PROJET STA201 ETUDE DE DONNEES HYDROLOGIQUES

KOEN ARISTOTE
MEDOUS CHARLES

Questions préliminaires

La p-valeur du test de Shapiro étant inferieure à 0.05 on peut donc rejeter l'hypothèse nulle, c'est à dire que les données sont normalement distribuées. Il est donc pertinent de modéliser l'échantillon avec une autre loi.

```
> shapiro.test(data$Debit)

        Shapiro-Wilk normality test

data:  data$Debit
W = 0.69259, p-value = 1.233e-09
```

Figure 1: Test de Shapiro sur abermule.csv

1 Propriétés de la loi de Cauchy

1 Etude de la loi de Cauchy

1.1 question (a)

Soit X une variable aléatoire suivant une loi de Cauchy $\mathscr{C}(0,1)$. Soit $(a,b) \in \mathbb{R} \times \mathbb{R}_+^*$. $\forall x \in \mathbb{R}$, on note:

$$\begin{aligned} F_{\mathbf{X}}(x) &= \mathbb{P}(a + b\mathbf{X} \leq x) \\ &= \mathbb{P}\left(\mathbf{X} \leq \frac{x-a}{b}\right) \\ &= \int_{-\infty}^{\frac{x-a}{b}} f_{0,1}(t)\,dt \\ &= \int_{-\infty}^{\frac{x-a}{b}} \frac{1}{\pi(t^2+1)}\,dt \end{aligned}$$

avec le changement de variable $z \leftarrow bt + a$, on a:

$$\begin{aligned} &= \int_{-\infty}^{x} \frac{1}{b\pi(\frac{z-a}{b}^2+1)}\,dz \\ &= \int_{-\infty}^{x} f_{a,b}(t)\,dt \\ &= \mathbb{P}(\mathbf{Y} \leq x) \end{aligned}$$

avec $\mathbf{Y} \hookrightarrow \mathscr{C}(a,b)$

Soit $k \in \mathbb{N}^*$, on a $g(x) := x^k f_{0,1}(x) = \frac{x^k}{\pi(x^2+1)} \underset{x\to+\infty}{\sim} \frac{x^{k-2}}{\pi}$.
Or $\frac{1}{x}$ n'est pas intégrable sur $[0,+\infty[$, donc l'intégrale $\int g$ diverge. Donc $\mathbf{X}$ n'admet aucun moment fini d'ordre ≥ 1.

1.2 question (b)

$$\begin{aligned} TF^{-1}[: x \mapsto e^{-|x|}] &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{-|x|}e^{-itx}dx \\ &= \frac{1}{2\pi}\left(\int_{-\infty}^{+\infty} e^{-|x|}cos(tx)dx + i\int_{-\infty}^{+\infty} e^{-|x|}sin(tx)dx\right) \\ &= \frac{1}{\pi}\int_{0}^{+\infty} e^{-x}cos(tx)dx \end{aligned}$$

Car l'intégrande de droite est impaire et celle de gauche est paire

On pose

$$I = \int_{0}^{+\infty} e^{-x}cos(tx)dx$$

par IPP avec $u = e^{-x}$ et $v' = cos(tx)$, on a

$$= 1 - t\int_{0}^{+\infty} e^{-x}sin(tx)dx$$

par IPP avec $u = e^{-x}$ et $v' = sin(tx)$, on a

$$\begin{aligned} &= 1 - t(0 + t\int_{0}^{+\infty} e^{-x}cos(tx)dx) \\ &= 1 - t^2 I \end{aligned}$$

Donc

$$I = \frac{1}{1+t^2}$$

Donc

$$TF^{-1}[: x \mapsto e^{-|x|}] = \frac{1}{\pi(1+t^2)} = f_{0,1}(t)$$

La fonction caractéristique de $\mathbf{X}$ est par définition $\Phi_{\mathbf{X}}(t) := \mathbb{E}(e^{it\mathbf{X}}) = TF[f_{0,1}](t)$ Donc $e^{-|x|} = TF[TF^{-1}[: x \mapsto e^{-|x|}]](x) = TF[f_{0,1}](x) = \Phi_{\mathbf{X}}(x)$

1.3 question (c)

Soit $\mathbf{X_1}, ..., \mathbf{X_n} \overset{i.i.d}{\hookrightarrow} \mathscr{C}(a,b)$
La fonction caractéristique de la variable aléatoire $\mathbf{S}_n := \sum_{i=1}^{n} \frac{\mathbf{X}_i}{n}$ est:

$$\begin{aligned} \Phi_{\mathbf{S}_n}(t) &= \mathbb{E}(e^{it\sum_{i=1}^{n}\frac{\mathbf{X}_i}{n}}) \\ &= \prod_{i=1}^{n}\mathbb{E}(e^{it\frac{\mathbf{X}_i}{n}}) \end{aligned}$$

D'après la question (a), et avec $\forall i, \mathbf{Y}_i \hookrightarrow \mathscr{C}(0,1)$ on a

$$\begin{aligned}\Phi_{\mathbf{S}_n}(t) &= \prod_{i=1}^{n} \mathbb{E}(e^{it\frac{a+b\mathbf{Y}_i}{n}}) \\ &= e^{ita}\prod_{i=1}^{n} \mathbb{E}(e^{i\frac{tb}{n}\mathbf{Y}_i}) \\ &= e^{ita}\prod_{i=1}^{n} \int_{-\infty}^{+\infty} e^{i\frac{tb}{n}x} f_{0,1}(x)dx\end{aligned}$$

avec le changement de variable $z \leftarrow \frac{b}{n}x$, on a

$$\begin{aligned}\Phi_{\mathbf{S}_n}(t) &= e^{ita}\prod_{i=1}^{n} \int_{-\infty}^{+\infty} e^{itz}\frac{n}{b} f_{0,1}(\frac{nz}{b})dx \\ &= e^{ita}\prod_{i=1}^{n} \int_{-\infty}^{+\infty} e^{itz}\frac{1}{\frac{b}{n}\pi(1+(\frac{nz}{b})^2}dx \\ &= e^{ita}\prod_{i=1}^{n} \int_{-\infty}^{+\infty} e^{itz} f_{0,\frac{b}{n}}(x)dx \\ &= e^{ita}\prod_{i=1}^{n} TF[f_{0,\frac{b}{n}}](t)\end{aligned}$$

