

Fiche 4

Emeline Schmisser, `emeline.schmisser@math.univ-lille1.fr`, bureau 314 (bâtiment M3).

Exercice 1 (Maximum de vraisemblance pour la loi Gamma)

Les lois Gamma $\Gamma(k, \theta)$ sont des lois de probabilité définies à partir :

1. d'un paramètre d'échelle $\theta > 0$
2. d'un paramètre de forme $k > 0$

par des densités de probabilité de la forme :

$$f(x; \theta, k) = \frac{x^{k-1} \exp\left(-\frac{x}{\theta}\right)}{\Gamma(k)\theta^k} \mathbf{1}_{]0, +\infty[}(x), \quad \text{où} \quad \Gamma(k) = \int_0^{+\infty} t^{k-1} e^{-t} dt. \quad (1)$$

Les lois Gamma sont utilisées pour modéliser des variables aléatoires positives, souvent des durées. Par exemple, pour des événements se réalisant suivant un processus de Poisson de paramètre θ , l'évènement N (pour N fixé) suit une loi $\Gamma(N, \theta)$.

Partie A (théorique)

1. Vérifier que lorsque $k = 1$, on retrouve une loi exponentielle.
2. Calculer l'espérance, le mode, la variance de la loi $\Gamma(k, \theta)$.
3. Soient $(X_i)_{i \in \llbracket 1, n \rrbracket}$ n variables aléatoires indépendantes de lois respectives $\Gamma(\alpha_i, \theta)$, $i \in \llbracket 1, n \rrbracket$. Montrer que $\sum_{i=1}^n X_i$ suit une loi $\Gamma(\sum_{i=1}^n \alpha_i, \theta)$ (Propriété d'infinie divisibilité).
4. Prouver que si X suit une loi $\Gamma(k, \theta)$ et si $\lambda > 0$, alors λX suit une loi $\Gamma(k, t\theta)$.

Partie B - Estimateur du maximum de vraisemblance (Manipulations sur \mathbf{R})

1. Dessiner la densité de la loi Gamma pour $k = 3$ et $\theta = 0.5$, avec `dgamma(, shape=3, scale=0.5)`.
2. On se donne n variables aléatoires $(X_i)_{i \in \llbracket 1, n \rrbracket}$ indépendantes identiquement distribuées de loi $\Gamma(k, \theta)$. Ecrire la log-vraisemblance renormalisée des observations $\ell(x_1, \dots, x_n; k, \theta)/n$ et programmer-la en tant que fonction \mathbf{R} (On utilisera la commande `gamma(k)` pour calculer $\Gamma(k)$).
3. Simuler un échantillon avec $n = 1000$, $k = 3$ et $\theta = 0.5$ avec `x<-rgamma(n=1000, shape=3, scale=0.5)`. Nous nous proposons dans cet exercice de déterminer les estimateurs du maximum de vraisemblance \hat{k} et $\hat{\theta}$.
4. Représenter par un graphique en 2D la fonction $(k, \theta) \in \mathbb{R}_+^2 \mapsto \ell(x_1, \dots, x_n; k, \theta)/n \in \mathbb{R}$. On calculera pour cela la valeur de la fonction sur une grille de $]0, 5]^2$ de pas $0,05$, puis on utilisera les commandes `persp`, `image` et `contour`. En particulier, pour `contour`, jouer avec les options `levels nlevels` et modifier le cadrage avec les options `xlim=c(,) ylim=c(,)`. Déterminer graphiquement l'estimateur du maximum de vraisemblance (EMV).

5. Déterminer par le calcul l'EMV $\widehat{\theta}_n$ du paramètre θ .

6. Peut-on obtenir une expression explicite pour l'EMV \widehat{k}_n du paramètre k ? Montrer qu'il vérifie l'équation suivante où ψ est la fonction digamma définie par $\psi(k) = \Gamma'(k)/\Gamma(k)$:

$$\ln(k) - \psi(k) = \ln\left(\frac{1}{n} \sum_{i=1}^n x_i\right) - \frac{1}{n} \sum_{i=1}^n \ln(x_i). \quad (2)$$

7. Etudier la convexité de la fonction $k \mapsto \ln(k) - \psi(k)$.

8. Nous souhaitons résoudre numériquement (2) en utilisant la méthode de Newton-Raphson.

8.1. Rappeler le principe de cette méthode.

8.2. Déterminer l'expression pour passer de l'approximation $k^{(m)}$ à l'approximation $k^{(m+1)}$ de la solution de (2) (on fera intervenir la fonction trigamma ψ' , dérivée de la fonction ψ).

