MAP311 - Statistiques non paramétriques

Fabrice Serret & Haris Sahovic

Lundi 3 juillet 2017

Test de Kolmogorov-Smirnov

1. On note F_X la fonction de répartition de X_i.
   Pour tout $x \in \mathbb{R}$ et i tel que $1 \leq i \leq n$, on a par définition de F_X :

   $$\mathbb{P}(X_i \leq x) = F_X(x)$$

   En notant $Y_i = 1_{X_i \leq x}$, Y_i est le résultat d'une experience à deux issues possibles et donc d'une expérience de Bernoulli de paramètre $p = F_X(x)$.
   Ainsi, $Y_i \sim \mathscr{B}(F_X(x))$. De plus, les Y_i sont i.i.d. en vertu du caractère i.i.d. des X_i.

   $(Y_i)_{1 \leq i \leq n}$ est donc une suite de variables aléatoires indépendantes, de même loi et de carré intégrable. D'après la loi forte des grands nombres, on a :

   $$\frac{\sum_{i=1}^n Y_i}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x} = \hat{F}_n(x) \underset{\text{p.s.}}{\longrightarrow} \mathbb{E}(Y_i) = F_X(x)$$

   D'où :

   $$\hat{F}_n(x) \underset{\text{p.s.}}{\longrightarrow} F_X(x)$$

   Il en résulte que la fonction de répartition empirique évaluée en x est un estimateur fortement consistant de la fonction de répartition des X_i évaluée en x.

   Ce résultat équivaut à montrer la convergence simple de $\hat{F}_n$ vers F_X tandis que le théorème de Glivenko-Cantelli montre la convergence uniforme de $\hat{F}_n$ vers F_X, ce qui constitue un résultat plus fort. □

2. En supposant F_X continue, on a $\forall x \in \mathbb{R}$:

   $$\mathbb{P}(X_i = x) = 0$$

   en particulier, $\forall i, j \in \mathbb{N}, 1 \leq i, j \leq n, i \neq j$, on a :

   $$\mathbb{P}(X_i = X_j) = 0$$

   D'où $\forall i \in \mathbb{N}, 1 \leq i \leq n-1$,

   $$X_{(i)} < X_{(i+1)} \quad p.s$$

   On peut donc supposer (presque sûrement) strictes les inégalités de la seconde condition. □

3. En supposant F_{ref} continue, on a :

   $$\begin{aligned} K_n =& \|\hat{F}_n - F_{ref}\|_\infty \\ =& \sup_{x \in \mathbb{R}} \{|\hat{F}_n(x) - F_{ref}(x)|\} \\ =& \sup_{x \in \mathbb{R}} \{|\frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x} - F_{ref}(x)|\} \end{aligned}$$

D'une part, on a :

$$\begin{aligned}&\sup_{x\leq X_{(1)}}\{|\frac{1}{n}\sum_{i=1}^{n}1_{X_{(i)}\leq x}-F_{ref}(x)|\}\\=&\sup_{x\leq X_{(1)}}\{|F_{ref}(x)|\}\\=&|F_{ref}(X_{(1)})|\end{aligned}$$

puisque $F_{ref}(0)=0$ et F_{ref} est croissante, positive et continue.

D'autre part, on a :

$$\begin{aligned}&\sup_{X_{(n)}\leq x}\{|\frac{1}{n}\sum_{i=1}^{n}1_{X_{(i)}\leq x}-F_{ref}(x)|\}\\=&\sup_{X_{(n)}\leq x}\{|1-F_{ref}(x)|\}\\=&|1-F_{ref}(X_{(n)})|\end{aligned}$$

puisque $\lim_{x\to\infty}F_{ref}(x)=1$ d'où $\lim_{x\to\infty}F_{ref}(x)-1=0$ et F_{ref} est croissante, positive et continue.

Enfin, on a :

$$\begin{aligned}&\sup_{X_{(1)}\leq x<X_{(n)}}\{|\frac{1}{n}\sum_{i=1}^{n}1_{X_i\leq x}-F_{ref}(x)|\}\\=&\max_{1\leq k\leq n-1}\sup_{X_{(k)}\leq x<X_{(k+1)}}\{|\frac{1}{n}\sum_{i=1}^{n}1_{X_{(i)}\leq x}-F_{ref}(x)|\}\\=&\max_{1\leq k\leq n-1}\sup_{X_{(k)}\leq x<X_{(k+1)}}\{|\frac{k}{n}-F_{ref}(x)|\}\\=&\max_{1\leq k\leq n-1}\max\{|\frac{k}{n}-F_{ref}(X_{(k)})|,|\frac{k}{n}-F_{ref}(X_{(k+1)})|\}\}\end{aligned}$$

puisque F_{ref} est croissante, positive et continue.

