UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 29 May, 2015 1:30 pm to 3:30 pm

PAPER 32

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

Let $X \in \mathbb{R}^{n\times p}$ be a matrix of predictors and Y an n-vector of responses. Assume that the columns of X have been centred and scaled and that Y has been centred.

Define the *Lasso* estimator $\hat{\beta}^L_\lambda$ with tuning parameter $\lambda > 0$ in this context. Write out the KKT conditions for $\hat{\beta}^L_\lambda$.

Now write out the steps of the *Least Angle Regression* (LAR) algorithm for regressing Y on X where the initial active set $A_1 = \emptyset$, $\lambda_0 = \infty$, and $\lambda_1, \lambda_2, \ldots$ are successive values of λ^{hit} where the active sets then change to $A_2, A_3, \ldots$. You may assume that the variable to enter the active set at λ^{hit} is always uniquely determined. Let $\hat{\beta}$ be the solution path produced by the LAR algorithm. Prove that for $m \geqslant 2$,

$$\frac{1}{n}|X_k^T\{Y - X\hat{\beta}(\lambda)\}| = \lambda \ \text{ for } k \in A_m,\ \lambda \in [\lambda_m, \lambda_{m-1}],$$

with λ_m taken as 0 in the above if m is the final step of the algorithm.

Now assume that the Lasso solution is unique at every $\lambda > 0$. Show that if for $m \geqslant 2$,

$$\text{sign}(X_k^T\{Y - X\hat{\beta}(\lambda)\}) = \text{sign}(\hat{\beta}_k(\lambda)) \ \text{ for } k \in A_m,\ \lambda \in [\lambda_m, \lambda_{m-1}],$$

again with $\lambda_m = 0$ in the above if m is the final step of the algorithm, then the Lasso solution path and the LAR path coincide so $\hat{\beta}(\lambda) = \hat{\beta}^L_\lambda$ for $\lambda > 0$.

2

Let $Y \in \mathbb{R}^n$ be a vector of responses and $X \in \mathbb{R}^{n\times p}$ a matrix of predictors with $\text{rank}(X) = p$. Suppose that the columns of X have been centred and scaled, and that Y is also centred. Consider the linear model (after centring),

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

where $\text{Var}(\varepsilon) = \sigma^2 I$ ($\sigma^2 > 0$), $\mathbf{1}$ is an n-vector of 1's and $\bar{\varepsilon} = \mathbf{1}^T\varepsilon/n$. Write down a formula for the ordinary least squares estimator $\hat{\beta}^{OLS}$ of β^0.

Write down a formula for the *ridge regression* estimator $\hat{\beta}^R_\lambda$ of β^0 when the tuning parameter is $\lambda > 0$.

Prove that there exists a $\lambda > 0$ depending on β^0 and σ^2, such that for all $x^* \in \mathbb{R}^p$ with $\|x^*\|_2 = 1$, we have

$$\mathbb{E}\{(x^{*T}\hat{\beta}^R_\lambda - x^{*T}\beta^0)^2\} < \mathbb{E}\{(x^{*T}\hat{\beta}^{OLS} - x^{*T}\beta^0)^2\}.$$

Finally show that for any fixed $\lambda > 0$ and fixed $\delta > 0$, there exist $x^* \in \mathbb{R}^p$ with $\|x^*\|_2 = 1$ and $\beta^0 \in \mathbb{R}^p$ such that

$$\mathbb{E}\{(x^{*T}\hat{\beta}^R_\lambda - x^{*T}\beta^0)^2\} > \mathbb{E}\{(x^{*T}\hat{\beta}^{OLS} - x^{*T}\beta^0)^2\} + \delta.$$

3

Suppose we have null hypotheses $H_1, \ldots, H_m$ and associated p-values $p_1, \ldots, p_m$. Let I_0 be the set of indices corresponding to true null hypotheses so that $H_i : i \in I_0$ are the true null hypotheses. What is the *family-wise error rate* (FWER)? Describe the *Bonferroni correction* and prove that it can be used to control the FWER at a desired level α.

What is an *intersection hypothesis*? What is the *closure* of the family $H_1, \ldots, H_m$ of hypotheses? Describe the *closed testing procedure*, introducing any other tests that are needed in order for it to work. Prove that the closed testing procedure can control the FWER at level α.

Now consider a family of intersection hypotheses $H_I : I \in \mathcal{I}$ that is hierarchical in the sense that for any $I,\ J \in \mathcal{I}$, we either have $I \cap J = \emptyset$ or $I \subseteq J$ or $J \subseteq I$. Suppose that for each H_I, $I \in \mathcal{I}$ we have a p-value p_I. Define the adjusted p-value of H_I to be

$$p_I^{\text{adj}} = \max_{J : J\in\mathcal{I},\, J\supseteq I} \frac{m}{|J|} p_J.$$

Consider the procedure that rejects all hypotheses H_I for which $p_I^{\text{adj}} \leqslant \alpha$. Show that with probability at least $1 - \alpha$, this procedure makes no false rejections.

4

Let n, p be integers greater than 1, and let $k \in \{1, \ldots, p\}$. In this question, we use the following notation. For a vector $z \in \mathbb{R}^p$, $z_{-k} \in \mathbb{R}^{p-1}$ is the vector z with its kth component removed. For a matrix $X \in \mathbb{R}^{n \times p}$, X_k is its kth column and $X_{-k} \in \mathbb{R}^{n \times (p-1)}$ is X with its kth column removed. Furthermore, for a matrix $A \in \mathbb{R}^{p \times p}$, we will write $A_{-k,k} \in \mathbb{R}^{p-1}$ for the kth column of A with its kth component A_{kk} removed. We will denote an n-vector of 1's by $\mathbf{1}$.

Let $Z \sim N_p(\mu, \Sigma)$ with Σ positive definite. Explain what is meant by a conditional independence graph for this distribution. [You need not explain what a graph is.]

Let $z \in \mathbb{R}^p$. Derive the distribution of $Z_k | Z_{-k} = z_{-k}$.

Suppose we have data $x_1, \ldots, x_n$ forming the rows of a matrix $X \in \mathbb{R}^{n \times p}$, which we can model as realisations of independent $N_p(\mu, \Sigma)$ random vectors. Motivate and explain the procedure of *nodewise regression* for estimating the conditional independence graph based on this data.

Now consider the following objective function over $\mu_1, \ldots, \mu_p \in \mathbb{R}$ and $\Theta \in \mathbb{R}^{p \times p}$, where we constrain $\Theta_{kk} = 0$ for $k = 1, \ldots, p$:

$$\frac{1}{2n} \sum_{k=1}^{p} \|X_k - \mu_k \mathbf{1} - X_{-k} \Theta_{-k,k}\|_2^2 + \lambda \sum_{j<k} \sqrt{\Theta_{jk}^2 + \Theta_{kj}^2}. \tag{1}$$

Explain how the minimiser of this objective can be used to estimate the conditional independence graph, discussing and motivating the form of the penalty function being used.

END OF PAPER