TD Estimation non paramétrique

Master 1 MIMSE – 2015/2016

Exercice 1 : *Fonction caractéristique empirique*

On cherche à estimer la fonction caractéristique ϕ_X d'un échantillon (X_n) constitué de n variables aléatoires indépendantes et identiquement distribuées. La fonction caractéristique empirique $\widehat{\phi}_X$ au point $t \in \mathbb{R}$ est donné par

$$\widehat{\phi}_X(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{\mathrm{i}tX_i}.$$

1) Montrer que $\widehat{\phi}_X(t)$ est un estimateur sans biais de $\phi_X(t)$.
2) Calculer la variance de $\widehat{\phi}_X(t)$.
3) En déduire l'ordre de grandeur de son risque quadratique ponctuel.

Exercice 2 : *Validation croisée pour l'histogramme*

Soit $X_1 \dots X_n$ un échantillon i.i.d de densité f appartenant à $L^2([0,1])$. L'estimateur par histogramme de f s'écrit pour tout $x \in [0,1]$:

$$\widehat{f}_h(x) = \frac{1}{h}\sum_{j=1}^{m} \widehat{p}_j \, \mathbf{1}_{C_j}(x) \qquad \text{où} \quad \widehat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{C_j}(X_i).$$

On note $R(h)$ le risque quadratique intégré de $\widehat{f}_h$ et $J(h) = R(h) - \|f\|_2^2$.

1) Calculer le biais et la variance de $\widehat{f}_h(x)$. Puis calculer $J(h)$.
2) Montrer que

$$\widehat{J}(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{j=1}^{m} \widehat{p}_j^{\,2}$$

   est un estimateur sans biais de $J(h)$.

Exercice 3 : *Vitesse de convergence pour l'histogramme*

Supposons que f est Hölderienne d'ordre α : pour tous x,y,

$$|f(x) - f(y)| \leq C_f |x-y|^{\alpha}$$

où C_f est une constante qui dépend uniquement de f.

1) Donner l'expression formelle dans la base d'histogrammes réguliers $\{\varphi_1 \dots \varphi_D\}$ du projeté orthogonal f_D de f.
2) Montrer que le biais vérifie $\|f - f_D\|^2 \leq C_f^2 D^{-2\alpha}$.
3) En déduire un majorant du risque quadratique puis déterminer la valeur D_{opt} de D qui minimise cette borne.

Exercice 4 : *Estimateur par projection dans un modèle multiplicatif*

On considère le modèle
$$Y_i = \sigma(X_i)\varepsilon_i, \quad i = 1 \dots n$$
où les (X_i) sont iid de densité commune f et les (ε_i) sont iid centrées de variance 1 et indépendantes des (X_i). On notera $m_4 = \mathbb{E}[\varepsilon_i]$. On observe les couples $(X_i, Y_i)_i$. On pose pour toute fonction t de $L^2([0,1])$
$$\gamma_n(t) = \|t\|_2^2 - \frac{2}{n}\sum_{i=1}^n Y_i^2 t(X_i).$$

1) Calculer $\mathbb{E}[\gamma_n(t)]$ et en déduire comment estimer $\psi := \sigma^2 f$.
2) On note $\widehat{\psi}_D = \arg\min_{t\in S_D} \gamma_n(t)$ où S_D est un sev de dimension D de $L^2([0,1])$ engendré par une base orthonormale $\varphi_1, \dots, \varphi_D$. Exprimer les coefficients de $\widehat{\psi}_D$ dans cette base.
3) Donner l'expression de ψ_D, la projection de ψ sur S_D, et montrer que $\mathbb{E}[\widehat{\psi}_D] = \psi_D$.
4) Calculer puis majorer le risque quadratique intégré de $\widehat{\psi}_D$.
5) Donner un estimateur de σ^2 déduit de $\widehat{\psi}_D$.

Exercice 5 : *Transformée de Fourier de la fonction caractéristique empirique*

On se place dans le cadre de l'exercice 1. On considère une pondération ψ telle que $\psi(0) = 1$ et qui décroît ensuite vers 0. On note $\psi_h(t) = \psi(h_n t)$, pour tout $t \in \mathbb{R}$. La transformée de Fourier de l'estimateur de ϕ_X associé à ψ_h est donnée, pour $x \in \mathbb{R}$, par
$$\widehat{f}_n(x) = \int_{\mathbb{R}} \widehat{\phi}_X(t)\, \psi(h_n t)\, \mathrm{e}^{-\mathrm{i}tx}\mathrm{d}t.$$
Montrer que $\widehat{f}_n$ peut s'écrire comme un estimateur à noyau, que l'on précisera, de la densité f.

Exercice 6 : *Estimation par déformation dans un modèle multiplicatif*

On reprend le cadre de l'exercice 4. On suppose de plus que la fonction de répartition commune aux X_i est bijective. Soit K un noyau, on se propose d'estimer $l := \sigma^2 \circ F^{-1}$ par
$$\widehat{l}_h(x) = \frac{1}{nh}\sum_{i=1}^n Y_i^2 K\left(\frac{F(X_i) - x}{h}\right).$$

1) Montrer que si σ^2 est bornée par M et $\|K\|_2^2 = N$ alors $\mathrm{Var}(\widehat{l}_h(x)) \leq \frac{M^2 N m_4}{nh}$.
2) Exprimer $\mathbb{E}[\widehat{l}_h(x)]$ en fonction de K, h, l et x uniquement.
3) On suppose l de classe $\mathcal{C}^3$ et de dérivée troisième bornée, et K d'ordre 2 à support dans $[-1,1]$ tel que $C_K := \int |u|^3 |K(u)|\, du < +\infty$. Déterminer C et β tels que
$$\forall h \in \,]0, 1/2[\,, \forall x \in [h, 1-h]\,, \quad |b_h(x)| \leq Ch^\beta$$
4) En déduire une majoration du risque ponctuel et déterminer h_{opt} qui la minimise.
5) En déduire un estimateur de σ^2 si F est connue.
6) Quelle solution peut-on proposer quand F est inconnue ?