École Normale Supérieure
FIMFA — Statistique 2015–2016

Examen

1. On définit la « distance » H entre deux lois P_1, P_2 de densités f_1, f_2 par rapport à une mesure σ-finie μ :

$$H^2(P_1, P_2) := \frac{1}{2} \int_{\mathcal{X}} \left(\sqrt{f_1(x)} - \sqrt{f_2(x)} \right)^2 \mathrm{d}\mu(x) \,.$$

On appelle affinité la quantité $\rho(P_1, P_2) := 1 - H^2(P_1, P_2) = \int_{\mathcal{X}} \sqrt{f_1(x) f_2(x)} \mathrm{d}\mu(x)$.

(a) (1 point) La quantité $H(P_1, P_2)$ dépend-t-elle du choix de la mesure dominante μ ?

(b) (1 point) La fonction H mérite-telle le nom de distance ?

(c) (1 point) Vérifier que

$$\rho(P_1^{\otimes n}, P_2^{\otimes n}) = (\rho(P_1, P_2))^n \,.$$

($P^{\otimes}$ désigne la loi produit sur $\mathcal{X}^n$ dont toutes les marginales sont égales à P).

(d) (2 points) Soit T un test entre l'hypothèse nulle $\{P_1\}$ et l'alternative $\{P_2\}$. Si $T = 2$, le test rejette l'hypothèse nulle, si $T = 1$, le test ne rejette pas l'hypothèse nulle. Montrer que

$$\rho\left(P_1, P_2\right)^n \leq \sqrt{P_1^{\otimes n}\{T = 2\}} + \sqrt{P_2^{\otimes n}\{T = 1\}}$$

(e) (2 points) Montrer que

$$P_1^{\otimes n} \left\{ \log \frac{\prod_{i=1}^n f_2(x_i)}{\prod_{i=1}^n f_1(x_i)} > \tau \right\} \leq \mathrm{e}^{-nH^2(P_1, P_2) - \frac{\tau}{2}}$$

(f) (1 point) On ne suppose plus que les ensembles de lois $\mathcal{P}_1$ et $\mathcal{P}_2$ sont des singletons. On note $\mathrm{conv}(\mathcal{P}_i)$ ($i \in \{1, 2\}$) l'ensemble des combinaisons convexes de lois de $\mathcal{P}_i$. Montrer que pour tout test T de $\mathcal{P}_1$ contre $\mathcal{P}_2$

$$\max_{i \in \{1,2\}} \sup_{P \in \mathcal{P}_i} P^{\otimes n}(T \neq i) \geq \frac{1}{2} \left(1 - 2n \inf_{Q_1 \in \mathrm{conv}(\mathcal{P}_1), Q_2 \in \mathrm{conv}(\mathcal{P}_2)} H^2(Q_1, Q_2) \right) \,.$$

(g) (2 points) Dans un certain modèle $(P_\theta, \theta \in \Theta \subseteq \mathbb{R})$, on a pu établir qu'il existe deux constantes c, C pour tous $\theta, \eta \in \Theta$,

$$c|\theta - \eta| \leq H(P_\theta, P_\eta) \leq C|\theta - \eta| \,.$$

Montrer qu'il existe une constante $\kappa > 0$, telle que pour tout estimateur $\widehat{\theta}_n$ de θ

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[|\widehat{\theta}_n - \theta| \right] \geq \frac{\kappa}{\sqrt{n}} \,.$$

Ne pas chercher à optimiser κ.

2. Dans cet exercice H^2 (et aussi H) est défini comme dans l'exercice précédent.

Dans cet exercice,

$$g(x) := \mathbb{I}_{[-1,1]}(x) C(\alpha) \left(|x|^{-\alpha} - 1 \right)$$

avec $C(\alpha) := \frac{1}{2} \frac{1-\alpha}{\alpha}$ pour $\alpha \in]0, 1[$. La constante α est fixée. On s'intéresse au modèle de translation défini par les densités $g_\theta(x) = g(x - \theta)$ pour $\theta \in \Theta = \mathbb{R}$. Le paramètre à estimer est θ. On dispose d'échantillons $X_1, \ldots, X_n$ identiquement, indépendamment distribués selon g_θ.

(a) (3 points) Montrer que $H^2(g, g_\theta) \sim_{\theta\to 0} C(\alpha)|\theta|^{1-\alpha}$.

(b) (2 points) Déterminer une suite $(r_n)_{n\in\mathbb{N}}$ telle que
— si $|\theta_n|/r_n \to \infty$, il existe une suite de tests T_n vérifiant

$$\max(P_0^{\otimes n}\{T_n = 2\}, P_{\theta_n}^{\otimes n}\{T_n = 1\}) \to 0$$

— si $|\theta_n|/r_n \to 0$, pour toute suite de tests $(T_n)_n$

$$\liminf_n \max(P_0^{\otimes n}\{T_n = 2\}, P_{\theta_n}^{\otimes n}\{T_n = 1\}) > 0\,.$$

(c) (1 point) Le maximum de vraisemblance est-il défini ? de manière unique ?

(d) (1 point) Proposer un estimateur $\overline{\theta}_n$ du paramètre de localisation θ construit à partir de la moyenne empirique. Décrire son biais, sa variance. Est-il asymptotiquement normal ?

