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

MAT3

MATHEMATICAL TRIPOS Part III

Thursday 30 May 2024 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) Define a *positive definite kernel* and a *Reproducing Kernel Hilbert Space.*

(b) Consider the Gaussian kernel on $\mathbb{R}^d$, $g_\sigma(x,y) = e^{-\|x-y\|^2/(2\sigma^2)}$. Prove that there is no feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ with finite p, such that $g_\sigma(x,y) = \phi(x)^T\phi(y)$ for all $x, y \in \mathbb{R}^d$. [*Hint: Consider the rank of the kernel matrix. A matrix*

$$\begin{bmatrix} 1 & r^{1^2} & r^{2^2} & \dots & r^{n^2} \\ r^{1^2} & 1 & r^{1^2} & \dots & r^{(n-1)^2} \\ r^{2^2} & r^{1^2} & 1 & \dots & r^{(n-2)^2} \\ & & \vdots & & \\ r^{n^2} & r^{(n-1)^2} & r^{(n-2)^2} & \dots & 1 \end{bmatrix}$$

with r small enough is strictly diagonally dominant, hence positive definite.]

(c) Show that the function $k_1(x,y) = (x+y)^{-1}$ is a positive definite kernel on $(0,\infty)$.

(d) Show that the function

$$k_2(x,y) = \frac{1}{1 - \frac{x^T y}{\|x\|^2 + \|y\|^2}} - 1$$

for $x, y \neq 0$ is a positive definite kernel on the set $\mathbb{R}^d \setminus \{0\}$, citing any necessary result from lectures. You may find the result from part (c) useful.

Deduce that $d(x,y) = \sqrt{2 - 2k_2(x,y)}$ is a metric on this set. [*Hint: Consider a metric in the RKHS.*]

2 Let $x_1, \ldots, x_n$ be i.i.d. $N(\mu, \Sigma)$ random variables with parameters $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$. The Graphical Lasso estimator $\hat{\Omega}$ for the precision matrix Σ^{-1} is the minimiser over positive definite matrices Ω of the objective

$$Q(\Omega) = -\log\det\Omega + \mathrm{Tr}(S\Omega) + \lambda \sum_{i,j=1}^{p} |\Omega_{i,j}|.$$

(a) What is the matrix S in the objective?

(b) Derive the Karush–Kuhn–Tucker conditions for the estimator $\hat{\Omega}$, citing any necessary result from lectures.

(c) Prove that $Q(\hat{\Omega}) = -\log\det\hat{\Omega} + p$.

(d) Suppose that there is some $p_0 < p$, such that for all $i \leqslant p_0$, $j > p_0$, we have $|S_{ij}| \leqslant \lambda$. Write S blockwise:

$$S = \begin{bmatrix} S^{(11)} & S^{(12)} \\ S^{(21)} & S^{(22)} \end{bmatrix},$$

where $S^{(11)} \in \mathbb{R}^{p_0 \times p_0}$; let $d_1 = p_0$ and $d_2 = p - p_0$.

Show that the Graphical Lasso estimator with parameter λ is

$$\hat{\Omega} = \begin{bmatrix} \hat{\Omega}^{(1)} & 0 \\ 0 & \hat{\Omega}^{(2)} \end{bmatrix},$$

where, for $k \in \{1, 2\}$, $\hat{\Omega}^{(k)}$ minimises the objective

$$-\log\det W + \mathrm{Tr}(S^{(kk)}W) + \lambda \sum_{i,j=1}^{d_k} |W_{i,j}|$$

over positive definite matrices $W \in \mathbb{R}^{d_k \times d_k}$.

3 Consider a linear model $Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1}$, where the design matrix $X \in \mathbb{R}^{n\times p}$ has centred columns with ℓ_2-norm $\sqrt{n}$. We assume that $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with $\mathbb{P}(\varepsilon_1 = 1) = \mathbb{P}(\varepsilon_1 = -1) = 1/2$. Let $S := \{j \in \{1, \ldots, p\} : \beta_j^0 \neq 0\}$ be the set of non-zero entries in β^0, and $N := \{1, \ldots, p\} \setminus S$ be its complement.

