École Normale Supérieure
FIMFA — Statistique 2015–2016

# Devoir 3 pour le 13 janvier

1. On considère dans cet exercice une variable aléatoire $X$ définie par $X = ZY_1 + (1-Z)Y_2$, où $Z \sim Bernoulli(\frac{1}{2})$, $Y_1 \sim Poisson(\lambda_1)$, $Y_2 \sim Poisson(\lambda_2)$ et où $Z$, $Y_1$ et $Y_2$ sont indépendantes. On dit que $X$ suit un *mélange de lois de Poisson*. On observe un échantillon $X_1, \dots, X_n$ de même loi que $X$.

   Dans le formalisme des *expériences statistiques* $(\Omega, \mathcal{F}, (P_\theta, \theta \in \Theta), \mathcal{X}, X)$, on peut choisir :
   - $\Omega = \{0,1\} \times \mathbb{N} \times \mathbb{N}$,
   - $\Theta = ]0,\infty)^2$, $P_\theta = \mathrm{Be}(1/2) \otimes \mathrm{Po}(\lambda_1) \otimes \mathrm{Po}(\lambda_2)$ avec $\theta = \begin{pmatrix}\lambda_1 \\ \lambda_2\end{pmatrix}$, chaque $P_\theta$ est la loi jointe de $(Z, Y_1, Y_2)$,
   - les observations sont à valeurs dans $\mathcal{X} = \mathbb{N}$,
   - $X : \Omega \to \mathcal{X}$ est défini par $X(z, y_1, y_2) = z \times y_1 + (1-z) \times y_2$.

   On suppose que le paramètre $(\lambda_1, \lambda_2)$ est inconnu.
   (a) Le paramètre $(\lambda_1, \lambda_2)$ est-il identifiable ? Dans la négative, donner le paramètre identifiable du modèle.
   (b) Proposer un estimateur de ce paramètre par la méthode des moments. Est-il toujours défini ? Est-il consistant ?

2. Dans cette question, on suppose que le paramètre $\gamma = \lambda_1 + \lambda_2$ est connu, et que $\lambda_1 \geq \lambda_2$. On souhaite tester l'hypothèse nulle $H_0 : \lambda_1 = \lambda_2$ contre l'hypothèse alternative $H_1 : \lambda_1 > \lambda_2$.
   (a) Donner la loi de $X$ dans le cas où $\lambda_1 = \lambda_2$.
   (b) Dans le cas général $\lambda_1 \geq \lambda_2$, donner une suite $(a_n)$ et une loi non dégénérée $\mu$ telles que
   $$a_n(\bar{X}_n - \frac{\gamma}{2}) \xrightarrow[n\to\infty]{\mathcal{L}} \mu$$
   (c) Construire un intervalle de confiance de niveau asymptotique $1-\alpha$ pour le paramètre $\nu = \lambda_1 - \lambda_2$.
   (d) En déduire un test de niveau asymptotique $\alpha$ de $H_0$ contre $H_1$.

3. ALGORITHME EM.

Dans cette question, on se place dans le cadre général d'une variable aléatoire $X$ dont la loi dépend de la réalisation d'une variable aléatoire *non observée/cachée/latente* $Z$ et d'un paramètre inconnu $\theta$. Pour simplifier, on supposera $Z$ discrète à valeur sans $\mathcal{Z}$ et $X$ à valeur dans $\mathcal{X} := \mathbb{R}$.

On supposera le modèle dominé, on notera $p(x, z \mid \theta)$ la densité de la loi conjointe $P_\theta$ du couple $(X, Z)$ en $(x, z)$, $p(x \mid \theta, z)$ la densité de la loi conditionnelle $P_{\theta,z}$ de $X$ sachant $Z = z$ sous $P_\theta$, $p(z \mid \theta, x)$ la densité de la loi conditionnelle $P_{\theta,x}$ de $Z$ sachant $X = x$ sous $P_\theta$, $p(x \mid \theta)$ la densité de la marginale en $X$, $P_\theta \circ X^{-1}$ et $p(z \mid \theta)$ la densité de la marginale en $Z$, $P_\theta \circ Z^{-1}$.

Etant donnée une observation $x$, on dispose d'un estimateur préliminaire $\widehat{\theta}$ de $\theta$ et on cherche un estimateur $\widetilde{\theta}$ qui soit de plus forte vraisemblance ($p(x \mid \widetilde{\theta}) > p(x \mid \widehat{\theta})$).

