Advanced Statistics

Exam (length 2h)

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

Throughout the exam subject, by $\mathbb{I}\{\mathcal{E}\}$ is meant the indicator function of any event $\mathcal{E}$, by $|z|$ the modulus of any complex number z.

Ex. 1 - Nonparametric estimation of a characteristic function

The goal is to estimate the characteristic function $t \in \mathbb{R} \mapsto \Phi_X(t) = \mathbb{E}[e^{itX}]$ of a real valued r.v. X based on the observation of $n \geq 1$ i.i.d. copies of X : $X_1, \ \ldots, \ X_n$.

1. Explain the statistical method leading to consider

$$\widehat{\Phi}_n(t) = \frac{1}{n}\sum_{k=1}^{n} \exp\left(itX_k\right)$$

as (nonparametric) estimator of $\Phi_X(t)$.

2. Fix $t \in \mathbb{R}$. Show that $\widehat{\Phi}_n(t)$ is an unbiased estimator of $\Phi_X(t)$ and compute its variance

$$var(\widehat{\Phi}_n(t)) = \mathbb{E}\left[\left|\widehat{\Phi}_n(t) - \mathbb{E}[\widehat{\Phi}_n(t)]\right|^2\right].$$

3. Deduce the order of magnitude of the pointwise quadratic risk of the estimator $\widehat{\Phi}_n(t)$, namely

$$\mathbb{E}\left[\left|\widehat{\Phi}_n(t) - \Phi_X(t)\right|^2\right].$$

Ex. 2 - Cross-validation for the histogram

Consider an i.i.d. sample $X_1, \ \ldots, \ X_n$ with support included in $[0,1]$ and density $f \in L_2([0,1])$ (*i.e.* such that $||f||_2^2 = \int_0^1 f^2(x)dx < +\infty$) w.r.t. Lebesgue measure on $[0,1]$. Let $h > 0$ and consider the histogram estimator of the density f with bin width $h = 1/m$, where $m \geq 1$:

$$\widehat{f}_{h,n}(x) = \frac{1}{h}\sum_{k=1}^{m} \widehat{p}_k \mathbb{I}\left\{x \in [(k-1)/m, \ k/m[\right\},$$

where, for $1 \leq k \leq m$, we set :

$$\widehat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{X_i \in [(k-1)/m, \ k/m[\right\}.$$

Denote by $R(h)$ the integrated quadratic risk (on $[0,1]$) of the estimator $\widehat{f}_{h,n}$ and set

$$J(h) = R(h) - ||f||_2^2.$$

1. Calculate the bias and the variance of $\widehat{f}_{h,n}$. Calculate next $J(h)$.
2. Show that
$$\widehat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{k=1}^{m}\widehat{p}_k^2$$
is an unbiased estimator of $J(h)$.

EX 3. - RATE FOR THE RISK OF THE HISTOGRAM

We place ourselves in the same setting as in the exercise above and re-use its notations. We assume in addition that the density f satisfies the Hölder property : there exists a constant $C > 0$ s.t.
$$\forall (x, x') \in [0,1]^2, \ \ |f(x) - f(x')| \leq C|x - x'|^{\alpha}$$

1. Compute the orthogonal projection f_h of f onto the subspace of the Hilbert space $L_2([0,1])$ composed of functions that are constant almost-everywhere on each interval $[(k-1)/m, \ k/m[$ for all $k \in \{1, \ \ldots, \ m\}$.
2. Prove that
$$||f - f_h||_2 \leq C^2 m^{-2\alpha}.$$
3. Deduce from the bound above an upper bound for the integrated quadratic risk of the estimator $\widehat{f}_{h,n}$ and propose a value for the parameter m so as to minimize the upper bound.

EX. 4 - MULTIPLICATIVE REGRESSION MODEL

Suppose we observe $(Y_1, \ X_1), \ \ldots, \ (Y_n, \ X_n)$ such that :
$$Y_i = \sigma(X_i)\varepsilon_i, \qquad i = 1, \ \ldots, n,$$
where the (X_i, ε_i)'s are independent and identically distributed random pairs, valued in $[0,1] \times \mathbb{R}$ and $\sigma : [0,1] \to \mathbb{R}_+$ is a bounded function : there exists $C < +\infty$ s.t. $\sup_{x\in[0,1]} \sigma^2(x) \leq C$. We suppose that $(X_1, \ \ldots, \ X_n)$ is independent from $(\varepsilon_1, \ \ldots, \ \varepsilon_n)$, as well as $\mathbb{E}[\varepsilon_1] = m < +\infty$ and $\mathbb{E}[\varepsilon_1^2] = 1$. Let $F : [0,1] \to [0,1]$ be the cumulative distribution function of X_1 (*i.e.* $F(x) = \mathbb{P}\{X_1 \leq x\}$) and assume it is bijective. The goal pursued here is to estimate $\ell = \sigma^2 \circ F^{-1}$ using a kernel smoothing method, when F is known. Let $K : \mathbb{R} \to \mathbb{R}$ be a Parzen-Rosenblatt kernel function, $h > 0$ a bandwidth and define
$$\widehat{\ell}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} Y_i^2 K\left(\frac{F(X_i) - x}{h}\right),$$
for any $x \in [0,1]$.

1. Show that the variance of the statistic defined above is bounded as follows
$$var\left(\widehat{\ell}_h(x)\right) \leq \frac{C^2 \int K^2(t)dtm}{nh}.$$
2. Express $\mathbb{E}[\widehat{\ell}_h(x)]$ depending on $K, \ h, \ \ell$ and x only.

3. Suppose now in addition that ℓ is of class $\mathcal{C}^3$ and that there exists $M < +\infty$ s.t. $|\ell'''(x)| \leq M$ for all x in $[0,1]$. Assume also that the kernel K is supported on $[-1,\ +1]$ and is of order 2 and set $C_K = \int |t|^3 |K(t)| dt < +\infty$. Find constants $\kappa > 0$ and $\beta > 0$ such that : $\forall h \in]0,\ 1/2[$, $\forall x \in [h,\ 1-h]$,

$$\left| \mathbb{E}[\widehat{\ell}_h(x)] - \ell(x) \right| \leq \kappa h^\beta.$$

4. Deduce an upper bound for the pointwise quadratic risk of the estimator $\widehat{\ell}_h(x)$ of $\ell(x)$ and find the bandwidth h^* that minimizes this upper bound.
5. Deduce an estimator of σ^2 when F is known.
6. Propose an estimator of σ^2 when F is unknown.