MAT3

MATHEMATICAL TRIPOS Part III

Monday 8 June 2026 9:00am to 12:00pm

PAPER 205

MODERN STATISTICAL METHODS

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

Attempt no more than **FOUR** questions.
There are **FIVE** 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 Consider the high-dimensional linear model

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

where $X \in \mathbb{R}^{n\times p}$ has its columns centred and scaled to each have ℓ_2-norm $\sqrt{n}$. Write down the KKT conditions for the Lasso estimator $\hat{\beta}_\lambda$ with tuning parameter $\lambda > 0$. Show that

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

Suppose that the components of ε are independent and each is mean-zero and sub-Gaussian with parameter $\sigma > 0$. For a choice of λ that you should specify, prove that with probability at least $1 - 2p^{-(A^2/2-1)}$, we have

$$\frac{1}{n}\|X(\beta^0 - \hat{\beta}_\lambda)\|_2^2 \leqslant 2A\sigma\sqrt{\frac{\log p}{n}}\|\beta^0\|_1.$$

[Standard properties of sub-Gaussian random variables may be used without proof.]

Now consider the so-called Dantzig selector, given by

$$\hat{\beta}^{\mathrm{D}}_\lambda := \mathrm{argmin}_{\beta: n^{-1}\|X^\top(Y - X\beta)\|_\infty \leqslant \lambda}\|\beta\|_1.$$

Suppose we are in the low-dimensional setting where $X^\top X$ is invertible and moreover $(X^\top X)^{-1}$ is strictly diagonal dominant in the sense that for each $j = 1, \ldots, p$,

$$\sum_{k:k\neq j} |\{(X^\top X)^{-1}\}_{jk}| < \{(X^\top X)^{-1}\}_{jj}.$$

Show that $\hat{\beta}^{\mathrm{D}}_\lambda = \hat{\beta}_\lambda$ for all $\lambda > 0$. [*Hint: Fixing λ, suppose, for a contradiction that $\tilde{\beta} := \hat{\beta}^{\mathrm{D}}_\lambda \neq \hat{\beta}_\lambda$. Argue that there exists some $j \in \{k : \tilde{\beta}_k \neq 0\}$ such that*

$$\frac{1}{n}X_j^\top(Y - X\tilde{\beta}) \neq \mathrm{sgn}(\tilde{\beta}_j)\lambda.$$

Next consider the ℓ_1-norm of

$$\tilde{\beta} - \eta\,\mathrm{sgn}(\tilde{\beta}_j)(X^\top X)^{-1}e_j$$

for $\eta > 0$ sufficiently small, and where $e_j \in \mathbb{R}^p$ is the jth standard basis vector.]

2 Suppose we have data $\mathcal{D}_n := (X_i, Y_i, T_i)_{i=1}^n$ formed of i.i.d. copies of the triple $(X, Y, T) \in \mathbb{R}^p \times \mathbb{R} \times \{0, 1\}$, where $\mathbb{E}Y^2 < \infty$ and $\mathrm{Var}(Y \mid T, X) < c$ almost surely for some $c > 0$. Let

$$\mu(x) := \mathbb{E}(Y \mid T = 1, X = x) \qquad \text{and} \qquad \pi(x) := \mathbb{P}(T = 1 \mid X = x).$$

Suppose that there exists $\eta > 0$ such that $\pi(x) > \eta$ for all x. Let $\theta := \mathbb{E}(\mu(X))$. Show that

$$\mathbb{E}\left(\frac{YT}{\pi(X)}\right) = \theta.$$

Suppose we have available estimators $\hat{\mu}_n$ and $\hat{\pi}_n$ of μ and π respectively, trained on an independent copy of $\mathcal{D}_n$ that satisfy $\mathcal{E}_\mu \to 0$, $\mathcal{E}_\pi \to 0$ and $\mathcal{E}_\mu \mathcal{E}_\pi = o(n^{-1})$, where

$$\mathcal{E}_\mu := \mathbb{E}\big[\{\hat{\mu}_n(X) - \mu(X)\}^2\big] \qquad \text{and} \qquad \mathcal{E}_\pi := \mathbb{E}\left[\left\{\frac{1}{\hat{\pi}_n(X)} - \frac{1}{\pi(X)}\right\}^2\right].$$

Find, with proof, an estimator $\hat{\theta}_n$ of θ that uses only data $\mathcal{D}_n$ and estimators $\hat{\mu}_n$ and $\hat{\pi}_n$, and that satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, v), \tag{*}$$

where

$$v = \mathrm{Var}\left(\frac{(Y - \mu(X))T}{\pi(X)} + \mu(X)\right).$$

Briefly explain how your estimator could be modified in the case where no auxiliary data is available such that (*) is still satisfied for your new estimator. [You need not prove that (*) is satisfied.]

