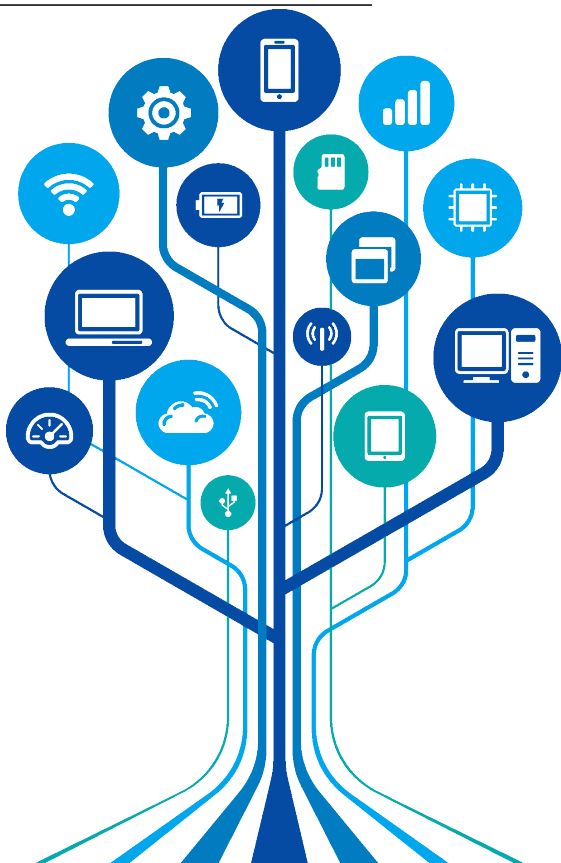
Mastère HPC-AI

Statistiques & Probabilités

Projet Proba & Stat Toolbox for ML

Élèves :
Manon Tourbier
Arthur Viens
Professeur :
Miguel Munoz Zuniga



6 janvier 2022

Table des matières

1 Théorie : Comparaison entre la méthode des moments et du maximum de vraisemblance

Soit la fonction suivante :

$$f(x) = \frac{1}{\alpha} e^{-\frac{x-m}{\alpha}} e^{-e^{-\frac{x-m}{\alpha}}}$$

1.1 Démontrer que f(x) est une loi de densité de probabilité

On a tout d'abord que $\alpha > 0$ et $\forall t \in \mathbb{R}, e^t > 0$. Donc $f(x) >= 0 \ \forall x \in \mathbb{R}$. De plus, la fonction exponentielle étant continue sur $\mathbb{R}$, il en suit facilement d'après les théorèmes généraux que $f(x)$ est continue sur $\mathbb{R}$ par produit de fonctions continues.
Il reste alors à montrer que

$$\int_{-\infty}^{\infty} \frac{1}{\alpha} e^{-\frac{x-m}{\alpha}} e^{-e^{-\frac{x-m}{\alpha}}} \, \mathrm{d}x = 1$$

On pose $u(x) = -e^{-\frac{x-m}{\alpha}}$, donc $u'(x) = \frac{1}{\alpha} e^{-\frac{x-m}{\alpha}}$

$$\int_{-\infty}^{\infty} \frac{1}{\alpha} e^{-\frac{x-m}{\alpha}} e^{-e^{-\frac{x-m}{\alpha}}} \, \mathrm{d}x = \int_{-\infty}^{\infty} u'(x) e^{u(x)}$$

$$= \left[e^{u(x)} \right]_{-\infty}^{+\infty}$$

$$= \left[e^{-e^{-\frac{x-m}{\alpha}}} \right]_{-\infty}^{+\infty}$$

Or $\lim_{x \to +\infty} -e^{-\frac{x-m}{\alpha}} = 0$ et $\lim_{x \to -\infty} -e^{-\frac{x-m}{\alpha}} = +\infty$ Donc

$$= e^0 - \lim_{y \to \infty} e^{-y} = 1 - 0$$

D'où

$$\int_{-\infty}^{\infty} \frac{1}{\alpha} e^{-\frac{x-m}{\alpha}} e^{-e^{-\frac{x-m}{\alpha}}} \, \mathrm{d}x = 1$$

1.2 Calcul des estimateurs des moments

Déterminons la moyenne et la variance de cette loi de probabilité grâce à la méthode des moments, pour ensuite déduire des estimateurs de m et α.
Calculons la fonction caractéristique de X.

$$\mathbb{E}\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itx} \frac{1}{\alpha} e^{-(\frac{x-m}{\alpha})} e^{-e^{-\frac{x-m}{\alpha}}} dx$$

On va effectuer un changement de variable

$$\text{Posons } \begin{cases} x = m - \alpha \ln(y) \\ y = e^{-(\frac{x-m}{\alpha})} \end{cases} \Rightarrow \begin{cases} \frac{dx}{dy} = \frac{-\alpha}{y} \\ dx = \frac{-\alpha}{y} dy \end{cases}$$

$$\begin{cases} \lim_{x \to +\infty}(y) = & 0 \\ \lim_{x \to -\infty}(y) = & +\infty \end{cases}$$

$$\begin{aligned}
\mathbb{E}\left[e^{itX}\right] &= \int_{+\infty}^{0} e^{it(m-\alpha\ln(y))}\frac{1}{\alpha}ye^{-y}\frac{-\alpha}{y}dy \\
&= \int_{0}^{+\infty} e^{it(m-\alpha\ln(y))}e^{-y}dy \\
&= e^{itm}\int_{0}^{+\infty} e^{-it\alpha\ln(y)}e^{-y}dy \\
&= e^{itm}\int_{0}^{+\infty} e^{\ln(y^{-it\alpha})}e^{-y}dy \\
&= e^{itm}\int_{0}^{+\infty} y^{-it\alpha}e^{-y}dy
\end{aligned}$$

On effectue le changement de variable trivial $z-1=-it\alpha \Rightarrow z=1-it\alpha$
Donc

$$\mathbb{E}\left[e^{itX}\right] = e^{itm}\int_{0}^{+\infty} y^{z-1}e^{-y}dy$$

En outre on sait que

$$\int_{0}^{+\infty} y^{z-1}e^{-y}dy = \Gamma(z)$$

où $\Gamma(z)$ est la fonction Gamma.
Il en suit que

$$\mathbb{E}\left[e^{itX}\right] = e^{itm}\times\Gamma(1-it\alpha) = \varphi(t)$$

Afin de trouver les moments d'ordre k, dérivons k fois $\varphi(t)$.

$$\varphi'(t) = \mathbb{E}\left[iXe^{itX}\right] = ime^{itm}\Gamma(1-it\alpha) - e^{itm}i\alpha T'(1-it\alpha)$$

$$\varphi'(0) = \mathbb{E}\left[iX\right] = im\underbrace{\Gamma(1)}_{=0!} - i\alpha\Gamma'(1)$$

En outre $\Gamma'(z)=\Gamma(z)\psi_0(z)$, $\psi_0(z)$ étant la fonction digamma et $\psi_0(1)=-\gamma$ la constante d'Euler-Mascheroni. Donc

$$\begin{aligned}
\varphi'(0) = \mathbb{E}\left[iX\right] &= im + i\alpha\gamma \\
-\mathbb{E}\left[X\right] &= -m - \alpha\gamma \\
\mathbb{E}\left[X\right] &= m + \alpha\gamma
\end{aligned}$$

