Statistique non-asymptotique - Examen

14 décembre 2016, 9h-12h.

Les appareils électroniques et les documents ne sont pas autorisés. Le sujet comporte **deux** pages.

Partie 1. Statistique paramétrique (S. Pergamenchtchikov)

1. Simple nonrandom regression (4 points)

1. Give the definition for the simple nonrandom regression model with unknown parameter λ.
2. Show that the least square estimator $\widehat{\lambda}_n$ is optimal in the mean square accuracy sense in the class of all unbiased linear estimators, i.e. for any $n \geq 1$

$$\mathbf{E}\,(\widehat{\lambda}_n - \lambda)^2 \leq \mathbf{E}\,(\widetilde{\lambda}_n - \lambda)^2\,,$$

where $\widetilde{\lambda}_n$ is a linear estimator constructed on the n observations.

2. Sequential estimation (4 points)

Let us $(y_j)_{j\geq 1}$ be the first order autoregressive process defined as

$$y_j = \lambda y_{j-1} + \xi_j\,, \quad y_0 = 0\,.$$

Here λ is an unknown constant parameter, $(\xi_j)_{j\geq 1}$ is i.i.d. sequence of random variables with $\mathbf{E}\,\xi_j = 0$ and $\mathbf{E}\,\xi_j^2 = 1$. We set $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_j = \sigma\{y_1, \ldots, y_j\}$ for any $j \geq 1$.

Let for some $H > 0$

$$\tau_H = \inf\{n \geq 1\,:\, \sum_{j=1}^{n} y_{j-1}^2 \geq H\}\,.$$

1. Show that $\mathbf{P}(\tau_H < \infty) = 1$ for any $H > 0$.
2. Give the definition for the stopping times. Show that for any $H > 0$ the moment τ_H is the stopping time with respect to $(\mathcal{F}_j)_{j\geq 0}$.
3. Write the sequential estimator for the parameter λ.

3. Tests (2 points)

Let us $(X_j)_{j\geq 1}$ be an i.i.d. sequence of random variables having a density.

1. Write the test problem in the sequential setting.
2. Write the Wald test.

Partie 2. Statistique non-paramétrique (G. Chagny)

Question de cours (1 point).

On observe $(X_i)_{i=1,\ldots,n}$ un échantillon de variables indépendantes identiquement distribuées de densité f inconnue appartenant à $L^2(I)$, pour I un intervalle de $\mathbb{R}$. Donner la définition de l'estimateur de f obtenu par la méthode de projection sur un sous espace vectoriel $S_D \subset L^2(I)$ engendré par une base orthonormée $\{\varphi_1, \ldots, \varphi_D\}$, (avec $\varphi_j : I \to \mathbb{R}$).

Exercice. Estimation d'une densité bivariée par noyaux (9 points).

On observe n couples de variables aléatoires réelles $(X_1, Y_1), \ldots, (X_n, Y_n)$ indépendants identiquement distribués de densité inconnue $f : \mathbb{R}^2 \to \mathbb{R}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$.

Pour tout $h = (h_1, h_2) \in (\mathbb{R}_+^*)^2$, on considère l'estimateur $\hat{f}_h$ de f défini, pour tout $(x, y) \in \mathbb{R}^2$, par

$$\hat{f}_h(x,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} Q\left(\frac{x - X_i}{h_1}, \frac{y - Y_i}{h_2}\right),$$

où $Q : \mathbb{R}^2 \to \mathbb{R}$ est un noyau *i.e.* une fonction intégrable telle que $\iint_{\mathbb{R}^2} Q(u,v)dudv = 1$. On supposera que

$$\|Q\|_2^2 := \iint_{\mathbb{R}^2} Q^2(u,v)dudv < \infty, \quad \iint_{\mathbb{R}^2} |Q(u,v)||v|dudv < \infty, \quad \iint_{\mathbb{R}^2} Q(u,v)|u|^{1/2}dudv < \infty.$$

On considère le risque quadratique ponctuel en $(x_0, y_0) \in \mathbb{R}^2$ de $\hat{f}_h$,

$$R_{(x_0,y_0)}(\hat{f}_h, f) = \mathbb{E}\left[\left(\hat{f}_h(x_0,y_0) - f(x_0,y_0)\right)^2\right].$$

On note, pour tout $(u, v) \in \mathbb{R}^2$,

$$Q_h(u,v) = \frac{1}{h_1h_2} Q\left(\frac{u}{h_1}, \frac{v}{h_2}\right).$$

Pour toutes fonctions f_1 et f_2 de deux variables, on note $\star$ le produit de convolution entre f_1 et f_2, dès que celui-ci a un sens,

$$(f_1 \star f_2)(x,y) = \iint_{\mathbb{R}^2} f_1(x-u, y-v)f_2'u,v)dudv, \quad (x,y) \in \mathbb{R}^2.$$

1. Justifier que si $K : \mathbb{R} \to \mathbb{R}$ est un noyau sur $\mathbb{R}$, alors la fonction définie pour tout $(u, v) \in \mathbb{R}^2$ par $Q(u, v) = K(u)K(v)$ est bien un noyau sur $\mathbb{R}^2$.

2. Démontrer que $R_{(x_0,y_0)}(\hat{f}_h, f) = \left(\mathbb{E}\left[\hat{f}_h(x_0,y_0)\right] - f(x_0,y_0)\right)^2 + \mathrm{Var}(\hat{f}_h(x_0,y_0))$.

3. **(a)** Montrer que $\mathbb{E}\left[\hat{f}_h(x_0,y_0)\right] = Q_h \star f(x_0,y_0)$.

(b) Montrer que, si la fonction f est supposée bornée par $M > 0$ sur $\mathbb{R}^2$,

$$\mathrm{Var}(\hat{f}_h(x_0,y_0)) \leq \frac{M\|Q\|_2^2}{nh_1h_2}.$$

4. On suppose de plus que la densité f vérifie la propriété suivante :

$$\forall((u,v),(u',v')) \in (\mathbb{R}^2)^2, \quad |f(u,v) - f(u',v')| \leq |u-u'|^{1/2} + |v-v'|.$$

Montrer qu'alors,

$$\left(\mathbb{E}\left[\hat{f}_h(x_0,y_0)\right] - f(x_0,y_0)\right)^2 \leq C(h_1 + h_2^2),$$

où C est une constante ne dépendant que du noyau Q.

5. **(a)** Soit la fonction g définie sur $(\mathbb{R}_+^*)^2$ par $g(u,v) = u + v^2 + 1/(nuv)$, $(u,v) \in (\mathbb{R}_+^*)^2$. Justifier que g admet sur $(\mathbb{R}_+^*)^2$ un minimum global dont on calculera la valeur.

(b) Déduire des questions précédentes que la vitesse de l'estimateur $\hat{f}_h$, pour le risque quadratique ponctuel, est $n^{-1/5}$.

6. On note f_X la densité marginale de X_1.

(a) Rappeler comment exprimer f_X en fonction de f.

(b) Proposer alors un estimateur $\hat{f}_X$ de f_X en fonction de l'estimateur $\hat{f}_h$ précédent.

(c) On suppose que Q s'écrit $Q(u,v) = K(u)K(v)$, $(u,v) \in \mathbb{R}^2$, où K est un noyau sur $\mathbb{R}$. Quelle est alors l'expression de l'estimateur proposé $\hat{f}_X$?