3 What does it mean for $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a *kernel*?

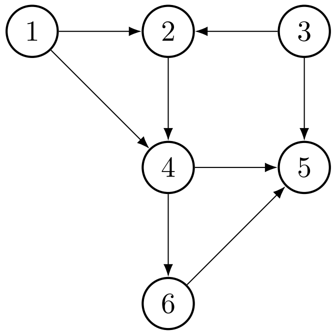
State and prove the *representer theorem*.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, and $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $(\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ are RKHSs with reproducing kernels $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Consider testing the independence of the X_i and Y_i using the test statistic

$$T := \sup_{\substack{f : \|f\|_{\mathcal{H}} \leqslant 1 \\ g : \|g\|_{\mathcal{G}} \leqslant 1}} \sum_{i=1}^n f(X_i)\, g(Y_i).$$

Show that T is given by the maximum singular value of a matrix you should specify. [You may assume there exists a solution $(\hat{f}, \hat{g}) \in \mathcal{H} \times \mathcal{G}$ to the constrained optimisation problem above.]

4 (a) Consider the following DAG $\mathcal{G}$:



Write down all sets that d-separate 1 and 6.

Now let $\mathcal{S}$ be the skeleton of the DAG $\mathcal{G}$, that is the undirected graph obtained by replacing every directed edge with an undirected edge. Suppose we know that another DAG $\mathcal{H}$ has $\mathcal{S}$ as its skeleton, and in $\mathcal{H}$, we still have $1 \to 2$ and $6 \to 5$, but $2 \perp\!\!\!\perp_{\mathcal{H}} 5 \mid 1, 6$. Draw the DAG $\mathcal{H}$.

(b) Prove that if nodes j and k in a DAG $\mathcal{G}$ are not adjacent, and π is a topological order for $\mathcal{G}$ with $\pi(j) < \pi(k)$, then they are d-separated by $\mathrm{pa}(k)$.

Suppose $Z \in \mathbb{R}^p$ has a distribution that is global Markov with respect to a DAG $\mathcal{G}$. Suppose we are able to test any number of conditional independence statements among the components of Z, such as the null hypothesis that $Z_1, Z_2 \perp\!\!\!\perp Z_3 \mid Z_4$, for instance. Find a collection of such null hypotheses which, if rejected, would allow us to conclude that Z_1 and Z_2 are adjacent in $\mathcal{G}$.

Now suppose $(I, X, Y) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ follows a distribution that is global Markov with respect to a DAG $\mathcal{G}$ where, identifying the vertices of $\mathcal{G}$ with the random variables, the only edges with I as endpoints are $I \to X_j$ for each $j = 1, \ldots, d$. Suppose we are able to test any number of conditional independence statements among X, I, Y, such as the null hypothesis that $X_1, X_2 \perp\!\!\!\perp Y \mid I, X_3$, for instance. Find a collection of such null hypotheses which, if rejected, would allow us to conclude that X_1 is a parent of Y.

5 Suppose $W \sim N_{p+2}(0, \Xi)$ where the covariance matrix Ξ is positive definite, and let $W_1, \ldots, W_n$ be i.i.d. copies of W. Let $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{n\times p}$ be the first, second and remaining columns of the $n \times (p+2)$ matrix with ith row $W_i^\top$ respectively. Let $\Sigma := \Xi_{-1,-1}$ and write $y = Y_1$ and $z \in \mathbb{R}^p$ for the first row of Z. Show that

$$y = z^\top \beta + \sigma e,$$

where $\beta := \Sigma_{-1,-1}^{-1}\Sigma_{-1,1}$, $\sigma > 0$ and $e \sim N(0,1)$ independently of z. [You may use the fact that for any symmetric positive-definite matrix $M \in \mathbb{R}^{d\times d}$, $M_{1,1} - M_{1,-1}M_{-1,-1}^{-1}M_{-1,1} > 0$.]

Let $\hat{\theta}$ be the vector of regression coefficients from a square-root Lasso regression of X on Z using tuning parameter $\gamma := \sqrt{2\log(p)/n}$, with no intercept term included. Write down the objective function minimised by $\hat{\theta}$.

Let $R := X - Z\hat{\theta} \in \mathbb{R}^n$. Assuming $R \neq 0$, show that

$$\frac{1}{\sqrt{n}} \frac{\|Z^\top R\|_\infty}{\|R\|_2} \leqslant \gamma.$$

In the following, we consider an asymptotic regime where as $n \to \infty$, p and β may change but σ remains constant. Suppose we have an estimator $\hat{\beta}$ of β such that writing $\hat{f} := Z\hat{\beta} \in \mathbb{R}^n$, we have $\|\hat{f} - f\|_2^2/n \xrightarrow{p} 0$, where $f := Z\beta$. Show that

$$\hat{\sigma} := \frac{1}{\sqrt{n}}\|Y - \hat{f}\|_2$$

satisfies $\hat{\sigma} \xrightarrow{p} \sigma$.

Now suppose that $\sqrt{\log p}\|\hat{\beta} - \beta\|_1 \xrightarrow{p} 0$ and we always have $R \neq 0$. Show that when $X \perp\!\!\!\perp Y \,|\, Z$, we have

$$\frac{R^\top(Y - Z\hat{\beta})}{\|R\|_2\hat{\sigma}} \xrightarrow{d} N(0,1).$$

END OF PAPER