Pour l'ordre 2 :

$$\varphi''(t) = \left[-m^2e^{itm}\Gamma(1-it\alpha) + \alpha me^{itm}\Gamma'(1-it\alpha) + me^{itm}\alpha\Gamma'(1-it\alpha) - e^{itm}\alpha^2\Gamma''(1-it\alpha)\right]$$

$$\begin{aligned}
\varphi''(0) &= -m^2\Gamma(1) + 2\alpha m\Gamma'(1) - \alpha^2\Gamma''(1) \\
\mathbb{E}\left[X^2\right] &= m^2 - 2\alpha m\Gamma'(1) + \alpha^2\Gamma''(1)
\end{aligned}$$

On a également que $\Gamma'(z) = \Gamma(z)\psi_0(z) \Rightarrow \Gamma''(z) = \Gamma'(z)\psi_0(z) + \Gamma(z)\psi_0'(z)$
$\Rightarrow \Gamma''(z) = \Gamma'(z)\psi_0(z) + \Gamma(z)\psi_1(z)$ avec $\psi_1(z)$ la fonction trigamma.

$$\Rightarrow \Gamma''(z) = \Gamma(z)\psi_0^2(z) + \Gamma(z)\psi_1(z)$$

On sait également que $\psi_1(1) = \frac{\pi^2}{6}$ D'où $\Gamma''(1) = \gamma^2 + \frac{\pi^2}{6}$

D'autre part $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

$$\mathrm{Var}(X) = m^2 - 2\alpha m\Gamma'(1) + \alpha^2\Gamma''(1) - (m - \alpha\Gamma'(1))^2$$

$$\begin{aligned}\mathrm{Var}(X) &= m^2 - 2\alpha m\Gamma'(1) + \alpha\Gamma''(1) - m^2 + 2\alpha m\Gamma'(1) - \alpha^2\Gamma'(1)^2 \\ &= \alpha^2\left(\Gamma(1)\dot{\psi_0'}(1) + \Gamma'(1)\psi_0(1)\right) - \alpha^2\Gamma(1)^2\psi_0(1)^2 \\ &= \alpha^2\left(\psi_0'(1) + \Gamma(1)\psi_0(1)^2\right) - \alpha^2\psi_0(1)^2 \\ &= \alpha^2\psi_0'(1) + \alpha^2\psi_0(1)^2 - \alpha^2\psi_0(1)^2\end{aligned}$$

$$\mathrm{Var}(X) = \alpha^2\frac{\pi^2}{6}$$

Soient $\bar{x_1}$ et $\bar{x_2}$ les moments empiriques d'ordre 1 et 2.

$$\text{On a : } \begin{cases} \mathbb{E}[X] = m + \alpha\gamma \\ \mathbb{E}[X^2] = m^2 + 2m\alpha\gamma + \alpha^2\Gamma''(1) \end{cases} \Rightarrow \begin{cases} \bar{x}_1 = \hat{m} + \hat{\alpha}\gamma \\ \bar{x}_2 = \hat{m}^2 + 2\hat{m}\hat{\alpha}\gamma + \alpha^2\Gamma''(1) \end{cases}$$

$$\Rightarrow \begin{cases} \hat{m} = \bar{x}_1 - \hat{\alpha}\gamma \\ \bar{x}_2 = \bar{x}_1^2 - 2\hat{\alpha}\gamma\bar{x}_1 + \hat{\alpha}^2\gamma^2 + 2\hat{\alpha}\gamma\bar{x}_1 - 2\hat{\alpha}^2\gamma^2 + \hat{\alpha}^2\Gamma''(1) \end{cases}$$

$$\Rightarrow \begin{cases} \hat{m} = \bar{x}_1 - \hat{\alpha}\gamma \\ \bar{x}_2 - \bar{x}_1^2 = \hat{\alpha}^2\Gamma''(1) - \hat{\alpha}^2\gamma^2 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{m} = \bar{x}_1 - \hat{\alpha}\gamma \\ \hat{\alpha} = \sqrt{\frac{\bar{x}_2 - \bar{x}_1^2}{\frac{\pi^2}{6}}} \end{cases}$$

1.3 Calcul des estimateurs du maximum de vraisemblance

3) Déterminons des estimateurs de m et α par la méthode du maximum de vraisemblance.
Soient $X_1, ..., X_n$ iid de même loi que X avec $X \sim \text{Gumbel}(m, \alpha)$.
X a pour densité $f(x) = \frac{1}{\alpha}e^{-\frac{x-m}{\alpha}}e^{-e^{-\frac{x-m}{\alpha}}}$.
On cherche max $f_{X_1,...,X_n}(x_1, ..., x_n|\theta)$ avec $\theta = \begin{bmatrix} m \\ \alpha \end{bmatrix}$

$$f_{X_1,...,X_n}(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

Soit $L(m, \alpha)$ la log-vraisemblance négative.

$$\ln(\prod_{i=1}^{n} f_{X_i}(x_i|\theta)) = n\ln\left(\frac{1}{\alpha}\right) - \sum_{i=1}^{n}\frac{x_i - m}{\alpha} - \sum_{i=1}^{n} e^{-\frac{x_i-m}{\alpha}}$$

Calculons les dérivées partielles de cette fonction :

$$\frac{\partial L(\alpha,m)}{\partial m} = n - \sum_{i=1}^{n} e^{-\frac{x_i-m}{\alpha}} = n - e^{\frac{m}{\alpha}} \sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}}$$

$$\frac{\partial L(\alpha,m)}{\partial \alpha} = \sum_{i=1}^{n} \frac{x_i-m}{\alpha^2} - \frac{n}{\alpha} - \sum_{i=1}^{n} \frac{(x_i-m)}{\alpha^2} e^{-\frac{x_i-m}{\alpha}}$$

Il faut ensuite résoudre le système avec ces deux équations :

$$\begin{cases} \frac{\partial L(\alpha,m)}{\partial m} = 0 \\ \frac{\partial L(\alpha,m)}{\partial \alpha} = 0 \end{cases} \Rightarrow \begin{cases} n - e^{\frac{m}{\alpha}} \sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}} & = 0 \\ \sum_{i=1}^{n} \frac{x_i-m}{\alpha^2} - \frac{n}{\alpha} - \sum_{i=1}^{n} \frac{(x_i-m)}{\alpha^2} e^{-\frac{x_i-m}{\alpha}} & = 0 \end{cases}$$

On résout pour m dans la première équation et on remplace $\frac{m}{\alpha^2} \sum_{i=1}^{n} e^{-\frac{x_i-m}{\alpha}}$ par $\frac{m}{\alpha^2} n$ dans la seconde

$$\Rightarrow \begin{cases} \frac{n}{\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}}} & = e^{\frac{m}{\alpha}} \\ \sum_{i=1}^{n} \frac{x_i-m}{\alpha^2} - \frac{n}{\alpha} + \frac{m}{\alpha^2} n - \sum_{i=1}^{n} \frac{(x_i)}{\alpha^2} e^{-\frac{x_i-m}{\alpha}} & = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \ln(n) - \ln(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}}) & = \frac{m}{\alpha} \\ \sum_{i=1}^{n} \frac{x_i-m}{\alpha^2} - \frac{n}{\alpha} + \frac{m}{\alpha^2} n - \sum_{i=1}^{n} \frac{(x_i)}{\alpha^2} e^{-\frac{x_i-m}{\alpha}} & = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \alpha \left(\ln(n) - \ln(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}}) \right) & = m \\ \sum_{i=1}^{n} \frac{x_i-m}{\alpha^2} - \frac{n}{\alpha} + \frac{m}{\alpha^2} n - \sum_{i=1}^{n} \frac{(x_i)}{\alpha^2} e^{-\frac{x_i-m}{\alpha}} & = 0 \end{cases}$$

