UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 8 June, 2012 1:30 pm to 3:30 pm

PAPER 36

SEMIPARAMETRIC STATISTICS

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 $\mathcal{P}$ be a statistical model, that is a collection of probability measures on a given space $\mathcal{X}$ equipped with a σ-algebra $\mathcal{A}$. All measures are supposed to be dominated by a common σ-finite measure μ. That is, for any P in $\mathcal{P}$, one can write $dP = pd\mu$, with p the density of the probability measure P with respect to μ. Let now P be a fixed element of $\mathcal{P}$, and denote by $L^2(P)$ be the space of real-valued functions on $\mathcal{X}$ that are square integrable with respect to P.

(a) Give the definition of a differentiable path at P through the model, with score function $g : \mathcal{X} \to \mathbb{R}$ at P.

(b) Prove that any score function g at P is centered, that is $\int gdP = 0$ (you may use without proof that g belongs to $L^2(P)$ and that for any densities p, q with respect to μ, one has $\int(\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})d\mu = 0$).

Consider a semiparametric model $\mathcal{P} = \{P_{\theta,\eta},\ \theta \in \Theta,\ \eta \in H\}$, where Θ is an open subset of $\mathbb{R}$ and H is a set of functions. For any a, t in $\mathbb{R}$ and η_t in H, suppose that there exist differentiable paths $t \to P_{\theta+ta,\eta_t}$ through the model at $P_{\theta,\eta}$ such that the scores can be written additively as $a\dot{\ell}_{\theta,\eta} + g$, where $\dot{\ell}_{\theta,\eta}$ and g are in $L^2(P_{\theta,\eta})$.

(c) Define the efficient score $\tilde{\ell}_{\theta,\eta}$ and the efficient information $\tilde{I}_{\theta,\eta}$.

(d) Prove that any efficient score function $\tilde{\ell}_{\theta,\eta}$ is centered and that $P_{\theta,\eta}(\dot{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}) = \tilde{I}_{\theta,\eta}$.

2

Consider the statistical model $\mathcal{P}$ of the probability measures P_f having probability density f with respect to Lebesgue measure on the interval $[0,1]$.

(a) Give the definition of a tangent set at P_f.

(b) Show that a tangent set $\dot{\mathcal{P}}_f$ at P_f consists of the set of all measurable and bounded functions g on $[0,1]$ such that $\int_0^1 g(u)f(u)du = 0$.

Let a be a bounded measurable function on $[0,1]$. For any f, define the functional $\psi(P_f) = \int_0^1 a(u)f(u)du$.

(c) Prove that this functional is differentiable at P_f relative to $\dot{\mathcal{P}}_f$.

(d) Define the efficient influence function for estimating a functional $\psi : \mathcal{P} \to \mathbb{R}$. Determine the efficient influence function for the functional $\psi(P_f)$ defined above.

[You may use without proof that the closure in $L^2(P_f)$ of $\dot{\mathcal{P}}_f$ consists of all g in $L^2[0,1]$ such that $\int_0^1 g(u)f(u)du = 0$.]

(e) Suppose now that the density f is bounded and consider the functional $\psi(P_f) = \int_0^1 f(u)^4 du$. Determine the efficient influence function for this new functional.

3

Consider a model $\mathcal{P} = \{P_\eta,\ \eta \in \mathcal{F}\}$ of probability measures P_η on a measurable space $(\mathcal{X}, \mathcal{A})$ indexed by a class of functions $\mathcal{F}$ defined on $\mathcal{X}$, dominated by a common σ-finite measure μ and let $dP_\eta = p_\eta d\mu$. Let $\psi : \mathcal{P} \to \mathbb{R}$ be a functional of interest and let $\hat{\eta}$ be any estimator of η based on a single observation X from the model.

(a) Let f, g be in $\mathcal{F}$ and let d be some metric on $\mathcal{F}$. Prove that

$$\inf_{\hat{\eta}} \sup_{\eta \in \mathcal{F}} P_\eta \left[d(\hat{\eta}(X), \eta)^2\right] \geqslant \frac{1}{4} d(f,g)^2 \int (p_f \wedge p_g) d\mu.$$

Let now $P_\eta^{(n)} = \otimes_{i=1}^n P_\eta$ denote the joint law of i.i.d. random variables $X_1, \ldots, X_n$ drawn from P_η and let $\hat{\psi}_n$ be any estimator of $\psi(P_\eta)$ based on observations $X_1, \ldots X_n$. Let $\mu^{(n)} = \otimes_{i=1}^n \mu$. Define, for any f, g in $\mathcal{F}$, $\|P_f^{(n)} - P_g^{(n)}\|_1 = \int_{\mathcal{X}^n} |p_f^{(n)} - p_g^{(n)}| d\mu^{(n)}$.

