

Modélisation statistique (STA9)

Examen (3h)

Olivier Rioul

5 octobre 2026

Contrôle avec **notes de cours manuscrites, UNIQUEMENT**. Calculatrices autorisées. Pour chaque partie, répondre dans l'ordre aux questions posées. Ne pas recopier l'énoncé sur la copie. On détaillera avec précision tous les raisonnements. Rédiger lisiblement, en faisant des phrases en langue humaine SVP. Le soin et la présentation comptent pour la note finale. **Si, au cours de l'épreuve, un élève repère ce qui lui semble être une erreur d'énoncé, il le signale sur sa copie et poursuit sa composition en expliquant les raisons des initiatives qu'il est amené à prendre.**

Rappels de probabilités

- La loi Gamma $\Gamma(\alpha, \lambda)$ ($\alpha > 0, \lambda > 0$) est définie par la densité

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \mathbb{1}_{x>0}$$

- Si $X \sim \Gamma(\alpha, \lambda)$, ses espérance, variance et mode sont respectivement

$$\mathbf{E}(X) = \frac{\alpha}{\lambda} \quad \text{Var}(X) = \frac{\alpha}{\lambda^2} \quad \text{mode}(X) = \frac{\alpha-1}{\lambda} \quad (\text{si } \alpha > 1)$$

- Si $X \sim \Gamma(\alpha, \lambda)$ alors $cX \sim \Gamma(\alpha, \lambda/c)$ pour tout $c > 0$.
- Si $X_1 \sim \Gamma(\alpha_1, \lambda)$ et $X_2 \sim \Gamma(\alpha_2, \lambda)$ sont indépendantes alors leur somme $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.
- La loi exponentielle $\text{Exp}(\lambda)$ est la loi $\Gamma(1, \lambda)$.
- Si $X \sim \Gamma(\alpha, \lambda)$ alors $Y = \frac{1}{X}$ suit la loi inverse Gamma $\Gamma^{-1}(\alpha, \lambda)$ définie par la densité

$$g(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{e^{-\lambda/y}}{y^{\alpha+1}} \mathbb{1}_{y>0}$$

- Si $Y \sim \Gamma^{-1}(\alpha, \lambda)$, ses espérance (si $\alpha > 1$), variance (si $\alpha > 2$) et mode sont respectivement

$$\mathbf{E}(Y) = \frac{\lambda}{\alpha-1} \quad \text{Var}(Y) = \frac{\lambda^2}{(\alpha-1)^2(\alpha-2)} \quad \text{mode}(Y) = \frac{\lambda}{\alpha+1}$$

Estimation du trafic internet

ON MODÉLISE LE TRAFIC internet à l'aide d'une distribution de la taille des paquets envoyés, lorsque celle-ci dépasse un seuil u connu à l'avance, donnée par la distribution de *Pareto* de seuil $u > 0$ et d'indice $\theta > 0$ définie par :

$$p_{\theta}(x) = \frac{\theta u^{\theta}}{x^{\theta+1}} \mathbb{1}_{x>u}$$

(densité par rapport à la mesure de Lebesgue). Le modèle statistique est défini par $X = (X_1, X_2, \dots, X_n)$ i.i.d. selon la loi de Pareto (u, θ) , avec un nombre d'observations $n \geq 3$. On cherche à estimer l'indice de forme $\theta > 0$.

1. Donner l'expression de la log-vraisemblance $\ell_X(\theta)$ pour le modèle.
2. En déduire (en le justifiant soigneusement) que l'estimateur ML est

