TOPICS IN STATISTICAL THEORY Part III.

Example Sheet 2 (of 3) TBB/Mich 2017

Comments and corrections to t.berrett@statslab.cam.ac.uk. Starred questions will be marked for the examples class.*

1. Let $X_1, \ldots, X_n \stackrel{i.i.d.}{\sim} \Gamma(3, \lambda)$. For a bounded, non-negative second order kernel K, compute the AMISE optimal bandwidth h_{AMISE} and, for a general estimator $\hat{\lambda}$ of λ, compare it with the normal scale bandwidth, $\hat{h}_{NS}$. Now let $\hat{\lambda}^2 = 3n / \sum_{i=1}^n (X_i - \bar{X}_n)^2$, where $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Determine the asymptotic distribution of $n^{1/2}(\hat{\lambda}^{-1} - \lambda^{-1})$. For this estimator $\hat{\lambda}$, describe the large-sample behaviour of $(\hat{h}_{NS} - h_{AMISE})/h_{AMISE}$.

2. [*In this question you may assume any required regularity conditions are satisfied.*]

A natural kernel estimator of the rth derivative, $f^{(r)}(x)$ of a density $f(x)$ is

$$\hat{f}_h^{(r)}(x) = \frac{1}{nh^{r+1}} \sum_{i=1}^n K^{(r)}\left(\frac{x - X_i}{h}\right).$$

Show that $K_h^{(r)} * f = K_h * f^{(r)}$, and hence that

$$MSE\big(\hat{f}_h^{(r)}(x)\big) = \frac{1}{nh^{2r+1}} R(K^{(r)}) f(x) + \frac{1}{4}\mu_2(K)^2 f^{(r+2)}(x)^2 h^4 + o\left(\frac{1}{nh^{2r+1}} + h^4\right).$$

Deduce that the optimal MSE is of order $n^{-4/(2r+5)}$ and argue informally that the optimal $MISE$ is of the same order.

3. Recall that the Epanechnikov kernel is a second-order kernel defined by

$$K_E(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right)\mathbb{1}_{\{|x| \leq \sqrt{5}\}},$$

and that $\mu_2(K_E) = 1$. Let K_0 be another non-negative second-order kernel with $\mu_2(K_0) = 1$. By considering $e(x) = K_0(x) - K_E(x)$, or otherwise, show that $R(K_0) \geq R(K_E)$.

4.* Let K be a symmetric kth order kernel. Explain why k must be even. Now suppose f has a bounded, continuous, square-integrable kth derivative $f^{(k)}(x)$, and that $h = h_n$ satisfies $h \to 0$ as $n \to \infty$ and $nh \to \infty$ as $n \to \infty$. Show that

$$\mathbb{E}\{\hat{f}_h(x)\} = f(x) + \frac{\mu_k(K)}{k!} f^{(k)}(x) h^k + o(h^k)$$

as $n \to \infty$, for each $x \in \mathbb{R}$. Assuming that term-by-term integration of this asymptotic expansion, and the corresponding expansion for $\mathrm{Var}\big(\hat{f}_h(x)\big)$, is valid, find the asymptotically optimal bandwidth. Deduce that

$$\inf_{h>0} AMISE(\hat{f}_h) = \frac{2k+1}{2k}\left\{\frac{2k}{(k!)^2}\mu_k(K)^2 R(K)^{2k} R(f^{(k)})\right\}^{1/(2k+1)} n^{-2k/(2k+1)}.$$

5. Suppose $X_1, \ldots, X_n$ taking values in $\mathbb{R}^d$ are i.i.d. with density function f, write P_X for the measure on $\mathbb{R}^d$ defined by $P_X(A) = \mathbb{P}(X_1 \in A)$ and write $B_x(r)$ for the closed Euclidean ball of radius r centred at $x \in \mathbb{R}^d$. Recalling the definition of the k-nearest neighbour distance $\rho_{(k)}(x)$ from lectures, show that the real random variable defined by $P = P_X\{B_x(\rho_{(k)}(x))\}$ satisfies

$$\mathbb{P}(P \leq p) = \sum_{j=k}^{n} \binom{n}{j} p^j (1-p)^{n-j}.$$

Deduce that $P \stackrel{d}{=} B \sim \text{Beta}(k, n+1-k)$. Show that $\mathbb{E}B = k/(n+1)$ and that in the asymptotic regime in which $k \to \infty$ and $k/n \to 0$ as $n \to \infty$ we have that the standard deviation $\text{sd}(B) := \sqrt{\text{Var}B} = o(k/n)$. Hence argue informally that

$$k/(n+1) \approx V_d \rho_{(k)}^d(x) f(x),$$

where $V_d = \pi^{d/2}/\Gamma(1+d/2)$ is the Lebesgue measure of $B_0(1)$.

