UNIVERSITY OF CAMBRIDGE

MAT3, MAMA

# MATHEMATICAL TRIPOS Part III

Thursday, 6 June, 2019 9:00 am to 12:00 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*
*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 $H_1, \ldots, H_m$ be a sequence of null hypotheses with associated $p$-values $p_1, \ldots, p_m$. Let $I \subseteq \{1, \ldots, m\}$ be the indices corresponding to the set of true null hypotheses. What does it mean for a multiple testing procedure to control the *familywise error rate* (FWER) at level $\alpha$?

Consider the procedure that sets $R = \min\{j : p_j > \alpha\}$ and rejects hypotheses $H_1, \ldots, H_{R-1}$ if $R > 1$, rejects all hypotheses if $R$ is not defined (so $p_j \leqslant \alpha$ for all $j$), and rejects no hypotheses if $R = 1$. Prove that the FWER is controlled at level $\alpha$.

Suppose random variables $Z_1, \ldots, Z_p$ have joint distribution $P$. What does it mean for $P$ to satisfy the *global Markov property* with respect to a DAG $\mathcal{G}$? [You need not define graph terminology such as $d$-separation in your answer.]

Define

$$\mathcal{S} = \{\text{DAGs } \mathcal{G} \text{ such that } P \text{ is global Markov with respect to } \mathcal{G}\}.$$

Fix $j, k \in \{1, \ldots, p\}$ with $j \neq k$ and let $H_0$ be the null hypothesis that there exists $\mathcal{G} \in \mathcal{S}$ where nodes $j$ and $k$ are not adjacent. Suppose that for each $S \subseteq \{1, \ldots, p\} \setminus \{j, k\}$ we have a $p$-value $p_S$ for the null hypothesis $H_S$ that $Z_j \perp\!\!\!\perp Z_k | Z_S$. Note that $H_\emptyset$ should be understood as the null hypothesis that $Z_j$ and $Z_k$ are independent. Give, with careful justification, a non-trivial procedure for testing $H_0$ that will falsely reject $H_0$ with probability at most $\alpha$. [You may assume without proof that every DAG has a topological order.]

**2** Given independent data $x_1, \ldots, x_n \sim N_p(\mu, \Sigma^0)$ where $\Sigma^0 \in \mathbb{R}^{p\times p}$ is positive definite, write down the maximum likelihood estimate $\hat{\Sigma}$ of $\Sigma^0$.

For a matrix $A \in \mathbb{R}^{r\times s}$, let $\|A\|_1 = \sum_{j,k} |A_{jk}|$ and $\|A\|_\infty = \max_{j,k} |A_{jk}|$. Also define $\|A\|_{L_1} = \max_j \sum_i |A_{ij}| = \max_j \|A_j\|_1$, where $A_j \in \mathbb{R}^r$ denotes the the $j$th column of $A$. Show that for two matrices $A \in \mathbb{R}^{r\times s}$ and $B \in \mathbb{R}^{s\times t}$, $\|AB\|_\infty \leqslant \|A\|_\infty \|B\|_{L_1}$. Show also that if $r = s$ and $A$ is symmetric then $\|AB\|_\infty \leqslant \|A\|_{L_1} \|B\|_\infty$.

Consider the following estimator for the precision matrix $\Omega^0 = (\Sigma^0)^{-1}$:

$$\hat{\Omega} := \operatorname*{arg\,min}_{\Omega \in \mathbb{R}^{p\times p}} \|\Omega\|_1 \text{ subject to } \|\hat{\Sigma}\Omega - I\|_\infty \leqslant \lambda,$$

for some $\lambda > 0$. Assuming there is a feasible solution to the constrained optimisation problem above, so $\hat{\Omega}$ exists, show that $\hat{\Omega}_j$ is a minimiser over $\beta$ of $\|\beta\|_1$ subject to $\|\hat{\Sigma}\beta - e_j\|_\infty \leqslant \lambda$, where $e_j \in \mathbb{R}^p$ is the $j$th standard basis vector.

Suppose for the remainder of this question that

$$\lambda \geqslant \|\hat{\Sigma} - \Sigma^0\|_\infty \|\Omega^0\|_{L_1}.$$

Show that $\|\hat{\Sigma}\Omega^0_j - e_j\|_\infty \leqslant \lambda$ for all $j$. What does this imply about how $\|\hat{\Omega}_j\|_1$ compares to $\|\Omega^0_j\|_1$?

Next show that $\|\Sigma^0(\hat{\Omega} - \Omega^0)\|_\infty \leqslant 2\lambda$. [*Hint: consider subtracting and adding* $\hat{\Sigma}\hat{\Omega}$.]

Finally show that $\|\hat{\Omega} - \Omega^0\|_\infty \leqslant 2\lambda \|\Omega^0\|_{L_1}$.

**3** Let $\mathcal{X}$ be a (non-empty) input space. What is a *positive definite kernel*? In the remainder of this question we will refer to a positive definite kernel as simply a kernel.

Show that if $\mathcal{H}$ is an inner product space and $\phi : \mathcal{X} \to \mathcal{H}$ is a feature map, then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$k(x, x') = \langle \phi(x), \phi(x') \rangle$$

is a kernel.

Show that if $k_1, k_2, \ldots$ are kernels on input space $\mathcal{X}$, then

1. $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel for $\alpha_1, \alpha_2 \geqslant 0$;
2. if $k(x, x') := \lim_{m\to\infty} k_m(x, x')$ exists for all $x, x' \in \mathcal{X}$ then $k$ is a kernel;
3. if $k(x, x') := k_1(x, x') k_2(x, x')$ for all $x, x' \in \mathcal{X}$ then $k$ is a kernel.

