UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 4 June, 2018 1:30 pm to 4:30 pm

PAPER 205

MODERN STATISTICAL METHODS

*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

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 $Y \in \mathbb{R}^n$ be a vector of observations with $Y = \mu^0 + \varepsilon$ for some fixed $\mu^0 \in \mathbb{R}^n$ and $\mathbb{E}(\varepsilon) = 0$. Define

$$S = \{i : 1 \leqslant i \leqslant n-1,\ \mu_i^0 \neq \mu_{i+1}^0\}$$

to be the set of indices where μ^0 changes, and suppose $s = |S|$ is small compared to n. Describe a method for estimating μ^0 using the minimiser of an appropriate penalised least squares objective function that you should specify.

Explain the closed testing procedure for multiple testing and state and prove a result concerning the error control it provides against falsely rejecting null hypotheses.

Now let

$$\mathcal{I} = \{[i, j] : i, j \in \{1, \ldots, n\} \text{ and } i < j\},$$

where $[i, j] = \{i, i+1, \ldots, j\}$. For each $I \in \mathcal{I}$, let H_I be the null hypothesis that μ^0 is constant on I, that is there exists $m \in \mathbb{R}$ with $\mu_i^0 = m$ for all $i \in I$. Suppose that for each null hypothesis H_I, we have an associated p-value q_I. Define the adjusted p-value p_I for H_I by

$$p_I = \max_{J \in \mathcal{I}: J \supseteq I} q_J \frac{n}{|J|}.$$

Consider the procedure that rejects all H_I where the adjusted p-values have $p_I \leqslant \alpha$. Writing $S = \{i_1, \ldots, i_s\}$ with $i_1 < \cdots < i_s$, let

$$\mathcal{T} = \{[1, i_1], [i_1 + 1, i_2], [i_2 + 1, i_3], \ldots, [i_{s-1} + 1, i_s], [i_s + 1, n]\}.$$

Explain why the procedure will make a false rejection if and only if H_I is rejected for some $I \in \mathcal{T}$.

Finally show that the procedure makes no false rejections with probability at least $1 - \alpha$.

2 Let $Y \in \mathbb{R}^n$ be a vector of centred responses and $X \in \mathbb{R}^{n \times p}$ a matrix of predictors that have been centred and scaled to have ℓ_2-norm $\sqrt{n}$. Write down the objective function optimised by the ridge regression estimator and give a closed form expression for the estimator with tuning parameter $\gamma > 0$. [You need not derive this expression.]

Now consider the penalised regression procedure defined by the objective function

$$(\hat{\delta}_\lambda, \hat{\beta}_\lambda) = \text{argmin}_{\delta,\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|Y - X(\delta + \beta)\|_2^2 + \lambda_1 \|\delta\|_1 + \lambda_2 \|\beta\|_2^2 \right\}$$

where $\lambda_1, \lambda_2 > 0$. Find an expression for $\hat{\beta}_\lambda$ involving Y, X, $\hat{\delta}_\lambda$, λ_2 and n.

Thus show that $\hat{\delta}_\lambda$ is the minimiser of a Lasso objective with transformed design matrix $\tilde{X} = AX$ and transformed response AY where A is a matrix you should specify. [*Hint: It may help to first multiply the objective function by* $2n$ *and define* $\theta = Y - X\delta$ *and* $Q = (X^T X + 2n\lambda_2)^{-1}$.]

Consider now the model

$$Y = X(\delta^0 + \beta^0) + \varepsilon$$

and let Ω be the event that $\|\tilde{X}^T(\tilde{X}\beta^0 + A\varepsilon)\|_\infty / n \leqslant \lambda_1$. Show that on Ω,

$$\frac{1}{n} \|\tilde{X}(\hat{\delta}_\lambda - \delta^0)\|_2^2 \leqslant 4\lambda_1 \|\delta^0\|_1.$$

Conclude that on Ω we have

$$\frac{1}{n} \|X(\hat{\delta}_\lambda - \delta^0)\|_2^2 \leqslant \frac{4\lambda_1 \|\delta^0\|_1}{1 - \kappa}$$

when the maximum eigenvalue of XQX^T is $\kappa < 1$.

3 Let $A \in \mathbb{R}^{d\times p}$ have i.i.d. standard normal entries. Show that

$$\mathbb{P}(|(A^TA)_{jj}/d - 1| \geqslant t) \leqslant 2e^{-dt^2/8}$$

for all $j = 1, \ldots, p$ and $t \in (0, 1)$. [*You may use without proof the facts that the moment generating function of a χ^2_1 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 $|\alpha| < 1/4$.*]

Now suppose V and W are independent standard normal random variables. By considering the identity $VW = (V+W)^2/4-(V-W)^2/4$, show that the moment generating function of VW is $1/\sqrt{1-\alpha^2}$ for $\alpha \in (-1, 1)$. Hence show that

$$\mathbb{P}(|(A^TA)_{jk}|/d \geqslant t) \leqslant 2e^{-dt^2/4}.$$

for $j \neq k$ and $t \in (0, 1)$. [*You may use without proof that $1/\sqrt{1-\alpha^2} \leqslant e^{\alpha^2}$ when $|\alpha| \leqslant 1/2$.*]

Finally show that provided $c\sqrt{\log(p)/d} < 1$, with probability at least $1-2p^{-(c^2/8-1)}-p^{-(c^2/4-2)}$, we have

$$\left|\frac{\|A(u-v)\|_2^2}{d\|u-v\|_2^2} - 1\right| \leqslant sc\sqrt{\log(p)/d}$$

for all $u, v \in \mathbb{R}^p$ with at most $s/2$ non-zero elements.

