UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 6 June, 2022 9:00 am to 12:00 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 $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of functions on an input space $\mathcal{X}$ with reproducing kernel k. Let $(x_i, Y_i)_{i=1}^n$ be i.i.d. and satisfy

$$Y_i = f^0(x_i) + \varepsilon_i.$$

Here, $f^0 \in \mathcal{H}$ with $\|f^0\|_{\mathcal{H}} \leqslant 1$, and writing $X \in \mathcal{X}^n$ for the collection $x_1, \ldots, x_n$, we have that $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)$ satisfies $\mathbb{E}(\varepsilon \mid X) = 0$ and $\mathbb{E}(\varepsilon\varepsilon^T \mid X) = \sigma^2 I$. For a tuning parameter value $\lambda > 0$, write down the objective function minimised over $f \in \mathcal{H}$ to produce the kernel ridge regression estimate $\hat{f}_\lambda \in \mathcal{H}$.

Let $K \in \mathbb{R}^{n \times n}$ be the matrix with ijth entry $K_{ij} = k(x_i, x_j)$ and let the eigenvalues of K/n be given by $\hat{\mu}_1 \geqslant \hat{\mu}_2 \geqslant \cdots \geqslant \hat{\mu}_n$ (we assume $\hat{\mu}_1 > 0$). Write down a closed form expression for $(\hat{f}_\lambda(x_i))_{i=1}^n \in \mathbb{R}^n$ involving K, λ and $Y := (Y_1, \ldots, Y_n)^T$.

Show that $(f^0(x_i))_{i=1}^n = K\alpha$ for some $\alpha \in \mathbb{R}^n$, and moreover that $\|f^0\|_{\mathcal{H}}^2 \geqslant \alpha^T K \alpha$.

Using the fact (which you need not prove) that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big([f^0(x_i) - \mathbb{E}\{\hat{f}_\lambda(x_i) \mid X\}]^2 \mid X\big) \leqslant \frac{\lambda}{4n},$$

show that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\{f^0(x_i) - \hat{f}_\lambda(x_i)\}^2 \mid X] \leqslant \frac{\sigma^2}{\lambda}\sum_{i=1}^n \min(\hat{\mu}_i/4, \lambda/n) + \frac{\lambda}{4n}. \qquad (*)$$

Assume that there exists a non-negative sequence $\mu_1, \mu_2, \ldots$ be such that $\sum_{j=1}^\infty \mu_j < \infty$ and

$$\mathbb{E}\left(\sum_{i=1}^n \min(\hat{\mu}_i/4, \gamma)\right) \leqslant \sum_{j=1}^\infty \min(\mu_j/4, \gamma)$$

for all $\gamma > 0$. Now let $\hat{\lambda}$ minimise the r.h.s. of (*) over $\lambda > 0$. Show that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\{f^0(x_i) - \hat{f}_{\hat{\lambda}}(x_i)\}^2] \leqslant \inf_{\gamma > 0}\left\{\frac{\sigma^2}{n\gamma}\sum_{j=1}^\infty \min(\mu_j/4, \gamma) + \frac{\gamma}{4}\right\}.$$

2 Suppose we have null hypotheses $H_1, \ldots, H_m$ and associated p-values $p_1, \ldots, p_m$. What is the *false discovery rate* (FDR)?

In all that follows, we will assume that with probability 1, the p-values $p_1, \ldots, p_m$ are distinct. Describe the *Benjamini–Hochberg (BH) procedure*.

Suppose I_0 is the set of indices of true null hypotheses and $m_0 = |I_0|$. Let $p_{-i} \in \mathbb{R}^{m-1}$ be the vector of p-values with the ith p-value removed. Consider the following condition.

A: For each $i \in I_0$, p_i is independent of p_{-i}.

By considering for each $i \in I_0$ a modified BH procedure applied to p_{-i} with R_i rejections, prove that the BH procedure controls the FDR at a given level α when Assumption A holds.

We say a set $D \in [0,1]^d$ is 'increasing' if whenever $x \in D$ and $y \in [0,1]^d$ is such that $y_i \geqslant x_i$ for all $i = 1, \ldots, d$, then $y \in D$. Explain why the set of p-values in $[0,1]^{m-1}$ resulting in at most $r-1$ rejections from your modified BH procedure is an increasing set (i.e. why $\{R_i \leqslant r-1\} = \{p_{-i} \in D\}$ for some increasing set D).

We no longer assume Assumption A, but instead assume Assumption B below.

B: For each $i \in I_0$ and any increasing set $D \in [0,1]^{m-1}$, $\mathbb{P}(p_{-i} \in D \,|\, p_i \leqslant x)$ is non-decreasing in $x \in [0,1]$.

