MAMA/205, MGM3/205, NST3AS/205, MAAS/205

MAT3

MATHEMATICAL TRIPOS Part III

Thursday 5 June 2025 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 **FIVE** 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 (a) For $f : \mathbb{R}^d \to \mathbb{R}$ convex, define the *subdifferential* $\partial f(x)$ of f at a point $x \in \mathbb{R}^d$.

Let $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times mK}$, and consider the Group Lasso objective,

$$Q: \ \mathbb{R}^{mK} \to \mathbb{R}$$
$$Q(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda \sum_{k=1}^{K} \sqrt{m}\|\beta^{(k)}\|_2$$

where for any vector $\beta \in \mathbb{R}^{mK}$, we define $\beta^{(k)} \in \mathbb{R}^{mK}$ as the vector with entries $\beta_i^{(k)} := \mathbb{1}_{\{(k-1)m < i \leqslant km\}}\, \beta_i$.

(b) Derive the subdifferential of the function $f_k : \beta \mapsto \|\beta^{(k)}\|_2$ from the definition.

(c) Write down the subdifferential of Q at β. [You may use any result from lectures, provided it is clearly stated].

(d) Let $\hat{\beta}$ be a minimiser of Q over $\beta \in \mathbb{R}^{mK}$. Prove that the fitted values $X\hat{\beta}$ are unique.

(e) Let $\hat{\nu} = X^T(Y - X\hat{\beta})$, and let $E = \{k \in \{1, \ldots, K\} : \|\hat{\nu}^{(k)}\|_2 = n\lambda\sqrt{m}\}$. Is the set E unique? Show that if the matrix $\tilde{X}$ with columns $\{X_i : (k-1)m < i \leqslant km, k \in E\}$ has column rank $m|E|$, then $\hat{\beta}$ is the unique minimiser of Q over $\mathbb{R}^{mK}$.

2 Consider the linear model $Y = X\beta^0 + \varepsilon$, with $X \in \mathbb{R}^{n\times p}$ and $\beta^0 \in \mathbb{R}^p$. We assume that $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ has entries which are independent, mean 0, and sub-Gaussian with parameter σ. Let $\hat{\beta}$ be a minimiser over $\beta \in \mathbb{R}^p$ of the function

$$Q(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

(a) Prove the basic inequality

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leqslant \frac{1}{n}\varepsilon^T X(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1.$$

Now suppose that the design matrix X is random, with i.i.d. rows distributed as $N_p(0, \Sigma^0)$ where $|\Sigma^0_{j,j}| \leqslant v^2$ for each $j = 1, \ldots, p$. Assume that X is independent of ε. Let $\lambda = A\sigma v\sqrt{\log(p)/n}$ for some constant A, and $n > \log p$.

(b) Show that, for a choice of A which you must specify, the event

$$\Omega_1 := \left\{ \frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leqslant 2A\sigma v\sqrt{\frac{\log(p)}{n}}\min(\|\beta^0\|_1, \|\hat{\beta} - \beta^0\|_1) \right\}$$

has $\mathbb{P}(\Omega_1) \to 1$ as $p \to \infty$. [You may quote properties of sub-Gaussian random variables and basic concentration inequalities without proof.]

(c) Define the event

$$\Omega_2 = \left\{ \left\| \frac{1}{n}X^T X - \Sigma^0 \right\|_\infty \leqslant \frac{\mu}{2(\|\hat{\beta}\|_0 + \|\beta^0\|_0)} \right\}$$

where μ is the smallest eigenvalue of Σ^0 and $\|z\|_0 := |\{j : z_j \neq 0\}|$ denotes the number of non-zero entries in a vector z. Show that on $\Omega_1 \cap \Omega_2$, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leqslant 8A^2\sigma^2 v^2 \frac{(\|\hat{\beta}\|_0 + \|\beta^0\|_0)}{\mu}\frac{\log p}{n}.$$

3 (a) Define a *positive definite kernel*. Define a *reproducing kernel Hilbert space* and its *reproducing kernel*.

Consider the function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, given by

$$k(x, y) = \frac{1}{e^{x-y} + e^{y-x}}.$$

(b) Show that, for random variables W and U which you must specify, and some constant $c > 0$,

$$k(x, y) = c\,\mathbb{E}[\cos(Wx + U)\cos(Wy + U)] \quad \text{for all } x, y \in \mathbb{R}.$$

[*Hint:* $\int_{-\infty}^{\infty} e^{-i2\pi z\xi} \frac{2}{e^{z}+e^{-z}} dz = \frac{2\pi}{e^{\pi^2\xi}+e^{-\pi^2\xi}}$ *for all* $\xi \in \mathbb{R}$].

(c) Prove that k is a positive definite kernel.