Finalement, on obtient :
$K_n=\max(|F_{ref}(X_{(1)})|,|1-F_{ref}(X_{(n)})|,\max_{1\leq k\leq n-1}\max\{|\frac{k}{n}-F_{ref}(X_{(k)})|,|\frac{k}{n}-F_{ref}(X_{(k+1)})|\})$

Que l'on peut réécrire :
$K_n=\max_{1\leq i\leq n}\{\max\{|F_{ref}(X_{(i)})-\frac{i-1}{n}|,|F_{ref}(X_{(i)})-\frac{i}{n}|\}\}$

D'où le résultat. □

4. On rappele que F_X est la fonction de répartition des X_i.
On définit ensuite $G:\ ]0;1[\to\mathbb{R}$ par $G(\omega)=\inf\{x\in\mathbb{R}|F_X(x)\geq\omega\}$.

Avec $U\sim\mathscr{U}(0,1)$, $Y=G(U)$ a pour fonction de répartition F_X (4.2.4 du cours).

Il en résulte en particulier qu'à chaque X_i correspond une variable aléatoire $U_i\sim\mathscr{U}(0,1)$ telle que $(X_i)=(G(U_i))$.

Par croissance de G, à la statistique d'ordre $(X_{(i)})$ correspond $(U_{(i)})$, telle que $G(U_{(i)})=X_{(i)}$ puisqu'il y a alors préservation de la relation d'ordre.

D'où, avec $F_{ref}=F_X$:

$$\begin{aligned}K_n&=\max_{1\leq i\leq n}\{\max\{|F_X(X_{(i)})-\frac{i-1}{n}|,|F_X(X_{(i)})-\frac{i}{n}|\}\}\\&=\max_{1\leq i\leq n}\{\max\{|F_X(G(U_{(i)}))-\frac{i-1}{n}|,|F_X(G(U_{(i)}))-\frac{i}{n}|\}\}\end{aligned}$$

On note alors que $\forall \omega \in]0;1[$:

$$\begin{aligned} F_X(G(\omega)) =& \mathbb{P}(X \leq G(\omega)) \\ =& \mathbb{P}(X \leq \inf\{x \in \mathbb{R} | F_X(x) \geq \omega\}) \\ =& \mathbb{P}(F_X(X) \leq \omega) \\ =& \omega \text{ par continuité de } F_X. \end{aligned}$$

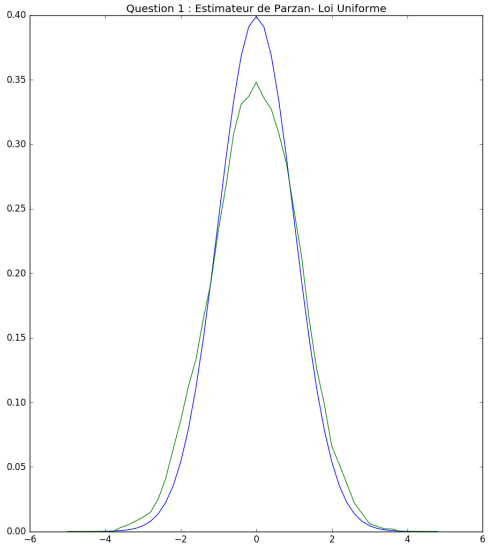
D'où finalement $K_n = \max\limits_{1 \leq i \leq n} \{\max\{|U_{(i)} - \frac{i-1}{n}|, |U_{(i)} - \frac{i}{n}|\}\}$

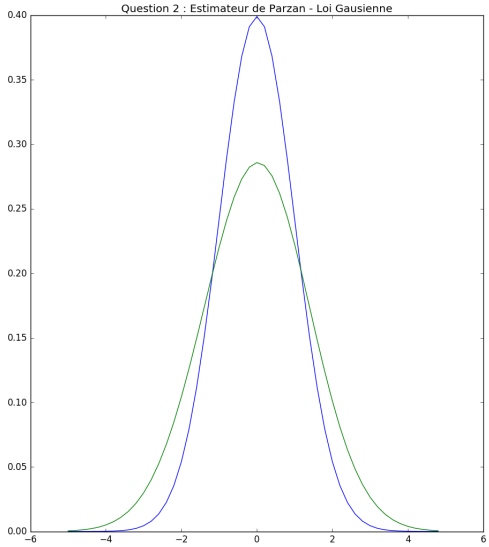
Ainsi, K_n ne dépend pas de la loi des X_i. □

Voilà le rendu du code joint par courriel (remis en forme pour LaTeX) :