Prove that the BH procedure controls the FDR at a given level α when Assumption B holds. [*Hint: Aim to use Assumption B to obtain a telescoping sum.*]

3 Suppose data $(X, Y, Z) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n \times p}$ is 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: $x_1 \perp\!\!\!\perp y_1 \,|\, z_1$. Show that under the null,

$$\mathbb{E}[\{x_1 - \mathbb{E}(x_1 \,|\, z_1)\}\{y_1 - \mathbb{E}(y_1 \,|\, z_1)\}s(z_1)] = 0$$

where $s : \mathbb{R}^p \to \{-1, 1\}$.

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 $g(\cdot) = \mathbb{E}(y_1 \,|\, z_1 = \cdot)$. Suppose we have an estimator τ_D of $\sqrt{\mathrm{Var}(\varepsilon_1 \xi_1)}$ such that $\tau_D \xrightarrow{p} \sqrt{\mathrm{Var}(\varepsilon_1 \xi_1)}$, and estimated regression functions $\hat{f}$ and $\hat{g}$ formed through regressing each of X and Y on Z respectively. Let

$$\tau_N := \frac{1}{n} \sum_{i=1}^{n} \{x_i - \hat{f}(z_i)\}\{y_i - \hat{g}(z_i)\}s(z_i).$$

Show that under H_0 and conditions (i)–(iii) below, test statistic $T := \sqrt{n}\tau_N / \tau_D$ has the property that $T \xrightarrow{d} N(0, 1)$.

(i) We have $0 < \mathrm{Var}(\varepsilon_1 \xi_1) < \infty$.

(ii) We have that $\mathrm{Var}(\varepsilon_1 \,|\, z_1) \leqslant c$ and $\mathrm{Var}(\xi_1 \,|\, z_1) \leqslant c$ for some $c > 0$.

(iii) Writing

$$\mathrm{MSPE}_f := \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\{f(z_i) - \hat{f}(z_i)\}^2\right) \quad \text{and} \quad \mathrm{MSPE}_g := \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\{g(z_i) - \hat{g}(z_i)\}^2\right),$$

we have $\mathrm{MSPE}_f \to 0$, $\mathrm{MSPE}_g \to 0$ and $n\mathrm{MSPE}_f\mathrm{MSPE}_g \to 0$.

Now show that even when H_0 is not true, we have

$$\mathbb{E}[\{x_1 - \mathbb{E}(x_1 \,|\, z_1)\}\{y_1 - \mathbb{E}(y_1 \,|\, z_1)\}s(z_1)] = \mathbb{E}[x_1\{\mathbb{E}(y_1 \,|\, x_1, z_1) - \mathbb{E}(y_1 \,|\, z_1)\}s(z_1)].$$

Consider an alternative (i.e. where H_0 does not hold) where x_1 and z_1 are independent, $\mathbb{E}x_1 = 0$, $\mathbb{E}x_1^2 > 0$ and

$$\mathbb{E}(y_1 \,|\, x_1, z_1) = x_1 h(z_1) + g(z_1)$$

for some $h : \mathbb{R}^p \to \mathbb{R}$ with $\mathbb{E}h(z_1) = 0$ and $\mathbb{E}|h(z_1)| > 0$. Explain why when s is the constant function always taking the value 1, the test corresponding to T is not expected to have power against this alternative. Give a choice of s (depending on h) such that we can expect the resulting test will have power against this alternative.

4 Consider a regression setting with response vector $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$ related via $Y = X\beta^0 + \varepsilon$ where $\varepsilon \sim N_n(0, I)$. Suppose non-empty groups $G_1, \ldots, G_q$ partition $\{1, \ldots, p\}$. Write down the group Lasso penalty $\lambda P(\beta)$ with tuning parameter $\lambda > 0$ and group multipliers $m_1, \ldots, m_q > 0$.

Show that for $u, v \in \mathbb{R}^p$,

$$|u^T v| \leqslant \max_{k=1,\ldots,q} (m_k^{-1} \|v_{G_k}\|_2) \sum_{j=1}^{q} m_j \|u_{G_j}\|_2.$$

(Here v_{G_k} is the sub-vector of v consisting of those components indexed by G_k.)