On montre comme précédemment que l'on a $\mathrm{TF}^{-1}[: x \mapsto e^{-|\frac{b}{n}x|}] = f_{0,n}, donc$

$$\begin{aligned}\Phi_{\mathbf{S}_n}(t) &= e^{ita}\prod_{i=1}^{n} TF[TF^{-1}[: x \mapsto e^{-|\frac{b}{n}x|}]](t) \\ &= e^{ita}\prod_{i=1}^{n} e^{-|\frac{b}{n}t|} \\ &= e^{iat-b|t|}\end{aligned}$$

Or cette expression est bien celle de la fonction caractéristique d'une variable aléatoire suivant une loi $\mathscr{C}(a,b)$, en effet:

$$\begin{aligned}TF^{-1}[: x^{iat-b|t|}] &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{i(a-x)t-b|t|}dt \\ &= \frac{1}{2\pi}\left(\left[\frac{e^{i(a-x)t+bt}}{i(a-x)t+b}\right]_{-\infty}^{0} + \left[\frac{e^{i(a-x)t-bt}}{i(a-x)t-b}\right]_{0}^{+\infty}\right) \\ &= \frac{1}{2\pi}\left(\frac{1}{i(a-x)+b} - \frac{1}{i(a-x)-b}\right) \\ &= \frac{1}{2\pi}\left(\frac{2b}{(a-x)^2+b^2}\right) \\ &= \frac{1}{b\pi}\frac{1}{1+(\frac{x-a}{b})^2} \\ &= f_{a,b}(x)\end{aligned}$$

On a donc, $\forall n \in \mathbb{N}^*$

$$\mathbf{X_1}, ..., \mathbf{X_n} \overset{i.i.d}{\hookrightarrow} \mathscr{C}(a,b) \Rightarrow \mathbf{S}_n := \sum_{i=1}^{n}\frac{\mathbf{X}_i}{n} \hookrightarrow \mathscr{C}(a,b)$$

1.4 question (d)

Soit $(a,b) \in \mathbb{R} \times \mathbb{R}_+^*$
Soit **X** une variable aléatoire suivant une loi de Cauchy $\mathscr{C}(a,b)$
On a

$$F_{\mathbf{X}}(x) := \int_{-\infty}^{x} \frac{1}{b\pi(1+(\frac{x-a}{b})^2)} dt$$

avec le changement de variable $z \leftarrow \frac{t-a}{b}$, on a

$$\begin{aligned} F_{\mathbf{X}}(x) &= \frac{1}{\pi} \int_{-\infty}^{\frac{x-a}{b}} \frac{1}{1+z^2} dz \\ &= \frac{1}{\pi} \left[Arctan(\frac{x-a}{b}) + \frac{\pi}{2} \right] \end{aligned}$$

La médiane x^* est définie comme suit

$$\begin{aligned} &\mathbb{P}(\mathbf{X} \leq x^*) = \frac{1}{2} \\ &\Leftrightarrow \frac{1}{\pi} Arctan(\frac{x^*-a}{b}) + \frac{1}{2} = \frac{1}{2} \\ &\Leftrightarrow Arctan(\frac{x^*-a}{b}) = 0 \\ &\Leftrightarrow x^* = a \end{aligned}$$

2 Méthode des moments

2.1 question (a)

Si $\overline{X}$ est consistant pour un paramètre $\theta \in \mathbb{R}$ alors $\forall \epsilon > 0$:
$\lim_{n\to\infty} \mathbb{P}(|\overline{X} - \theta| > \varepsilon) = 0$

Or, $\mathbb{P}(|\overline{X} - \theta| > \varepsilon) = 1 - \mathbb{P}(|\overline{X} - \theta| \leq \varepsilon) = 1 - \mathbb{P}(\overline{X} \leq \theta + \varepsilon) + \mathbb{P}(\overline{X} \leq \theta - \varepsilon)$

On sait que $\overline{X}$ suit une $\mathscr{C}(a,b)$ ainsi:

$= 1 - F_{\mathscr{C}(a,b)}(\varepsilon + \theta) + F_{\mathscr{C}(a,b)}(\theta - \varepsilon) \not\to 0$ (ne dépend pas de n)
La moyenne empirique n'est donc pas un estimateur consistant.

De plus, on ne pourra pas construire un estimateur de a avec la méthode des moments car les moments d'ordre 1 ou plus ne sont pas définis.

2.2 question (b)

Le code suivant génère un échantillon iid de $(X_n)_{1\leq n\leq 500}$ variables aléatoires suivant une $\mathscr{C}(0,1)$.

