MAP433 Statistique
PC3: maximum de vraisemblance

11 septembre 2015

Exercice 1 (Durées de connection)**.** On peut modéliser la durée d'une connection sur le site `www.Cpascher.com` par une loi gamma(a, b) ($a > 0$, $b > 0$) de densité

$$p_{a,b}(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \mathrm{e}^{-bx} \mathbb{1}_{[0,+\infty)}(x).$$

On note $\theta = (a, b)$. Pour fixer vos tarifs publicitaires, vous voulez estimer le paramètre θ à partir d'un échantillon $X_1, \ldots, X_n$ de n durées de connexion.

1. Proposez un estimateur par la méthode des moments.
2. Ecrire les équations de vraisemblance. Déterminer pour une valeur de a fixée, le maximum $\hat{b}_n(a)$ de la fonction de vraisemblance.
3. Montrer que l'estimateur du maximum de vraisemblance est $\hat{\theta}_n = (\hat{a}_n, \hat{b}(\hat{a}_n)$ où $\hat{a}_n$ est le maximum de la fonction $a \mapsto L_n(a)$

   $$L_n(a) = na \ln(a) - na \ln(\bar{X}_n) - n \ln(\Gamma(a)) + n(a-1)\overline{\ln(X)}_n - na$$

   où $\overline{\ln(X)}_n = n^{-1} \sum_{i=1}^n \ln(X_i)$.
4. Proposez une méthode numérique pour déterminer l'estimateur du maximum de vraisemblance.
5. Question bonus : vous trouverez sur moodle un fichier de données (format texte). Implémentez la méthode pour calculer l'estimateur du maximum de vraisemblance. Vous proposerez aussi une méthode pour calculer des régions de confiance par bootstrap (voir la méthode du bootstrap sur moodle).

Corrigé :

1. On rappelle que l'espérance et la variance d'une loi Gamma de paramètres (a, b) sont resp. donnés par a/b et a/b^2. On définit

   $$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \qquad \hat{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^n \left(X_k - \bar{X}_n\right)^2.$$

   Estimateur proposé :

   $$\hat{\theta}_n = \frac{\bar{X}_n}{\hat{\sigma}_n^2} \times \begin{bmatrix} \bar{X}_n \\ 1 \end{bmatrix}$$

2. La vraisemblance est donnée par (on note $Z_n = (X_1, \cdots, X_n)$)

$$\theta \mapsto L_n(Z_n, \theta) = \prod_{k=1}^n p_\theta(X_k) = \left(\frac{b^a}{\Gamma(a)}\right)^n \left(\prod_{k=1}^n X_k\right)^{a-1} \exp(-b\sum_{k=1}^n X_n) \prod_{k=1}^n 1_{\mathbb{R}^+}(X_k).$$

Dont on déduit les équations de vraisemblance (ψ désigne la fonction Digamma)

$$\ln b - \psi(a) + \frac{1}{n}\sum_{k=1}^n \ln X_k = 0, \qquad \frac{a}{b} - \bar{X}_n = 0.$$

puis $\hat{b}_n(a) = \frac{a}{\bar{X}_n}$.

3. $L_n(a)$ correspond à ce que l'on note $L_n(Z_n, (a, \hat{b}_n(a)))$. Par définition de $\hat{b}_n(a)$ et $\hat{a}_n$ on a pour tout $a, b > 0$,

$$L_n(Z_n, (a,b)) \le L_n(Z_n, (a, \hat{b}_n(a)) \qquad L_n(Z_n, (a, \hat{b}_n(a)) \le L_n(Z_n, (\hat{a}_n, \hat{b}_n(\hat{a}_n))$$

dont on déduit que $(\hat{a}_n, \hat{b}_n(\hat{a}_n))$ est estimateur MV.

4. On fait successivement
   (i) une méthode numérique (par exemple, un algorithme de gradient) pour calculer $\hat{a}_n^{MV}$.
   (ii) puis poser $\hat{b}_n^{MV} = \hat{b}_n(\hat{a}_n^{MV})$.

Exercice 2 (Modèle exponentiel)**.** Considérons une famille de fonctions de répartition $\{F_\theta,\ \theta \in \Theta\}$ ayant une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ de la forme

$$p_\theta(x) = c(\theta) \exp\left(\theta f(x) + h(x)\right).$$

On suppose que Θ est un intervalle ouvert de $\mathbb{R}$, et $c(\cdot) \in C^2$, $c(\theta) > 0$ pour tout $\theta \in \Theta$. On note $\varphi(\theta) := \mathbb{E}_\theta\left(f(X)\right) = -\frac{d}{d\theta}\log(c(\theta))$.