6. Let $x_1 < \ldots < x_n$ be known real numbers, let $Y = (Y_1, \ldots, Y_n)^T$ denote a random vector with independent components and let K_h denote a scaled kernel. Show that the weighted least squares estimator $\hat{\beta}$ of $\beta = (\beta_0, \ldots, \beta_p)^T$, which minimises

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1(x_i - x) - \ldots - \beta_p(x_i - x)^p\}^2 K_h(x_i - x),$$

is of the form $\hat{\beta} = (X^T W X)^{-1} X^T W Y$, where X and W are matrices that you should specify.

7. In the random design nonparametric regression model for independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, observe that the regression function m may be expressed as

$$m(x) = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)}\, dy,$$

where $f_{X,Y}$ is the joint density of (X_1, Y_1) and f_X is the marginal density of X_1. Find the estimator of $m(x)$ that results from estimating f_X and $f_{X,Y}$ using kernel density estimators with symmetric kernel K (and the corresponding product kernel in the latter case) and a common bandwidth.

8.* Consider the random design nonparametric regression model for independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, where X_1 has marginal density f supported and continuous on $[0,1]$, and continuously differentiable on $(0,1)$. Assume the standard conditions on the regression function m, variance function v, kernel K and bandwidth h from lectures. For $r \in \mathbb{N}_0$ and $x \in (0,1)$, write

$$\hat{s}_{r,h}(x) = \frac{1}{n} \sum_{i=1}^{n} (X_i - x)^r K_h(X_i - x).$$

Show that $\hat{s}_{r,h}(x)$ has the same distribution as $\frac{h^{r-1}}{n} \sum_{i=1}^{N} Y_{ni}^r K(Y_{ni})$, where $Y_{n1}, Y_{n2}, \ldots$ are, for each n, independent and identically distributed random variables, with density supported on $[-1, 1]$ that you should specify, and are independent of N, whose distribution you should also specify. Use the tower rule of conditional expectation, the conditional variance formula $\text{Var}(S) = \mathbb{E}\{\text{Var}(S|N)\} + \text{Var}\{\mathbb{E}(S|N)\}$ and Chebychev's inequality to deduce that

$$\hat{s}_{r,h}(x) = \begin{cases} h^r \mu_r(K) f(x) + o_p(h^r) & \text{if } r \text{ is even} \\ o_p(h^r) & \text{if } r \text{ is odd.} \end{cases}$$

9. (Continuation) Deduce expressions for the conditional mean squared error of $\hat{m}_h(x; 1)$ given $X_1, \ldots, X_n$, and the conditional weighted mean integrated squared error, defined by

$$CWMISE\{\hat{m}_h(\cdot; 1)\} = \mathbb{E}\biggl(\int_{-\infty}^{\infty} \{\hat{m}_h(x; 1) - m(x)\}^2 f(x)\, dx \;\bigg|\; X_1, \ldots, X_n \biggr).$$

10. In the random design nonparametric regression model, compute the bias of the local constant estimator at an interior point, and contrast its order with that of the bias at a sequence of boundary points.

11. Consider the truncated power series representation for cubic splines with N interior knots. Let

$$g(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{N} \theta_k (x - \eta_k)_+^3.$$

Prove that the natural boundary conditions for natural cubic splines imply the following linear constraints on the coefficients:

$$\beta_2 = \beta_3 = 0, \quad \sum_{k=1}^{N} \theta_k = 0, \quad \sum_{k=1}^{N} \eta_k \theta_k = 0.$$

Hence derive the reduced basis

$$N_1(x) = 1, \quad N_2(x) = x, \quad N_{k+2}(x) = d_k(x) - d_{N-1}(x), \quad k = 1, \ldots, N-2$$

where

$$d_k(x) = \frac{(x - \eta_k)_+^3 - (x - \eta_N)_+^3}{\eta_N - \eta_k}.$$

12. Let $a \le x_1 < \ldots < x_n \le b$, and let $h_i = x_{i+1} - x_i$ for $i = 1, \ldots, n-1$. Given $\mathbf{g} = (g_1, \ldots, g_n)^T$ and $\gamma = (\gamma_2, \ldots, \gamma_{n-1})^T$, show that if there is a natural cubic spline g with $g(x_i) = g_i$ for $i = 1, \ldots, n$ and $g''(x_i) = \gamma_i$ for $i = 2, \ldots, n-1$ then

$$g(x) = \frac{(x - x_i) g_{i+1} + (x_{i+1} - x) g_i}{h_i} - \frac{1}{6}(x - x_i)(x_{i+1} - x) \biggl\{ \Bigl(1 + \frac{x - x_i}{h_i}\Bigr) \gamma_{i+1} + \Bigl(1 + \frac{x_{i+1} - x}{h_i}\Bigr) \gamma_i \biggr\}$$

for $x \in [x_i, x_{i+1}]$ and $i = 1, \ldots, n-1$. Find the corresponding expressions for g on $[a, x_1]$ and $[x_n, b]$.