Statistics 583, Second Midterm Exam

Wellner; 6/5/2009

Instructions: This is an "in class" and "closed-book" exam. Please do it completely on your own with no books or notes.

1. (36 points) **Define any three of** the following terms.
   (a) A Fréchet - differentiable functional $T : \mathcal{F} \to \mathbb{R}$ with respect to a metric d_* on $\mathcal{F}$.
   (b) A Hadamard - differentiable functional $T : \mathcal{F} \to \mathbb{R}$ with respect to a metric d_* on $\mathcal{F}$.
   (c) A metric d between distribution functions which is *compatible with respect to the empirical distribution function*. Give one example of such a metric.
   (d) The kernel estimator of a density function f on $\mathbb{R}$.
   (e) The *influence function* corresponding to a *Gateaux - differentiable functional* $T(F)$.
   (f) The *Prohorov metric* between two probability measures P and Q.

2. (30 points). Give a complete *statement* of **two** of the following results or theorems:
   (a) An example of a functional $T(F)$ which is *not* weakly continuous.
   (b) A limit theorem for the the bootstrap empirical process $\sqrt{m}(\mathbb{F}_m^* - \mathbb{F}_n)$ when $m \wedge n \to \infty$.
   (c) Any theorem about asymptotic normality of an estimator via differentiability of the corresponding statistical functional.
   (d) A result concerning the bias reduction property of the jackknife (for a functional $T(F)$ with a corresponding estimator $T_n \equiv T(\mathbb{F}_n)$).
   (e) The Politis-Romano without replacement bootstrap limit theorem.
   (f) Your favorite theorem about the consistency of a kernel density estimator.

   Do **either** problem 3 **or** problem 4 (but **not both**).

3. (40 points). Let $X_1, \ldots, X_n$ be i.i.d. with unknown density function f. Consider the kernel estimator

$$\widehat{f_n}(x) = \int \frac{1}{h_n} K\left(\frac{x-y}{h_n}\right) d\mathbb{F}_n(y)$$

   of an unknown density f at a point x with the "box-car" or uniform kernel $K(x) = 2^{-1}1_{[-1,1]}(x)$. Assume that f' exists at x and is continuous in a neighborhood of x.

(a) Compute $E\widehat{f}_n(x)$ at x explicitly in terms of F and h after taking advantage of the given kernel K.
(b) Use the result of (a) and Taylor expansion to give an expression for $\text{bias}_n(x) = E\widehat{f}_n(x) - f(x)$ in terms of $f'(x)$ assuming that $h_n \to 0$.
(c) Give a formula for $Var(\widehat{f}_n(x))$ in terms of K, F, and h_n, and then use it to give an asymptotic expression for the variance assuming that $h_n \to 0$ and $nh_n \to \infty$.
(d) Combine the results of (b) and (c) to give an expression for the risk (at the point x) for the squared error loss $R(f(x), \hat{f}_n(x)) = E_f(f(x) - \hat{f}_n(x))^2$.
(e) Based on the expression for $R(f(x), \hat{f}_n(x))$ in (d), what is the optimal choice of $h_n = Cn^{-r}$ in this case?
(f) For the optimal choice of r in (d), sketch how you would prove that

$$\sqrt{nh_n}(\widehat{f}_n(x) - f(x)) \to_d N(b(x), \sigma^2(x))$$

and identify the functions $b(x)$ and $\sigma^2(x)$ that will appear here, including their dependence on C.

4. (40 points). Suppose that $X_1, \ldots, X_n$ are i.i.d. F with density function f having $p > 2$ continuous derivatives at a point x. Suppose that we use a "kernel estimator" $\widehat{f}_n(x)$ based on the bandwidth $h = h_n$ and kernel k *of order* p: i.e. k satisfies

$$\int k(z)dz = 1, \qquad \int zk(z)dz = 0, \ldots, \int z^{p-1}k(z)dz = 0,$$
$$\int |z|^p k(z)dz < \infty, \qquad \int k^2(z)dz < \infty.$$

Such a kernel cannot be a probability density function since the condition $\int z^2 k(z)dz = 0$ forces k to take negative values; thus such kernels are sometimes called "higher-order kernels". For example, $k(x) = 8^{-1}(9 - 15x^2)1_{[-1,1]}(x)$ is a kernel of order $p = 4$.
(a) Use the same method as in class and homework to show that the resulting estimator $\widehat{f}_n(x)$ has bias given by

$$E\{\widehat{f}_n(x)\} - f(x) = \frac{h_n^p}{p!}(-1)^p \int z^p k(z) f^{(p)}(x - h_n z)dz$$

(b) Use the same methods as in class and homework to show that $\widehat{f}_n(x)$ has variance

$$Var(\widehat{f}_n(x)) = \frac{f(x)}{nh_n} \int k^2(z)dz + o((nh_n)^{-1}).$$

(c) Combine (a) and (b) to find an asymptotic expression for $E(f(x) - \widehat{f}_n(x))^2$, and hence show that $\widehat{f}_n(x)$ achieves the optimal rate of convergence $n^{p/(2p+1)}$ with optimal bandwidth choice $h_{n,opt} = n^{-1/(2p+1)}$.

Do **either** problem 5 **or** problem 6 (but **not both**).

5. (42 points). Let h be a fixed function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ and let $T(P) = \int\int h(x,y)dP(x)dP(y)$. If $X_1, \ldots, X_n$ are i.i.d. P on $(\mathcal{X}, \mathcal{A})$ with empirical measure $\mathbb{P}_n$, then $T(\mathbb{P}_n)$ is a $V-$statistic.
(a) Find the influence function of $T(P)$.
(b) What do you expect for the asymptotic variance of $\sqrt{n}(T(\mathbb{P}_n) - T(P))$?
(c) Describe how you would use the bootstrap to estimate $nVar_P(T(\mathbb{P}_n))$ and

$$H_n(x,P) \equiv Pr_P(\sqrt{n}(T(\mathbb{P}_n) - T(P)) \leq x)$$

distinguishing clearly in your description between the "ideal bootstrap" and the Monte-Carlo implementation thereof.

6. (42 points). Let F be a bivariate d.f. with marginal d.f.'s F_1 and F_2 respectively. Let $a : [0,2] \mapsto \mathbb{R}$ be differentiable, and define

$$T(F) = \int a(F_1(x) + F_2(y))dF(x,y).$$

Consider $\sqrt{n}(T(\mathbb{F}_n) - T(F))$ where $\mathbb{F}_n$ is the (bivariate) empirical d.f. of $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. F on $\mathbb{R}^2$. (a) Find the influence function of $T(F)$.
(b) What asymptotic variance do you expect to be able to demonstrate for $\sqrt{n}(T(\mathbb{F}_n) - T(F))$ when $a(z) = z^2$?
(c) Will the bootstrap work for estimation of $H_n(x,F) \equiv Pr_F(\sqrt{n}(T(\mathbb{F}_n) - T(F)) \leq x)$ when a is differentiable?