(a) Letting $\hat{\beta}$ be the Lasso estimator with regularisation parameter λ, prove the *basic inequality*:

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

(b) Show that, on the event $\Omega := \{\|X^T\varepsilon\|_\infty/n \leqslant \lambda/2\}$, we have

$$\|\hat{\beta}_N - \beta_N^0\|_1 \leqslant 3\|\hat{\beta}_S - \beta_S^0\|_1.$$

(c) Let $\lambda = A\sqrt{\log(p)/n}$ for some constant $A > 0$. Prove a lower bound on $\mathbb{P}(\Omega)$ in terms of p and A, giving conditions on A under which $\mathbb{P}(\Omega) \to 1$ as $p \to \infty$. [You may cite Hoeffding's lemma and basic properties of sub-Gaussian random variables without proof.]

(d) Now suppose that X satisfies the *γ-restricted eigenvalue condition*:

$$\inf_{\delta\in\mathbb{R}^p:\delta\neq 0, \|\delta_N\|_1\leqslant 3\|\delta_S\|_1} \frac{\frac{1}{n}\delta^T X^T X\delta}{\|\delta\|_2^2} > \gamma,$$

for some constant $\gamma > 0$. Prove that, on the event Ω,

$$\|\hat{\beta} - \beta^0\|_2 \leqslant \frac{3\lambda\sqrt{|S|}}{2\gamma}.$$

4 Consider the linear model

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

where the design matrix $X \in \mathbb{R}^{n\times p}$ has centred columns with ℓ_2-norm $\sqrt{n}$ and compatibility factor $\phi^2 > 0$, and suppose that $\varepsilon \sim N_n(0, \sigma^2 I)$.

(a) Define the square-root Lasso estimator with regularisation parameter γ.

(b) Define an approximate test of size α for the null hypothesis $H_j^\beta : \beta_j^0 = 0$. State without proof conditions on γ, n, p, and $s := |\{j : \beta_j^0 \neq 0\}|$ under which the size tends to α as $n \to \infty$.

Suppose we observe a second response vector,

$$Z = X\eta^0 + \xi - \bar{\xi}\mathbf{1},$$

where $\xi \sim N_n(0, v^2 I)$. We wish to estimate the set of coefficients which are non-zero for both β^0 and η^0:

$$B := \{j : \beta_j^0 \neq 0\} \cap \{j : \eta_j^0 \neq 0\}.$$

Let q_j^β be the p-value for a test of the null hypothesis $H_j^\beta : \beta_j^0 = 0$. Let q_j^η be the p-value in a similar test for the null hypothesis $H_j^\eta : \eta_j^0 = 0$. In the following parts, you may assume that these tests have exact size α. Define

$$q_j = \max\{q_j^\beta, q_j^\eta\} \quad \text{for } j = 1, \ldots, p.$$

(c) Show that the test which rejects the null hypothesis $H_0 : B = \emptyset$ when $\min_{j=1}^p q_j \leqslant \alpha/p$ has size at most α.

(d) Let τ be a permutation of $\{1, \ldots, p\}$ such that $q_{\tau(1)} \leqslant q_{\tau(2)} \leqslant \ldots \leqslant q_{\tau(p)}$. We estimate B with the set $\hat{B} = \{\tau(1), \ldots, \tau(k)\}$, where

$$k := \min\left\{1 \leqslant j \leqslant p : q_{\tau(j)} > \frac{\alpha}{(p - j + 1)}\right\} - 1,$$

and $\hat{B} = \emptyset$ if $k = 0$. Prove that the probability that $\hat{B}$ contains an element which is not in B is smaller or equal to α.

5 Consider the model $Y_i = f^0(x_i) + \varepsilon_i$ where $x_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$, $\varepsilon \sim N_n(0, \sigma^2 I)$, and f^0 lies in a reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel k.

(a) Define the *kernel ridge regression estimator* $\hat{f}_\lambda$ and, applying the representer theorem, derive an expression for the fitted values $\hat{Y} = (\hat{f}_\lambda(x_1), \ldots, \hat{f}_\lambda(x_n))^T$.

(b) Consider the leave-one-out cross-validation estimator

$$\hat{T}_\lambda = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}_{\lambda,-i}(x_i))^2,$$

where $\hat{f}_{\lambda,-i}$ denotes the ridge regression estimator fitted to all the samples except (x_i, Y_i). We wish to compute $\hat{T}_\lambda$ for all values of λ in the set $\{\lambda_1, \ldots, \lambda_L\}$. Describe an algorithm to do this in $O(n^3 + Ln^2)$ iterations. [*Hint: Recall the formula for blockwise matrix inversion:*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \;].$$

END OF PAPER