La log-vraisemblance en $\theta, x$ s'écrit

$$\ell(\theta, x) \quad := \quad \log p(x \mid \theta) = \log \left( \sum_{z \in \mathcal{Z}} p(x, z \mid \theta) \right)$$

La log-vraisemblance complète en $\theta, x, z$ est simplement $\ell_c(\theta, x, z) := \log p(x, z \mid \theta)$.

(a) Pour tout paramètre $\theta$ et toute loi $P$ sur $\mathcal{Z}$, on note $Q(\theta \mid P) := \mathbb{E}_P [\log p(x, Z \mid \theta)] = \sum_{z \in \mathcal{Z}} P(z) \times \ell_c(\theta, x, z)$ (la fonction $Q$ est parfois appelée *fonction auxiliaire*) et on note

$$H(P) := -\sum_{s \in \mathcal{Z}} P(z) \log P(z) \,,$$

l'entropie de Shannon de la loi $P$.
Montrer que pour tous $\theta, x$, et toute loi $P$ sur $\mathcal{Z}$ on peut écrire

$$\ell(\theta, x) \quad \geq \quad Q(\theta \mid P) + H\,(P) \;.$$

Pour quel choix de $P$ a-t-on égalité ?

(b) On propose de choisir

$$\widetilde{\theta} := \arg\max_\eta \left\{ Q(\eta \mid P_{\widehat{\theta},x}) \right\}.$$

Vérifier qu'on a

$$\sup_\theta \ell(\theta, x) \geq \ell(\widetilde{\theta}, x) \geq \ell(\widehat{\theta}, x)$$

(cette procédure correspond à une itération de l'algorithme EM).

4. APPLICATION DE EM.

On souhaite maintenant appliquer la méthode développée à la question 3 à l'estimateur obtenu à la question 1. On note $\theta = (\lambda_1, \lambda_2)$ le paramètre à estimer et $\widehat{\theta} = (\widehat{\lambda_1}, \widehat{\lambda_2})$ l'estimateur des moments.

(a) Donner la loi conditionnelle de $Z_i$ sachant $X_i$ et $\widehat{\theta}$ et l'expression de $\tau_i = \mathbb{E}_{\widehat{\theta}}[Z_i | X_i]$

(b) En déduire l'expression de $\widetilde{\theta} = \arg\max_\eta \left\{ Q(\eta \mid P_{\widehat{\theta},x}) \right\}$ en fonction des $x_i$ et des $\tau_i$.

5. ONE-STEP

La loi de Cauchy admet comme densité sur $\mathbb{R}$

$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2}\,.$$

On considère le modèle des translatées définies par

$$p_\theta(x) = p(x-\theta) \qquad \text{pour } x, \theta \in \mathbb{R}\,.$$

On note $m(\theta)$ la médiane de la loi de densité $p_\theta$.
On note $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ les statistiques d'ordre d'un $n$-échantillon d'une loi $p_\theta$.

(a) L'estimateur au maximum de vraisemblance de $\theta$ est il bien défini dans ce modèle ? Est-il unique ? Est il calculable (pour toute taille d'échantillon) ?

(b) Cacluler l'information de Fisher dans ce modèle.

(c) On estime $m(\theta)$ par la médiane empirique $\widehat{m}_{2n+1}$ dans un échantillon de taille $2n+1$. Quel est le biais de cet estimateur.

(d) La suite $(\widehat{m}_{2n+1})_n$ définit-elle une suite consistante d'estimateurs de $m(\theta)$ ?

(e) La suite $(\widehat{m}_{2n+1})_n$ définit-elle une suite asymptotiquement normale d'estimateurs de $m(\theta)$ ?

(f) Proposez un intervalle de niveau de confiance asymptotique $\alpha \in ]0,1[$ pour l'estimation de $m(\theta)$.

(g) On souhaite maintenant construire un nouvel estimateur de $\theta$, en applicant une étape de « descente de gradient » à $(\widehat{m}_{2n+1})_n$. On note $\dot{\ell}_{2n+1}(\theta')$ la dérivée par rapport à $\theta$ de la log-vraisemblance calculée sur l'échantillon de taille $2n+1$ en $\theta'$, et $\ddot{\ell}_{2n+1}(\theta')$ la dérivée seconde. L'estimateur $\widetilde{\theta}_{2n+1}$ est défini par

$$\widetilde{\theta}_{2n+1} := \widehat{m}_{2n+1} - \frac{\dot{\ell}_{2n+1}(\widehat{m}_{2n+1})}{\ddot{\ell}_{2n+1}(\widehat{m}_{2n+1})}\,.$$

Calculer la loi limite de

$$\sqrt{2n+1}\left(\widetilde{\theta}_{2n+1} - \theta\right)\,.$$

Pensez-vous qu'il soit pertinent d'itérer la descente de gradient ?