8.3. Pour déterminer la condition initiale, nous allons utiliser l'approximation suivante, qui fait l'objet de la partie **D** (estimation par la méthode des moments) :

$$\tilde{\theta}_n = \frac{s_n^2}{\bar{x}_n}, \quad \tilde{k}_n = \frac{\bar{x}_n^2}{s_n^2}, \quad (3)$$

où \bar{x}_n et s_n^2 sont respectivement la moyenne et la variance empirique (commandes `mean` et `var`).

8.4. Rassembler ces résultats dans une fonction EMV1 d'argument (x_1, \dots, x_n) et retournant :

- l'EMV $(\widehat{k}_n, \widehat{\theta}_n)$,
- le nombre d'itérations faites dans l'algorithme de Newton-Raphson pour obtenir le résultat ainsi,
- la valeur de $\ln(k) - \psi(k) - \ln(\bar{x}_n) + \overline{\ln(x)}$.

On utilisera les commandes `digamma(k)` ou `psigamma(k, 1)` pour calculer $\psi(k)$ et les commandes `trigamma(k)` ou `psigamma(k, 2)` pour calculer $\psi'(k)$.

9. Nous souhaitons comparer les résultats précédents à ceux retournés par les méthode `nlm` et `mle` pré-implémentée dans **R**.

9.1. Montrer que \widehat{k}_n minimise la fonction suivante :

$$\tilde{\ell}(x_1, \dots, x_n; k) = -k \left(\sum_{i=1}^n \frac{\ln(x_i)}{n} - 1 - \ln(\bar{x}) + \ln(k) \right) + \ln(\Gamma(k)) \quad (4)$$

9.2. Tracer la courbe de la fonction $k \mapsto \tilde{\ell}(x_1, \dots, x_n; k)$. Déterminer graphiquement son minimum.

9.3. Calculer $\widehat{k}_n^{(2)}$ en minimisant la fonction $k \mapsto \tilde{\ell}(x_1, \dots, x_n; k)$ à l'aide de la commande `nlm`. On prendra les mêmes conditions initiales qu'en (3).

9.4. Calculer $\widehat{k}_n^{(3)}$ en utilisant la commande `mle`, après avoir préalablement appelé la librairie `stats4` (`Taper library(stats4)`).

9.5. Rassembler les résultats précédents dans une fonction EMV2 qui prend en argument (x_1, \dots, x_n) et retourne les EMV $(\widehat{k}_n^{(2)}, \widehat{k}_n^{(3)})$.

9.5. Pour comparer les trois estimateurs de k :

- Générer $N = 1000$ échantillons de $n = 100$ observations indépendantes et de loi $\Gamma(k = 3, \theta = 0.5)$.
- Pour chacun des échantillons générés, calculer \widehat{k}_n , $\widehat{k}_n^{(2)}$ et $\widehat{k}_n^{(3)}$ en utilisant les fonctions obtenues aux questions 8 et 9. Conserver les valeurs trouvées dans un tableau.
- A la fin de la manœuvre, nous avons trois séries de $N = 1000$ réalisations (couplées) des estimateurs \widehat{k}_n , $\widehat{k}_n^{(2)}$ et $\widehat{k}_n^{(3)}$. Utiliser-les pour comparer ces deux distributions (moyennes, QQ-plots ...)

