UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 30 May, 2013 1:30 pm to 3:30 pm

PAPER 31

STATISTICAL THEORY

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 Y be a real-valued random variable with $\mathbb{E}(|Y|) < \infty$, and with continuous, strictly increasing distribution function F. For $\tau \in (0,1)$, let $\rho_\tau(y) = \tau y \mathbb{1}_{\{y \geqslant 0\}} + (1-\tau)|y|\mathbb{1}_{\{y<0\}}$. Find the unique minimiser, q_τ, of $\mathbb{E}\{\rho_\tau(Y-q)\}$ over $q \in \mathbb{R}$.

Now suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed pairs taking values in $\mathbb{R}^m \times \mathbb{R}$ and satisfying

$$Y_i = g(X_i, \theta_0) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\theta_0 \in \Theta \subseteq \mathbb{R}^p$. Assume that $\mathbb{E}(|\epsilon_1|) < \infty$ and that the conditional distribution function of ϵ_1 given X_1 is continuous and strictly increasing with τth quantile zero. Suppose further that g is a known, bounded, continuous function and that $\mathbb{P}\{g(X_1,\theta) = g(X_1,\theta_0)\} = 1$ if and only if $\theta = \theta_0$. Let

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho_\tau\bigl(Y_i - g(X_i,\theta)\bigr).$$

Prove that if Θ is compact and if

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n \rho_\tau\bigl(Y_i - g(X_i,\theta)\bigr) - \mathbb{E}\bigl\{\rho_\tau\bigl(Y_1 - g(X_1,\theta)\bigr)\bigr\}\right| \xrightarrow{p} 0,$$

then $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$.

[General results about M-estimators should not be used without proof.]

2 Consider the linear model $Y = \beta_0 \mathbf{1}_n + X\beta + \epsilon$, where the columns of the deterministic design matrix $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$ are centred and have $\|x_j\|_2^2 = n$ for $j = 1, \ldots, p$, and where $\epsilon \sim N_n(0, \sigma^2 I)$. Define the *Lasso estimator* $\hat{\beta}_\lambda^L$ of β with regularisation parameter $\lambda > 0$.

Let $S = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$, let $N = \{1, \ldots, p\} \setminus S$, and let $s = |S|$. For an arbitrary $A \subseteq \{1, \ldots, p\}$ and $b \in \mathbb{R}^p$, we 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 there exists $\phi_0 > 0$ such that for all $b \in \mathbb{R}^p$ with $\|b_N\|_1 \leqslant 3\|b_S\|_1$, we have

$$\|b_S\|_1^2 \leqslant \frac{s\|Xb\|_2^2}{n\phi_0^2}.$$

Prove that if $\lambda = A\sigma\sqrt{\frac{\log p}{n}}$ for some $A > 0$, then with probability at least $1 - p^{-(A^2/8-1)}$, we have

$$\frac{1}{n}\|X(\hat{\beta}_\lambda^L - \beta)\|_2^2 + \lambda\|\hat{\beta}_\lambda^L - \beta\|_1 \leqslant \frac{16A^2}{\phi_0^2}\frac{\sigma^2 s \log p}{n}.$$

[Tail probability bounds for normal random variables should not be used without proof.]

3 Consider the linear model $Y = \beta_0 \mathbf{1}_n + X\beta + \epsilon$, where the columns of the deterministic design matrix $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$ are centred and satisfy $\|x_j\|_2^2 = n$ for $j = 1, \ldots, p$, and where $\epsilon \sim N_n(0, \sigma^2 I)$. For a given regularisation parameter $\lambda > 0$, write down the optimisation problem used to obtain the *ridge regression estimator* $\hat{\beta}_\lambda^R$ of β. After having replaced each Y_i with $Y_i - \bar{Y}$, where $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$, write down a closed-form expression for $\hat{\beta}_\lambda^R$.

Assume for now that $n > p$ and that $\text{rank}(X) = p$. Give a brief, intuitive explanation of how the ridge regression estimator differs from the maximum likelihood estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$, and also explain why the ridge regression estimator might be preferable when some of the columns of X are nearly collinear.

Without assuming that $n > p$ or that $\text{rank}(X) = p$, show that

$$\lim_{\lambda \searrow 0} \hat{\beta}_\lambda^R = (X^T X)^+ X^T Y,$$

where $(X^T X)^+$ denotes the Moore–Penrose pseudoinverse of $X^T X$.

4 In the context of testing null hypotheses $H_1, \ldots, H_m$, define the *Familywise Error Rate* (FWER). Letting P_i denote the p-value corresponding to H_i, define the *Bonferroni correction* for controlling the FWER at level $\alpha \in (0,1)$. Assuming that the p-values corresponding to true null hypotheses have a $U(0,1)$ distribution, prove that the procedure does indeed control the FWER at level α.

Now denote the ordered p-values as $P_{(1)} \leqslant \ldots \leqslant P_{(m)}$, and let $H_{(i)}$ denote the null hypothesis corresponding to $P_{(i)}$. *Holm's step-down procedure* rejects $H_{(1)}, \ldots, H_{(k)}$, where

$$k = \max\left\{i \in \{1, \ldots, m\} : P_{(1)} \leqslant \frac{\alpha}{m}, P_{(2)} \leqslant \frac{\alpha}{m-1}, \ldots, P_{(i)} \leqslant \frac{\alpha}{m-i+1}\right\}.$$

Assuming again that the p-values corresponding to true null hypotheses have a $U(0,1)$ distribution, prove that Holm's step-down procedure controls the FWER at level α.

Now define the *False Discovery Rate* (FDR). Define the Benjamini–Hochberg (BH) procedure for controlling the FDR at level $\alpha \in (0,1)$. Under conditions that you should specify, prove that the BH procedure does indeed control the FDR at level α.

END OF PAPER