $(\hat{\mu}, \hat{\beta}) \in \mathbb{R} \times \mathbb{R}^p$ with tuning parameter $\lambda > 0$. Show that $\hat{\mu} = \bar{Y} := \sum_{i=1}^n Y_i/n$ and $\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY = X^T(XX^T + \lambda I)^{-1}Y$.

(b) Prove that if $A \subseteq \{1, \ldots, p\}$ is non-empty, then for each $j \in A$,

$$X_j^T(X_AX_A^T + \lambda I)^{-1}X_j < 1.$$

(c) Consider the following algorithm for producing a sequence of variable indices $j_1, \ldots, j_p$. We initialise $A_1 = \{1, \ldots, p\}$ and then repeat for $k = 1, \ldots, p$:

1. Perform ridge regression but enforcing that all coefficients whose indices are not in A_k are set to 0. This gives estimate $\hat{\beta}^{(k)} \in \mathbb{R}^p$ with $\hat{\beta}_j^{(k)} = 0$ for $j \notin A_k$.

2. Set $j_k := \arg\min_{j\in A_k} |\hat{\beta}_j^{(k)}|$ and update $A_{k+1} = A_k \setminus \{j_k\}$.

Throughout we fix the ridge regression parameter $\lambda > 0$ and in step 2 above, we assume the minimiser is unique. Assume that the computational complexity of inverting $M \in \mathbb{R}^{m\times m}$ is $O(m^3)$, and forming BC where $B \in \mathbb{R}^{a\times b}$ and $C \in \mathbb{R}^{b\times c}$ is $O(abc)$. Show that in the case where $p \geqslant n$, the computational complexity of the algorithm above can be made to be $O(p^2n)$.

[*Hint: If $M \in \mathbb{R}^{m\times m}$ is non-singular and $b \in \mathbb{R}^m$ satisfies $b^TM^{-1}b \neq 1$, then*

$$(M - bb^T)^{-1} = M^{-1} + \frac{M^{-1}bb^TM^{-1}}{1 - b^TM^{-1}b}.$$

]

6 Let $Y \in \mathbb{R}^n$ be a vector of responses and $X \in \mathbb{R}^{n \times p}$ a matrix of predictors. Suppose that the columns of X have been centred and scaled to have ℓ_2-norm $\sqrt{n}$, and that Y is also centred. Consider the linear model (after centring),

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1},$$

where $\mathbf{1}$ is an n-vector of 1's and $\bar{\varepsilon} := \mathbf{1}^T\varepsilon/n$. Let $S := \{j : \beta^0_j \neq 0\}$, $s := |S| \in [1, p-1]$ and $N := \{1, \ldots, p\} \setminus S$. Define the *Lasso estimator* $\hat{\beta}$ of β^0 with regularisation parameter $\lambda > 0$ (here and throughout we suppress the dependence of the Lasso solution on λ).

Suppose $\varepsilon_1, \ldots, \varepsilon_n$ are independent, mean-zero and sub-Gaussian with parameter $\sigma = 1$. Set $\lambda = A\sqrt{\log p/n}$ for $A > 0$. Prove that

$$\mathbb{P}(2\|X^T\varepsilon\|_\infty/n \leqslant \lambda) \geqslant 1 - 2p^{-(A^2/8-1)}.$$

[You may use standard results about sub-Gaussian random variables without proof.]

Write down the KKT conditions for the Lasso.

Suppose $\hat{\Sigma} := X^T X/n$ has the following property: there exists $\psi > 0$ such that for all $\delta \in \mathbb{R}^p$ with $\|\delta_N\|_1 \leqslant 3\|\delta_S\|_1$,

$$\psi\|\delta_S\|_\infty \leqslant \|\hat{\Sigma}\delta\|_\infty.$$

Prove that on an event with probability at least $1 - 2p^{-(A^2/8-1)}$, the following hold:

(a) If $\min_{j \in S} |\beta^0_j| > \frac{3A}{2\psi}\sqrt{\log(p)/n}$ then $\operatorname{sgn}(\hat{\beta}_S) = \operatorname{sgn}(\beta^0_S)$.

(b) $\|\hat{\beta} - \beta^0\|_1 \leqslant \frac{6sA}{\psi}\sqrt{\log(p)/n}$.

END OF PAPER