On injecte la formule de m dans la deuxième équation, prenons les termes un à un.
Le premier

$$\sum_{i=1}^{n} \frac{x_i - m}{\alpha^2} = \sum_{i=1}^{n} \frac{x_i - \alpha \left(\ln(n) - \ln\left(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}} \right) \right)}{\alpha^2}$$

$$= -\frac{n}{\alpha} \ln(n) + \sum_{i=1}^{n} \frac{x_i + \alpha \ln\left(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}} \right)}{\alpha^2}$$

$$= -\frac{n}{\alpha} \ln(n) + \sum_{i=1}^{n} \frac{x_i}{\alpha^2} + \frac{n}{\alpha} \ln\left(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}} \right) \quad (1)$$

Le deuxième

$$\frac{m}{\alpha^2} n = \frac{1}{\alpha} n \left(\ln(n) - \ln\left(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}} \right) \right)$$

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

Le troisième

$$\sum_{i=1}^{n} \frac{x_i}{\alpha^2} e^{-\frac{x_i - \alpha \ln(n) - \alpha \ln\left(\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}} \right)}{\alpha}}$$

$$= \frac{n}{\alpha^2} \sum_{i=1}^{n} x_i e^{-\frac{x_i}{\alpha}} \times \frac{1}{\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}}} \quad (3)$$

En additionnant tout, $(1) - \frac{n}{\alpha} + (2) - (3)$ puis en simplifiant on arrive à

$$n\frac{\bar{x}}{\alpha^2} - \frac{n}{\alpha} - \frac{n}{\alpha^2}\frac{\sum_{i=1}^{n} x_i e^{-\frac{x_i}{\alpha}}}{\sum_{i=1}^{n} e^{-\frac{x_i}{\alpha}}} = 0$$

$$\bar{x} - \alpha - \frac{\sum x_i e^{-\frac{x_i}{x}}}{\sum e^{-\frac{x_i}{\alpha}}} = 0$$

On pose ensuite

$$f(\alpha) = \bar{x} - \alpha - \frac{\sum x_i e^{-\frac{x_i}{x}}}{\sum e^{-\frac{x_i}{\alpha}}} = 0$$

puis on peut ensuite utiliser la méthode de Newton pour trouver le α qui est solution de $f(\alpha) = 0$ avec une précision arbitraire. Dans la suite des implémentations nous utilisons $\epsilon = 1e - 7$.

1.4 Implémentation numérique des estimateurs calculés théoriquement

Une fois que les formules théoriques des estimateurs ont été calculés, il est possible de simuler la convergence de ces estimateurs. En effet, la méthode serait d'échantillonner n réalisations tirés de la loi de Gumbel, d'utiliser les formules des estimateurs pour les calculer puis de voir à quel point l'estimation est proche de la vraie valeur.

Effectuons ceci pour l'estimateur $\hat{m}$ de m sur la figure 1 pour m = 5.

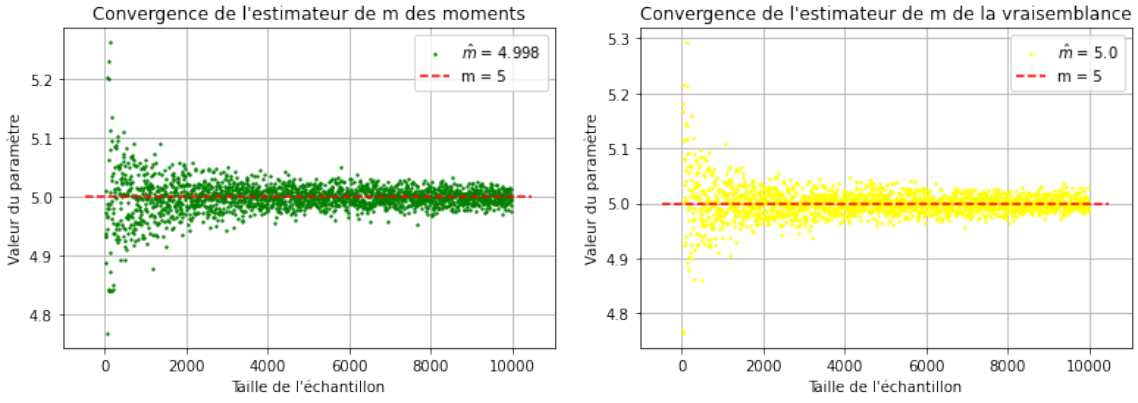


FIGURE 1 – Estimateurs de m

Voici les valeurs simulées pour l'estimateur $\hat{\alpha}$ de α pour $\alpha = 1$ sur la figure 2

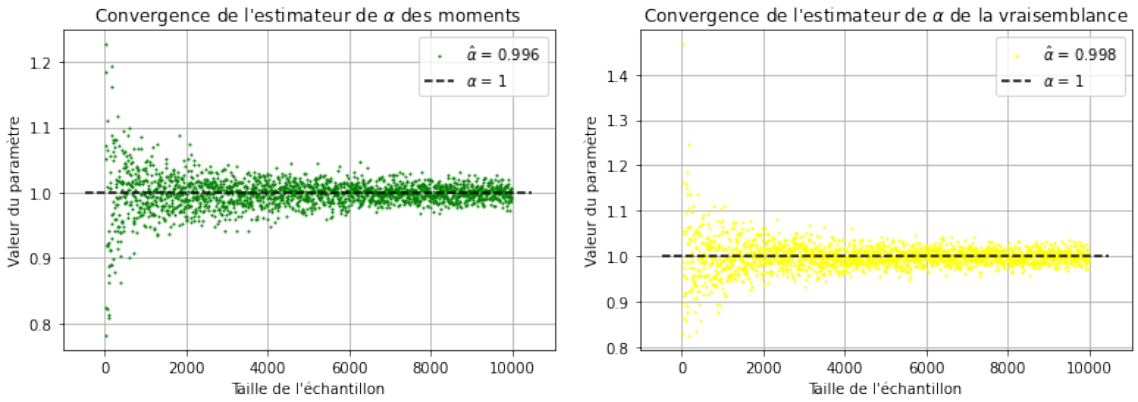


FIGURE 2 – Estimateurs de α

Étant donné que les échelles sont différentes, représentons les estimateurs de m (figure 3) et de α (figure 4) sur le même graphique afin de comparer leur convergence.

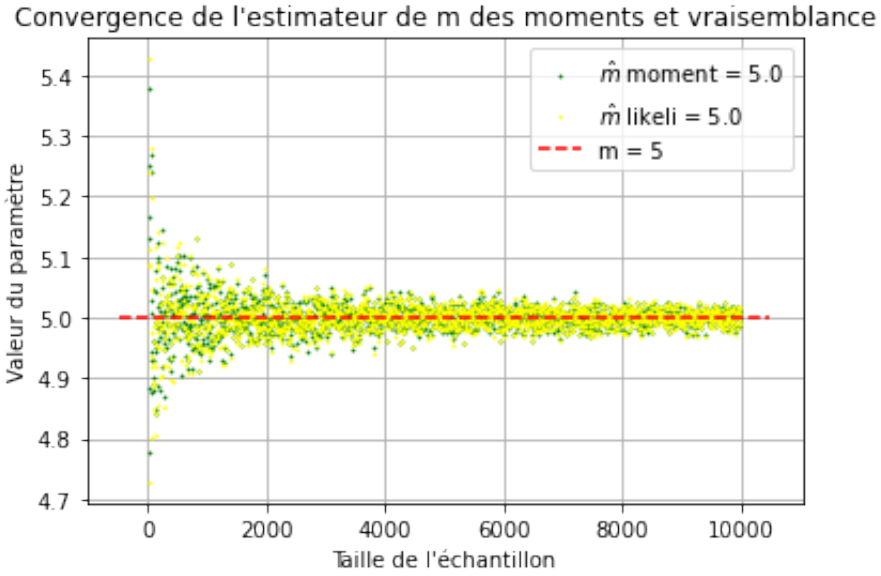


FIGURE 3 – Estimateurs de m

On voit que pour m, les estimateurs semblent très similaires et avoir une vitesse de convergence quasiment identique car les points sont à la même distance de la vraie valeur selon la taille de l'échantillon.