(d) Let $(W_i, U_i)_{i=1}^{\ell}$ be i.i.d. copies of the pair (W, U) defined in part (b). Let $\phi(x) = \sqrt{c/\ell}(\cos(W_1 x + U_1), \ldots, \cos(W_\ell x + U_\ell))^T$. Show that there are positive constants C_1 and C_2, such that for all $0 < \varepsilon \leqslant 1$,

$$\mathbb{P}\left(\sup_{x,y\in[-L,L]} |k(x, y) - \phi(x)^T\phi(y)| \geqslant \varepsilon \right) \leqslant \frac{C_1 L^2}{\varepsilon^2} \exp\left(-C_2 \ell \varepsilon^2\right)$$

[*Hint: Approximate the supremum by the supremum over* x, y *in a grid of evenly spaced points in* $[-L, L]$. *You may assume* $|\partial k(x, y)/\partial x| < 1$.]

4 (a) What does it mean to say that p_i is a *p-value* for the null hypothesis H_i? Define the *Benjamini–Hochberg* multiple testing procedure with parameter α for a family of null hypotheses $H_1, \ldots, H_m$ with p-values $p_1, \ldots, p_m$.

Throughout this problem, assume that $p_1, \ldots, p_m$ are independent.

(b) Show that the Benjamini–Hochberg procedure has false discovery rate less than or equal to α.

(c) Suppose that under the hypothesis H_i, p_i has a Uniform$(0, 1)$ distribution, for $i = 1, \ldots, m$. Show that, under the intersection hypothesis $\cap_{i=1}^{m} H_i$, the Benjamini–Hochberg procedure has familywise error rate α. [*Hint: If* $X_1 \leqslant X_2 \leqslant \ldots \leqslant X_n$ *are order statistics of* n *i.i.d. Uniform*$(0, 1)$ *random variables, then for any* $m \leqslant n$, $(\frac{X_1}{X_m}, \frac{X_2}{X_m}, \ldots, \frac{X_{m-1}}{X_m})$ *is equal in distribution to the order statistics of* $m - 1$ *i.i.d. Uniform*$(0, 1)$ *random variables and independent of* X_m.]

5 Consider a model $Y = X\beta^0 + \varepsilon$, where $\beta^0 \in \mathbb{R}^p$, $X \in \mathbb{R}^{n\times p}$, and $\varepsilon \sim N_n(0, \sigma^2 I)$.

(a) Define the *ridge regression estimator* $\hat{\beta}_\lambda$ of β^0, and show that it is equal to $(X^TX + \lambda I)^{-1}X^TY$.

(b) For any matrix A with thin singular value decomposition $A = UDV^T$, write $A^+ = UD^+V^T$ where D^+ is the diagonal matrix with

$$D^+_{ii} = \begin{cases} D^{-1}_{ii} & \text{if } D_{ii} \neq 0 \\ 0 & \text{if } D_{ii} = 0. \end{cases}$$

The *ridgeless* estimator is defined as $\lim_{\lambda\to 0}\hat{\beta}_\lambda$. Show that it is equal to $(X^TX)^+X^TY$.

(c) Let $W \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. $N(0, \lambda/d)$ entries. Let $\tilde{\beta}$ be the ridgeless estimator fit to the response vector Y, with design matrix $[X, W] \in \mathbb{R}^{n\times(p+d)}$. Let $\tilde{\beta}_{1:p}$ denote the first p entries of $\tilde{\beta}$. Show that

$$\tilde{\beta}_{1:p} \stackrel{\text{a.s.}}{\to} \hat{\beta}_\lambda \quad \text{as } d \to \infty.$$

(d) Consider a model with random design matrix X in which the rows $(x_1, \ldots, x_n)$ are i.i.d. $N_p(0, I)$. Let $x^* \sim N_p(0, I)$ be independent from the training data (X, Y). Show that

$$R_X(\hat{\beta}_\lambda) := \mathbb{E}((x^{*T}\beta^0 - x^{*T}\hat{\beta}_\lambda)^2 \mid X) = \ell^2(\beta^0)^T(\hat{\Sigma} + \ell I)^{-2}\beta^0 + \frac{\sigma^2}{n}\mathrm{tr}(\hat{\Sigma}(\hat{\Sigma} + \ell I)^{-2}),$$

where $\hat{\Sigma} = X^TX/n$ and $\lambda = n\ell$.

(e) Consider an asymptotic regime where $p/n \to \gamma \in (0, \infty)$ as $n, p \to \infty$, whilst $\|\beta^0\|_2 = r$ and $\lambda = \ell n$ for constants r, ℓ. Let $(A_p), (B_p)$ be sequences of matrices in $\mathbb{R}^{p\times p}$; we write $A_p \asymp B_p$ if $\mathrm{tr}(\Theta_p(A_p - B_p)) \to 0$ as $p \to \infty$ for every sequence of positive definite matrices (Θ_p) with $\mathrm{tr}(\Theta_p) \leqslant 1$ for all p. We are told that there is a differentiable function $m : (0, \infty] \to \mathbb{R}$, such that, as $n, p \to \infty$, a.s.

$$(\hat{\Sigma} + \ell I)^{-1} \asymp m(\ell)I \qquad \text{and} \qquad (\hat{\Sigma} + \ell I)^{-2} \asymp -m'(\ell)I.$$

Show that

$$R_X(\hat{\beta}_\lambda) \stackrel{\text{a.s.}}{\to} \ell^2 r^2 m'(\ell) + \sigma^2\gamma\left(m(\ell) - \ell m'(\ell)\right).$$

END OF PAPER