Lecture course *Statistics II*
Winter semester 2016/17
Ruprecht-Karls-Universität Heidelberg

Prof. Dr. Jan JOHANNES
Xavier LOIZEAU

Exercise sheet 8

Exercise 1. Consider i.i.d. r.v.'s $(Y, Z), (Y_1, Z_1), (Y_2, Z_2), \dots$ obeying a non-parametric regression model $\mathbb{E}_f(Y|Z) = f(Z)$ and satisfying the Assumptions §5.3.1. Denote by $\mathbb{F}$ the c.d.f. of Z and assume that $\mathbb{F}$ is continuous and admits an inverse denoted by $\mathbb{F}^{-1}$. We set $\ell := f \circ \mathbb{F}^{-1}(z)$ and note, that $f = \ell \circ \mathbb{F}$. Given a kernel K and a bandwidth h consider the kernel estimator $\widehat{\ell}_h(z) = \frac{1}{nh} \sum_{i=1}^n Y_i K\big(\frac{\mathbb{F}(Z_i) - z}{h}\big)$ of f.

(a) Derive an upper bound for $\mathrm{bias}(z) = \mathbb{E}(\widehat{\ell}_h(z)) - \ell(z)$ if f belongs to an Hölder class and K is a kernel of appropriate order.

(b) Derive an upper bound for $\mathbb{V}\mathrm{ar}(\widehat{\ell}_h(z))$ assuming that $\|L^\infty\| \, f < \infty$ and $\|K\|_{L^2} < \infty$.

(c) Find an upper bound for $\mathrm{MSE}(z) = \mathbb{E}|\widehat{\ell}_h(z) - \ell(z)|^2$ depending on the bandwidth. Select an optimal value for the bandwidth and derive the associated upper bound.

(d) Propose an estimator of f if $\mathbb{F}$ is known and if it isn't.

(4 points)

Exercise 2. Consider i.i.d. r.v.'s $(X, U), (X_1, U_1), (X_2, U_2), \dots$ where U is uniformly distributed on the interval $[0, 1]$, i.e., $U \sim \mathfrak{U}([0, 1])$ and X is non-negative with unknown density $\mathbb{p}$. Moreover, X and U are independent. Let $\mathbb{p}^y$ denote the common density of the r.v.'s $Y := XU, Y_1 = X_1 U_1, \dots$. Given a kernel K with derivative $\dot{K}$ and a bandwidth h consider the kernel estimator $\widehat{\mathbb{p}}_h(z) = \frac{1}{nh} \sum_{i=1}^n \{\frac{Y_i}{h} \dot{K}\big(\frac{Y_i - x}{h}\big) + K\big(\frac{Y_i - x}{h}\big)\}$ of $\mathbb{p}$.

(a) Let g be a function with derivative $\dot{g}$ such that g, $y \mapsto \mathrm{Id}(y) g(y) := y g(y)$ and $\mathrm{Id}\, \dot{g}$ are bounded. Show that, $\mathbb{E}(Y \dot{g}(Y) + g(Y)) = \mathbb{E}(g(X))$.
Hint: First show $\mathbb{E}(g(Y)) = \int_0^\infty g(v) \int_v^\infty \frac{\mathbb{p}(x)}{x} dx) dv$ *and conclude* $\mathbb{p}^y(y) = \int_y^\infty \frac{\mathbb{p}(x)}{x} dx$ *and* $\dot{\mathbb{p}}^y(y) = -\frac{\mathbb{p}(y)}{y}$.

(b) Derive an upper bound for $\mathrm{bias}(x)$ if $\mathbb{p}$ is three-times differentiable with bounded third derivative and the kernel is of order 2 such that $\int |u|^3 |K(u)| du < \infty$.

(c) Derive an upper bound for $\mathbb{V}\mathrm{ar}(\widehat{\mathbb{p}}(x))$ assuming that $\|\mathbb{p}^y\|_{L^\infty} < \infty$, $\|K\|_{L^2} < \infty$, $\|\mathrm{Id}^2\, \mathbb{p}^y\|_{L^\infty} < \infty$, $\|\dot{K}\|_{L^2} < \infty$.

(d) Find an upper bound for the $\mathrm{MSE}(x)$ depending on the bandwidth. Select an optimal value for the bandwidth and derive the associated upper bound of the $\mathrm{MSE}(x)$. What do you notice?