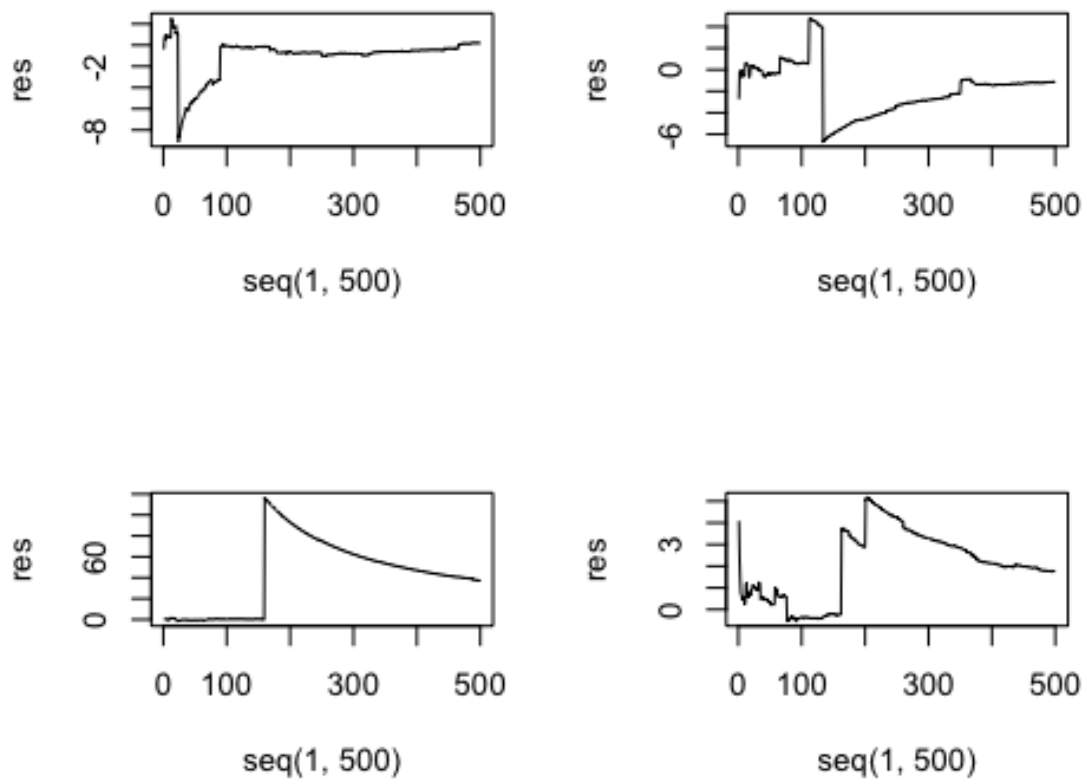

Figure 1.1: La moyenne empirique ne converge pas

Ensuite dans le but de visualiser le fait que la loi des grands nombres ne s'applique pas dans le cas d'une loi de Cauchy, on crée une liste V_i de taille $n = 500$ telle que: $V_i = \frac{1}{i}\sum_{n=1}^{i} X_n$. La dernière commande trace V_i en fonction de i. Chaque composante i de cette liste correspond donc à la moyenne empirique de l'échantillon de taille i et on observe que en augmentant le nombre de réalisations on n'observe aucune convergence. En effet, en réalisant l'expérience quatre fois, on observe sur la figure 1.1 que ces valeurs ont un comportement aléatoire et qu'il n'y a pas de convergence de la moyenne empirique pour de grandes valeurs de l'échantillon ce qui confirme notre analyse précédente.

2 Maximum de vraisemblance pour estimer a et b

On suppose que $\mathbf{X_1}, ..., \mathbf{X_n} \overset{i.i.d}{\hookrightarrow} \mathscr{C}(a, b)$ avec $(a, b) \in \mathbb{R} \times \mathbb{R}_+^*$

1 On suppose ici $b = 1$ et on propose d'estimer a

1.1 question(a)

La log-vraisemblance du modèle est défini comme suit

$$\begin{aligned} l(a) &:= L(\mathbf{X}_1, ..., \mathbf{X}_n; a) \\ &= \sum_{i=1}^{n} log(f_{a,b}(\mathbf{X}_i)) \end{aligned}$$

Comme b=1, on a

$$\begin{aligned} l(a) &= \sum_{i=1}^{n} log\left(\frac{1}{\pi(1 + (\mathbf{X}_i - a)^2))}\right) \\ &= -log(\pi) - \sum_{i=1}^{n} log\left(1 + (\mathbf{X}_i - a)^2\right) \end{aligned}$$

On observe que $l(a = \pm\infty) = +\infty$, mais

$$\frac{\partial l}{\partial a}(a_m) = \sum_{i=1}^{n} \frac{2(\mathbf{X}_i - a_m)}{1 + (\mathbf{X}_i - a_m)^2} = 0$$

Or les termes sommés sont de signe quelconque, on ne pourra donc pas retrouver d'expression explicite de $\widehat{a}$

1.2 question (b)

On admet que $\widehat{a}$ suit le régime asymptotique classique:

$$\sqrt{n}(\widehat{a}-a) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0, \frac{1}{I(a)})$$

Pour obtenir la variance asymptotique de $\widehat{a}$, on va donc devoir calculer l'inverse de l'information de Fisher $I(a)$.
Le modèle étant régulier, on a la relation suivante:

$$I(a) = -\mathbb{E}_a\left[\partial_a V_1(a)\right]$$

Où le vecteur du score de la loi est:

$$\begin{aligned} V_1(a) &:= \partial_a log\left(p(\mathbf{X}_1;a)\right) \\ &= \partial_a\left[-log(\pi) - log\left(1+(\mathbf{X}_1-a)^2\right)\right] \\ &= \frac{2(\mathbf{X}_1-a)}{1+(\mathbf{X}_1-a)^2} \end{aligned}$$

