RPM4 2023: correction du TD statistique

Clément Gauchy

Novembre 2023

Exercice 1

1) Le modèle statistique de ce problème est $\mathcal{M} = \{\mathcal{U}([-\alpha, \alpha]),\ \alpha > 0\}$.

2) La densité de la loi $\mathcal{U}([-\alpha, \alpha])$ s'écrit

$$f(x) = \frac{1}{2\alpha}\mathbf{1}_{x\in[-\alpha,\alpha]}\ ,$$

Par définition, la vraisemblance s'écrit

$$\mathcal{L}(\alpha, X_1, \dots, X_n) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \frac{1}{2\alpha}\mathbf{1}_{X_i\in[-\alpha,\alpha]} = \left(\frac{1}{2\alpha}\right)^n \prod_{i=1}^{n} \mathbf{1}_{X_i\in[-\alpha,\alpha]}$$

3) On remarque que la vraisemblance est maximisé quand tout les $\mathbf{1}_{X_i\in[-\alpha,\alpha]}$ sont égaux à 1 et que $(1/2\alpha)^n$ est maximisé. On a donc $\widehat{\alpha}_{\mathrm{MV}} = \max_{1\leqslant i\leqslant n} |X_i|$

Exercice 2

1) Le modèle statistique est $\mathcal{M} = \{f_\lambda(t) = \lambda e^{-\lambda t}\ ,\ \lambda > 0\}$, le paramètre à estimer est λ.
2) D'après la loi des grands nombres, on sait que

$$\frac{1}{n}\sum_{i=1}^{n} T_i \to \mathbb{E}[T_1] = \frac{1}{\lambda}\ ,$$

ainsi on peut poser l'estimateur des moments suivant pour λ:

$$\frac{1}{\widehat{\lambda}_n} = \frac{1}{n}\sum_{i=1}^{n} T_i \iff \widehat{\lambda}_n = \frac{n}{\sum\limits_{i=1}^{n} T_i}$$

3) On a par définition

$$\begin{aligned}\mathcal{L}(\lambda; T_1, \dots, T_n) &= \prod_{i=1}^{n} f_\lambda(T_i)\\ \mathcal{L}(\lambda; T_1, \dots, T_n) &= \lambda^n e^{-\lambda \sum\limits_{i=1}^{n} T_i}\end{aligned}$$

La log vraisemblance s'écrit alors

$$\ell(\lambda; T_1, \ldots, T_n) = n \log(\lambda) - \lambda \sum_{i=1}^{n} T_i$$

le maximum de vraisemblance $\widehat{\lambda}_{\mathrm{MV}}$ annule la dérivée de la log vraisemblance:

$$\frac{\partial \ell(\widehat{\lambda}_{\mathrm{MV}}; T_1, \ldots, T_n)}{\partial \lambda} = 0 \iff \frac{n}{\widehat{\lambda}_{\mathrm{MV}}} - \sum_{i=1}^{n} T_i = 0$$

Ainsi,

$$\widehat{\lambda}_{\mathrm{MV}} = \frac{n}{\sum_{i=1}^{n} T_i}$$

L'estimateur du maximum de vraisemblance est donc le même que celui des moments.

4) On écrit l'espérance $\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} T_i}\right]$ à l'aide du théorème de transport

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} T_i}\right] &= \int_{\mathbb{R}^+} \frac{\lambda^n}{\Gamma(n)} t^{n-1} e^{-\lambda t} \times \frac{1}{t} dt \\
\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} T_i}\right] &= \int_{\mathbb{R}^+} \frac{\lambda^n t^{n-2}}{\Gamma(n)} e^{-\lambda t} dt \\
\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} T_i}\right] &= \frac{\lambda}{n-1} \underbrace{\int_{\mathbb{R}^+} \frac{\lambda^{n-1} t^{n-2} e^{-\lambda t}}{\Gamma(n-1)} dt}_{=1 \text{ car c'est la densité de probabilité de la loi } \mathcal{G}(n-1, \lambda)}
\end{aligned}$$

On peut désormais calculer le biais

$$\mathbb{E}[\widehat{\lambda}_{\mathrm{MV}}] = n \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} T_i}\right] = \frac{n}{n-1} \lambda$$

L'estimateur du maximum de vraisemblance est donc biaisé mais asymptotiquement sans biais.

6) On applique la règle de Bayes:

$$\begin{aligned}
\pi(\lambda | T_1, \ldots, T_n) &\propto \mathcal{L}(\lambda; T_1, \ldots, T_n) \pi(\lambda) \\
\pi(\lambda | T_1, \ldots, T_n) &\propto \lambda^n e^{-\lambda \sum_{i=1}^{n} T_i} \lambda^{\alpha} e^{-\beta \lambda} \\
\pi(\lambda | T_1, \ldots, T_n) &\propto \lambda^{n+\alpha} e^{-\lambda(\beta + \sum_{i=1}^{n} T_i)}
\end{aligned}$$

On reconnait la loi $\mathcal{G}(n + \alpha, \beta + \sum_{i=1}^{n} T_i)$, qui est donc la loi *a posteriori* pour ce modèle.