(4 points)

Exercise 3. Consider r.v.'s $Y_i = f(x_i) + \varepsilon_i$, $i \in [\![1, n]\!]$, where $x_1, \dots, x_n$ are $\mathbb{R}^d$-valued deterministic covariates, $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. real-valued centred r.v.'s with finite variance

σ^2, and $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function. Denote by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$. Given a bandwidth $h > 0$ and the kernel $K := \mathbb{1}_{[0,1]}$ for each $x \in \mathbb{R}^d$ such that $\sum_{i=1}^n K\big(\frac{\|x_i - x\|}{h}\big) > 0$, define a locally constant estimator of $f(x)$ by:

$$\widehat{f}_h(x) = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^n (Y_i - a)^2 K\big(\tfrac{\|x_i - x\|}{h}\big).$$

(a) Give an explicit form for $\widehat{f}_h(x)$.

(b) Let f be Lipschitz with constant $L > 0$, i.e. $|f(x_1) - f(x_2)| \leqslant L\|x_1 - x_2\|$, for all $x_1, x_2 \in \mathbb{R}^d$. Show that $|\mathbb{E}\,[\widehat{f}_h(x)] - f(x)| \leqslant Lh$, for all $x \in \mathbb{R}^d$.

(c) Given a ball $B_h := \{u \in \mathbb{R}^d : \|u\| \leqslant h\}$ in $\mathbb{R}^d$ and denoting by $\mathrm{Vol}(B_h)$ its volume suppose that there exist a constant $C > 0$ such that $\sum_{i=1}^n \mathbb{1}_{B_h}(x_i - x) \geqslant C\,n\,\mathrm{Vol}(B_h)$. Show that there is D depending on C and d such that $\mathbb{V}\mathrm{ar}[\widehat{f}_h(x)] \leqslant (nh^d)^{-1} D\sigma^2$.

(d) Deduce from (b) and (c) an upper bound for the $\mathrm{MSE}(x)$ depending on the bandwidth. Select an optimal value for the bandwidth and compute the value of the associated $\mathrm{MSE}(x)$. How does d influences this bound ? Give an interpretation for this.

Hint : you may show $\{\mathbb{E}\,[\widehat{f}_h(x)] - f(x)\} \sum_{i=1}^n K\big(\frac{\|x_i - x\|}{h}\big) = \sum_{i=1}^n \{f(x_i) - f(x)\} K\big(\frac{\|x_i - x\|}{h}\big)$ *and* $\mathbb{V}\mathrm{ar}[\widehat{f}_h(x)] \sum_{i=1}^n K\big(\frac{\|x_i - x\|}{h}\big) \leqslant \sigma^2$. (4 points)

Exercise 4. Consider r.v.'s $Y_i = f(X_i) + \varepsilon_i$, $i \in [\![1, n]\!]$, where $X_1, \dots, X_n$ are $\mathbb{R}^d$-valued r.v.'s, $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. real-valued centred r.v.'s with finite variance σ^2 and $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown regression function. Suppose that f admits at least l derivatives with Lipschitz property with constant $L > 0$. Given a bandwidth $h > 0$, a kernel K and $U(x) := (1, x, \dots, \frac{x^l}{l!})$ define a local polynomial estimator of degree l by $\widehat{f}_h(x) := \widehat{\theta}^t U(0)$ where

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^{l+1}} \sum_{i=1}^n |Y_i - \theta^t U\big(\tfrac{X_i - x}{h}\big)|^2 K\big(\tfrac{X_i - x}{h}\big).$$

(a) Give an explicit form for $W_i(x)$ such that $\widehat{f}_h(x) = \sum_{i=1}^n Y_i W_i(x)$.

(b) Show that $\widehat{f}_h(x)$ reproduces polynomials of order lower or equal to l, that is to say, if Q is a polynomial of order lower that l, then $\sum_{i=1}^n Q(X_i)\,W_i(x) = Q(x)$.

(c) Deduce from this that $\sum_{i=1}^n W_i(x) = 1$ and $\sum_{i=1}^n (X_i - x)^k\,W_i(x) = 0$ for all $k \in [\![1, l]\!]$.

(d) Suggest an estimator of the k^{th} derivative of f depending on U, $\widehat{\theta}$ and h. (4 points)

Handing in during the lecture on **Friday, January 20, 2017** in **fixed groups of two**.