Fix $\lambda > 0$ and let $\hat{\beta} \in \mathbb{R}^p$ be a minimiser of

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda P(\beta)$$

over $\beta \in \mathbb{R}^p$. For a non-empty set $G \subseteq \{1, \ldots, p\}$, let X_G be the sub-matrix of X consisting of those columns indexed by G. Show that on the event $\Omega := \{\max_{k=1,\ldots,q}(m_k^{-1}\|X_{G_k}^T \varepsilon\|_2) \leqslant n\lambda\}$, we have that

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta})\|_2^2 \leqslant 4\lambda P(\beta^0).$$

Now suppose $|G_j| = r$ and $m_j = \sqrt{r}$ for all j (so $p = qr$). Suppose further that $X_{G_j}^T X_{G_j} = nI$ for all j. Show that when λ is such that

$$(n\lambda^2 - 1)^2 = \frac{8(A+1)\log q}{r},$$

for $A > 0$ and such that the above is less than 1, we have that $\mathbb{P}(\Omega) \geqslant 1 - q^{-A}$. [You may use the facts that the mgf of a χ_1^2 random variable is $1/\sqrt{1-2\alpha}$ for $\alpha < 1/2$, and $e^{-\alpha}/\sqrt{1-2\alpha} \leqslant e^{2\alpha^2}$ when $|\alpha| < 1/4$.]

5 Let $Z \sim N_p(\mu, \Sigma^0)$ with Σ^0 positive definite. For a non-empty set $A \subseteq \{1, \ldots, p\}$, let Z_A be the sub-vector of Z consisting of components indexed by A. Derive the conditional distribution $Z_A \,|\, Z_B = z_B$, where A and B are non-empty subsets of $\{1, \ldots, p\}$.

Writing $\Omega^0 = (\Sigma^0)^{-1}$ for the precision matrix, show that $\mathrm{Var}(Z_A \,|\, Z_{A^c}) = (\Omega^0_{A,A})^{-1}$ and derive an expression for $\mathrm{Cov}(Z_j, Z_k \,|\, Z_{-jk})$ involving only Ω^0_{jj}, Ω^0_{kk} and Ω^0_{jk}. Here $\Omega^0_{A,A}$ is the submatrix of Ω^0 consisting of those rows and columns indexed by A. Hence or otherwise show that

$$Z_j \perp\!\!\!\perp Z_k \,|\, Z_{-jk} \Leftrightarrow \Omega^0_{jk} = 0.$$

[You may use without proof the following facts. Let $M \in \mathbb{R}^{p \times p}$ be a symmetric positive definite matrix and suppose

$$M = \begin{pmatrix} P & Q^T \\ Q & R \end{pmatrix}$$

with P and R square matrices. Writing $S := P - Q^T R^{-1} Q$, we have that S is positive definite and

$$M^{-1} = \begin{pmatrix} S^{-1} & -S^{-1} Q^T R^{-1} \\ -R^{-1} Q S^{-1} & R^{-1} + R^{-1} Q S^{-1} Q^T R^{-1} \end{pmatrix}.$$

]

Suppose $x_1, \ldots, x_n$ are independent random vectors with each $x_i \sim N_p(\mu, \Sigma^0)$. Write $X \in \mathbb{R}^{n \times p}$ for the matrix with ith row x_i and suppose that X has full column rank. Show that the maximum likelihood estimator for Ω^0 minimises

$$-\log\det(\Omega) + \mathrm{tr}(S\Omega)$$

over $\Omega \succ 0$ (i.e. symmetric positive definite Ω) where

$$S := \frac{1}{n}\sum_{i=1}^n (x_i - \bar{X})(x_i - \bar{X})^T, \qquad \bar{X} := \frac{1}{n}\sum_{i=1}^n x_i.$$

Give the optimisation problem solved by the *graphical Lasso* estimator $\hat{\Omega}_\lambda$ of the precision matrix with tuning parameter $\lambda > 0$.

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 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).

Write down the KKT conditions for the Lasso and show that

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

Show that on the event

$$\Omega = \{2\|X^T\varepsilon\|_\infty/n < \lambda\},$$

we have

$$\frac{1}{n\lambda}\|X(\hat{\beta} - \beta^0)\|_2^2 + \frac{1}{2}\|\hat{\beta}_N - \beta^0_N\|_1 < \frac{3}{2}\|\beta^0_S - \hat{\beta}_S\|_1.$$

Let $B := \{\beta \in \mathbb{R}^p : \|\beta_N\|_1 \leqslant 3\|\beta_S\|_1\}$. Suppose there exists $\kappa > 0$ such that for all $\beta \in B$, we have

$$\kappa\|\beta\|_2 \leqslant \frac{1}{\sqrt{n}}\|X\beta\|_2.$$

Show that on Ω,

$$\|\beta^0 - \hat{\beta}\|_2 < \frac{3\lambda\sqrt{s}}{2\kappa^2}.$$

Suppose that $\min_{j \in S} |\beta^0_j| > 3\lambda\sqrt{s}/\kappa^2$. Give, with justification, a choice of τ such that on Ω, $\hat{S}^\tau := \{j : |\hat{\beta}_j| > \tau\}$ satisfies $\hat{S}^\tau = S$.

END OF PAPER