Sorbonne Université
Master Mathématiques et Applications
MU4MA015 : Statistique

9 Janvier 2025
Aucun document autorisé
Durée : 3 heures

Examen

Rappels

- Il sera tenu compte du soin apporté à la présentation et à la rédaction.
- Si vous n'arrivez pas à démontrer un résultat, vous pouvez l'admettre pour la suite de l'exercice.

Exercice 1 (8 points)
Soient $\theta \in \mathbb{R}$ un paramètre inconnu et X de densité

$$f_\theta(x) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{((\log x)-\theta)^2}{2}} \mathbb{1}_{x>0}.$$

On rappelle que si $Y \sim \mathcal{N}(0,1)$, alors pour tout réel s, on a $\mathbb{E}[e^{sY}] = e^{\frac{s^2}{2}}$.

1. Montrer que si $Y \sim \mathcal{N}(0,1)$, alors X a même loi que la variable aléatoire $e^{\theta+Y}$. Grâce au rappel, en déduire $\mathbb{E}_\theta[X]$ et $\mathrm{Var}_\theta(X)$.
   Correction. Méthode de la fonction muette ou fonction de répartition. Puis $\mathbb{E}_\theta[X] = e^{\theta+\frac{1}{2}}$ et $\mathrm{Var}_\theta(X) = e^{1+2\theta}(e-1)$.
2. On suppose désormais disposer d'un échantillon i.i.d. $(X_1, \ldots, X_n)$ de même loi que X. Déduire de la méthode des moments un estimateur $\widehat{\theta}_n$ de θ et établir sa normalité asymptotique.
   Correction. On pose $\widehat{\theta}_n := \log(\overline{X}_n) - \frac{1}{2}$, où $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, et on obtient
   $$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, e-1).$$
3. On note comme d'habitude Φ la fonction de répartition de la loi $\mathcal{N}(0,1)$. Grâce à la première question, donner l'expression de la fonction de répartition $F_\theta(x)$ de la variable X en fonction de Φ, θ et x. En déduire la médiane de la loi de X.
   Correction. Puisque $F_\theta(x) = \Phi((\log x) - \theta)\mathbb{1}_{x>0}$, bijection de $]0,\infty[$ vers $]0,1[$, on a $x_{1/2} = e^\theta$.
4. Déduire de la question précédente un nouvel estimateur $\widetilde{\theta}_n$ de θ et établir sa normalité asymptotique.
   Correction. On pose $\widetilde{\theta}_n := \log x_{1/2}(n)$ et on obtient
   $$\sqrt{n}\left(\widetilde{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \frac{\pi}{2}\right).$$
5. Déterminer l'estimateur du maximum de vraisemblance $\check{\theta}_n$. Grâce à la première question, préciser sa loi.
   Correction. On obtient $\check{\theta}_n = \frac{1}{n}\sum_{i=1}^n \log X_i \sim \mathcal{N}(\theta, 1/n)$ puisque $\log X_i = \theta + Y_i$ avec les Y_i i.i.d. selon une loi normale standard.

6. Le modèle $(f_\theta)_{\theta\in\mathbb{R}}$ est-il régulier ? Si oui, calculer l'information de Fisher $I(\theta)$ pour une donnée. Préciser si l'un des estimateurs précédents est asymptotiquement efficace.
Correction. Pour tout $x > 0$ fixé, la fonction $\theta \mapsto f_\theta(x)$ est C^1 sur $\mathbb{R}$ donc les deux premiers points de la définition d'un modèle régulier sont vérifiés. De plus, le score vaut $\ell'_\theta(X) = (\log X) - \theta$, d'où $\theta \mapsto \mathbb{E}_\theta[\ell'_\theta(X)^2] = \mathbb{E}[Y^2] = 1$ est une fonction continue sur $\mathbb{R}$, donc le modèle est régulier avec $I(\theta) = 1$. Puisque $\sqrt{n}(\breve{\theta}_n - \theta) \sim \mathcal{N}(0,1)$, l'EMV est asymptotiquement efficace (il est même efficace).