Partie C Consistence et vitesse de convergence (R)

1. (théorique) L'EMV dans cet exercice est-il convergent ? asymptotiquement normal ?
2. Ecrire une fonction qui prend un entier n en argument et :
 - génère un échantillon *iid* de taille n dans la loi $\Gamma(k = 3, \theta = 0.5)$,
 - calcule $\widehat{k}_n - 3$ à l'aide de la fonction EMV1 obtenue à la partie B.8,(c'est une fonction aléatoire, puisque \widehat{k}_n est une variable aléatoire qui dépend de l'échantillon généré)
3. Tracer le graphe de $n \mapsto \widehat{k}_n - 3$ pour $n \in \llbracket 10, 3000 \rrbracket$, n variant de 10 en 10.
4. On cherche à montrer que $\sqrt{n}(\widehat{k} - k)$ est normalement asymptotique.
 - Générer $N = 1000$ échantillons de $n = 100$ observations indépendantes et de loi $\Gamma(k = 3, \theta = 0.5)$.
 - Pour chacun de ces échantillons, calculer $\sqrt{n}(\widehat{k} - k)$. On obtient $N = 1000$ réalisations de cette variable aléatoire.
 - Faire un QQ-plot pour la comparer avec une loi normale.
5. Par quelle quantité doit-on renormaliser $\sqrt{n}(\widehat{k} - k)$ pour obtenir une loi asymptotique $\mathcal{N}(0, 1)$? Reprendre la question 4 avec une version renormalisée de $\sqrt{n}(\widehat{k} - k)$ et comparer avec la loi $\mathcal{N}(0, 1)$.

Partie D Estimateur des moments (R)

1. (théorique) Calculer k et θ en fonction de l'espérance et de la variance obtenus en 2.
2. (théorique) Justifier les estimateurs proposés en (3). Ces estimateurs sont-ils convergents ? Sont-ils asymptotiquement normaux ?
3. Comparer les distributions des estimateurs \widetilde{k}_n et \widehat{k}_n de la même façon qu'on avait comparé les estimateurs \widehat{k}_n et $\widehat{k}_n^{(2)}$ en B.8.5 (superposition des densités approchées, calcul des moyennes et variances de ces estimateurs, QQ-plots et boxplots...) Conclusion ?

Exercice 2 (EMV d'une borne)

Soient X_1, \dots, X_n des variables aléatoires *iid* de loi uniforme sur l'intervalle $[a, b]$ où $a < b$ sont des paramètres réels inconnus que l'on cherche à estimer.

1. Ecrire la vraisemblance des observations $\ell(X_1, \dots, X_n; a, b)$.
2. Dessiner la log-vraisemblance renormalisée $(a, b) \mapsto \ln(\ell(X_1, \dots, X_n; a, b))/n$ pour un échantillon de $n = 100$ variables aléatoires que vous aurez simulé avec $a = 0$ et $b = 1$.
3. Déterminer les estimateurs du maximum de vraisemblance $(\widehat{a}_n, \widehat{b}_n)$. (Graphiquement puis théoriquement).
4. Ecrire une fonction qui trace la courbe de la fonction aléatoire $n \mapsto \sqrt{n}(\widehat{a}_n - a)$ pour $a = 0$ et $n \in \llbracket 10, 10000 \rrbracket$.
5. Faire de même avec les fonctions $n \mapsto n^2(\widehat{a}_n - a)$ et $n \mapsto n(\widehat{a}_n - a)$. Qu'en déduire ?

6. (théorique) Déterminer la loi limite de $n(\hat{a} - a, b - \hat{b})$. Pour cela, on pourra calculer

$$\mathbb{P}\left(\{n(\hat{a}_n - a) > t\} \cap \{n(b - \hat{b}_n) > u\}\right).$$

Que peut-on dire de la dépendance asymptotique entre \hat{a} et \hat{b} ?

7. Pour vérifier le résultat précédent :

- Générer $N = 1000$ échantillons de $n = 100$ observations *iid* de loi $\mathcal{U}[a, b]$, avec $a = 0$ et $b = 1$.
- Calculer l'EMV (\hat{a}_n, \hat{b}_n) pour chacun d'eux. On dispose de $N = 1000$ réalisations de l'EMV.
- Faire un QQ-plot pour comparer la distribution des $n\hat{a}_n$ avec la loi exponentielle de paramètre $1/(b - a)$. De même pour $n(1 - \hat{b})$.
- Calculer la corrélation des \hat{a}_n avec les \hat{b}_n .