On calcule la dérivé du vecteur du score par rapport au paramètre a:

$$\begin{aligned} \partial_a V_1(a) &= 2\frac{-1-(\mathbf{X}_1-a)^2+2(\mathbf{X}_1-a)(\mathbf{X}_1-a)}{(1+(\mathbf{X}_1-a)^2)^2} \\ &= 2\frac{-1+(\mathbf{X}_1-a)^2}{(1+(\mathbf{X}_1-a)^2)^2} \end{aligned}$$

L'espérance de la variable aléatoire $\partial_a V_1(a)$ est donc:

$$\begin{aligned} \mathbb{E}[\partial_a V_1(a)] &= \frac{2}{\pi}\int_{-\infty}^{+\infty} \frac{-1+(x-a)^2}{(1+(x-a)^2)^2} f_{a,1}(x)\,dx \\ &= \frac{2}{\pi}\int_{-\infty}^{+\infty} \frac{-1+(x-a)^2}{(1+(x-a)^2)^3}\,dx \end{aligned}$$

Avec le changement de variable $x \mapsto x-a$, on a

$$\mathbb{E}[\partial_a V_1(a)] = \frac{2}{\pi}\int_{-\infty}^{+\infty} \frac{-1+x^2}{(1+x^2)^3}\,dx$$

Par parité, on a

$$\mathbb{E}[\partial_a V_1(a)] = \frac{4}{\pi}\int_{0}^{+\infty} \frac{-1+x^2}{(1+x^2)^3}\,dx$$

En ajoutant 1 au numérateur, on a

$$\begin{aligned} \mathbb{E}[\partial_a V_1(a)] &= \frac{4}{\pi}\left(\int_0^{+\infty} \frac{1+x^2}{(1+x^2)^3}\,dx - \int_0^{+\infty} \frac{2}{(1+x^2)^3}\,dx\right) \\ &= \frac{4}{\pi}\left(\int_0^{+\infty} \frac{1}{(1+x^2)^2}\,dx - \int_0^{+\infty} \frac{2}{(1+x^2)^3}\,dx\right) \end{aligned}$$

On note I l'intégrale de droite et on fait le changement de variable $u \mapsto Arctan(x)$

$$\begin{aligned} I &= \int_0^{\frac{\pi}{2}} \frac{1}{1+tan(u)^2} du \\ &= \int_0^{\frac{\pi}{2}} cos(u)^2 du \\ &= \frac{\pi}{4} \end{aligned}$$

On note J l'intégrale de droite et on utilise la formule de réduction avec $a = 1, b = 1, n = 3$:

$$\begin{aligned} J &= 2\left(\left[\frac{x}{2(x^2+1)}\right]_0^{+\infty} + \frac{3}{4}\int_0^{+\infty} \frac{1}{(1+x^2)^2} dx\right) \\ &= 2\left(0 + \frac{3}{4}I\right) \\ &= \frac{6\pi}{16} \end{aligned}$$

Finalement

$$\begin{aligned} \mathbb{E}[\partial_a V_1(a)] &= \frac{4}{\pi}(\frac{\pi}{4} - \frac{6\pi}{16}) \\ &= -\frac{1}{2} \end{aligned}$$

Et donc

$$Var_a(\widehat{a}) = \frac{1}{I(a)} = -\frac{1}{\mathbb{E}_a[\partial_a V_1(a)]} = 2$$

On a donc:

$$\sqrt{n}(\widehat{a} - a) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,2)$$

Ou encore:

$$\sqrt{\frac{n}{2}}(\widehat{a} - a) \xrightarrow[n\to+\infty]{loi} \mathcal{N}(0,1)$$

Et on peut calculer un intervalle de confiance de niveau $\alpha \in [0,1]$ en utilisant les quantiles de la loi normale:

$$\mathbb{P}(q_{\frac{\alpha}{2}} \leq \sqrt{\frac{n}{2}}(\widehat{a} - a) \leq q_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

$$\mathbb{P}(\frac{q_{\frac{\alpha}{2}}}{\sqrt{\frac{n}{2}}} \leq \widehat{a} - a \leq \frac{q_{1-\frac{\alpha}{2}}}{\sqrt{\frac{n}{2}}}) = 1 - \alpha$$

$$\mathbb{P}(\widehat{a} - \sqrt{\frac{2}{n}} q_{1-\frac{\alpha}{2}} \leq a \leq \widehat{a} + \sqrt{\frac{2}{n}} q_{\frac{\alpha}{2}}) = 1 - \alpha$$

Donc l'intervalle de confiance IC est

$$IC = \left[\widehat{a} - \sqrt{\frac{2}{n}} q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}, \widehat{a} + \sqrt{\frac{2}{n}} q_{\frac{\alpha}{2}}^{\mathcal{N}(0,1)}\right]$$

1.3 question (C)

On admet le résultat suivant:
Si $\mathbf{X}_1,...,\mathbf{X}_n$ sont i.i.d de fonction de répartition F avec F dérivable de dérivée f jamais nulle, alors:

$$\sqrt{n}\left(\widehat{M}-F^{-1}(1/2)\right)\xrightarrow[n\to+\infty]{loi}\mathcal{N}\left(0,\frac{1}{4f[F^{-1}(1/2)]^2}\right)$$

Ici on a que les $\mathbf{X}_1,...,\mathbf{X}_n$ sont i.i.d de fonction de répartition $F(x)=\frac{1}{\pi}Arctan(x_a)+\frac{1}{2}$ et F est dérivable de dérivée $f_{a,1}(x)=\frac{1}{\pi}\frac{1}{1+(x-a)^2}$ jamais nulle. On va donc appliquer le résultat admis pour avoir un estimateur de la médiane $\widehat{M}$, qui, on l'a vu en partie 1, est un estimateur de a quelque soit b.