7. Soit $\alpha \in]0,1[$. Donner un intervalle de confiance de niveau $(1-\alpha)$ pour θ.
Correction. On considère bien entendu $\breve{\theta}_n$, ce qui donne l'intervalle de confiance de niveau exactement $(1-\alpha)$ et non asymptotique :
$$\left[\breve{\theta}_n - \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}} \ ; \ \breve{\theta}_n + \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}\right].$$

8. Proposer un test de niveau α pour décider entre $H_0 : \theta = 0$ et $H_1 : \theta \neq 0$. Déterminer la p-value $\alpha_0(\mathbf{x})$ associée à une réalisation $\mathbf{x} = (x_1, \dots, x_n)$.
Correction. On rejette H_0 au niveau α ssi $|\breve{\theta}_n| > \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}$. On en déduit la p-value
$$\alpha_0(\mathbf{x}) = 2\left(1 - \Phi\left(\frac{|\sum_{i=1}^n \log x_i|}{\sqrt{n}}\right)\right).$$

Exercice 2 (9 points)
Pour $a > 0$ et $b > 0$, la loi inverse-Beta(a,b), notée IB(a,b), est la loi dont la densité est donnée par
$$\theta \mapsto \frac{1}{B(a,b)}\theta^{-a-b}(\theta-1)^{b-1}\mathbb{1}_{]1,+\infty[}(\theta),$$
où $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ et Γ est la fonction gamma, vérifiant, pour tout $x > 0$, $\Gamma(x+1) = x\Gamma(x)$. La distribution IB(a,b) est la loi de l'inverse d'une variable aléatoire de loi Beta(a,b). Par ailleurs, si $Y \sim$ IB(a,b) avec $a > 2$, on a
$$\mathbb{E}[Y] = \frac{a+b-1}{a-1} \quad \text{et} \quad \mathrm{Var}(Y) = \frac{(a+b-1)b}{(a-1)^2(a-2)}.$$
On rappelle que si X suit la loi géométrique de paramètre $p \in]0,1[$, notée $\mathcal{G}(p)$, alors pour tout $k \in \mathbb{N}^*$, $\mathbb{P}(X = k) = (1-p)^{k-1}p$. De plus, $\mathbb{E}[X] = \frac{1}{p}$ et $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

On fixe maintenant $a > 0$ et un entier $n \geq 2$, et l'on considère le modèle :
$$\boldsymbol{\theta} \sim \Pi := \mathrm{IB}(a,2)$$
$$\mathbf{X} = (X_1, \dots, X_n) \,\big|\, \boldsymbol{\theta} = \theta \sim \mathcal{G}\left(\frac{1}{\theta}\right)^{\otimes n}.$$

1. Montrer que la loi a posteriori est donnée par $\Pi[\cdot \,|\, \mathbf{X}] = \mathrm{IB}(a+n, n\overline{X}_n - n + 2)$, où $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Correction. On a
$$\begin{aligned}\pi(\theta|\mathbf{X}) &\propto \theta^{-a-2}(\theta-1)\prod_{i=1}^n\left(\left(1-\frac{1}{\theta}\right)^{X_i-1}\frac{1}{\theta}\right)\mathbb{1}_{]1,+\infty[}(\theta) \\ &\propto \theta^{-a-n-2}(\theta-1)\left(1-\frac{1}{\theta}\right)^{n\overline{X}_n-n}\mathbb{1}_{]1,+\infty[}(\theta) \\ &\propto \theta^{-a-n\overline{X}_n-2}(\theta-1)^{n\overline{X}_n-n+1}\mathbb{1}_{]1,+\infty[}(\theta).\end{aligned}$$

On reconnaît le terme général de la loi $\mathrm{IB}(a+n, n\overline{X}_n - n + 2)$.

2. Donner l'espérance a posteriori, notée $m_{\mathbf{X}}$, et la variance a posteriori, notée $v_{\mathbf{X}}$.
Correction. On a

$$m_{\mathbf{X}} = \frac{a + n\overline{X}_n + 1}{a+n-1} \qquad \text{et} \qquad v_{\mathbf{X}} = \frac{(a+n\overline{X}_n+1)(n\overline{X}_n - n + 2)}{(a+n-1)^2(a+n-2)} \cdot$$

On considère la fonction de perte ℓ définie par

$$\forall \theta > 1\,,\; t > 1\,,\; \ell(\theta, t) = \frac{(t-\theta)^2}{\theta(\theta-1)} \cdot$$