8. En déduire un intervalle asymptotique de confiance de niveau 95% pour a .

Exercice 3 (EMV pour une double-exponentielle)

Soient X_1, \dots, X_n des variables aléatoires *iid* de loi double-exponentielle de paramètre de translation $\theta \in \mathbb{R}$, dont la densité est donnée par :

$$f(x, \theta) = 0.5 \exp(-|x - \theta|). \quad (5)$$

1. Ecrire une fonction qui :

- prend deux arguments $(n, a) \in \mathbb{N}^* \times \mathbb{R}$,
- génère n variables aléatoires *iid* Y_i de loi $\mathcal{E}(1)$,
- génère n variables aléatoires *iid* Z'_i de loi $\mathcal{B}(1, 0.5)$ et leur associe $Z_i = 2Z'_i - 1$,
- calcule $X_1 = a + Y_1 Z_1, \dots, X_n = a + Y_n Z_n$ qu'elle retournera,
- trace l'histogramme (en utilisant les options `freq=F`, `breaks=50`, `ylim=c(0, 0.5)`) de l'échantillon X_1, \dots, X_n ,
- superpose (sans effacer l'histogramme) a courbe de la fonction $x \mapsto f(x, a)$. Pour cela, on calculera les valeurs de cette fonction pour une grille allant de $-5a$ à $5a$ avec un pas 0.01.

2. Appeler cette fonction avec différentes valeurs de n et a . Quelles conclusions en tirez-vous? Montrer par le calcul que X_1 suit bien la loi double-exponentielle de paramètre a .

3. (théorique) Trouver l'EMV $\hat{\theta}_n$ de θ .

4. Comment peut-on montrer la convergence de $\hat{\theta}_n$ vers θ numériquement? Mettre numériquement en oeuvre la réponse pour $a = 1.5$.

5. Ecrire une fonction qui :

- prend trois arguments $(n, N, a) \in \mathbb{N}^* \times \mathbb{N}^* \times \mathbb{R}$,
- génère N échantillons de taille n de variables *iid* de loi double-exponentielle de paramètre a ,
- calcule pour chacun de ces échantillons l'EMV $\hat{\theta}_n$. On obtient un échantillon $(\hat{\theta}_n^{(i)})_{i \in [1, N]}$ de variables aléatoires indépendantes et de la loi de $\hat{\theta}_n$.
- calcule la variance s_T^2 de l'échantillon $(\hat{\theta}_n^{(i)})_{i \in [1, N]}$,
- trace l'histogramme de l'échantillon $(\sqrt{n}(\hat{\theta}_n^{(i)} - a))_{i \in [1, N]}$,
- superpose la courbe de la densité de la loi normale centrée de variance ns_T^2 ,
- affiche la valeur de ns_T^2 .

6. Appeler la fonction précédente en donnant des valeurs différentes à n , pour $N = 1000$ et $a = 1.5$. Que constatez-vous? Ces résultats suggèrent-ils la convergence :

- en probabilité de la suite ns_T^2 ?
 - en loi de la suite $(\widehat{\theta}_n - a)/s_T$?
- 7.** Appeler la fonction de la question **5** pour $n = N = 1000$ et en donnant différentes valeurs à a . Que constatez-vous ? La limite de ns_T^2 semble-t-elle dépendre de a ?
- 8.** En admettant la convergence en loi de $\sqrt{n}(\widehat{\theta}_n - \theta)$ vers $\mathcal{N}(0, 1)$ lorsque $n \rightarrow +\infty$, déterminer un intervalle de confiance asymptotique de niveau 95% pour θ .
- 9.** Ecrire une fonction qui :
- prend trois arguments (n, N, a) ,
 - génère comme précédemment un échantillon $(\widehat{\theta}_n^{(i)})_{i \in [1, N]}$ de variables aléatoires indépendantes et de la loi de $\widehat{\theta}_n$,
 - calcule pour chaque $i \in [1, N]$ les extrémités de l'intervalle de confiance associé $\widehat{\theta}_n^{(i)} - 1.96/\sqrt{n}, \widehat{\theta}_n^{(i)} + 1.96/\sqrt{n}$,
 - compte le nombre S de i tels que $a \in [\widehat{\theta}_n^{(i)} - 1.96/\sqrt{n}, \widehat{\theta}_n^{(i)} + 1.96/\sqrt{n}]$,
 - affiche S/N .
- 10.** Appeler plusieurs fois la fonction précédente en faisant varier n pour $N = 1000$ et $a = 1.5$. Que constatez-vous ?