$$\begin{aligned}
F(x)&=\frac{1}{\pi}Arctan(x-a)+\frac{1}{2}\\
&\Rightarrow\pi[F(x)-\frac{1}{2}]=Arctan(x-a)\\
&\Rightarrow tan(\pi[F(x)-\frac{1}{2}])=x-a\\
&\Rightarrow tan(\pi[F(x)-\frac{1}{2}])+a=x\\
&\Rightarrow F^{-1}(x)=tan(\pi[x-\frac{1}{2}])+a
\end{aligned}$$

En particulier:

$$\begin{aligned}
F^{-1}(\frac{1}{2})&=tan(\pi[\frac{1}{2}-\frac{1}{2}])+a\\
&=a
\end{aligned}$$

Et donc:

$$\begin{aligned}
f_{a,1}(F^{-1}(\frac{1}{2}))&=f_{a,1}(a)\\
&=\frac{1}{\pi}\frac{1}{1+(a-a)^2}\\
&=\frac{1}{\pi}
\end{aligned}$$

Finalement:

$$\frac{1}{4f_{a,1}[F^{-1}(\frac{1}{2})]^2}=\frac{\pi^2}{4}$$

On a donc le régime asymptotique suivant pour l'estimateur $\widehat{M}$:

$$\sqrt{n}\left(\widehat{M}-a\right)\xrightarrow[n\to+\infty]{loi}\mathcal{N}(0,\frac{\pi^2}{4})$$

L'intervalle de confiance de niveau α pour a en utilisant la médiane est:

$$IC=\left[\widehat{M}-\frac{\pi}{2\sqrt{n}}q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}},\widehat{M}+\frac{\pi}{2\sqrt{n}}q^{\mathcal{N}(0,1)}_{\frac{\alpha}{2}}\right]$$

On rappelle que l'intervalle de confiance de même niveau α pour a en utilisant $\widehat{a}$ est:

$$IC = \left[\widehat{a} - \sqrt{\frac{2}{n}} q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}, \widehat{a} + \sqrt{\frac{2}{n}} q^{\mathcal{N}(0,1)}_{\frac{\alpha}{2}}\right]$$

On voit donc que l'intervalle de confiance est plus large en utilisant l'estimateur de la médiane, car $\frac{\pi}{2} \simeq 1.57 > 1.41 \simeq \sqrt{2}$. Donc l'estimateur $\widehat{a}$ est asymptotiquement préférable à l'estimateur de la médiane $\widehat{M}$.

1.4 question (d)

La fonction est infinie à l'infini ($\lim_{a \to \pm\infty} l(a) = -\infty$)et continue sur $\mathbb{R}$ ainsi elle admet au moins un maximum global(donc l'opposé de la log-vraisemblance admet un minimum global). De plus, la Hessienne étant égale à $-\frac{n}{2b^2} I_d$ elle est définie négative et donc la log vraisemblance admet un unique maximum global. Cependant on peut observer sur la figure 2.1 qu'elle admet un maximum local autour de -4. Ainsi, notre maximum n'étant pas explicite, il est important de choisir un point d'initialisation qui n'entraînera pas de convergence de l'algorithme vers ce minimum local en considérant l'opposé de la log-vraisemblance. On peut donc voir en se plaçant sur une large fenêtre [-500,500] qu'en initialisant l'algorithme d'optimisation sur une grande valeur positive, notre algorithme d'optimisation pour l'opposé de la log-vraisemblance convergera vers le minimum global de cette fonction ce qui nous donnera le maximum de la log vraisemblance.

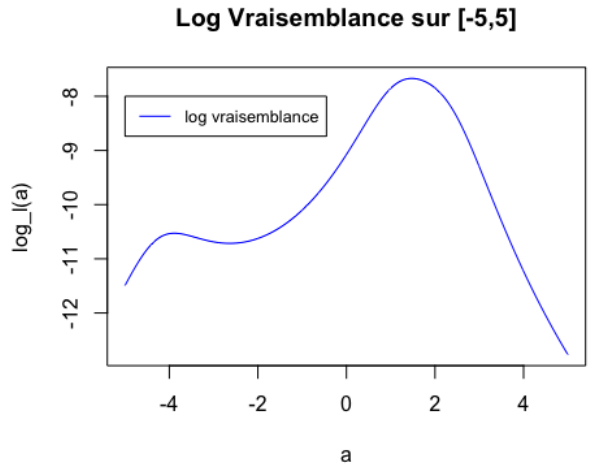

Figure 2.1: Fenêtre sur [-5,5]

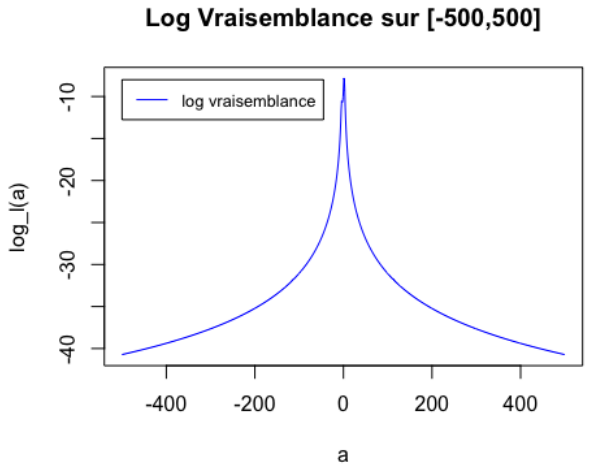

Figure 2.2: Fenêtre sur [-500,500]

