École Normale Supérieure
FIMFA — Statistique 2014–2015

Partiel

1. Dans cet exercice, $\begin{pmatrix} X \\ Y \end{pmatrix}$ est un vecteur Gaussien d'espérance $\mu = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$ et de covariance connue $C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. L'objectif est d'estimer θ_1. On dispose de n tirages indépendants distribués selon $\mathcal{N}(\mu, C)$: $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$.
   Dans le scenario 1, θ_2 est connu du statisticien, dans le scenario 2, θ_2 n'est pas connu (c'est un paramètre de nuisance par opposition à θ_1 qui est un paramètre d'intérêt).
   (a) Proposer un estimateur de θ_1 au maximum de vraisemblance dans le scenario 1. Est il biaisé ? Quel est son risque quadratique ?
   (b) Proposer un estimateur de θ_1 au maximum de vraisemblance dans le scenario 2. Est il biaisé ? Quel est son risque quadratique ?
   (c) Calculer l'information de Fisher dans les deux scenarios. Les deux estimateurs réalisent-ils la borne de Cramer-Rao ?
   (d) Comparer les risques des deux estimateurs. Le paramètre θ_2 est il vraiment une nuisance ?
2. Dans cet exercice, le modèle est constitué par les distributions de Laplace dont la densité sur $\mathbb{R}$ est donnée par $f_{\mu,\sigma}(x) = \exp(-|x-\mu|/\sigma)/(2\sigma)$ avec $\mu \in \mathbb{R}$ (paramètre de localisation) et $\sigma \in]0, \infty)$ (paramètre d'échelle). Les données $X_1, \ldots, X_n$ sont supposées indépendamment identiquement distribuées selon une loi de Laplace de paramètres inconnus.
   Dans un premier temps on suppose σ fixé et égal à 1.
   (a) Quelle est l'espérance de la loi de densité $f_{\mu,1}$? En déduire un estimateur $\widetilde{\mu}$ de μ par la méthode des moments. Est il biaisé ? Quel est son risque quadratique (pour une taille d'échantillon donnée) ? Est-il asymptotiquement normal ?
   (b) Ecrire la log-vraisemblance en μ d'un n-échantillon $x_1, \ldots, x_n$. Le maximum de vraisemblance est il bien défini ? unique ? Proposer un estimateur au maximum de vraisemblance $\widehat{\mu}_n$ pour les échantillons de taille impaire.
   (c) Si $(\widehat{\mu}_{2n+1})_{n\in\mathbb{N}}$ est une suite d'estimateurs au maximum de vraisemblance, est-elle consistante ?
   (d) Si oui, la suite $\sqrt{2n+1}(\widehat{\mu}_{2n+1} - \mu)$ est elle asymptotiquement gaussienne et quelle est la variance de la loi limite ?
   (e) Proposer un intervalle de niveau de confiance asymptotique $1-\alpha$ ($\alpha \in]0,1[$) pour μ.
   (f) Comparer les risques quadratiques de $\widehat{\mu}_{2n+1}$ et de $\widetilde{\mu}_{2n+1}$ lorsque $n \to \infty$.
   (g) On ne suppose plus que σ est connu, proposer un estimateur $\widehat{\sigma}$ de σ (par exemple en maximisant la vraisemblance mais ce n'est pas obligatoire). Montrer qu'il est consistant. Proposer un intervalle de niveau de confiance asymptotique $1-\alpha$ pour μ dans ce contexte.
3. Dans cet exercice $p_1, \ldots, p_n$ sont une suite de réels connus pris dans $]0,1[$. On veut distinguer deux lois $P = \otimes_{i=1}^n \mathcal{B}(p_i)$ et $Q = \otimes_{i=1}^n \mathcal{B}(1-p_i)$. On observe $X_1, \ldots, X_n$. Si H_0 est vraie, les X_i sont indépendantes et $X_i \sim \mathcal{B}(p_i)$, sous H_1, les X_i sont indépendantes et $X_i \sim \mathcal{B}(1-p_i)$ ($\mathcal{B}(p)$ est la loi de Bernoulli de paramètre p).
   (a) Ecrire la statistique du rapport de vraisemblance. Proposer un test de niveau au plus α ($\alpha \in]0,1[$). Votre test est-il le plus puissant parmi les tests de même niveau ?
   (b) On définit $p = \frac{1}{n}\sum_{i=1}^n p_i$. Montrer que l'information de Kullback entre P et Q est minorée par $n(2p-1)\log\frac{p}{1-p}$. Cette minoration est elle toujours fine ?
   (c) On considère maintenant que n tend vers l'infini. $P_n = \otimes_{i=1}^n \mathcal{B}(p_n)$ et $Q_n = \otimes_{i=1}^n \mathcal{B}(1-p_n)$. A quelle condition sur $(p_n - 1/2)_n$ peut on garantir l'existence d'une suite de tests dont les probabilités d'erreur de première et de seconde espèce convergent toutes les deux vers 0 ?

Les questions 4, 5 peuvent être traitées indépendamment.

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 $X_1, \ldots, X_n$ de même loi que X.

4. (a) On suppose que le paramètre (λ_1, λ_2) est inconnu. 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 ?

5. 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 μ 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 α de H_0 contre H_1.

6. 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 Z et d'un paramètre inconnu θ. On dispose d'un estimateur $\hat{\theta}$ de θ et on cherche un estimateur η qui soit de plus forte vraisemblance, autrement dit tel que $P_\eta[X] \geq P_{\hat{\theta}}[X]$.
   (a) Montrer que pour tout θ, on peut écrire
   $$\log P_\theta[X] = Q(\theta|\hat{\theta}) + H(\theta|\hat{\theta})$$
   où
   $$Q(\theta|\hat{\theta}) = E_\theta[\log P[X,Z]] \qquad H(\theta|\hat{\theta}) = -E_\theta[\log P[Z|X]].$$
   (b) Montrer que si $(p_1, \ldots, p_k)$ et $(q_1, \ldots, q_k)$ sont deux vecteurs de réels positifs avec $\sum p_i = \sum q_i = 1$, alors
   $$\sum p_i \log p_i \geq \sum p_i \log q_i.$$
   (c) Montrer que pour tout θ,
   $$H(\theta|\hat{\theta}) \geq H(\hat{\theta}|\hat{\theta})$$
   et donner le(s) cas d'égalité.
   (d) On propose de prendre
   $$\eta = \arg\max_\theta Q(\theta|\hat{\theta}).$$
   Vérifier qu'on a bien $P_\eta[X] \geq P_{\hat{\theta}}[X]$.
   (Cette procédure correspond à une itération de l'algorithme EM.)
7. On souhaite maintenant appliquer la méthode développée à la question 6 à l'estimateur obtenu à la question **??**. On note $\theta = (\lambda_1, \lambda_2)$ le paramètre à estimer et $\hat{\theta} = (\hat{\lambda}_1, \hat{\lambda}_2)$ l'estimateur des moments.
   (a) Donner la loi conditionnelle de Z_i sachant X_i et $\hat{\theta}$ et l'expression de $\tau_i = E_{\hat{\theta}}[Z_i|X_i]$
   (b) En déduire l'expression de $\tilde{\theta} = \arg\max_\theta Q(\theta|\hat{\theta})$ en fonction des x_i et des τ_i.