3. Donner un estimateur de Bayes, noté T^*, pour la loi a priori Π et la perte ℓ. On pourra introduire la loi $\widetilde{\Pi}_{\mathbf{X}} = \mathrm{IB}(a+n+2, n\overline{X}_n - n + 1)$.
Correction. On procède par minimisation du risque a posteriori. Pour un estimateur T, on a

$$\begin{aligned}
\rho(\Pi, T \,|\, \mathbf{X}) &= \int_1^{+\infty} \frac{(T(\mathbf{X}) - \theta)^2}{\theta(\theta-1)} \pi(\theta \,|\, \mathbf{X}) d\theta \\
&= \frac{1}{B(a+n, n\overline{X}_n - n + 2)} \int_1^{+\infty} (T(\mathbf{X}) - \theta)^2\, \theta^{-a-n\overline{X}_n - 3} (\theta - 1)^{n\overline{X}_n - n} d\theta \\
&= \frac{B(a+n+2, n\overline{X}_n - n + 1)}{B(a+n, n\overline{X}_n - n + 2)} \int_1^{+\infty} (T(\mathbf{X}) - \theta)^2\, d\widetilde{\Pi}_{\mathbf{X}}(\theta)\,,
\end{aligned}$$

où $\widetilde{\Pi}_{\mathbf{X}} = \mathrm{IB}(a+n+2, n\overline{X}_n - n + 1)$. Cette quantité est minimale pour $T(\mathbf{X})$ donné par l'espérance de la loi $\widetilde{\Pi}_{\mathbf{X}}$, soit

$$T^*(\mathbf{X}) = \frac{a + n\overline{X}_n + 2}{a+n+1} \cdot$$

4. Calculer le risque a posteriori de T^*, et en déduire que le risque de Bayes $\mathbf{R}_B(\Pi)$ vaut $\frac{1}{n+a+1}$.
Correction. En remarquant que $\int_1^{+\infty} (T^*(\mathbf{X}) - \theta)^2\, d\widetilde{\Pi}_{\mathbf{X}}(\theta)$ correspond à la variance de la loi $\widetilde{\Pi}_{\mathbf{X}}$, et en utilisant la relation $\Gamma(x+1) = x\Gamma(x)$, on a

$$\begin{aligned}
\rho(\Pi, T^* \,|\, \mathbf{X}) &= \frac{B(a+n+2, n\overline{X}_n - n + 1)}{B(a+n, n\overline{X}_n - n + 2)} \frac{(a+n\overline{X}_n + 2)(n\overline{X}_n - n + 1)}{(a+n+1)^2(a+n)} \\
&= \frac{\Gamma(a+n+2)\Gamma(n\overline{X}_n - n + 1)}{\Gamma(a+n\overline{X}_n + 3)} \frac{\Gamma(a+n\overline{X}_n+2)}{\Gamma(a+n)\Gamma(n\overline{X}_n - n + 2)} \frac{(a+n\overline{X}_n + 2)(n\overline{X}_n - n + 1)}{(a+n+1)^2(a+n)} \\
&= \frac{1}{n+a+1} \cdot
\end{aligned}$$

On a donc $\mathbf{R}_B(\Pi) = \mathbb{E}[\rho(\Pi, T^* \,|\, \mathbf{X})] = \frac{1}{n+a+1}$.

5. On pose $T'(\mathbf{X}) = \frac{n\overline{X}_n + 1}{n+1}$.
   (a) Pour $\theta > 1$, calculer le risque ponctuel $\mathbf{R}(\theta, T')$ de T' (toujours pour la perte ℓ).

Correction. On a

$$\begin{aligned}
\mathbf{R}(\theta, T') &= \frac{1}{\theta(\theta-1)} \mathbb{E}_\theta \left[\left(\frac{n\overline{X}_n + 1}{n+1} - \theta \right)^2 \right] \\
&= \frac{1}{\theta(\theta-1)} \mathbb{E}_\theta \left[\left(\frac{n}{n+1} (\overline{X}_n - \theta) - \frac{(\theta-1)}{n+1} \right)^2 \right] \\
&= \frac{1}{\theta(\theta-1)} \left(\left(\frac{n}{n+1} \right)^2 \mathrm{Var}_\theta(\overline{X}_n) + \frac{(\theta-1)^2}{(n+1)^2} \right) \\
&= \frac{1}{\theta(\theta-1)} \left(\left(\frac{n}{n+1} \right)^2 \frac{\theta(\theta-1)}{n} + \frac{(\theta-1)^2}{(n+1)^2} \right) \\
&= \frac{n}{(n+1)^2} + \frac{\theta-1}{\theta(n+1)^2} \\
&= \frac{1}{n+1} - \frac{1}{\theta(n+1)^2} \,,
\end{aligned}$$

où l'on a utilisé que $\mathrm{Var}_\theta(\overline{X}_n) = \frac{1}{n} \frac{1-1/\theta}{(1/\theta)^2} = \frac{\theta(\theta-1)}{n}$.

