Ecole Normale Supérieure
Gérard Biau
FIMFA — Statistique

CONTRÔLE TERMINAL — 13/01/2011

Durée 3h — Aucun document autorisé

Exercice 1

Rappel. Les intégrales $\int_0^{+\infty} x^{t-1}e^{-x}\mathrm{d}x$ sont notées $\Gamma(t)$ pour $t>0$. On sait que $\Gamma(t+1)=t\,\Gamma(t)$ et que si $t\in\mathbb{N}$, $\Gamma(t+1)=t!$

On observe n variables aléatoires $X_1,\ldots,X_n$ i.i.d. réelles de densité $\frac{1}{2\lambda}e^{-|x|/\lambda}$, où λ est un paramètre réel strictement positif. On veut estimer λ en utilisant diverses méthodes.

1. Préciser, pour $q\in\mathbb{N}^*$, l'espérance et la variance de $|X_i|^q$.
2. On pose
$$S_n=\frac{1}{nq!}\sum_{i=1}^{n}|X_i|^q.$$
Donner la loi limite de $a_n(S_n-b)$ où a_n et b sont des réels convenablement choisis. On pourra utiliser la notation
$$\sigma_q^2=\frac{(2q)!}{(q!)^2}-1.$$
3. 3.a Utiliser la question 2) pour fabriquer, à q fixé, un estimateur $\hat{\lambda}_n$ convergent de λ.
   3.b Donner la loi limite de $\hat{\lambda}_n$.
   3.c Proposer un test asymptotique de niveau $\alpha\in]0,1[$ de l'hypothèse $H_0:\lambda=1$ contre $H_1:\lambda>1$.
4. On note $X_{(n)}$ la plus grande valeur des X_i.

4.a Calculer la fonction de répartition de $X_{(n)}$ ainsi que

$$\mathbb{P}_\lambda[X_{(n)} - \lambda \ln n \leq u] \quad \text{pour } u \in \mathbb{R}.$$

4.b Montrer que lorsque n tend vers l'infini, $X_{(n)} - \lambda \ln n$ a une loi limite dont on donnera la fonction de répartition.

4.c En déduire un second estimateur $\bar{\lambda}_n$ convergent de λ. Que pensez-vous de sa qualité ?

Exercice 2

On considère le modèle linéaire

$$y_i = \alpha w_i + \beta z_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

où $\underline{w} = (w_1, \ldots, w_n)^T$ et $\underline{z} = (z_1, \ldots, z_n)^T$ sont deux vecteurs **déterministes** de $\mathbb{R}^n$ et $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ est un vecteur aléatoire gaussien à composantes i.i.d. de loi normale $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. On note $\underline{y} = (y_1, \ldots, y_n)^T$.

Les paramètres α, $\beta \in \mathbb{R}$ et $\sigma^2 > 0$ sont **inconnus**. Le but de l'exercice consiste à comparer la qualité de l'estimation de ces paramètres lorsque les vecteurs $\underline{w}$ et $\underline{z}$ sont orthogonaux et lorsque ils ne le sont pas.

Dans tout l'exercice, on suppose que

$$\langle \underline{w}, \underline{z} \rangle^2 \neq \|\underline{w}\|^2 \, \|\underline{z}\|^2.$$

1. Montrer que l'estimateur des moindres carrés $\hat{\underline{b}}_n = (\hat{\alpha}_n, \hat{\beta}_n)^T$ du paramètre $\underline{b} = (\alpha, \beta)^T$ est donné par

$$\hat{\underline{b}}_n = \frac{1}{\|\underline{w}\|^2 \, \|\underline{z}\|^2 - \langle \underline{w}, \underline{z} \rangle^2} \begin{pmatrix} \|\underline{z}\|^2 \langle \underline{w}, \underline{y} \rangle - \langle \underline{w}, \underline{z} \rangle \langle \underline{z}, \underline{y} \rangle \\ \|\underline{w}\|^2 \langle \underline{z}, \underline{y} \rangle - \langle \underline{w}, \underline{z} \rangle \langle \underline{w}, \underline{y} \rangle \end{pmatrix}$$

   et préciser sa loi.