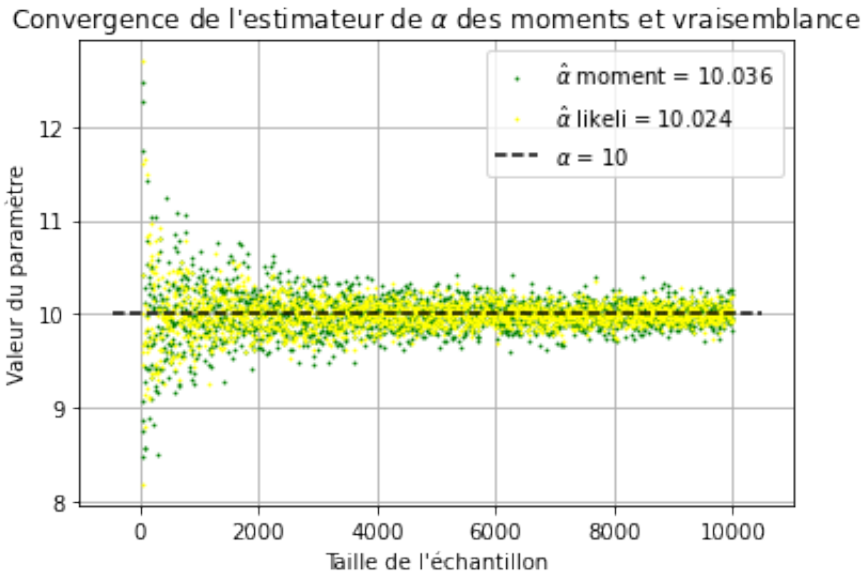


FIGURE 4 – Estimateurs de α

On voit ici que pour $\alpha = 10$ l'estimateur du maximum de vraisemblance semble plus rapide à converger que celui des moments.

2 Calibration des hyper-paramètres d'un processus gaussien par MLE

2.1 Réalisations d'un processus Gaussien

Soit le processus Gaussien $Z \sim \mathscr{PG}(m, k)$ de vecteur moyenne $m(x)$ et de matrice de covariance Σ.
Soient N valeurs équi-distribuées sur [0, 1] tel que $x_1 < ... < x_N$. Soit L la matrice de Cholesky de la matrice de covariance Σ et soit $G = (G_1, ..., G_N)$ un vecteur dont les coordonnées sont des gaussiennes centrées réduites et indépendantes.

Calculons la moyenne et la covariance d'un vecteur aléatoire $Y = m + L^T G$.
Tout d'abord on a

$$\mathbb{E}[Y] = \mathbb{E}[m + L^T G] = \mathbb{E}[m] + L^T \mathbb{E}[G]$$

Or on sait que la moyenne de chaque composante de G est 0, donc $\mathbb{E}[G] = O_{\mathbb{R}^N}$ et m est déjà une moyenne donc $\mathbb{E}[m] = m$. D'où on a

$$\mathbb{E}[Y] = m$$

Calculons maintenant la matrice de covariance de Y.

$$Cov(Y) = \mathbb{E}[(Y - \mathbb{E}[Y]) \times (Y - \mathbb{E}[Y])^T]$$

$$= \mathbb{E}[(m + L^T G - m) \times (m + L^T G - m)^T]$$

$$= \mathbb{E}[L^T G \times (L^T G)^T]$$

$$= L^T \mathbb{E}[G G^T] L$$

Or on sait que $\mathbb{E}[GG^T] = \mathbb{E}[(G - \mathbb{E}[G])(G - \mathbb{E}[G])^T] = Cov(G)$ car $\mathbb{E}[G] = 0_{\mathbb{R}^N}$.
De même, $\forall (i, j) \in \{1, ..., N\}^2, i \neq j, Cov(G_i, G_j) = 0$ car les variables sont indépendantes, et $\forall i \in \{1, ..., N\}, Var(G_i) = 1$. On a donc que $Cov(G) = I_N$.
Il ensuit que

$$Cov(Y) = L^T I_N L = L^T L = \Sigma$$

On a alors $\mathbb{E}[m + L^T G] = m$ et $Cov(m + L^T G) = \Sigma$

Il est donc possible de simuler un processus Gaussien suivant la loi $\mathbb{N}(m, \Sigma)$ grâce à un tirage de vecteur de gaussiennes centrées réduites.
C'est ce que nous allons désormais faire dans la suite du projet.

1. Calculer Σ
2. Déterminer L de la décomposition de Cholesky de Σ
3. Tirer G : vecteur N variables gaussiennes centrées réduites
4. Calculer $Z = L^T \times G$

Voici quelques exemples, pour $\lambda \in 0.01, 0.05, 0.1, 0.5, 1, 5, 10$ sur les figures 5 6 7 :

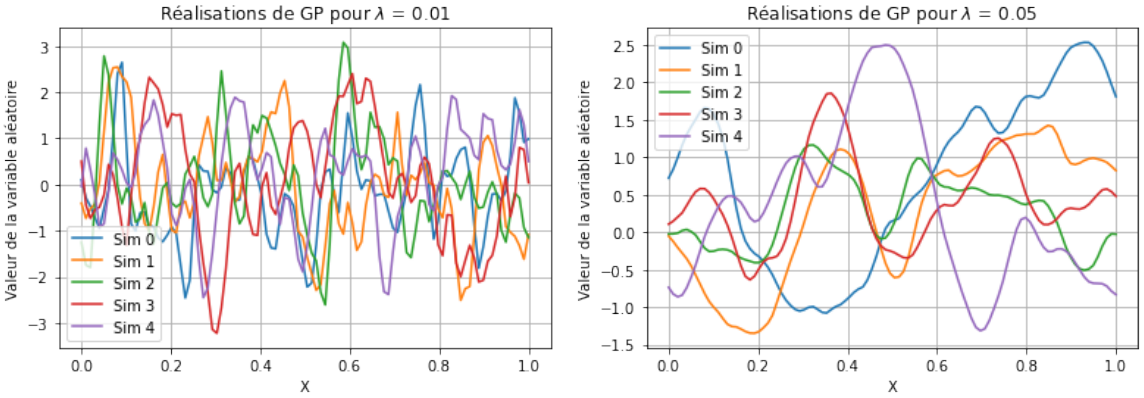


FIGURE 5 – Simulation de quelques processus Gaussiens $\lambda = 0.01$ et $\lambda = 0.05$