À l'aide du logiciel de votre choix, et pour les valeurs des paramètres de votre choix, effectuez une simulation numérique pour étudier le comportement des estimateurs obtenus dans l'exercice précédent, pour des valeurs de $n$ entre 3 et 10 000. Étudiez également le comportement de la moyenne empirique et de l'écart-type empirique d'un échantillon suivant la loi de Cauchy. Commentez.

6. ONE-STEP dans les modèles exponentiels

Dans cet exercice, nous considérons un modèle exponentiel identifiable, en forme canonique. L'espace des paramètres est noté $\Theta \subset \mathbb{R}^k$. $\theta^0 \in \Theta$ désigne la valeur du paramètre sous laquelle on effectue l'échantillonnage. La statistique suffisante est dénotée par $T(X)$. Elle est à valeur dans $\mathbb{R}^k$. On note $\ell_n(\theta)$ la log-vraisemblance en $\theta$ pour un $n$-échantillon (elle n'est pas normalisée). Lorsqu'on écrit $\nabla\ell_n(\theta')$, on désigne le gradient de la log-vraisemblance par rapport au paramètre (la fonction *score*) pris en $\theta'$.

Dans l'énoncé $\widetilde{\theta}_n$ désigne un estimateur (pas forcément un estimateur au maximum de vraisemblance). On suppose que la suite $(\widetilde{\theta}_n)_n$ est consistante et asymptotiquement normale. On note $J(\theta^0)$ la covariance asymptotique de $\sqrt{n}(\widetilde{\theta}_n - \theta^0)$.
On définit l'estimateur à un pas (*one-step*) $\overline{\theta}_n$ comme l'estimateur obtenu en appliquant un pas de la méthode de Newton pour approcher le maximum de vraisemblance en partant de $\widetilde{\theta}_n$ :

$$\overline{\theta}_n := \widetilde{\theta}_n - (\nabla^2\ell_n(\widetilde{\theta}_n))^{-1}\nabla\ell_n(\widetilde{\theta}_n)\,.$$

On note l'estimateur au maximum de vraisemblance $\widehat{\theta}_n$. On note $I(\theta^0)$ la matrice d'information de Fisher en $\theta^0$ ($I(\theta^0) = \nabla^2 \log Z(\theta^0)$).

(a) Quelle est la loi limite de $\frac{1}{\sqrt{n}}\nabla\ell_n(\theta^0)$ ?

(b) Vérifier que pour toute constante $M$

$$\sup_{\sqrt{n}\|\theta-\theta^0\|\leq M}\frac{1}{\sqrt{n}}\left\|\nabla\ell_n(\theta^0)-\nabla\ell_n(\theta)+\nabla^2\ell_n(\theta^0)\left(\theta-\theta^0\right)\right\|\xrightarrow{P}0$$

(est-il nécessaire de préciser en probabilité ?).

(c) Montrer que

$$\sqrt{n}\left(\overline{\theta}_n-\theta^0\right)+I(\theta^0)^{-1}\frac{1}{\sqrt{n}}\nabla\ell_n(\theta^0)\xrightarrow{P}0\,.$$

(d) La suite $\sqrt{n}\left(\overline{\theta}_n-\theta^0\right)$ admet-elle une limite en loi ? Si oui, quelle est elle ?

(e) Pensez-vous qu'il soit pertinent d'itérer la descente de gradient ?

(f) Les lois Gamma définissent un modèle exponentiel. La densité de la loi Gamma de paramètre de forme $p>0$ et de paramètre d'intensité $\lambda>0$ est donnée par

$$\mathbb{I}_{x>0}\mathrm{e}^{-\lambda x}\frac{\lambda^p x^{p-1}}{\Gamma(p)}\,.$$

Proposer un estimateur de $p,\lambda$, facile à calculer via la méthode des moments. Vérifier si la suite d'estimateurs définie ainsi est consistante et asymptotiquement normale. Peut-on utiliser la méthode à un pas dans ce contexte ? Si oui, comparer les matrices de covariance de l'estimateur obtenu par la méthode des moments et de l'estimateur obtenu par la méthode à un pas.