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de densité p_θ, avec θ inconnu. Calculez l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ (s'il existe).

Corrigé :
L'estimateur MV est donné comme la solution de l'équation en θ

$$\mathbb{E}_\theta\left[f(X)\right] = \frac{1}{n}\sum_{k=1}^n f(X_k)$$

Exercice 3 (Modèle d'autorégression)**.** On considère les observations $X_1, \ldots, X_n$, où les X_i sont issus du *modèle d'autorégression* :

$$X_i = \theta X_{i-1} + \xi_i, \quad i = 1, \ldots, n, \qquad X_0 = 0,$$

avec ξ_i i.i.d. de loi normale $\mathcal{N}(0, \sigma^2)$ et $\theta \in \mathbb{R}$. Calculez l'estimateur du maximum de vraisemblance $(\hat{\theta}_n, \hat{\sigma}_n^2)$ de (θ, σ^2).
Corrigé :
On a établi en PC2 que la densité de $(X_1, \cdots, X_n)$ est donnée par

$$(x_1, \cdots, x_n) \mapsto (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta x_{i-1})^2\right)$$

avec la convention $x_0 = 0$. On en déduit les estimateurs MV

$$\hat{\theta}_n = \frac{C_n}{S_{n-1}} \qquad \hat{\sigma}_n^2 = \frac{S_n S_{n-1} - C_n^2}{S_{n-1}}$$

en ayant posé

$$S_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2 \qquad C_n = \frac{1}{n}\sum_{k=1}^{n} X_k X_{k-1} = \frac{1}{n}\sum_{k=2}^{n} X_k X_{k-1}.$$

Exercice 4. Soient $\{(Y_i, Z_i)\}_{i=1}^n$ n vecteurs aléatoires i.i.d. ; On suppose que Y_1 et Z_1 sont indépendants et distribués suivant des lois exponentielles de paramètres $\lambda > 0$ and $\mu > 0$.

1. On observe $\{(Y_i, Z_i)\}_{i=1}^n$. Donnez le modèle statistique et déterminez l'estimateur du MV de λ et μ.
2. Déterminer la distribution limite de cet estimateur.
3. On observe $\{(X_i, \Delta_i)\}_{i=1}^n$ où $X_i = \min(Y_i, Z_i)$ et $\Delta_i = 1$ si $Y_i = X_i$ $\Delta_i = 0$ autrement. Donnez le modèle statistique et l'estimateur de vraisemblance de λ et μ dans ce modèle.
4. Déterminer la distribution limite de cet estimateur.

Corrigé :

1. On munit $(\mathbb{R}^+ \times \mathbb{R}^+)^n$ de sa tribu borélienne ; et cet espace mesurable de la famille de lois

$$\mathbb{P}_\theta(dy_1, dz_1, \cdots, dy_n, dz_n) = \lambda^n \mu^n \exp(-\lambda \sum_{k=1}^{n} y_k) \exp(-\mu \sum_{k=1}^{n} z_k) \prod_{k=1}^{n} 1_{\mathbb{R}^+ \times \mathbb{R}^+}(y_k, z_k)\, dy_1 dz_1 \cdots dy_n dz_n,$$

   où $\theta = (\lambda, \mu)$. L'estimateur MV $(\hat{\lambda}_n, \hat{\mu}_n)$ est donné par

$$\hat{\lambda}_n = \frac{1}{\bar{Y}_n} \qquad \hat{\mu}_n = \frac{1}{\bar{Z}_n}$$

   en ayant posé $\bar{Y}_n = \frac{1}{n}\sum_{k=1}^n Y_k \qquad \bar{Z}_n = \frac{1}{n}\sum_{k=1}^n Z_k$.
2. On commence par écrire le TCL pour le couple $(\bar{Y}_n, \bar{Z}_n)$. Puis on applique la méthode delta avec $g(x, y) = (1/x, 1/y)$ et on obtient

$$\sqrt{n}\left(\begin{bmatrix} 1/\bar{Y}_n \\ 1/\bar{Z}_n \end{bmatrix} - \begin{bmatrix} \lambda \\ \mu \end{bmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}_2\left(0; \begin{bmatrix} \lambda^2 & 0 \\ 0 & \mu^2 \end{bmatrix}\right).$$