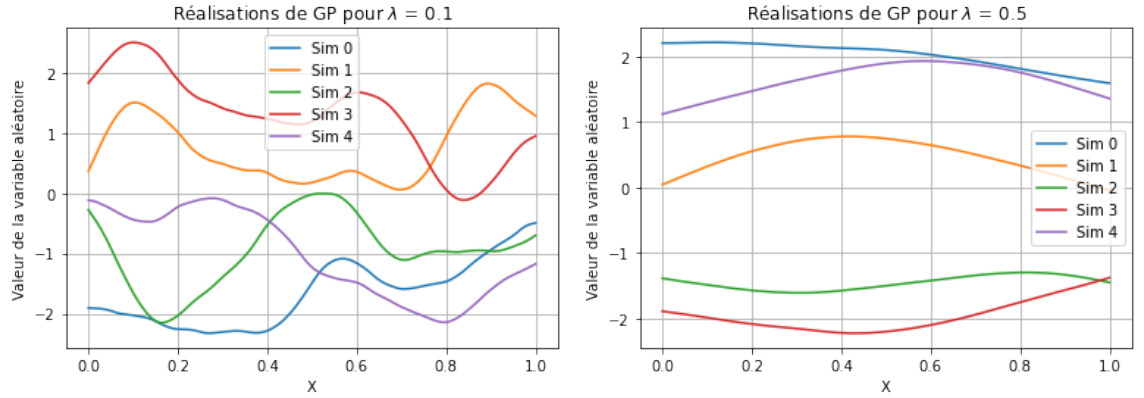


FIGURE 6 – Simulation de quelques processus Gaussiens $\lambda = 0.1$ et $\lambda = 0.5$

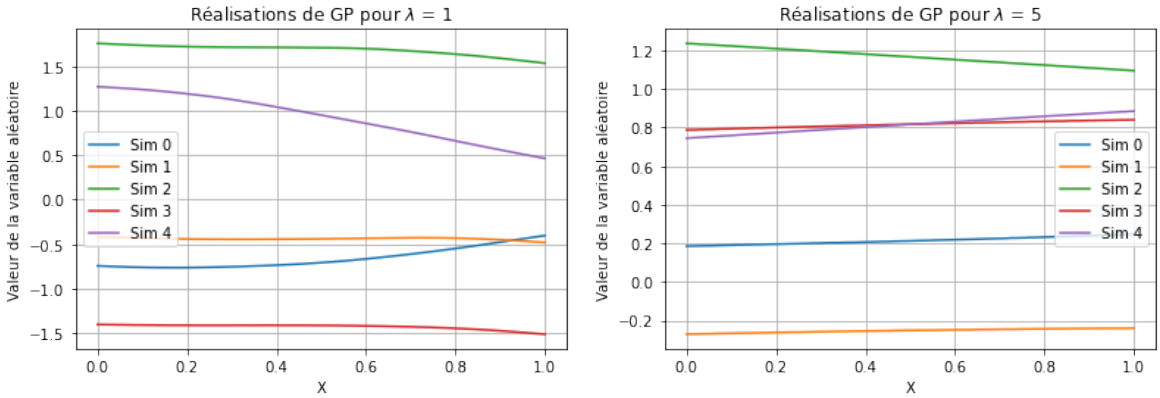


FIGURE 7 – Simulation de quelques processus Gaussiens $\lambda = 1$ et $\lambda = 5$

Pour expliquer ces résultats, regardons le profil de la matrice de covariance Σ selon λ sur la figure 8 :

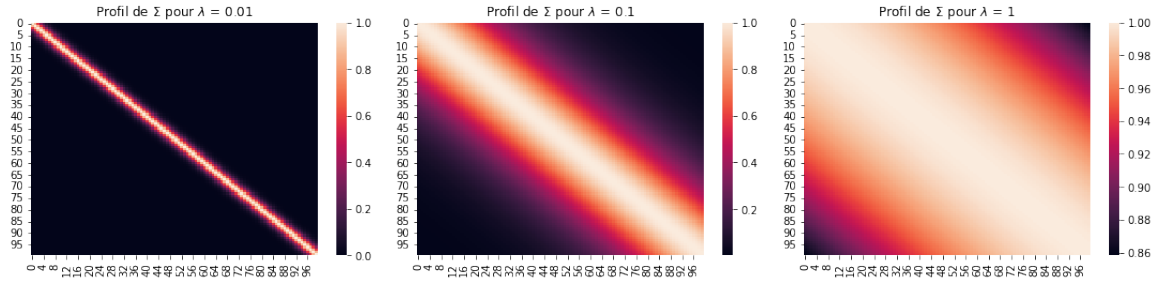


FIGURE 8 – Profil de Σ pour $\lambda \in \{0.01, 0.1, 1\}$

Plus λ est élevé, plus les valeurs sont corrélées. C'est pour cette raison que dans les figure 5 6 7, plus λ est grand, moins les courbes sont erratiques car chaque deux points de la courbe sont plus corrélés. Plus λ est petit, moins les points sont corrélés et plus la courbe est libre de varier

2.2 Estimation du paramètre λ par maximum de vraissemblance

La log-vraisemblance négative a pu être calculée de cette manière ci-après.

$$-ln(\mathscr{L}(Z,\lambda)) = -ln(\frac{1}{\sqrt{(2\pi)^d det(\Sigma)}})e^{-\frac{1}{2}z^T\Sigma^{-1}z}$$

$$= \frac{1}{2}dln(2\pi) + \frac{1}{2}ln(det(\Sigma)) + \frac{1}{2}z^T\Sigma^{-1}z$$

Dans cette formule, Σ dépend de λ.

Étant donné que le kernel utilisé dans le projet est plutôt instable (la matrice de covariance est numériquement pas définie positive pour certaines valeurs de λ), nous avons dû rajouter un bruit à la diagonale de la matrice de covariance de l'ordre de 10^{-7} afin de la rendre numériquement définie positive.

Cependant, cela a eu un impact sur la valeur de la log-vraisemblance négative qui est devenue bruitée également.

Graphiquement, en traçant $-ln(\mathscr{L}(Z,\lambda))$ en fonction de λ, nous nous retrouvons avec une courbe de la sorte (figure 9) :

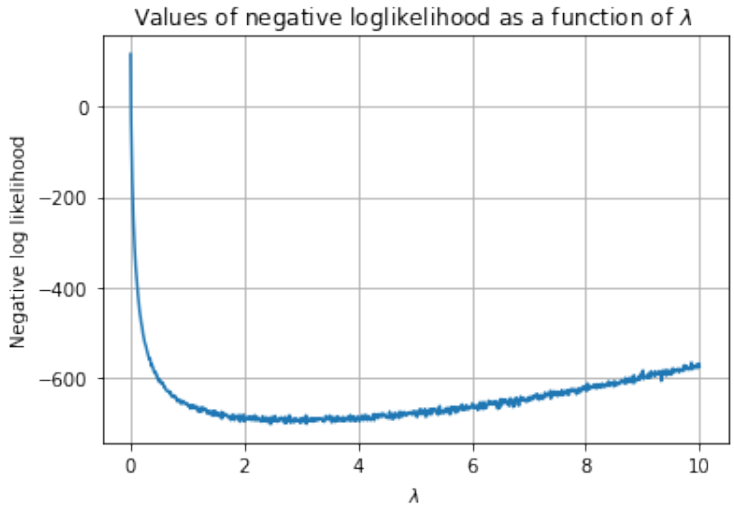


FIGURE 9 – Log-vraisemblance négative en fonction de λ

Les algorithmes de minimisations n'ont donc pas pu estimer correctement la meilleur valeur de lambda.