$$\hat{\theta}(X) = \frac{n}{\sum_{i=1}^n \log \frac{X_i}{u}}$$

3. Calculer $\mathbf{P}\{X_i > x\}$ pour $x \geq u$ et en déduire que $Y_i = \log \frac{X_i}{u}$ suit une loi exponentielle de paramètre θ .
 4. En déduire la loi de $\sum_{i=1}^n Y_i$
 5. En déduire la loi de l'estimateur ML de θ , noté $\hat{\theta}(X)$.
 6. Calculer le biais de l'estimateur ML.
 7. En déduire le risque quadratique de l'estimateur ML.
 8. Montrer que l'estimateur corrigé $\hat{\theta}'(X) = \frac{n-1}{\sum_{i=1}^n \log \frac{X_i}{u}}$ est non biaisé et calculer sa variance.
 9. Quel est le meilleur estimateur entre $\hat{\theta}$ et $\hat{\theta}'$?
 10. Montrer que $S = \sum_{i=1}^n Y_i$ est une statistique suffisante pour le modèle.
 11. Montrer que S est une statistique complète (totale). On pourra utiliser l'injectivité de la transformée de Laplace $\hat{f}(p) = \int_0^{\infty} f(t) e^{-pt} dt$ ($p > 0$).
 12. Les estimateurs $\hat{\theta}$ et $\hat{\theta}'$ sont-ils MVU?
 13. Calculer l'information de Fisher $J_n(\theta)$ pour le modèle statistique considéré (dont on admettra qu'il est régulier).
 14. Les estimateurs $\hat{\theta}$ et $\hat{\theta}'$ sont-ils efficaces?
 15. Les estimateurs $\hat{\theta}$ et $\hat{\theta}'$ sont-ils asymptotiquement efficaces?
 16. Les estimateurs $\hat{\theta}$ et $\hat{\theta}'$ sont-ils asymptotiquement normaux? (Si oui, on ne cherchera pas à déterminer les paramètres de la loi normale limite).
- On suppose maintenant que θ suit la distribution *a priori* $\theta \sim \Gamma(\alpha, \lambda)$
17. Calculer la loi *a posteriori*.
 18. En déduire l'estimateur MAP.
 19. En déduire l'estimateur MMSE.
 20. Peut-on retrouver le cas limite sans *a priori*? Commenter

Solutions

-1) $\ell_X(\theta) = \log p_\theta(X) = \sum_{i=1}^n \log p_\theta(X_i) = \sum_{i=1}^n \log\left(\frac{\theta u^\theta}{X_i^{\theta+1}} \mathbb{1}_{X_i > u}\right) = n \log \theta + n \theta \log u - (\theta + 1) \sum_{i=1}^n \log X_i$ pour tout $X \in]u, +\infty[^n$.

-2) La log-vraisemblance étant concave sur le domaine considéré, elle est maximale lorsque la dérivée s'annule :

$$\frac{\partial}{\partial \theta} \ell_X(\theta) = \frac{n}{\theta} + n \log u - \sum_{i=1}^n \log X_i = 0$$

d'où le résultat.

-3) On a $\mathbf{P}\{X_i > x\} = \int_x^{+\infty} \frac{\theta u^\theta}{t^{\theta+1}} dt = \left(\frac{u}{x}\right)^\theta$ d'où pour $y > 0$, $\mathbf{P}\{Y_i > y\} = \mathbf{P}\{X_i > ue^y\} = e^{-\theta y}$ ce qui montre que $Y_i \sim \text{Exp}(\theta)$.

-4) D'après les rappels, $\sum_{i=1}^n Y_i \sim \Gamma(n, \theta)$.

-5) Puis $\frac{1}{n} \sum_{i=1}^n Y_i \sim \Gamma(n, n\theta)$ et donc $\hat{\theta}(X) \sim \Gamma^{-1}(n, n\theta)$.

-6) D'après les rappels, on a (comme $n > 1$) $\mathbf{E} \hat{\theta}(X) = \frac{n\theta}{n-1}$ d'où un biais $= \frac{n\theta}{n-1} - \theta = \frac{\theta}{n-1}$ non nul.

-7) De plus (comme $n > 2$) $\text{Var}(\hat{\theta}(X)) = \frac{n^2 \theta^2}{(n-1)^2(n-2)}$. On en déduit le MSE=biais²+variance :

$$\text{MSE} = \frac{\theta^2}{(n-1)^2} + \frac{n^2 \theta^2}{(n-1)^2(n-2)} = \frac{\theta^2}{(n-1)^2} \frac{(n-1)(n+2)}{n-2} = \frac{n+2}{(n-1)(n-2)} \theta^2.$$

-8) Puisque $\hat{\theta}'(X) = \frac{n-1}{n} \hat{\theta}(X)$, on a $\mathbf{E} \hat{\theta}'(X) = \theta$ d'où un biais nul et une variance

$$\text{MSE}' = \text{Var}(\hat{\theta}'(X)) = \frac{(n-1)^2}{n^2} \frac{n^2 \theta^2}{(n-1)^2(n-2)} = \frac{\theta^2}{n-2}.$$

-9) Puisque $\text{MSE} > \text{MSE}'$ c'est $\hat{\theta}'$ qui est meilleur.