1.5 Question 2

Le code sur la figure 2.3 fait appel à la fonction nlm (intégrée à R) qui applique un algorithme d'optimisation itératif de type Newton afin de calculer le minimum d'une fonction.

```
f=function(theta,x){
  a=theta[1]; b=theta[2]
  res = sum((x-a)^2) + sum((x^2-b)^2)
  attr(res,"gradient")=c(-2*sum(x-a), -2*sum(x^2-b))
  res
}
x=rnorm(10)
nlm(f,p=c(0,0),x=x,hessian=TRUE)
mean(x);mean(x^2)
```

Figure 2.3: Algorithme de type Newton

Dans ce cas, nlm trouve le minimum de:

$$f(a,b) = \sum_{i=1}^{10} (X_i - a)^2 + \sum_{i=1}^{10} (X_i^2 - b)^2$$

En utilisant la formule du gradient de f (qui a été attribuée à f dans sa définition: "(attr(res,"gradient"))") et pour une initialisation au point (0,0). Ainsi nlm retournera la valeur du minimum et le point en lequel le min est atteint, la valeur du gradient en ce minimum ainsi que celle de la Hessienne puis le nombre d'itérations effectuées par l'algorithme.

1.6 Question b

Afin d'observer que notre méthode d'optimisation fonctionne et calcule bien l'EMV $(\hat{a},\hat{b})$, nous générons 50 échantillons de taille 100 pour la loi $\mathscr{C}(0,1)$ puis nous calculons l'EMV à travers la fonction nlm pour chaque échantillon afin de pouvoir tracer l'histogramme de ces 50 EMVs.
On observe bien sur les figures 2.4 et 2.5 que nos histogrammes sont concentrés autour de 0 pour l'EMV de a et autour de 1 pour l'EMV de b.

1.7 Question (c)

On calcule maintenant l'estimateur du maximum de vraisemblance $(\hat{a},\hat{b})$ sur le jeu de données 'abermule.csv' et on observe en superposant la densité d'une $\mathscr{C}(\hat{a},\hat{b})$ et une Gaussienne de paramètre la moyenne empirique et la variance empirique sur ce jeu de donnée que la densité d'une loi de Cauchy s'aligne bien mieux avec les données que la Gaussienne.
Le modèle Gaussien est bien plus applatie et donne ainsi bien moins de poids aux valeurs autour de la médiane de notre échantillon, ainsi la loi de Cauchy permet donc de prendre en compte les grandes valeurs de débit.