2.3 Simuler plusieurs réalisations dun processus Gaussien conditionné à un ensemble de données

Reprenons les variables de l'exercice, à savoir le processus Gaussien $Z \sim \mathscr{PG}(m, \Sigma)$ de vecteur moyenne $m(x)$ et de matrice de covariance Σ. et N valeurs $(x_1, ..., x_n) = X$ distribuées sur [0, 1] tel que $x_1 < ... < x_n$ et que chacune soit tirée uniformément dans une partition de n intervalles de même taille consécutifs de [0, 1]. On tire un autre ensemble de N valeurs $x_1^{'}, ..., x_N^{'} = X'$ dans [0, 1].
Par définition, à chaque point x_i est assignée une valeur $Z(x_i) \in \mathbb{R}$. On peut noter que

$$p(Z|X) = \mathscr{N}(Z|m, \Sigma)$$

Avec m le vecteur moyenne de Z : $(m(x_1), ..., m(x_n))$ et Σ sa fonction de covariance entre chaque x_i de X : $\Sigma_{i,j} = k_\lambda(x_i, x_j)$.

Notons Σ_n la matrice de covariance entre les n points d'apprentissages X, Σ_N la matrice de covariance entre les N nouveaux points X' et $\Sigma_{n,N}$ la matrice de covariance entre les n points d'apprentissage X et les N nouveaux points X'. Si l'on note $Z_X = (Z(x_1), ..., Z(x_n))$ et $Z_{X'} = (Z(x_1^{'}), ..., Z(x_n^{'}))$, par définition d'un processus Gaussien chaque ensemble fini de $Z(x_i)$ est Gaussien.

D'après les règles de conditionnement des lois normales multivariées, le vecteur $Z_{X'}|Z_X$ suit une loi normale $\mathscr{N}(\mu_{X'}, \Sigma_{N|n})$ avec $\mu_{X'} = \Sigma_{n,N}^T \Sigma_n Z_X$ et $\Sigma_{N|n} = \Sigma_N - \Sigma_{n,N}^T \Sigma_n^{-1} \Sigma_{n,N}$

De plus, le vecteur

$$\begin{bmatrix} Z_X \\ Z_{X'} \end{bmatrix} \sim \mathscr{N}\left(\begin{bmatrix} \mu_X \\ \mu_{X'} \end{bmatrix}, \begin{bmatrix} \Sigma_n & \Sigma_{n,N} \\ \Sigma_{n,N}^T & \Sigma_N \end{bmatrix} \right)$$

Pour un $x \in [0, 1]$ fixé, $Z(x) \sim \mathscr{N}(0, 1)$ car $\mathbb{E}[Z] = 0$ et $k_\lambda(x, x) = (1 + 0 + 0)e^{-0} = 1$.

Pour $Z(x)|Z_X$, On a alors

$$Z(x)|Z_X \sim \mathcal{N}(\mu_n(x), \Sigma_{1|n})$$

Il s'agit de la formule précédente pour N = 1. $\mu_n(x) = \Sigma_{n,1}^T \Sigma_n Z_X$ où $\Sigma_{n,1}$ est le vecteur de covariance entre x et chaque x_i de X.
De même, $\Sigma_{1|n} = 1 - \Sigma_{n,1}^T \Sigma_n^{-1} \Sigma_{n,1}$.

On peut donc calculer la fonction $\mu_n(x)$ ainsi que $\Sigma_{1|n}(x)$. Ensuite, il est possible d'afficher les paramètres ainsi que la fonction $sin(4\pi x)$ sur un même graphique (figure 10).

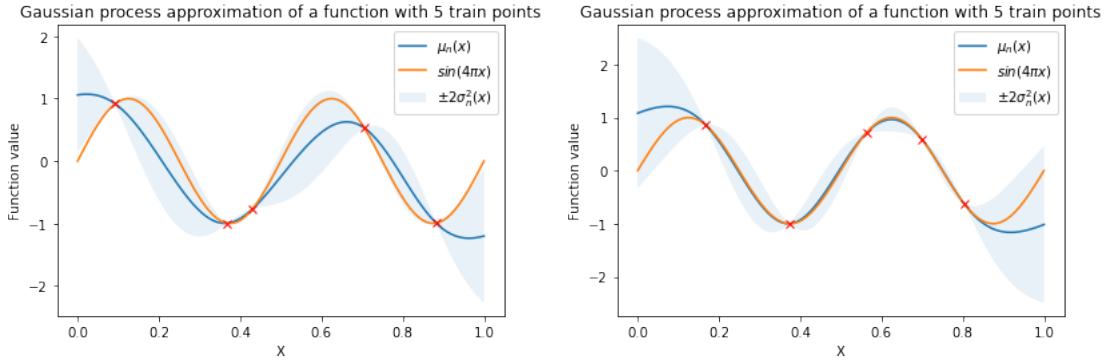


FIGURE 10 – $\mu_n(x)$ ainsi que $\Sigma_{1|n}(x)$ conditionné à 5 points de $sin(4\pi x)$

Générons maintenant grâce à la technique vue précedemment un grand nombre de processus gaussiens avec 5 (figure 11) puis 10 (figure 12) points d'entraînements de $sin(4\pi x)$. Le code reprend la méthode vue plus haut afin de générer ces processus gaussiens :

```
def generate_GP_samples(mu, cov, size, num_samples):
    """ Generate n gaussian process samples of size k
    according to mu and cov
    Args:
        mu: Mean vector
        cov: Covariance vector
        size: Size of samples
        num_samples: Number of GP to sample
    """
    L = np.linalg.cholesky(cov)
    samples = np.zeros((num_samples, size))
    for i in range(num_samples):
        g = np.random.normal(size=size)
        samples[i, :] = mu + L @ g
    return samples
```

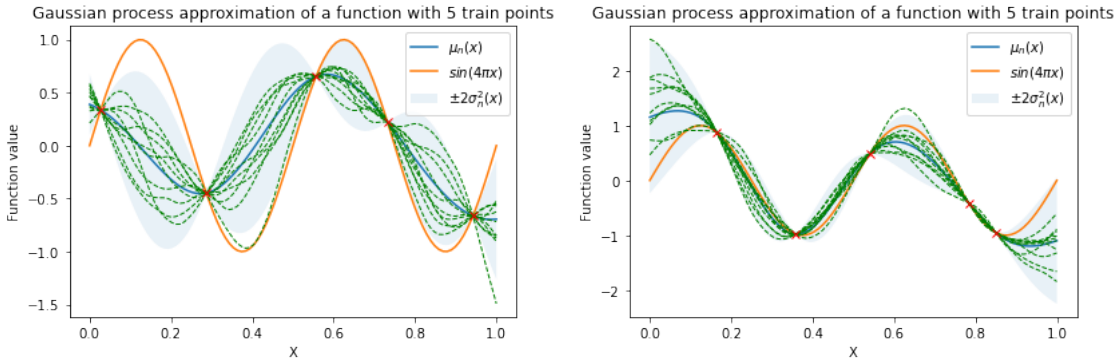


FIGURE 11 – Processus gaussiens régressés selon 5 points d'entraînement

On voit que les processus sont plutôt imprécis lorsque la moyenne ne suit pas bien la fonction sinus d'entraînement.

Naturellement, lorsqu'on augmente la taille du Z_X, on obtient des processus bien plus proches de la fonctions sinus (figure 12).