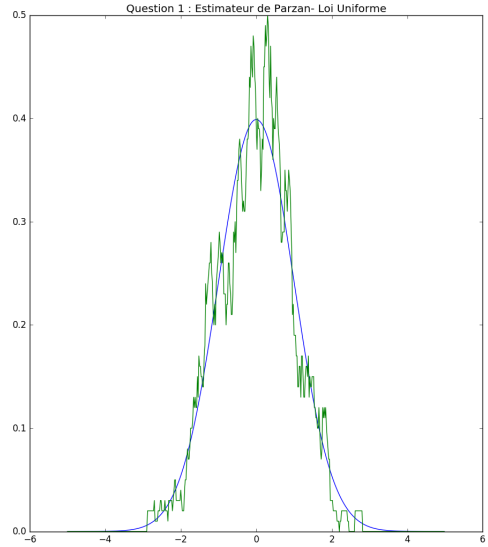
```
Partie 1 - Test de Kolmogorov-Smirnov

Question 5 :
Pour n =   10, le quantile à 0.90 de la loi de Kn est estimé à 0.367370 (échantillon de 10000 tirages)
Pour n =   10, le quantile à 0.95 de la loi de Kn est estimé à 0.406352 (échantillon de 10000 tirages)
Pour n =   10, le quantile à 0.99 de la loi de Kn est estimé à 0.484104 (échantillon de 10000 tirages)
Pour n =  100, le quantile à 0.90 de la loi de Kn est estimé à 0.119394 (échantillon de 10000 tirages)
Pour n =  100, le quantile à 0.95 de la loi de Kn est estimé à 0.132583 (échantillon de 10000 tirages)
Pour n =  100, le quantile à 0.99 de la loi de Kn est estimé à 0.156391 (échantillon de 10000 tirages)
Pour n = 1000, le quantile à 0.90 de la loi de Kn est estimé à 0.038570 (échantillon de 10000 tirages)
Pour n = 1000, le quantile à 0.95 de la loi de Kn est estimé à 0.042776 (échantillon de 10000 tirages)
Pour n = 1000, le quantile à 0.99 de la loi de Kn est estimé à 0.052164 (échantillon de 10000 tirages)

Question 6 :
On a estimé le quantile k_95 à 0.132583 dans la question précédente, dans le cas où n = 100.
Ici, on rejeterais l'hypothèse de normalité dans 132 cas sur 1000, soit 0.132 ou 13.2 %

Partie 2 - Estimateur de Parzen

Question 1 & 2 :
NB : en fonction du type de terminal et de l'environnement utilisé pour executer le programme,
l'affichage des graphes peut se faire sur la ligne de commande ou non.
On constate que les valeur de h trop élevées mènent à un trop grand étalement des fonctions estimées,
tandis que les valeurs de h trop petites mènent à une surrepresentation des fluctuations stochastiques
liées au tirage.
On constate aussi que l'estimateur avec loi normale est plus lissé et de meilleure qualité pour les
petites valeurs de h que celui basé sur une repartition uniforme.
```