(b) Let f, g be in $\mathcal{F}$, with $\psi(P_f) = \theta$ and $\psi(P_g) = \tau$, for some reals θ, τ. Prove that for any such f, g,

$$\inf_{\hat{\psi}_n} \sup_{\eta \in \mathcal{F}} P_\eta^{(n)} [(\hat{\psi}_n - \psi(P_\eta))^2] \geqslant \frac{1}{4}(\theta - \tau)^2 \left(1 - \frac{1}{2}\|P_f^{(n)} - P_g^{(n)}\|_1\right).$$

(c) Consider the functional $\psi(f) = \int_0^1 u f(u) du$. Let $\mathcal{F}$ be the set of all continuous densities on $[0,1]$. By an appropriate choice of alternatives f and g, prove that there exists a constant finite constant $C > 0$ such that

$$\inf_{\hat{\psi}_n} \sup_{\eta \in \mathcal{F}} P_\eta^{(n)} [(\hat{\psi}_n - \psi(\eta))^2] \geqslant C/n.$$

[*Hints: Take $f = 1$ and $g = 1 + a_n(x - 1/2)$, for $a_n \to 0$ to be chosen. You may further use without proof that if P_g has density $1 + \Delta$ with respect to P_f, then*

$$\|P_f^{(n)} - P_g^{(n)}\|_1^2 \leqslant \left(1 + \int \Delta^2 dP_f\right)^n - 1.]$$

4

Suppose that one observes $(X, Y) \in \mathbb{R} \times \mathbb{R}$ with

$$Y = g_\theta(X) + \varepsilon,$$

where g_θ is a given set of functions from $\mathbb{R}$ to $\mathbb{R}$ depending smoothly on a parameter $\theta \in \mathbb{R}$ (you may assume any differentiability or moment condition on $\theta \to g_\theta(x)$ suitable for your needs when computing scores). Assume that the variables X and ε are independent, that X has an unknown probability density η, and that the law of ε is Gaussian $\mathcal{N}(0, \sigma^2)$, with $\sigma^2 > 0$ known.

(a) Define and find the expression of the parametric score $\dot{\ell}_{\theta,\eta}$ in terms of the derivative $\dot{g}_\theta$ of the map $\theta \to g_\theta$. [It is *not* required to establish that the model is differentiable in quadratic mean (DQM).]

(b) Propose a non-parametric tangent set $\dot{\mathcal{P}}^N_{\theta,\eta}$ (again, establishing the DQM property is not required). Compute the efficient score and the efficient information. Is there a loss of information with respect to the parametric case when η is known ?

Now assume the pair (X, ε) has a joint probability density of the general form $\eta(x, e) = v(x)f(e)$, where v, f are unknown, with f sufficiently smooth so that derivatives and moments are well-defined. Suppose ε is square-integrable and satisfies $\int ef(e)de = 0$, so ε is of zero mean. Let $P_{\theta,\eta}$ denote the law of (X, Y) for some fixed (θ, η). Let $L^2(P_{\theta,\eta})$ denote the set of all square-integrable functions with respect to $P_{\theta,\eta}$.

(c) Find the parametric score. Consider paths through the model of the form $t \to P_{\theta,\eta_t}$ with $\eta_t(x, e) = v(x)f(e)(1 + t\gamma(e))$, with γ a bounded measurable function and $\int \gamma(y - g_\theta(x))dP_{\theta,\eta}(x, y) = 0$. Prove that the path has score γ [establishing the DQM property is not required] and that $\gamma(e)$ must be orthogonal in $L^2(P_{\theta,\eta})$ to the set of all functions of the form $e \cdot \psi(x)$, with $(x, y) \to \psi(x)$ in $L^2(P_{\theta,\eta})$.

(d) We admit that the efficient score $\tilde{\ell}_{\theta,\eta}$ can be written as $\tilde{\ell}_{\theta,\eta}(x, e) = e\zeta(x)$, for some square-integrable function ζ. Deduce that

$$\tilde{\ell}_{\theta,\eta}(x, e) = -e\frac{\int ef'(e)de}{\int e^2 f(e)de}\dot{g}_\theta(x).$$

END OF PAPER