4 What is the *subdifferential* of a convex function at a point x in its domain? State a result concerning the minimisers of convex functions and their subdifferentials. Write down an expression for the subdifferential of the ℓ_1-norm.

Let $Y \in \mathbb{R}^n$ be a vector of responses and $X \in \mathbb{R}^{n \times p}$ a matrix of predictors. Consider the objective function

$$Q_\gamma(\beta; Y) = \frac{1}{\sqrt{n}}\|Y - X\beta\|_2 + \gamma\|\beta\|_1$$

where $\gamma > 0$. Now let $\hat{\beta}_\lambda$ be a minimiser of

$$\frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

and define $\hat{\sigma}_\lambda = \|Y - X\hat{\beta}_\lambda\|_2/\sqrt{n}$. Show that writing $\gamma(\lambda) = \lambda/\hat{\sigma}_\lambda$, we have that $\hat{\beta}_\lambda$ is a minimiser of $Q_{\gamma(\lambda)}(\cdot; Y)$ when $\hat{\sigma}_\lambda > 0$. [Standard results stated in lectures concerning subdifferentials may be used without proof.]

Assume the response Y follows a normal linear model of the form

$$Y = X\beta^0 + \varepsilon,$$

where $\varepsilon \sim N_n(0, \sigma^2 I)$. Suppose $Z \in \mathbb{R}^n$ is an additional predictor and let $\tilde{\beta}_\gamma$ be a minimiser of $Q_\gamma(\cdot; Z)$. Define $R_\gamma = Z - X\tilde{\beta}_\gamma$. Finally define

$$T = R_\gamma^T(Y - X\hat{\beta}_\lambda)/\|R_\gamma\|_2$$

where it is assumed that $R_\gamma \neq 0$. Show that $T = W + \Delta$ where $W \sim N(0, \sigma^2)$ and

$$|\Delta| \leqslant \sqrt{n}\gamma\|\beta^0 - \hat{\beta}_\lambda\|_1.$$

5 Let $\mathcal{G}$ be a directed acyclic graph (DAG). What is meant by the *moralised graph* of $\mathcal{G}$? What does it mean for A to be *d-separated from* B by S, where A, B, S are disjoint sets of nodes of $\mathcal{G}$? State a result concerning presence or absence of an edge between two vertices and their d-separation. [You need not define standard graph terminology such as *descendant*, *parent*, *topological order*, *collider* or *path* in your answer.]

What does it mean for $\mathcal{G}$ to be *faithful* to a distribution P? What is the conditional independence graph of a distribution P? Prove that if $\mathcal{G}$ is faithful to P then the moralised graph of $\mathcal{G}$ is the conditional independence graph of P.

Now suppose $Z \sim N_p(0, \Sigma)$ with Σ positive definite. Briefly describe the methods of *neighbourhood selection* and the *graphical Lasso* for estimating the conditional independence graph using independent data $x_1, \ldots, x_n \sim N_p(0, \Sigma)$. [You need not motivate or justify why the methods work.]

Briefly explain how one could modify the PC algorithm to make use of an estimate of the conditional independence graph to reduce computation.

6 Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of functions on an input space $\mathcal{X}$ with reproducing kernel k. Let $Y \in \mathbb{R}^n$ be a response vector satisfying $Y_i = f^0(x_i) + \varepsilon_i$ with $x_i \in \mathcal{X}$ for $i = 1, \ldots, n$, $f^0 \in \mathcal{H}$, $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Write down the optimisation problem solved by kernel ridge regression to produce an estimated regression function $\hat{f}_\lambda \in \mathcal{H}$ when the tuning parameter is $\lambda > 0$.

Let $K \in \mathbb{R}^{n \times n}$ be the matrix with ijth entry $K_{ij} = k(x_i, x_j)$ and eigenvalues $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_n$. Prove that

$$\frac{1}{n}\mathbb{E}\Big\{ \sum_{i=1}^{n} \{f^0(x_i) - \hat{f}_\lambda(x_i)\}^2 \Big\} \leqslant \frac{\sigma^2}{n} \frac{1}{\lambda} \sum_{i=1}^{n} \min(d_i/4, \lambda) + \frac{\lambda}{4n} \|f^0\|_{\mathcal{H}}^2.$$

[You may use standard properties of positive definite kernels without proof, and you may use the representer theorem without proof.]

END OF PAPER