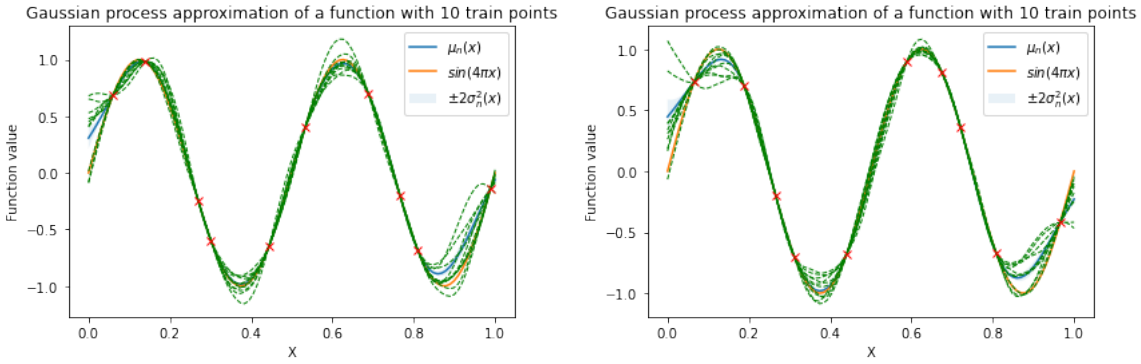


FIGURE 12 – Processus gaussiens régressés selon 10 points d'entraînement

3 Simuler une loi de distribution a posteriori avec un algorithme MCMC

Le but dans cette partie est d'échantillonner de nombreux éléments selon une certaine loi de probabilité par une méthode de MCMC. Grâce à une loi à priori sur un paramètre, ainsi qu'avec la connaissance de la vraisemblance des valeurs selon ce paramètre, il est possible de calculer la probabilité à posteriori grâce à la formule de bayes, à une constante multiplicative près.
En général, nous ne connaissons pas la « vraie loi » à posteriori que nous souhaitons échantillonner, mais ici nous la connaissons afin de pouvoir comparer nos résultats à la formule analytique.

Posons le problème :
Soit X une variable aléatoire suivant une loi de poisson de paramètre θ inconnu. La loi de distribution à priori de θ est une loi Gamma de paramètres k et λ. Afin de travailler avec une notation conventionnelle, nous appellerons $k : \alpha$ et $\lambda : \beta$.

$$X \sim \mathscr{P}(\theta)$$

$$\theta \sim \mathscr{G}(\alpha, \beta) \quad (\text{avec } \gamma = 0)$$

De même, lorsque $X = x$ est fixé, on connaît la loi de $\theta|X = x$.

$$\theta|X \sim \mathscr{G}(\alpha + x, \beta + 1) \quad (\text{avec } \gamma = 0)$$

On sait d'après le cours qu'il est possible de construire une séquence d'échantillons d'une variable aléatoire X_n : $X_n \sim p(x|d) \propto L(d|x)\pi(x)$ par MCMC.
L'algorithme de Metropolis-Hastings permet de créer une chaîne de Markov dont la distribution stationnaire est celle recherchée à une constante multiplicative près. Ici, on va donc chercher à modéliser

$$p(\theta|x) \propto L(x|\theta)\pi(\theta)$$

en fixant x dès le départ en entier naturel. On sait d'ailleurs que $L(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}$ car X suit une loi de Poisson, et que $\pi(\theta)$ suit une loi Gamma de paramètres α et β.
Les deux termes sont donc calculables pour tout $\theta \in \mathbb{R}^{+*}$.

Fixons, par exemple, $\alpha = 0.5$, $\beta = 0.01$ et $x = 10$. On connaît alors la distribution de $\theta|X$: une loi Gamma de paramètres $\alpha = 10.5$ et $\beta = 1.01$. Cela donne donc une courbe de ce type (figure 13) :

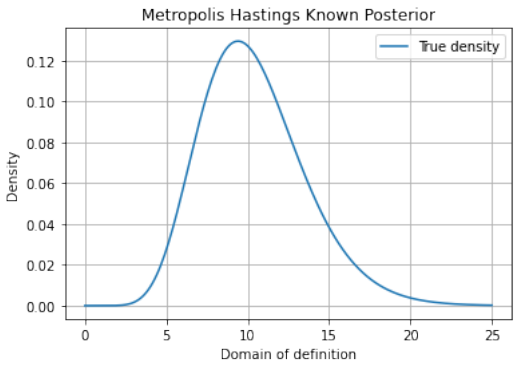


FIGURE 13 – Distribution de $\theta|X$ a posteriori (analytique)

Après avoir effectué un échantillonnage de la loi à posteriori, en ignorant que l'on connaît la vraie loi, il est possible de comparer l'histogramme obtenu avec la fonction de densité de la vraie loi. Nous avons utilisé l'algorithme de Metropolis-Hastings, dont une partie de l'implémentation vient du site MoonBooks que nous avons modifié selon nos besoins.

Nous avons créé une chaîne de Markov de longueur 100 000, dont nous n'avons pris qu'une valeur sur 10. En effet, un des défauts de l'échantillonnage par cette méthode est que les échantillons successifs sont très corrélés entre eux. Un des moyens de réduire cette corrélation est de n'en garder que quelques-uns. Ensuite, la méthode de Métropolis-Hastings a besoin d'un temps de *burn-in* afin d'atteindre un état de la distribution stationnaire. En outre, on sait que la chaîne est ergodique avec une distribution de proposition Gaussienne, car tout état de la chaîne peut être sélectionné en une seule transition. En effet, si l'on démarre à θ fixé, il existe une probabilité non nulle de choisir, pour tout $a \in \mathbb{R}$, une transition de $a - \theta$, pour arriver à $\theta + (a - \theta) = a$. Si l'on commence la chaîne de Markov dans une région dense de la distribution (par exemple près du mode), la période de *burn-in* n'est, en théorie, même pas nécessaire. Cependant pour une distribution de proposition uniforme entre $\theta - \Delta/2$ et $\theta + \Delta/2$ avec $\Delta \in \mathbb{R}$, cela n'est pas vrai. La chaîne serait ergodique car il serait possible d'atteindre tout état en un nombre de transitions N fini, mais le *burn-in* serait nécessaire.
Comme demandé dans le sujet, nous n'avons gardé que les 50% derniers échantillons de la chaîne, ce qui représente au final 5 000 échantillons.

