UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 30 May, 2016 9:00 am to 11:00 am

PAPER 205

MODERN STATISTICAL METHODS

Attempt no more than ***THREE*** *questions.*

There are ***FOUR*** *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 In answering the following questions, you need not explain what a graph is nor define any graph terminology such as d-separation. Let $\mathcal{G}$ be a directed acyclic graph (DAG) with vertex set $\{1, \ldots, p\}$, and let $Z \in \mathbb{R}^p$ be a random vector with distribution P. What does it mean for P to satisfy the *global Markov property* with respect to $\mathcal{G}$? What does it mean for P to satisfy *causal minimality* with respect to $\mathcal{G}$? What does it mean for P to be *faithful* to $\mathcal{G}$?

Give, with brief justification, an example of a distribution P and DAG $\mathcal{G}$ where P satisfies causal minimality with respect to $\mathcal{G}$ but where P is not faithful to $\mathcal{G}$.

Suppose that P satisfies the global Markov property with respect to $\mathcal{G}$. For any $A \subseteq \{1, \ldots, p\}$, by A^c we mean $\{1, \ldots, p\} \setminus A$. The *Markov blanket* of a node k, denoted $\mathrm{mb}(k)$, is defined to be the set of all nodes adjacent to k together with the parents of all of its children. Show that if $(\mathrm{mb}(k) \cup \{k\})^c \neq \emptyset$ then $Z_k \perp\!\!\!\perp Z_{(\mathrm{mb}(k)\cup\{k\})^c} | Z_{\mathrm{mb}(k)}$.

Now suppose P is faithful to $\mathcal{G}$. Show then that any set of nodes $A \subseteq \{1, \ldots, p\} \setminus \{k\}$ such that $Z_k \perp\!\!\!\perp Z_{(A\cup\{k\})^c} | Z_A$ must have $A \supseteq \mathrm{mb}(k)$.

2

Let $Y \in \mathbb{R}^n$ be a vector of responses and let $X \in \mathbb{R}^{n\times p}$ be a matrix of predictors, and suppose Y and the columns of X have been centred. Write down the optimisation problem solved by *Ridge regression* when the tuning parameter is $\lambda > 0$ and show that the fitted values are given by

$$X(X^T X + \lambda I)^{-1} X^T Y.$$

Show that the fitted values also equal

$$K(K + \lambda I)^{-1} Y$$

where $K = XX^T$.

Let $\mathcal{X}$ be a (non-empty) input space. What is a *positive definite kernel*? Show that if k is a positive definite kernel then

$$k(x, x')^2 \leqslant k(x, x)k(x', x')$$

for all $x, x' \in \mathcal{X}$.

Prove that for every positive definite kernel k there exists an inner product space $\mathcal{H}$ and feature map $\phi : \mathcal{X} \to \mathcal{H}$ with

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

for all $x, x' \in \mathcal{X}$.

3

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 scaled to have ℓ_2-norm $\sqrt{n}$, 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 $\bar{\varepsilon} = \mathbf{1}^T\varepsilon/n$. Define the *Lasso estimator* $\hat{\beta}^L_\lambda$ of β^0 with regularisation parameter $\lambda > 0$.

Write down the KKT conditions for the Lasso and show that

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

Let $S = \{k \in \{1, \ldots, p\} : \beta^0_k \neq 0\}$, let $N = \{1, \ldots, p\} \setminus S$, and let $s = |S|$. Suppose $0 < s < p$. For an arbitrary $A \subseteq \{1, \ldots, p\}$ and $b \in \mathbb{R}^p$, write b_A for the vector in $\mathbb{R}^{|A|}$ obtained by extracting the components of b with indices that are in A. Assume that for some $c \in (0, 1)$ there exists $\phi > 0$ such that for all $b \in \mathbb{R}^p$ with $(1-c)\|b_N\|_1 \leqslant (1+c)\|b_S\|_1$, we have

$$\|b_S\|_1^2 \leqslant \frac{s\|Xb\|_2^2}{n\phi^2}.$$

Define the event $\Omega = \{\|X^T\varepsilon\|_\infty/n \leqslant c\lambda\}$. Show that if $\varepsilon \sim N_n(0, \sigma^2 I)$ and $\lambda = A\sigma\sqrt{\log(p)/n}$ with $A > 0$, then $\mathbb{P}(\Omega) \geqslant 1 - p^{-(A^2c^2/2-1)}$. [You may assume a tail bound for a standard normal random variable provided you state it clearly.]

Show that on Ω,

$$\frac{1}{n}\|X(\hat{\beta}^L_\lambda - \beta^0)\|_2^2 + (1-c)\lambda\|\hat{\beta}^L_{\lambda,N}\|_1 \leqslant (1+c)^2\lambda^2\frac{s}{\phi^2}.$$

Finally show that on Ω, if $|\beta^0_k| > (1+c)\lambda s/\phi^2$ then $\operatorname{sgn}(\hat{\beta}^L_{\lambda,k}) = \operatorname{sgn}(\beta^0_k)$.

4

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 $\Omega^0 = (\Sigma^0)^{-1}$ minimises

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

over $\Omega \succ 0$ (i.e. 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 $\lambda > 0$, and write down its KKT conditions.

For any matrix $M \in \mathbb{R}^{p \times p}$ and $j \in \{1, \ldots, p\}$, let $M_{-j,-j} \in \mathbb{R}^{(p-1)\times(p-1)}$ be the submatrix of M excluding its jth row and column, and let $M_{-j,j} \in \mathbb{R}^{p-1}$ be the jth column of M excluding its jth component. Fix $j \in \{1, \ldots, p\}$ and $\lambda > 0$. Write $\hat{\Sigma} = \hat{\Omega}_\lambda^{-1}$ and let W be a symmetric positive definite matrix with $W^2 = \hat{\Sigma}_{-j,-j}$. Let b^* be a minimiser over $b \in \mathbb{R}^{p-1}$ of

$$\frac{1}{2}\|Wb - W^{-1}S_{-j,j}\|_2^2 + \lambda\|b\|_1.$$

Explain very briefly why b^* is unique. By comparing the KKT conditions of the optimisation problem above to those for the graphical Lasso, show that

$$\hat{\Sigma}_{-j,j} = \hat{\Sigma}_{-j,-j}b^*.$$

[You may use the fact that if $M \in \mathbb{R}^{p \times p}$ is a symmetric positive definite matrix and

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

with P and R square matrices, then writing $T = P - QR^{-1}Q^T$, we have that T is positive definite and

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

END OF PAPER