2. Montrer que si $\underline{w}$ et $\underline{z}$ sont orthogonaux, alors les deux estimateurs $\hat{\alpha}_n$ et $\hat{\beta}_n$ de α et β sont indépendants.

A partir de maintenant, on abandonne l'hypothèse d'orthogonalité de $\underline{w}$ et $\underline{z}$, et on note $\theta \in]0, \pi/2[$ l'angle formé par ces deux vecteurs. **On suppose, pour simplifier les calculs, que $\|\underline{w}\| = \|\underline{z}\| = 1$.**

3. Préciser la loi de la variable aléatoire

$$\frac{|\sin\theta|(\hat{\beta}_n - \beta)}{\sqrt{\hat{\sigma}_n^2}},$$

où $\hat{\sigma}_n^2$ est l'estimateur habituel (sans biais) de σ^2 dans le modèle de régression.

4. Construire un intervalle de confiance (exact et symétrique) de niveau $1-\alpha$ pour β.

5. Soit $L^2(\theta)$ la variable aléatoire égale au carré de la longueur de l'intervalle de confiance trouvé dans la question précédente. Etudier la fonction

$$g(\theta) = \mathbb{E}\left[L^2(\theta)\right], \quad \theta \in \left]0, \frac{\pi}{2}\right[$$

et interpréter les résultats obtenus à la lumière du comportement de g à la frontière de son domaine de définition.

Exercice 3

Soit (X, Y) un vecteur gaussien tel que $\mathbb{E}(X) = \mathbb{E}(Y) = 0$, $\mathbb{V}(X) = \mathbb{V}(Y) = 1$ et $\mathrm{Cov}(X, Y) = \rho$, avec $-1 \leq \rho \leq 1$. Montrer que

$$\mathbb{E}\left(\max\{X, Y\}\right) = \sqrt{\frac{1-\rho}{\pi}}.$$

Problème

Dans tout le devoir, $\mathbb{R}^d$ est muni de la norme euclidienne $\|.\|$.

Préliminaires

Etant donné un ensemble $\mathcal{C} = \{y_1, \ldots, y_k\}$ formé de k points distincts de $\mathbb{R}^d$, on appelle k-quantificateur associé à $\mathcal{C}$ toute application mesurable

$$q : \mathbb{R}^d \to \mathcal{C}.$$

1. Pourquoi un k-quantificateur peut-il être assimilé à une fonction de "compression" ?

2. Montrer qu'un k-quantificateur q est entièrement caractérisé par l'ensemble $\mathcal{C} = \{y_1, \ldots, y_k\}$ et les boréliens $S_i = \{x \in \mathbb{R}^d : q(x) = y_i\}$, $i = 1, \ldots, k$, via la règle

$$q(x) = y_i \quad \text{si et seulement si} \quad x \in S_i.$$

Les $\{y_1, \ldots, y_k\}$ sont appelés *centres* et les $\{S_1, \ldots, S_k\}$ sont appelées *cellules*.

3. Que peut-on dire des cellules $\{S_1, \ldots, S_k\}$?

Partie I

On considère dans cette partie une variable aléatoire X à valeurs dans $\mathbb{R}^d$, de loi μ et telle que $\mathbb{E}\|X\|^2 < \infty$. On se donne également un k-quantificateur q défini par l'ensemble $\mathcal{C} = \{y_1, \ldots, y_k\}$ et les cellules $\{S_1, \ldots, S_k\}$. Dans ce contexte, on appelle distorsion (relativement à μ et q) la quantité

$$D(\mu, q) = \mathbb{E}\left[\|X - q(X)\|^2\right].$$

1. Donner une interprétation de $D(\mu, q)$.
2. Montrer que $D(\mu, q) < +\infty$.
3. Soit alors q' le k-quantificateur défini par l'ensemble $\mathcal{C} = \{y_1, \ldots, y_k\}$ (le même que celui de q) et la relation

$$q'(x) = \arg\min_{y_i \in \mathcal{C}} \|x - y_i\|^2.$$

3.a Expliquer pourquoi un tel k-quantificateur est appelé k-quantificateur de type plus proches voisins (faire un dessin).