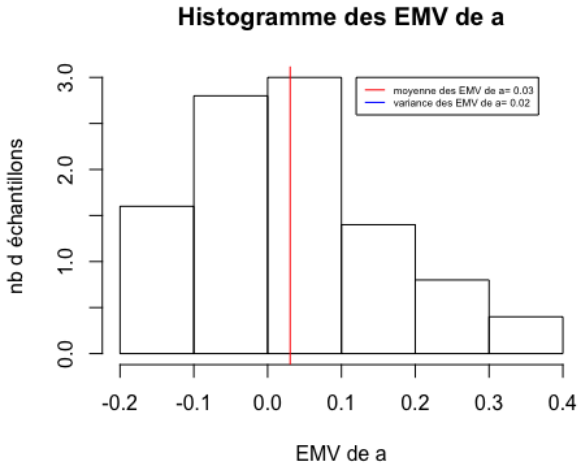

Figure 2.4: Histogramme des EMV de a

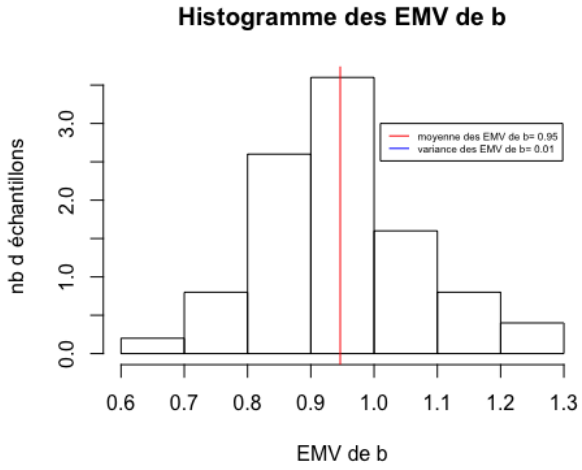

Figure 2.5: Histogramme des EMV de b

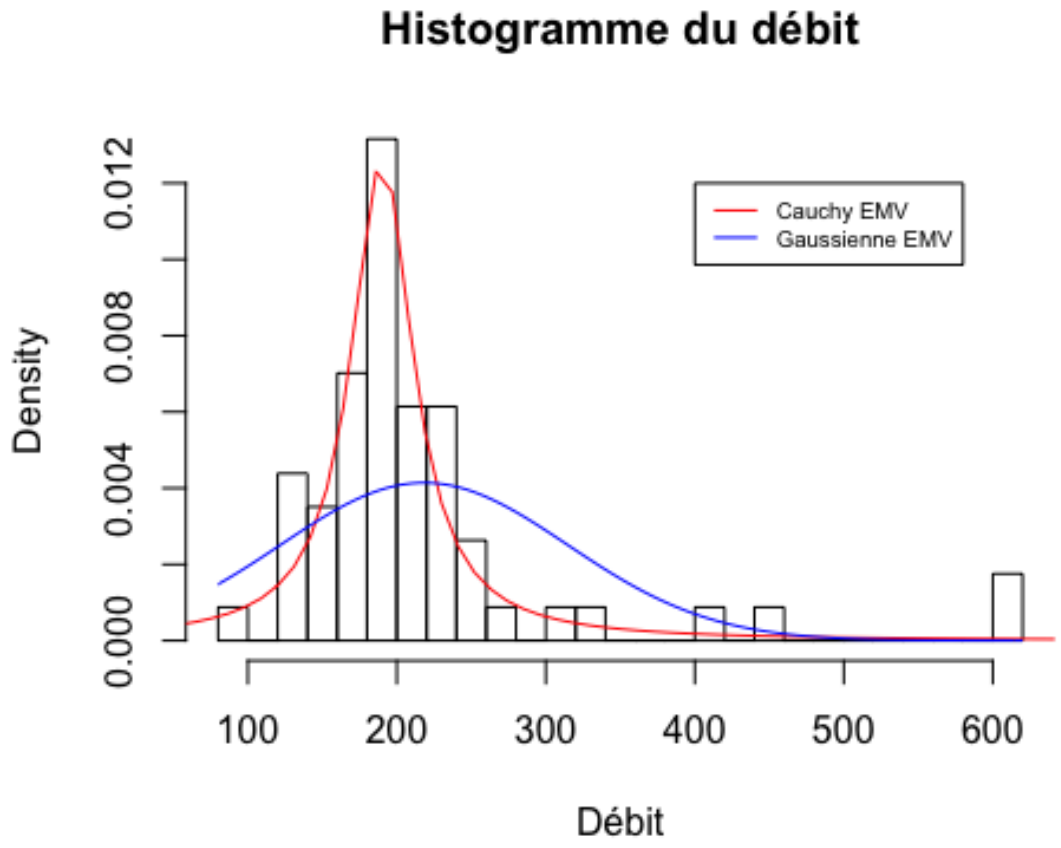

Figure 2.6: Histogramme des EMV de a pour 50 echantillons de Cauchy(0,1

3

Modèlisation par la loi de Cauchy pour $b = 1$ ou pour $b \neq 1$

0.1 Estimation de la variance de l'EMV

L'information de Fisher d'une loi de Cauchy de paramètre $\theta = (a, b)$ est:

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 l(X,\theta)}{\partial \theta^2} \right]$$

avec $\frac{\partial^2 l(X,\theta)}{\partial \theta^2}$, la Hessienne de la log vraisemblance.
Ainsi:

$$\begin{aligned} I(\hat{a}, \hat{b}) &= -\mathbb{E}_\theta \left[\frac{\partial^2 l(X, (\hat{a}, \hat{b}))}{\partial \theta^2} \right] \\ &= \begin{pmatrix} 0.0438 & -0.0092 \\ -0.0092 & 0.0460 \end{pmatrix} \end{aligned}$$

Notre estimateur suivant le régime asymptotique classique, on peut considérer l'inverse de l'information Fisher comme une estimation de la variance
Ainsi en inversant on obtient:

$$\begin{pmatrix} 22.8336 & 4.7558 \\ 4.7558 & 22.6613 \end{pmatrix}$$

0.2 Construction du test de Wald

(H_0) : g(θ) = 0 contre (H_1) : $g(\theta) \neq 0$
où g(θ)=g(a,b)=b-1.

La fonction g est différentiable et donc $\nabla g = (0, 1)$

par la delta méthode:

$$g(\hat{a}) - g(a) \xrightarrow{L} N(0, \nabla g(a) I_n(a,b)^{-1} \nabla g(a,b)^T) \tag{3.1}$$

Ainsi, par le lemme de Slutzky:

$$(\nabla g(a) I_n(a,b)^{-1} \nabla g(a,b)^T)^{-\frac{1}{2}} (g(\hat{a}) - g(a)) \xrightarrow{L} N(0,1) \tag{3.2}$$

On estime $I_n(a,b) par I_n(\hat{a},\hat{b})$ et on a donc:
$(\nabla g(a) I_n(\hat{a},\hat{b})^{-1} \nabla g(a,b)^T)^{-\frac{1}{2}} (g(\hat{a}) - g(a)) = -\left(\frac{\partial^2 logl(\hat{a},\hat{b})}{\partial b^2}\right)^{\frac{1}{2}} \left(g(\hat{a}) - g(a)\right) \xrightarrow{L} N(0,1)$

Pour construire le test nous considérons donc la statistique sous H0:

$$\zeta_n = -\left(\frac{\partial^2 logl(\hat{a},\hat{b})}{\partial b^2}\right)^{\frac{1}{2}} \left(g(\hat{a}) - g(a)\right) \tag{3.3}$$

Ainsi, pour obtenir un test de niveau alpha on définit la région de rejet :

$$R = \{|\zeta_n| \geq q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}\}$$

Calcul de la p-value (sur R):

$$\begin{aligned} p_{val} &= \mathbb{P}_{\theta_0}(|\zeta_n| \geq \zeta_{57}^{obs})\} \\ &= \mathbb{P}_{\theta_0}(\zeta_n \geq \zeta_{57}^{obs}) + \mathbb{P}_{\theta_0}(\zeta_n \leq -\zeta_{57}^{obs}) \\ &= 1 - \mathbb{P}_{\theta_0}(\zeta_n \leq \zeta_{57}^{obs}) + 1 - \mathbb{P}_{\theta_0}(\zeta_n \leq \zeta_{57}^{obs}) \\ &= 2 - 2\mathbb{P}_{\theta_0}(\zeta_n \leq \zeta_{57}^{obs}) = 2.106339e-07 \end{aligned}$$

La p-valeur étant quasi nulle (inférieure à alpha=0.05 par exemple), on rejette l'hypothèse nulle (b=1) ce qui justifie la nécessité de choisir un paramètre d'échelle b pour modéliser correctement des données hydrologiques. En effet les données étant très resserrées autour de la médiane, le paramètre b permet de prendre ce fait en considération.

4 BONUS: Statistique exhaustive

On cherche une statistique exhaustive minimale pour le modèle de Cauchy.

