TOPICS IN STATISTICAL THEORY Part III.

Example Sheet 2 (of 3) YZ/Lent 2020

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

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

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

Let $r \geq 0$ be an integer. A natural kernel estimator of the rth derivative, $f^{(r)}(x)$ of a density $f(x)$ is

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

Show that $K_h^{(r)} * f = K_h * f^{(r)}$.

Now let $\beta > r$ be a real number and let l be the unique integer such that $l - 1 < \beta \leq l$ and consider the class of functions

$$C^{\beta}_{den}(M) = \{f \text{ density} : f \in C^{l-1}, |f^{(l-1)}(x) - f^{(l-1)}(y)| \leq M|x - y|^{\beta - l + 1} \forall x, y \in \mathbb{R}\}.$$

Using a kernel of order $l - r$ such that

$$|\mu|_{\beta - r}(K) = \int |z|^{\beta - r}|K(z)|dz < \infty,$$

show that, for an appropriate choice of kernel K,

$$\inf_{h>0} \sup_{f \in C^{\beta}_{den}(M)} MSE\big(\widehat{f}_h^{(r)}(x)\big) \leq C(M, \beta, r, K) n^{-\frac{2(\beta - r)}{2\beta + 1}}.$$

You may use (without needing to provide a proof) that such a kernel exists and that

$$C_0(\beta, M) = \sup_{f \in C^{\beta}_{den}(M)} \|f\|_\infty < \infty.$$

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. a) Let $f(x)$ be the uniform density on $[0, 1]$ and let $g(x) = 2x$ on $[0, 1]$. Let the kernel K be uniform on $[-1, 1]$. For all $x \in \mathbb{R}$ and all $h < 1/2$, compute the bias of the kernel density estimators $\widehat{f}_h(x)$ and $\widehat{g}_h(x)$. What do you observe?

b) Define $\tilde{X}_i = -X_i$ and compute the bias of

$$\widehat{f}_h^{\dagger}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) + K_h(x - \tilde{X}_i)$$

as an estimator of $f(x)$, for all x and all $h < 1/2$. What do you observe? What extra knowledge one needs about the density f in order to apply this correction method?

5. 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 $\widehat{\beta}$ 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 $\widehat{\beta} = (X^T W X)^{-1} X^T W Y$, where X and W are matrices that you should specify.

6. 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.

7*. 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 $[0, 1]$. 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

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

Estimate the expectation and variance of $\widehat{s}_{r,h}(x)$ and use Chebichev's inequality to deduce that

$$\widehat{s}_{r,h}(x) = h^r \mu_r(K) f(x) + o_p(h^r).$$

Assuming that K is symmetric, what can be said about the magnitude of $\widehat{s}_{r,h}(x)$ when r is odd and when r is even?

8. (Continuation) Deduce expressions for the conditional mean squared error of $\widehat{m}_h(x;1)$ given $X_1,\dots,X_n$, and, informally, the conditional weighted mean integrated squared error

$$CWMISE\{\widehat{m}_h(\cdot;1)\} = \mathbb{E}\left(\int_{-\infty}^{\infty} \{\widehat{m}_h(x;1) - m(x)\}^2 f(x)\,dx \;\middle|\; X_1,\dots,X_n\right).$$

9. 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.

10. Consider a fixed design nonparametric regression model with assumptions (B0)–(B3) from lectures, and assume further that K is Lipschitz continuous on $[-1,1]$. Show, with full justification, that

$$\int_0^1 MSE(\widehat{m}_h(x,1))dx = \frac{R(K)\int_0^1 v(x)dx}{nh} + \frac{1}{4}h^4\mu_2^2(K)R(m'') + o\left(\frac{1}{nh} + h^4\right).$$

A possible strategy is as follows. a) Show that for $r \geq 0$ and $g_r(u) = u^r K(u)$, one has

$$h^{-r}s_{r,h}(x) = A_{r,h}(x) + B_{r,h}(x), \qquad A_{r,h}(x) := \int_{-x/h}^{(1-x)/h} g(t)dt,$$

where $B_{r,h}(x) \to 0$ uniformly in x, $|A_{r,h}(x)| \leq C(r,K)$ for all $x \in [0,1]$ and $A_{r,h}(x)$ is the same for all $h \leq x \leq 1-h$.

b) Deduce that to conclude, it suffices to bound the denominator $s_{2,h}(x)s_{0,h}(x) - s_{1,h}^2(x)$ from below by a constant times h^2, uniformly over $x \in [0,1]$.

c) Define the matrix

$$B_x = \begin{pmatrix} \int_{-x/h}^{(1-x)/h} K(u)du & \int_{-x/h}^{(1-x)/h} uK(u)du \\ \int_{-x/h}^{(1-x)/h} uK(u)du & \int_{-x/h}^{(1-x)/h} u^2K(u)du \end{pmatrix}.$$

Show that for all $x \in [0,1]$, either $B_x - B_0$ is positive semidefinite, or $B_x - B_1$ is positive semidefinite. Show that both B_0 and B_1 are positive definite and conclude that the determinant of B_x is bounded below by a positive constant, uniformly in $x \in [0,1]$.

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

$$g(x) = \frac{(x-x_i)g_{i+1} + (x_{i+1}-x)g_i}{h_i} - \frac{1}{6}(x-x_i)(x_{i+1}-x)\left\{\left(1+\frac{x-x_i}{h_i}\right)\gamma_{i+1} + \left(1+\frac{x_{i+1}-x}{h_i}\right)\gamma_i\right\}$$

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 $[x_n, b]$.