3.b Prouver que

$$\mathbb{E}\left[\|X - q(X)\|^2\right] \geq \mathbb{E}\left[\|X - q'(X)\|^2\right].$$

4. Un k-quantificateur $q^\star$ est dit optimal (pour μ) si

$$D(\mu, q^\star) = \inf_q D(\mu, q),$$

où l'infimum est pris sur tous les k-quantificateurs. Que nous enseigne l'analyse précédente ?

5. **Question de niveau recherche, hors barême. Ne l'abordez qu'à la fin, si vous avez encore du temps et une partie de l'après-midi devant vous**. Montrer que l'infimum de la question précédente est en fait un minimum.

Partie II

Etant données deux lois de probabilité μ et ν sur $\mathbb{R}^d$ admettant un moment d'ordre deux, on appelle distance L_2-Wasserstein entre μ et ν la quantité

$$\rho(\mu,\nu) = \inf_{X\sim\mu, Y\sim\nu} \left(\mathbb{E}\left[\|X-Y\|^2\right]\right)^{1/2},$$

où l'infimum est pris sur tous les couples aléatoires (X,Y) tels que X a pour loi μ et Y a pour loi ν. On admet que l'infimum est atteint et que ρ définit bien une distance entre μ et ν.

On considère dans cette partie un k-quantificateur q de type plus proches voisins, défini par l'ensemble $\mathcal{C} = \{y_1, \ldots, y_k\}$ et la relation

$$q(x) = \arg\min_{y_i\in\mathcal{C}} \|x-y_i\|^2.$$

On se donne également deux lois de probabilité μ et ν sur $\mathbb{R}^d$ admettant chacune un moment d'ordre deux.

1. Montrer que
$$D(\mu,q)^{1/2} \leq \rho(\mu,\nu) + D(\nu,q)^{1/2}.$$
2. Prouver alors que
$$\left|D(\mu,q)^{1/2} - D(\nu,q)^{1/2}\right| \leq \rho(\mu,\nu).$$
3. Soit $q^\star$ un k-quantificateur (de type plus proches voisins) optimal pour μ et $p^\star$ un k-quantificateur (de type plus proches voisins) optimal pour ν. Etablir finalement que
$$\left|D(\mu,q^\star)^{1/2} - D(\nu,p^\star)^{1/2}\right| \leq \rho(\mu,\nu).$$

Partie III

On considère désormais n variables aléatoires $X_1, \ldots, X_n$ à valeurs dans $\mathbb{R}^d$, i.i.d., de loi commune μ admettant un moment d'ordre deux. On note μ_n la mesure empirique associée à cet échantillon, c'est-à-dire, pour tout borélien A de $\mathbb{R}^d$,

$$\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[X_i \in A]}.$$

1. **Attention, cette question n'est pas facile...** Montrer que, lorsque $n \to +\infty$,
$$\rho(\mu, \mu_n) \to 0 \quad \text{p.s.}$$

2. Soit $q_n^\star$ un k-quantificateur de type plus proches voisins, optimal pour μ_n et associé à l'ensemble $\mathcal{C} = \{y_{1,n}^\star, \ldots, y_{k,n}^\star\}$. Expliquer pourquoi
$$\{y_{1,n}^\star, \ldots, y_{k,n}^\star\} = \arg \min_{\{y_1,\ldots,y_k\}} \frac{1}{n} \sum_{i=1}^{n} \min_{y \in \{y_1,\ldots,y_k\}} \|X_i - y\|^2.$$

3. Donner une interprétation de $q_n^\star$. Pourquoi dit-on que $q_n^\star$ est une fonction de compression des données ?

4. Prouver que, lorsque $n \to +\infty$,
$$D(\mu, q_n^\star) \to D(\mu, q^\star) \quad \text{p.s.}$$

5. Interpréter le résultat précédent.