Advanced Statistics

Exam (length 1h30)

Notes and duplicated copy are allowed, computers and tablets are prohibited.

Kernel estimation of a smooth density

Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ be an i.i.d. sample of probability distribution $F(dx)$ on $\mathbb{R}$. We suppose that F has a bounded density $f(x)$ with respect to Lebesgue measure, twice differentiable and with a Lipschitz second derivative : $\forall (x, y) \in \mathbb{R}^2$,

$$|f''(x) - f''(y)| \leq L|x - y|.$$

The estimation problem considered here is that of the density f at a given point $x_0 \in \mathbb{R}$ for the *quadratic risk*.

1. Let $K : \mathbb{R} \to \mathbb{R}$ be a continuous symmetric kernel (*i.e.* even, $K(x) = K(-x)$) such that :

$$\int_{x\in\mathbb{R}} K(x)dx = +1, \ \int_{x\in\mathbb{R}} x^2 K(x)dx = 0, \ C_K = \int_{x\in\mathbb{R}} |x^3 K(x)|dx < +\infty$$

   Give the explicit form of the kernel estimator of $f(x_0)$, denoted by $\widehat{f}_h$ based on the sample $\mathcal{D}_n$, kernel K and a window size $h > 0$. Prove that its bias at the point $x_0 \in \mathbb{R}$ is bounded by Ch^3, where C is a constant that depends on L and C_K only.

2. Let α and α' be two strictly positive real numbers and set : $\forall u \in \mathbb{R}$,

$$\kappa(u) = \frac{\alpha}{2}\mathbf{1}\{u \in [-1, 1]\} + \frac{\alpha'}{2}\mathbf{1}\{u \in [-2, 2]\}.$$

   Find α and α' so that κ fulfills the conditions stipulated in the previous question.

3. We suppose in addition that $K(0)f(x_0) > 0$ and $\int_{x\in\mathbb{R}} K^2(x)dx < +\infty$. Show that the variance of the random variable $\widehat{f}_h(x_0)$ (random, as a function of the random observations $\mathcal{D}_n$) is lower/upper bounded as follows :

$$\frac{c_1}{nh} \leq var\left(\widehat{f}_h(x_0)\right) \leq \frac{c_2}{nh},$$

   c_1 and c_2 being appropriate constants.

4. Deduce that an optimal choice for the bandwidth is $h_n \sim n^{-1}$. What is the order of the rate of convergence of the resulting estimator $\widehat{f}_{h_n}(x_0)$?

Multiplicative regression model

Suppose we observe $(Y_i, X_i)_{i=1,\dots,n}$ according to the (multiplicative) model :

$$Y_i = \sigma(X_i)\varepsilon_i, \qquad i = 1, \dots, n,$$

where $(X_i, \varepsilon_i)_{i=1,\dots,n}$ is an independent and identically distributed sequence of random variables valued in $[0,1] \times \mathbb{R}$ and $\sigma : [0,1] \to \mathbb{R}_+$ is a bounded function. We suppose that for each $i = 1, \dots, n$, X_i is independent from ε_i and that $\mathbb{E}[\varepsilon_1^2] = 1$. Let $f : [0,1] \to \mathbb{R}_+$ be the density of X_1 and define

$$\hat{\gamma}_n(g) = \int_0^1 g(x)^2 dx - \frac{2}{n}\sum_{i=1}^n Y_i^2 g(X_i)$$

1. Express $\mathbb{E}[\hat{\gamma}_n(g)]$ with the help of g, f and σ. Explain how $\hat{\gamma}_n(g)$ can be used to estimate $\sigma^2 f$.
2. Let S_D be the linear subspace of $L^2([0,1], d\lambda)$ generated by $(\varphi_1, \dots, \varphi_D)$ where for any (k, ℓ), $\int \varphi_k(x)\varphi_\ell(x)dx$ equal 0 if $k \neq \ell$ and equal 1 if $k = \ell$. Any function in $\psi \in S_D$ can be written as
$$\psi(x) = \sum_{k=1}^D c_k \varphi_k(x), \quad \forall x \in [0,1]$$
where the c_k are unique and are called the coefficients of ψ in S_D. Express $\hat{\psi}_{n,D} = \arg\min_{\psi \in S_D} \hat{\gamma}_n(\psi)$ using its coefficients, $\hat{\alpha}_k$, in S_D.
3. Let $x \in [0,1]$, compute $\mathbb{E}[\hat{\psi}_{n,D}(x)]$ in terms of the coefficients $\alpha_k = \int \sigma(x)^2 \varphi_k(x) f(x) dx$, $k = 1, \dots, D$.
4. Let $e_D(x) = \left(\sum_{k=1}^D \alpha_k \varphi_k(x)\right) - \sigma(x)^2 f(x)$. Compute the integrated error : $I_e = \mathbb{E}\int(\hat{\psi}_{n,D}(x) - \sigma(x)^2 f(x))^2 dx$ with the help of e_D, $v_k = \mathrm{Var}(Y_1^2 \varphi_k(X_1))$, $k = 1, \dots, D$, and n.
5. Suppose that $\mathbb{E}[\varepsilon_1^4] < \infty$. Show that $I_e \leq CD/n + B_D$ where C and B_D shall be specified. Conclude on the choice of D with respect to n.