(e) (3 points) Proposer un estimateur $\widetilde{\theta}_n$ du paramètre de localisation construit à partir de la médiane empirique M_n. Calculer son biais. Essayez de prouver que $\mathbb{E}_\theta\left[|M_n - \mathbb{E}_\theta M_n|\right]$ décroit plus vite que $n^{-1/2}$ avec n.
Ne pas chercher à déterminer la loi asymptotique de M_n. Une bonne inégalité de déviation suffit à prouver des résultats non triviaux.

3. On considère un modèle statistique $(\Omega, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$. Pour $m \in \mathbb{N}^*$, on considère $2m$ sous-ensembles de Θ notés $\Theta_{01}, \Theta_{11}, \Theta_{02}, \Theta_{12}, \ldots, \Theta_{0m}, \Theta_{1m}$ avec pour tout $i \in \{1, \ldots, m\}$

$$\Theta_{0i} \cap \Theta_{1i} = \emptyset$$

et on veut réaliser simultanément m tests

$$H_{0i} : \theta \in \Theta_{0i} \quad \text{contre} \quad H_{1i} : \theta \in \Theta_{1i}, \qquad i = 1, \ldots, m.$$

On note I_0 l'ensemble des indices i pour lesquels H_{0i} est vraie :

$$I_0 = \{i \in \{1, \ldots, m\} : \quad H_{0i} \text{ est vraie}\}\,.$$

On cherche à construire une procédure de tests multiples qui retourne un ensemble $\hat{R} \subset \{1, \ldots, m\}$ correspondant aux indices i pour lesquels H_{0i} est rejetée. On note FP le cardinal de l'ensemble des indices correspondant aux hypothèses nulles rejetées à tort et TP le cardinal de l'ensemble des indices correspondant aux hypothèses nulles rejetées à raison :

$$\mathrm{FP} = \mathrm{card}(\hat{R} \cap I_0), \quad \mathrm{TP} = \mathrm{card}(\hat{R} \setminus I_0).$$

FP est le cardinal des *faux positifs* et TP celui des *vrais positifs*. Idéalement, on cherche une procédure de tests de sorte que FP soit petit et TP soit grand. Pour cela, pour tout i, on suppose que pour tester H_{0i} contre H_{1i}, on dispose d'une statistique $\hat{p}_i$ satisfaisant que pour tout $u \in [0, 1]$,

$$\sup_{\theta\in\theta_{0i}} \mathbb{P}_\theta(\hat{p}_i \le u) \le u. \tag{1}$$

(a) (2 points) On propose tout d'abord la procédure de Bonferroni qui permet le contrôle de FP en posant pour $\alpha \in [0, 1]$:

$$\hat{R} = \left\{i \in \{1, \ldots, m\} : \quad \hat{p}_i \le \frac{\alpha}{m}\right\}.$$

Montrer que

$$\mathbb{P}(\mathrm{FP} > 0) \le \alpha.$$

La procédure de Bonferroni contrôle le nombre de faux positifs mais produit en général un petit nombre de vrais positifs. Aussi, on propose l'alternative suivante. On se donne une fonction $f : \mathbb{N} \longrightarrow \mathbb{R}$ supposée croisante et on ordonne les statistique $\hat{p}_i$ par ordre croissant :

$$\hat{p}_{(1)} \le \hat{p}_{(2)} \le \ldots \le \hat{p}_{(m)}.$$

On cherche à contrôler le rapport FDR défini par

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP}} \mathbb{I}_{\{\mathrm{FP}+\mathrm{TP}\ge 1\}}\right]$$

avec la convention $0/0 = 0$. On pose pour $\alpha \in]0,1[$,

$$\hat{R} = \left\{ i \in \{1,\dots,m\} : \quad \hat{p}_i \leq \frac{\alpha f(\hat{k})}{m} \right\}$$

avec

$$\hat{k} = \max \left\{ k \in \{1,\dots,m\} : \quad \hat{p}_{(k)} \leq \frac{\alpha f(k)}{m} \right\}.$$

En particulier, on pose $\hat{k} = 0$ et $\hat{R} = \emptyset$ si pour tout entier k, $\hat{p}_{(k)} > \frac{\alpha f(k)}{m}$.

(b) (½ point) Que vaut $\sum_{j \geq k}^{\infty} \frac{1}{j(j+1)}$?

(c) (2 points) Montrer que

$$\mathrm{FDR} = \sum_{i \in I_0} \mathbb{E}\left[\mathbb{I}_{\{\hat{p}_i \leq \alpha f(\hat{k})/m\}} \times \frac{\mathbb{I}_{\{\hat{k} \geq 1\}}}{\hat{k}} \right].$$

(d) (4 points) En déduire que

$$\mathrm{FDR} \leq \alpha \frac{\mathrm{card}(I_0)}{m} \sum_{j=1}^{+\infty} \frac{f(\min(j,m))}{j(j+1)}.$$

(e) (2 points) Donner une condition sur f qui permette de garantir que

$$\mathrm{FDR} \leq \alpha$$

et donner un exemple de fonction f convenable.

Question:	1	2	3	Total
Points:	10	10	10½	30½
Score:				