1 Rapport de vraisemblance de la loi de Cauchy

Pour tout $x \in \mathbb{R}^n$, On note $p_\theta(x)$ la densité jointe de $x = (x_1, ..., x_n)$.
Faisons la réciproque pour commencer:
Soit $(x, y) \in (\mathbb{R}^n)^2$, supposons qu'il existe une permutation σ de $\{1, ..., n\}$ tel que pour tout $1 < i < n$, on ait $x_i = y_{\sigma(i)}$ alors:

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\prod_{i=1}^n \pi b(1 + (\frac{y_i - a}{b})^2)}{\prod_{i=1}^n \pi b(1 + (\frac{x_i - a}{b})^2)}$$

En réindexant selon l'ordre des permutations, on a:

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\prod_{j=1}^n (1 + (\frac{y_{\sigma(j)} - a}{b})^2)}{\prod_{i=1}^n (1 + (\frac{x_i - a}{b})^2)}$$

En appliquant l'hypothèse sur y on a donc:

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\prod_{j=1}^n (1 + (\frac{x_j - a}{b})^2)}{\prod_{i=1}^n (1 + (\frac{x_i - a}{b})^2)}$$
$$\frac{p_\theta(x)}{p_\theta(y)} = 1 = \alpha(a, b)$$

On suppose maintenant que le rapport des vraisemblances est constant pour tout couple $(x, y) \in (\mathbb{R}^n)^2$.
On va raisonner par récurrence sur la dimension de l'espace, n.
si $n = 1$, on a:

$$\forall (x, y) \in \mathbb{R}^2, \frac{p_\theta(x)}{p_\theta(y)} = \frac{1 + (\frac{y - a}{b})^2}{1 + (\frac{x - a}{b})^2} = \alpha(a, b)$$

Donc en posant $X, Y \in \mathbb{R}^+$, tels que $Y = (\frac{y-a}{b})^2$ et $X = (\frac{x-a}{b})^2$, on a:

$$\forall (X,Y) \in (\mathbb{R}^+)^2, \frac{p_\theta(x)}{p_\theta(y)} = \frac{1+Y}{1+X} = \alpha(a,b)$$

Cela implique nécessairement:

$$Y = \alpha(a,b)X + 1 - \alpha(a,b)$$

Montrons maintenant que α vaut 1:

$$\forall (x,y) \in \mathbb{R}^2, (\frac{y-a}{b})^2 = \alpha(a,b)(\frac{x-a}{b})^2 + 1 - \alpha(a,b)$$

Donc:

$$(\frac{y_{(x=a)}-a}{b})^2 = 1 - \alpha(a,b) \geq 0 \Rightarrow \alpha(a,b) \leq 1$$

Et:

$$(\frac{x_{(y=a)}-a}{b})^2 = \alpha(a,b) - 1 \geq 0 \Rightarrow \alpha(a,b) \geq 1$$

Donc on a bien $x = y$, i.e qu'il existe une permutation σ de $\{1\}$ tel que pour tout $1 < i < 1$, on ait $x_i = y_{\sigma(i)}$.
Supposons maintenant le résultat vrai au rang $n \geq 1$.
$\forall (x,y) \in (\mathbb{R}^{n+1})^2$, $x = (X_n, x_{n+1})$ et $y = (Y_n, y_{n+1})$ avec $(X_n, Y_n) \in (\mathbb{R}^n)^2$
On a par hypothèse:

$$\forall (x,y) \in (\mathbb{R}^{n+1})^2, \frac{p_\theta(x)}{p_\theta(y)} = \alpha(a,b)$$

Donc, en prenant $x_{n+1} = y_{n+1} = a$, on a:

$$\forall (x,y) \in (\mathbb{R}^n)^2, \frac{p_\theta(X_n)}{p_\theta(Y_n)} = \alpha(a,b)$$

Donc, par hypothèse de récurrence, il existe une permutation σ de $\{1,...,n\}$ tel que pour tout $1 < i < n$, on ait $x_i = y_{\sigma(i)}$.
On a donc:

$$\forall (x,y) \in (\mathbb{R}^{n+1})^2, \frac{p_\theta(x)}{p_\theta(y)} = \frac{p_\theta(X_n)}{p_\theta(Y_n)} \frac{p_\theta(x_{n+1})}{p_\theta(y_{n+1})} = \frac{p_\theta(x_{n+1})}{p_\theta(y_{n+1})} = \alpha(a,b)$$

i.e:

$$\forall (x_{n+1}, y_{n+1}) \in \mathbb{R}^2, \frac{1 + (\frac{y_{n+1}-a}{b})^2}{1 + (\frac{x_{n+1}-a}{b})^2} = \alpha(a,b)$$

On a finalement que $x_{n+1} = y_{n+1}$ et donc il existe une permutation σ de $\{1,...,n+1\}$ tel que pour tout $1 < i < n+1$, on ait $x_i = y_{\sigma(i)}$.

2 Statistique exhaustive minimale

La statistique d'ordre $S(X) = X^{(1)}, ..., X^{(n)}$ est exhaustive par le critère de factorisation de Fisher-Neyman. De plus, une statistique S est suffisante minimale si:
$\forall (x, y) \in (\mathbb{R}^n)^2$, $\frac{p_\theta(x)}{p_\theta(y)}$ est indépendant du paramètre $\theta \Leftrightarrow S(x) = S(y)$
Or on a montré à la question précédente que le rapport des vraisemblance est égal à une constante (ici 1, indépendant du paramètre), si et seulement si Y est donnée par les permutations de X, i.e ssi $S(X) = S(Y)$. Donc S est bien une statistique exhaustive (suffisante) minimale pour le modèle de Cauchy.