(b) En déduire le risque maximal de T', noté $\mathbf{R}_{\max}(T')$.
Correction. On a $\mathbf{R}_{\max}(T') = \sup_{\theta>1} \mathbf{R}(\theta, T') = \frac{1}{n+1}$.

(c) Montrer que T' est minimax.
Correction. On remarque que $\frac{1}{n+1}$ correspond à la limite, quand a tend vers 0, du risque de Bayes $\frac{1}{n+a+1}$ pour l'a priori $\mathrm{IB}(a, 2)$. Par un résultat du cours, T' est minimax.

6. Soit $\theta_0 > 1$.

(a) Montrer que, sous $\mathcal{G}\left(\frac{1}{\theta_0}\right)^{\otimes n}$, on a

$$n v_{\mathbf{X}} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta_0(\theta_0 - 1) \quad \text{ et } \quad \sqrt{n}(m_{\mathbf{X}} - \theta_0) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \theta_0(\theta_0-1)) \,.$$

Correction. Par la loi des grands nombres, on a, sous $\mathcal{G}\left(\frac{1}{\theta_0}\right)^{\otimes n}$, $\overline{X}_n \xrightarrow{\mathbb{P}} \theta_0$. Ainsi, par continuité, $m_{\mathbf{X}} \xrightarrow{\mathbb{P}} \theta_0$ et

$$n v_{\mathbf{X}} = \frac{\left(\overline{X}_n + \frac{a+1}{n}\right)\left(\overline{X}_n - 1 + \frac{2}{n}\right)}{\left(1 + \frac{a-1}{n}\right)^2 \left(1 + \frac{a-2}{n}\right)} \xrightarrow{\mathbb{P}} \theta_0(\theta_0 - 1) \,.$$

De plus, on a

$$\sqrt{n}(m_{\mathbf{X}} - \theta_0) = \sqrt{n} \frac{n}{a+n-1} (\overline{X}_n - \theta_0) + \sqrt{n} \frac{a+1-(a-1)\theta_0}{a+n-1} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \theta_0(\theta_0-1)) \,,$$

par le lemme de Slutsky.

(b) En déduire que pour toute suite positive $(M_n)_{n\geq 1}$ avec $M_n \xrightarrow[n\to\infty]{} +\infty$, on a, sous $\mathcal{G}\left(\frac{1}{\theta_0}\right)^{\otimes n}$,

$$\mathbb{P}\left(|\boldsymbol{\theta} - \theta_0| > \frac{M_n}{\sqrt{n}} \,\Big|\, \mathbf{X} \right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \,.$$

Correction. En utilisant l'inégalité triangulaire et une borne union, on a

$$\mathbb{P}\left(|\boldsymbol{\theta} - \theta_0| > \frac{M_n}{\sqrt{n}} \,\Big|\, \mathbf{X} \right) \leq \mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{M_n}{2\sqrt{n}} \,\Big|\, \mathbf{X} \right) + \mathbb{1}_{\left\{ |m_{\mathbf{X}} - \theta_0| > \frac{M_n}{2\sqrt{n}} \right\}} \,.$$

Par la question précédente, on a $\sqrt{n}(m_{\mathbf{X}} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta_0(\theta_0 - 1))$. Ainsi, comme $M_n \to +\infty$, on a

$$\mathbb{P}_{\theta_0}(\sqrt{n}|m_{\mathbf{X}} - \theta_0| > M_n/2) \underset{n\to\infty}{\longrightarrow} 0\,,$$

soit $\mathbb{1}_{\{\sqrt{n}|m_{\mathbf{X}} - \theta_0| > M_n/2\}} \xrightarrow{\mathbb{P}} 0$. D'autre part, par l'inégalité de Tchebytchev, on a

$$\mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{M_n}{2\sqrt{n}} \,\middle|\, \mathbf{X}\right) \leq \frac{4nv_{\mathbf{X}}}{M_n^2}\,.$$