3. On a établi en PC2 exercice 5 que pour tout $t > 0$ et $u \in \{0, 1\}$,

$$\mathbb{P}_\theta\left(X_i > t, \Delta_i = u\right) = \exp(-(\lambda + \mu)t)\frac{\lambda^u \mu^{1-u}}{\lambda + \mu} = \mathbb{P}(X_i > t)\mathbb{P}(\Delta_i = u).$$

   On pose $\theta = (\lambda, \mu)$ et $\Theta = \mathbb{R}_\star^+ \times \mathbb{R}_\star^+$. Et le modèle statistique sur $(\mathbb{R}_\star^+ \times \{0, 1\})^n$,

$$\begin{aligned}\mathbb{P}_\theta(A_1 \times \{b_1\} \times \cdots \times A_n \times \{b_n\}) &= \prod_{i=1}^{n}\left\{\frac{\lambda^{b_i}\mu^{1-b_i}}{\lambda+\mu} \int_{A_i} (\lambda+\mu)\exp(-(\lambda+\mu)x_i)dx_i\right\} \\ &= \prod_{i=1}^{n}\left\{\lambda^{b_i}\mu^{1-b_i} \int_{A_i} \exp(-(\lambda+\mu)x_i)dx_i\right\}\end{aligned}$$

   L'estimateur MV est donné par

$$\hat{\lambda}_n = \frac{\bar{\Delta}_n}{\bar{X}_n} \qquad \hat{\mu}_n = \frac{1 - \bar{\Delta}_n}{\bar{X}_n}$$

   en ayant posé $\bar{X}_n = n^{-1}\sum_{k=1}^n X_k$ et $\bar{\Delta}_n = n^{-1}\sum_{k=1}^n \Delta_k$.

4. On commence par écrire le TCL pour le couple $(\bar{X}_n, \bar{\Delta}_n)$. Puis on applique la méthode delta avec $g(x,y) = (x/y, (1-x)/y)$ et on obtient

$$\sqrt{n}\left(\begin{bmatrix} \bar{\Delta}_n/\bar{X}_n \\ (1-\bar{\Delta}_n)/\bar{X}_n \end{bmatrix} - \begin{bmatrix} \lambda \\ \mu \end{bmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0; (\lambda+\mu)\begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}\right)$$

Exercice 5 (Répartition de génotypes dans une population)**.** Quand les fréquences de gènes sont en équilibre, les génotypes AA, Aa et aa se manifestent dans une population avec probabilités $(1-\theta)^2$, $2\theta(1-\theta)$ et θ^2 respectivement, où θ est un paramètre inconnu. Plato *et al.* (1964) ont publié les données suivantes sur le type de haptoglobine dans un échantillon de 190 personnes :

Type de haptoglobine :

$Hp-AA$	$Hp-Aa$	$Hp-aa$
10	68	112

1. Comment interpréter le paramètre θ ? Proposez un modèle statistique pour ce problème.
2. Calculez l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ.
3. Donnez la loi asymptotique de $\sqrt{n}(\hat{\theta}_n - \theta)$.

Corrigé :

1. On peut modéliser les observations comme une réalisation $(X_1, \cdots, X_n)$ où $X_i \in \{1,2,3\}$ et $(X_k)_k$ sont indépendants et de même loi

$$\mathbb{P}_\theta(\{1\}) = (1-\theta)^2 \qquad \mathbb{P}_\theta(\{2\}) = 2\theta(1-\theta)$$

de sorte que le modèle statistique est
– un espace probabilisable : $\{1,2,3\}^n$ muni de la tribu de ses parties.
– une famille de lois $\{\mathbb{P}_\theta, \theta \in]0,1[\}$ définies comme ci-dessus.

2. La log-vraisemblance normalisée est donnée par

$$\theta \mapsto \frac{2N_1}{n}\ln(1-\theta) + \frac{N_2}{n}\ln(2\theta(1-\theta)) + \frac{2N_3}{n}\ln\theta \qquad N_j = \sum_{k=1}^{n} 1_{X_k=j}$$

dont on déduit l'estimateur MV

$$\hat{\theta}_n = \frac{N_3}{n} + \frac{N_2}{2n}$$

3. Par le TCL pour des v.a. i.i.d. on a

$$\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n}(1_{X_k=3} + \frac{1}{2}1_{X_k=2}) - \theta\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, v^2)$$

avec

$$v^2 = \mathrm{Var}\left(1_{X_k=3} + \frac{1}{2}1_{X_k=2}\right) = \frac{\theta-\theta^2}{2} = \frac{\theta(1-\theta)}{2}.$$