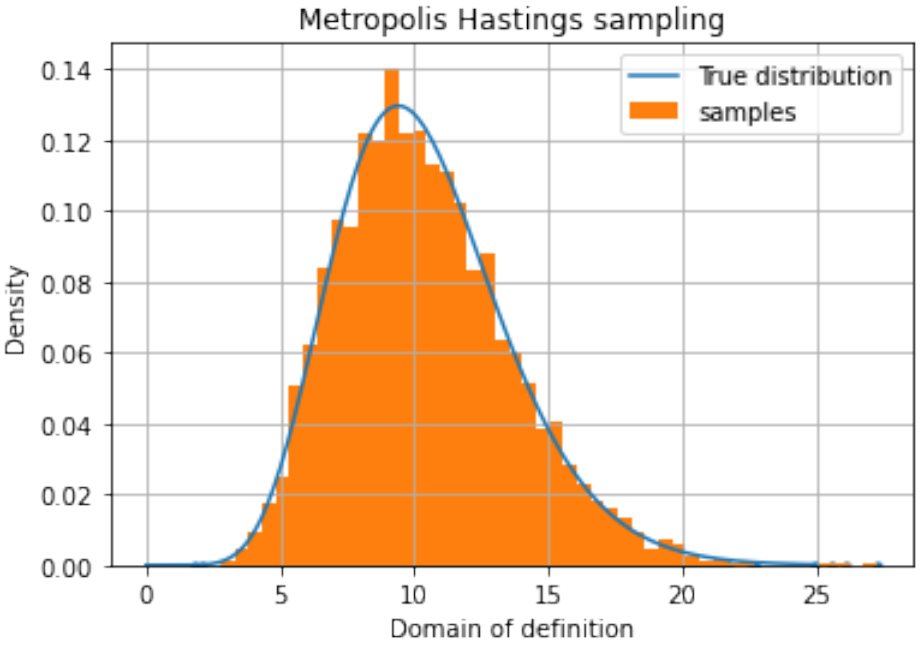


FIGURE 14 – Comparaison de la distribution échantillonnée à la vraie

On remarque sur la figure 14 que la distribution échantillonnée par MCMC s'approche fortement de la distribution réelle.
On pourrait affiner encore plus la distribution empirique, en augmentant la taille de la chaîne de markov par exemple, et en affiner la représentation en jouant sur le nombre de *bins* dans l'histogramme ou bien en utilisant une estimation à noyau de la densité.

Voici les résultats obtenus grâce à différentes estimation à noyau de la densité sur la figure 15.

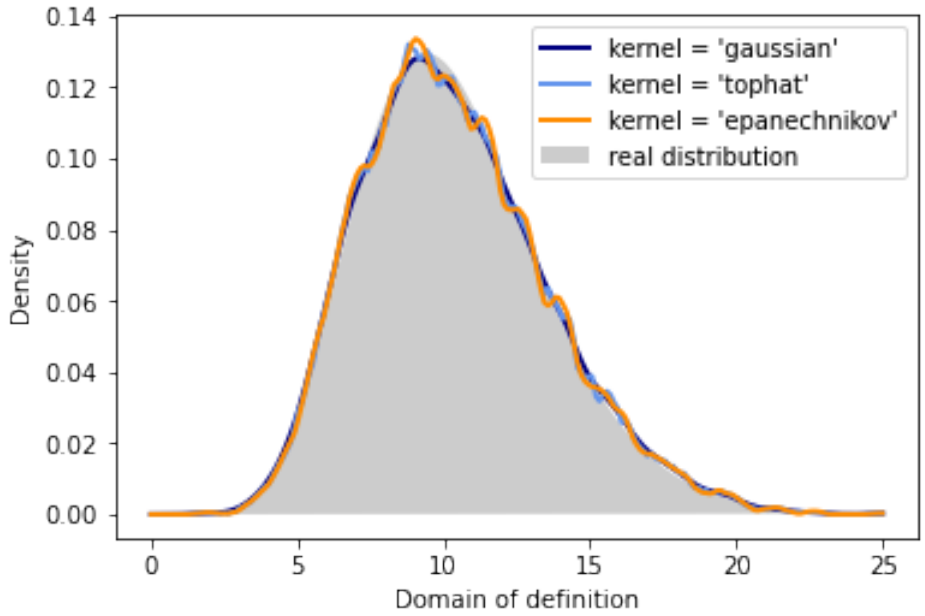


FIGURE 15 – Estimation à noyau de la densité des échantillons obtenus

L'estimation à noyau de la densité, selon les méthodes Gaussian, Tophat et Epanechnikov (respectivement en bleu foncé, bleu clair et orange) se rapprochent très fortement de la vraie densité (en gris).

Voici les mêmes figures, sans les mêmes commentaires, pour d'autres valeurs de paramètres. Ici sur les figures 16 et 17, $\alpha = 5$, $\beta = 1$, $x = 5$.

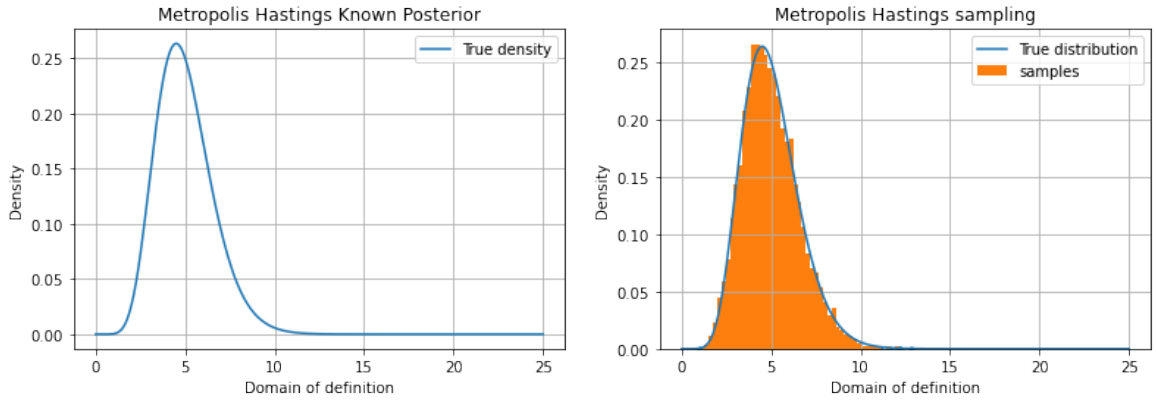


FIGURE 16 – Figures précédentes pour d'autres paramètres

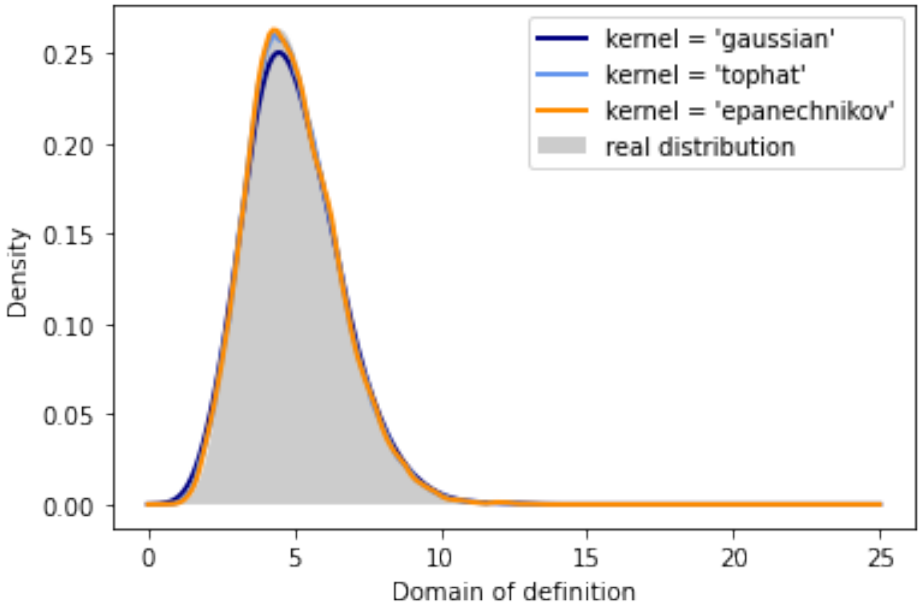


FIGURE 17 – Estimation à noyau pour d'autres paramètres

4 Conclusion

Ce projet, plutôt difficile, fut très long à réaliser étant donné le nombre de questions théoriques à rechercher et à rédiger.
La partie sur les processus gaussiens fut particulièrement difficile étant donné que l'on n'a vu que la définition de ces objets mathématiques en cours, sans les comprendre.

L'intégralité du code (avec historique git) peut être récupéré sur ce lien GitHub