7. CLASSIFICATION, MINIMISATION DU RISQUE EMPIRIQUE.

Dans cet exercice, $\mathcal{X}=\mathbb{R}^d$, et $\mathcal{C}$ est une classe de parties (mesurables) de $\mathcal{X}$. L'espace $\mathcal{X}$ est muni d'une loi de probabilité $P_X$ inconnue. La variable aléatoire $X$ est distribuée selon $P_X$. Dans la suite $C^*$ est une partie mesurable mais inconnue de $\mathcal{X}$. La variable aléatoire $Y$ est définie par :

$$Y:=\begin{cases}\mathbb{I}_{X\in C^*} & \text{si } \eta=1\\ 1-\mathbb{I}_{X\in C^*} & \text{si } \eta=0\,.\end{cases}$$

où $\eta$ est une variable aléatoire indépendante de $X$ qui vaut 1 avec probabilité $1/2<h\leq 1$ et 0 avec probabilité $1-h$. On note $P$ la loi jointe de $(X,Y)$. La variable aléatoire $\eta$ n'est pas observée.

On dispose d'un ensemble d'apprentissage (échantillon) formé de couples $(X_i,Y_i)_{i\leq n}$ indépendamment distribués selon $P$.

On cherche à choisir $\widehat{C}\in\mathcal{C}$ de façon à minimiser

$$\mathrm{err}(C):=P\left\{\mathbb{I}_{X\in C}\neq Y\right\}=\mathbb{E}\left[|\mathbb{I}_{X\in C}-Y|\right]\,.$$

On note

$$\mathrm{err}_n(C):=\frac{1}{n}\sum_{i=1}^n|\mathbb{I}_{X_i\in C}-Y_i|$$

et

$$Z=\sup_{C\in\mathcal{C}}|\mathrm{err}(C)-\mathrm{err}_n(C)|$$

et

$$\mathrm{tr}(\mathcal{C},x_1,\ldots,x_n):=\{A:A\subseteq\{x_1,\ldots,x_n\},\quad\exists C\in\mathcal{C}A=C\cap\{x_1,\ldots,x_n\}\}\,.$$

(a) Montrer que si $\mathrm{err}(\overline{C}) = \min_{C\in\mathcal{C}} \mathrm{err}(C)$ alors

$$\overline{C} = \arg\min_{C\in\mathcal{C}} P_X\{C^*\Delta C\}$$

où $A\Delta B = (A\cup B)\setminus(A\cap B)$.

(b) Montrer que

$$P^{\otimes n}\{Z\geq\epsilon\} \leq 2\mathbb{E}\left[|\mathrm{tr}(\mathcal{C}, X_1, \ldots, X_n)|\right] \mathrm{e}^{-\frac{n\epsilon^2}{2}}\,.$$

(c) On dit que $\mathcal{C}$ est une classe de Vapnik-Chervonenkis s'il existe un entier $v$ tel que

$$|\mathrm{tr}(\mathcal{C}, x_1, \ldots, x_n)| \leq \sum_{i=0}^{v} \binom{n}{i}$$

pour tout $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$. Le plus petit entier $v$ convenable est appelé dimension de Vapnik-Chervonenkis de $\mathcal{C}$ (les demi-espaces forment une classe de Vapnik-Chervonenkis de dimension $d+1$, il en va de même pour les boules).
Montrer que

$$\mathbb{E}Z \leq \sqrt{\frac{2v}{n}\log\frac{\mathrm{e}n}{v}} + \sqrt{\frac{8\pi}{n}}$$

(les constantes ne sont pas optimisées, elles ne sont données qu'à titre indicatif).

(d) La minimisation du risque empirique consiste à choisir $\widehat{C}$ comme un minimisant de $\mathrm{err}(C)$ dans $\mathcal{C}$. Montrer que si $\mathcal{C}$ est une classe de Vapnik-Chervonenkis de dimension $v$, alors avec probabilité plus grande que $1 - 2\left(\frac{\mathrm{e}n}{v}\right)^v \mathrm{e}^{-n\epsilon^2/2}$

$$\mathrm{err}(\widehat{C}) \leq \mathrm{err}(\overline{C}) + 2\epsilon\,.$$