Write down the equation for the Gaussian kernel on $\mathbb{R}^d$ with bandwidth $\sigma^2$. Show that it is a positive definite kernel.

**4** Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. What is meant by a *subgradient* of $f$ at a point $x \in \mathbb{R}^d$? State a result concerning minimisation of $f$ and subgradients. Write down the subdifferential of the absolute value function $|\cdot|$ at each $x \in \mathbb{R}$.

Explain the procedure of *coordinate descent* for minimising $f$.

Consider performing a Lasso regression with centred response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$ with centred columns scaled to have $\ell_2$-norm $\sqrt{n}$. Write down the optimisation problem solved by the Lasso with tuning parameter $\lambda > 0$.

Given an initialiser $\hat{\beta}^{(0)} \in \mathbb{R}^p$ for coordinate descent minimisation of the Lasso objective function, show that the next iterate $\hat{\beta}^{(1)} \in \mathbb{R}^p$ satisfies

$$\hat{\beta}_1^{(1)} = S_\lambda(R/n),$$

where $S_\lambda(t) = \max(|t| - \lambda, 0)\mathrm{sgn}(t)$ and $R \in \mathbb{R}$ is a function of $Y, X$ and $\hat{\beta}^{(0)}$ that you should specify.

Now consider minimising

$$Q(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} \rho(\beta_j)$$

where

$$\rho(t) = |t|\mathbb{1}_{\{|t|\leqslant\delta\}} + \frac{t^2+\delta^2}{2\delta}\mathbb{1}_{\{|t|>\delta\}}.$$

By considering the KKT conditions of the coordinatewise minimisation or otherwise, show that given an initialiser $\hat{\beta}^{(0)} \in \mathbb{R}^p$ for coordinate descent minimisation of $Q$, the next iterate $\hat{\beta}^{(1)} \in \mathbb{R}^p$ satisfies

$$\hat{\beta}_1^{(1)} = \begin{cases} S_\lambda(R/n) & \text{if } |S_\lambda(R/n)| \leqslant \delta \\ \frac{R/n}{1+\lambda/\delta} & \text{otherwise.} \end{cases}$$

[*In this question you may use standard results about subgradients and subdifferentials without proof.*]

**5** For a symmetric positive semi-definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ and non-empty set $S \subset \{1,\ldots,p\}$ (where the inclusion is strict), we define the compatibility factor

$$\phi_\Sigma^2(S) = \inf_{\beta:\|\beta_S\|_1\neq 0, \|\beta_N\|_1 \leqslant 3\|\beta_S\|_1} \frac{\beta^T\Sigma\beta}{\|\beta_S\|_1^2/|S|},$$

where $N := \{1,\ldots,p\} \setminus S$. Prove that if symmetric positive semi-definite matrices $\Theta, \Sigma \in \mathbb{R}^{p\times p}$ have $\max_{j,k}|\Sigma_{jk} - \Theta_{jk}| \leqslant \phi_\Sigma^2(S)/(32|S|)$ then $\phi_\Theta^2(S) \geqslant \phi_\Sigma^2(S)/2$.

What does it mean for a random variable $W \in \mathbb{R}$ to be sub-Gaussian with parameter $\sigma > 0$? State an upper bound on $\mathbb{P}(W > t)$ for $t > 0$ in the case where additionally $\mathbb{E}W = 0$.

Now suppose matrix $X \in [-1,1]^{n\times p}$ has independent rows with $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}X_{ik}) = \Sigma_{jk}$ for all $i,j,k$ and positive definite matrix $\Sigma$. Let $\hat\Sigma = X^TX/n$. Show that

$$\mathbb{P}(\max_{j,k}|\hat\Sigma_{jk} - \Sigma_{jk}| > 4\sqrt{2\log(p)/n}) \leqslant \frac{2}{p^2}.$$

[You may use without proof the fact that if random variable $W$ with $\mathbb{E}W = 0$ takes values in $[-2,2]$ then $W$ is sub-Gaussian with parameter 2.]

Let $c_{\min}$ be the minimum eigenvalue of $\Sigma$. State a result concerning the relative sizes of $c_{\min}$ and $\phi_\Sigma^2(S)$ for non-empty $S \subseteq \{1,\ldots,p\}$. From the results above, give a condition on $c_{\min}$ involving $s$, $n$ and $p$ such that when this holds, we have with probability at least $1 - 2p^{-2}$ that $\phi_{\hat\Sigma}^2(S) \geqslant c_{\min}/2$ for all $S$ with $0 < |S| \leqslant s < p$.

**6** Given a response $Y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n \times p}$ consider the regression estimator

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 + \frac{\gamma}{2} \|\beta\|_2^2, \qquad (*)$$

where $\lambda > 0$ and $\gamma > 0$ are tuning parameters. Briefly explain why the minimising $\hat{\beta}$, which you may assume exists, is unique. In the case where $X$ has two duplicate columns, argue that the corresponding coefficient estimates will be equal.

Write down the KKT conditions for the optimisation problem $(*)$.

Now consider the noiseless linear model, $Y = X\beta^0$. Let $S = \{j : \beta_j^0 \neq 0\}$. Show that if $\text{sgn}(\beta^0) = \text{sgn}(\hat{\beta})$, then

$$\|X_N^T X_S (X_S^T X_S + n\gamma I)^{-1} \{\gamma \beta_S^0 / \lambda + \text{sgn}(\beta_S^0)\}\|_\infty \leqslant 1. \qquad (**)$$

Show further that if $(**)$ holds and also

$$\text{sgn}(\beta_S^0) = \text{sgn}\left( (X_S^T X_S + n\gamma I)^{-1} (X_S^T X_S \beta_S^0 - \lambda \text{sgn}(\beta_S^0)) \right),$$

then we have $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^0)$.

## END OF PAPER