Comme $nv_{\mathbf{X}} \xrightarrow{\mathbb{P}} \theta_0(\theta_0 - 1)$ et $M_n \to +\infty$, on obtient $\mathbb{P}\left(|\boldsymbol{\theta} - m_{\mathbf{X}}| > \frac{M_n}{2\sqrt{n}} \,\middle|\, \mathbf{X}\right) \xrightarrow{\mathbb{P}} 0$.

Exercice 3 (3 points)
Soit la fonction $f : \mathbb{R}^2 \to \mathbb{R}_+$, définie par :

$$\forall (u,v) \in \mathbb{R}^2, \quad f(u,v) = \frac{(\cos|u+v|)^2}{100 + u^4 + v^4}.$$

On admet que $\iint_{\mathbb{R}^2} f(u,v)\,\mathrm{d}u\mathrm{d}v < \infty$.

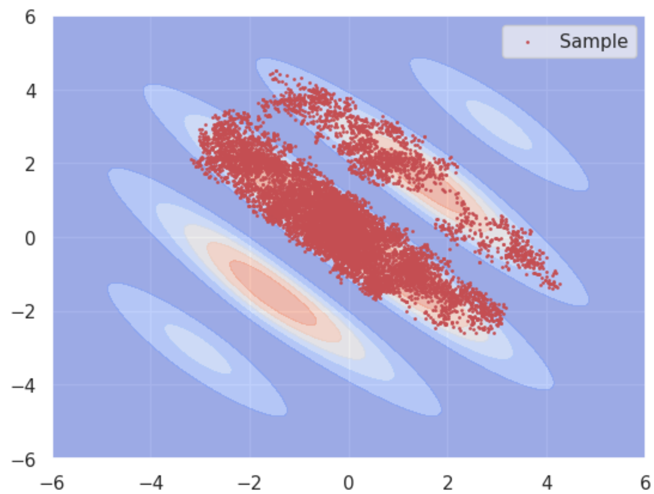
1. Par quoi faut-il remplacer `???` et `$$$` pour que l'algorithme de Metropolis-Hastings ci-dessous, où `f` représente la fonction f, simule une suite de variables aléatoires (X_i) dont la loi limite est de densité proportionnelle à f ?

[3]:
```
unif = stats.uniform()
kernel = stats.multivariate_normal(cov=1e-2 * np.eye(2))

n = 10000

threshold = unif.rvs(size=n)  # Array of the n thresholds U_i
move = kernel.rvs(size=n)  # Array of the n moves from the current position

sample = np.zeros((n+1, 2))
x = sample[0]
for i in range(n):
    y = x + move[i]
    crit = f(y) / f(x)
    dec = ???
    if dec:
        sample[i+1] = y
        x = y
    else:
        $$$
```

   Correction. On doit remplacer `???` par `crit >= threshold[i]` et `$$$` par `sample[i+1] = x`.

2. Expliquer cet algorithme grâce à des éléments théoriques.
   Correction. Au vu du script à trous, et en particulier de la ligne `crit = f(y) / f(x)`, il est clair qu'il s'agit d'un algorithme de Metropolis. Le noyau de proposition est ici, pour tout $x \in \mathbb{R}^2$, $\mathcal{N}(x, 10^{-2}I_2)$, dont la densité $q(x,\cdot)$ produit bien une fonction q symétrique (d'où le cas particulier de Metropolis-Hastings et le ratio qui devient $\frac{f(y)}{f(x)}$, où $y \in \mathbb{R}^2$ est le candidat).

3. Voici une représentation des courbes de niveau de f et d'une réalisation de la suite $(X_i)_{0\le i\le n}$ construite par l'algorithme précédent (`Sample` sur la figure). Quelle est votre analyse de ce résultat ? Quelle correction peut-on apporter ?



Correction. Le noyau de proposition est de variance trop faible, ce qui empêche la chaîne de Markov de bien explorer l'ensemble des endroits où la densité est élevée. Il suffirait par exemple de prendre une variance I_2 au lieu de $10^{-2}I_2$.