-10) On avait $\ell_X(\theta) = \log p_\theta(X) = n \log \theta + n \theta \log u - (\theta + 1) \sum_{i=1}^n \log X_i = n \log \theta + n(2\theta + 1) \log u - (\theta + 1)S$ qui vérifie bien la factorisation de Neyman-Fisher.

-11) Puisque $S \sim \Gamma(n, \theta)$, si pour tout θ on a $\mathbf{E} f(S) = 0$, il vient

$$\int_0^\infty s^{n-1} f(s) e^{-\theta s} ds = 0$$

qui est la transformée de Laplace en $\theta > 0$ de la fonction $s^{n-1} f(s)$, d'où $f(s) = 0$.

-11) $\hat{\theta}$ n'est pas MVU car biaisé; $\hat{\theta}'$ est non biaisé et MVU d'après le théorème de Lehmann-Scheffé. En effet S est une statistique suffisante totale et $\mathbf{E}(\hat{\theta}'|S) = \hat{\theta}'$ puisque $\hat{\theta}' = \frac{n-1}{S}$ est une fonction de S .

-13) On a $J_n(\theta) = nJ_1(\theta)$ car le modèle est i.i.d., et pour une observation le score (précédemment calculé) vaut $S_X(\theta) = \frac{1}{\theta} - \log \frac{X}{u}$ d'où en prenant l'opposé de la dérivée par rapport à θ , $J_1(\theta) = \frac{1}{\theta^2}$. Ainsi $J_n(\theta) = \frac{n}{\theta^2}$.

-14) $\hat{\theta}$ n'est pas efficace car biaisé; $\hat{\theta}'$ est non biaisé mais non efficace car sa variance est strictement supérieure à $1/J_n(\theta) = \theta^2/n$.

-15) Ils sont tous deux asymptotiquement efficaces ($\hat{\theta}$ est asymptotiquement non biaisé et ils sont tous deux de variance $\sim \theta^2/n$ lorsque $n \rightarrow +\infty$).

-16) On peut écrire $\hat{\theta} = \frac{1}{\bar{Y}_n}$ et $\hat{\theta}' = \frac{n-1}{n} \frac{1}{\bar{Y}_n}$, le TCL et la δ -méthode appliquée aux Y_i (donc à $\bar{Y}_n$) donne la normalité asymptotique.

-17) À une constante multiplicative (indépendante de θ) près,

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) = \theta^n u^{n\theta} \left(\prod_{i=1}^n x_i \right)^{-(\theta+1)} \theta^{\alpha-1} e^{-\lambda\theta} \\ &\propto \theta^{n+\alpha-1} e^{-\theta(\lambda + \sum_i \log \frac{x_i}{u})} \end{aligned}$$

qui suit la loi $\Gamma(n + \alpha - 1, \lambda + \sum_i \log \frac{x_i}{u})$.

-18) On en déduit le MAP

$$\hat{\theta}_{\text{MAP}}(X) = \mathbf{E}(\theta|X) = \frac{n + \alpha - 2}{\lambda + \sum_i \log \frac{x_i}{u}}$$

-19) On en déduit le MMSE

$$\hat{\theta}_{\text{MMSE}}(X) = \mathbf{E}(\theta|X) = \frac{n + \alpha - 1}{\lambda + \sum_i \log \frac{x_i}{u}}$$

-20) Pour le MAP et le MMSE, le cas limite $\lambda \rightarrow 0$ (aucun a priori) et $\alpha = 0$ ou 1 redonne les estimateurs précédents. Le cas limite $n \rightarrow +\infty$ donne un équivalent

$$\frac{1}{\bar{Y}_n} = \hat{\theta}_{\text{ML}}$$