7) D'après les informations de l'énoncé, on a l'estimateur de la moyenne *a posteriori* $\widehat{\lambda}_{\text{moy}}$ qui est égale à

$$\widehat{\lambda}_{\text{moy}} = \frac{\alpha + n}{\beta + \sum_{i=1}^{n} T_i} .$$

L'estimateur du maximum *a posteriori* $\widehat{\lambda}_{\text{MAP}}$ s'écrit quant à lui

$$\widehat{\lambda}_{\text{MAP}} = \frac{\alpha + n - 1}{\beta + \sum_{i=1}^{n} T_i} .$$

Exercice 3

1) X suit une loi exponentielle de paramètre μ, $X \sim \mathcal{E}(\mu)$.

2) p représente la proportion de particules qui transervent le matériau sans interaction avec celui ci. Cette quantité est importante car elle représente le rayonnement que l'opérateur recevra. Si le rayonnement est ionisant, cela peut avoir des conséquences de santé sur l'opérateur.

3) On utilise la propriété $\mathbb{P}(X > L) = \mathbb{E}[\mathbf{1}_{X>L}]$. On peut écrire l'estimateur Monte-Carlo à l'aide de la loi des grands nombres.

$$\widehat{p}_{\text{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{X_i > L}$$

4) On a $\mathbb{E}[\widehat{p}_{\text{MC}}] = p$ et $\text{Var}(\widehat{p}_{\text{MC}}) = \frac{p(1-p)}{N}$ ainsi

$$cv = \frac{\sqrt{p(1-p)}}{p\sqrt{N}} .$$

Vu que $p = 10^{-6}$ on peut faire l'approximation $1 - p \approx 1$ et ainsi

$$cv \approx \frac{1}{\sqrt{Np}}$$

Si on souhaite $cv \approx 1\% = 10^{-3}$ alors $N = 10^{12}$. On remarque que c'est un nombre de simulations aléatoires très important, on peut attendre les limites des moyens de calcul. De plus, dans un modèle plus réaliste (exemple: matériau non homogène) chaque simulation d'une trajectoire de particule peut être couteuse en temps de calcul, l'estimateur Monte-Carlo va donc être très compliqué à calculer.

5) On remarque que h_θ est la densité de la loi exponentielle $\mathcal{E}(\mu - \theta)$. On peut voir que le libre parcours moyen pour une particule de section efficace macroscopique $\mu - \theta$ est plus grand que pour une particule ayant μ pour section efficace macroscopique. La variable aléatoire Y simule donc la trajectoire d'une particule "virtuelle" qui interagit moins avec le matériau.

6) Par définition, l'estimateur par échantillonage d'importance s'écrit

$$\widehat{p}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^{N} \frac{h_0(Y_i)}{h_\theta(Y_i)} \mathbf{1}_{X_i > L}$$

7)

$$
\begin{aligned}
\mathrm{Var}(\widehat{p}_{\mathrm{IS}}) &= \frac{1}{N}\left(\mathbb{E}_{Y\sim h_\theta}\left[\frac{h_0(Y)^2}{h_\theta(Y)^2}\mathbf{1}_{Y>L}\right] - \mathbb{E}_{Y\sim h_\theta}\left[\frac{h_0(Y)}{h_\theta(Y)}\mathbf{1}_{Y>L}\right]^2\right)\\
\mathrm{Var}(\widehat{p}_{\mathrm{IS}}) &= \frac{1}{N}\left(\mathbb{E}_{X\sim h_0}\left[\frac{h_0(X)}{h_\theta(X)}\mathbf{1}_{X>L}\right] - p^2\right)\\
\mathrm{Var}(\widehat{p}_{\mathrm{IS}}) &= \frac{1}{N}\left(\mathbb{E}_{X\sim h_0}\left[\frac{\mu e^{-\theta X}}{\mu-\theta}\mathbf{1}_{X>L}\right] - p^2\right)\\
\mathrm{Var}(\widehat{p}_{\mathrm{IS}}) &= \frac{1}{N}\left(\frac{\mu}{\mu-\theta}\mathbb{E}_{X\sim h_0}\left[e^{-\theta X}\mathbf{1}_{X>L}\right] - p^2\right)
\end{aligned}
$$

On utilise la décroissance de $e^{-\theta x}$ pour borner supérieurement la variance

$$
\mathrm{Var}(\widehat{p}_{\mathrm{IS}}) < \frac{1}{N}\left(\frac{\mu e^{-\theta L}}{\mu-\theta}p - p^2\right)
$$

On doit donc étudier la fonction $g\ :\ \theta \mapsto \frac{\mu e^{-\theta L}}{\mu-\theta}$ pour vérifier que la variance est plus basse que celle de l'estimateur Monte-Carlo.

$$
g'(\theta) = \frac{e^{-\theta L}}{\mu-\theta}\left(\frac{\mu}{\mu-\theta} - L\mu\right)
$$

On doit donc vérifier $\theta < \mu - \frac{1}{L}$ et également $\mu - \frac{1}{L} > 0$ (étant donné que θ doit être positif) donc $\frac{1}{\mu} < L$. Le libre parcours moyen doit être plus petit que la longueur L pour l'estimateur par échantillonage d'importance ait une variance plus faible que celle de l'estimateur Monte-Carlo.