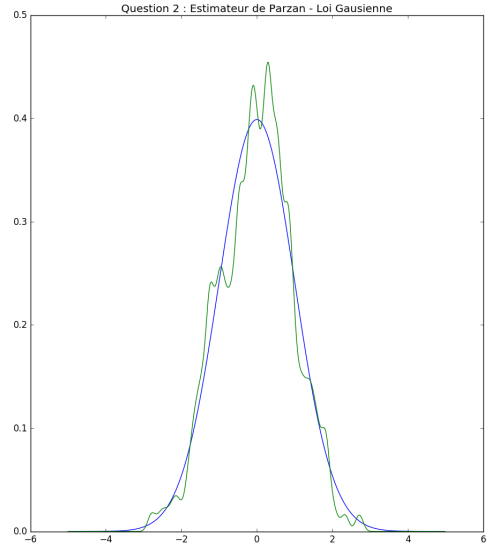
Simulations pour h = 1.000000

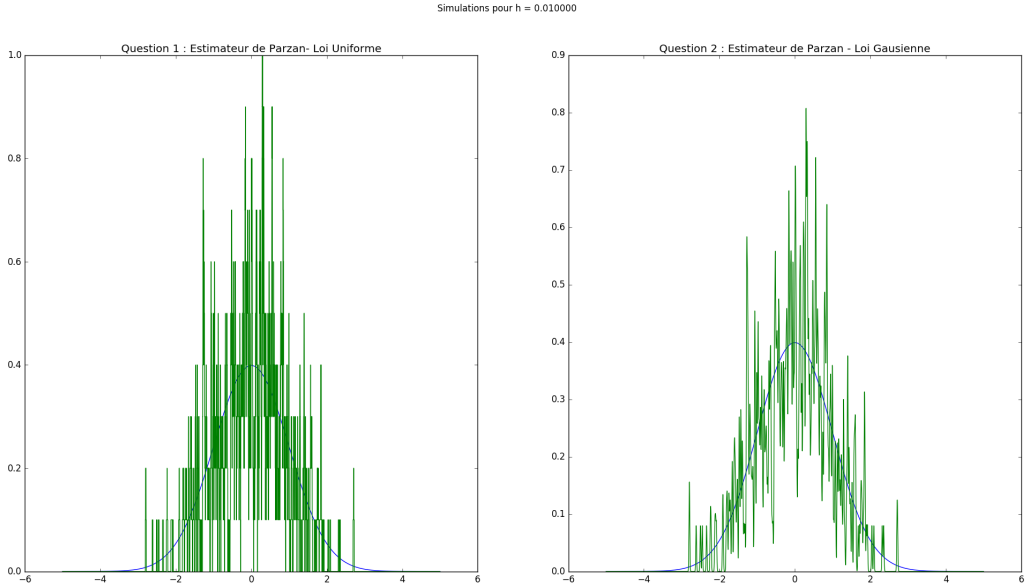




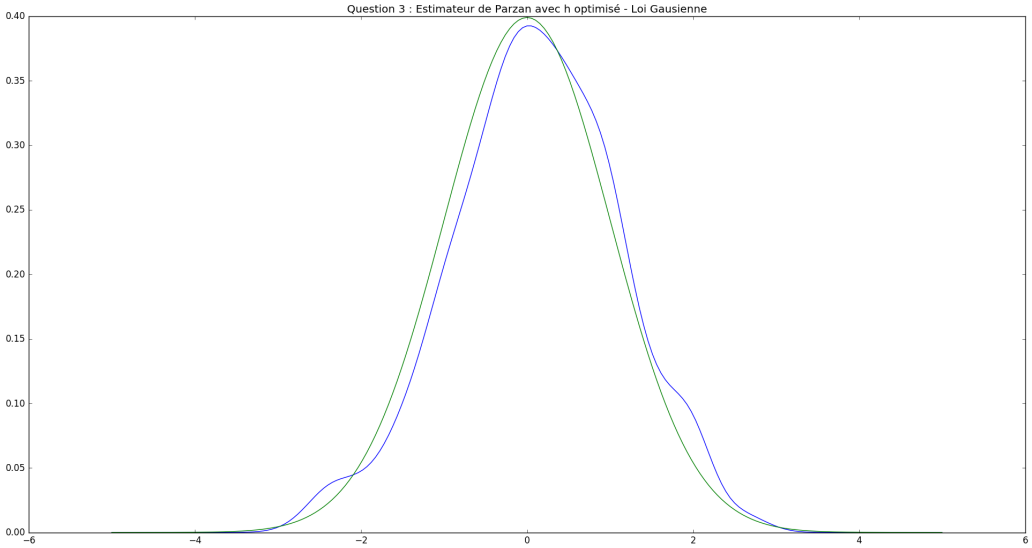
Simulations pour h = 0.100000

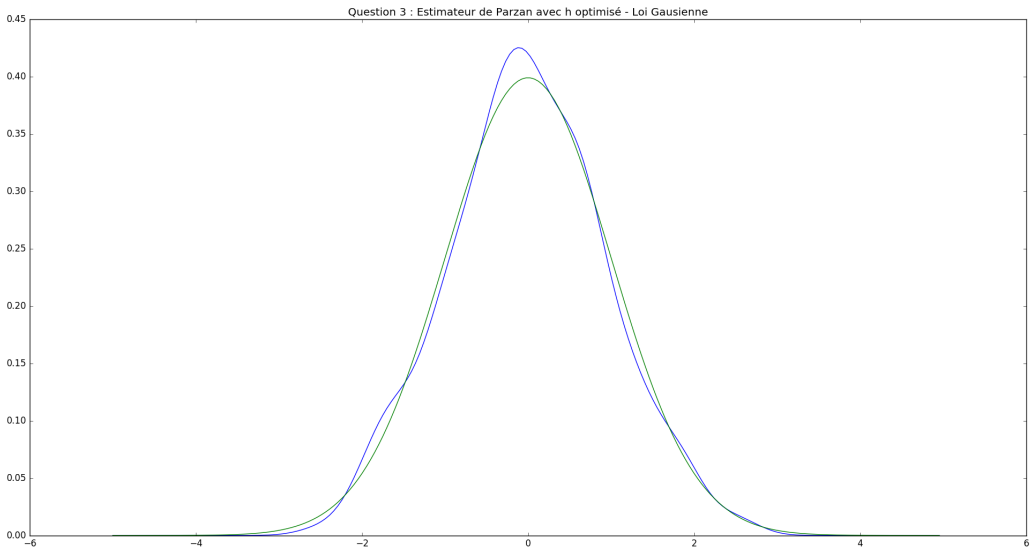


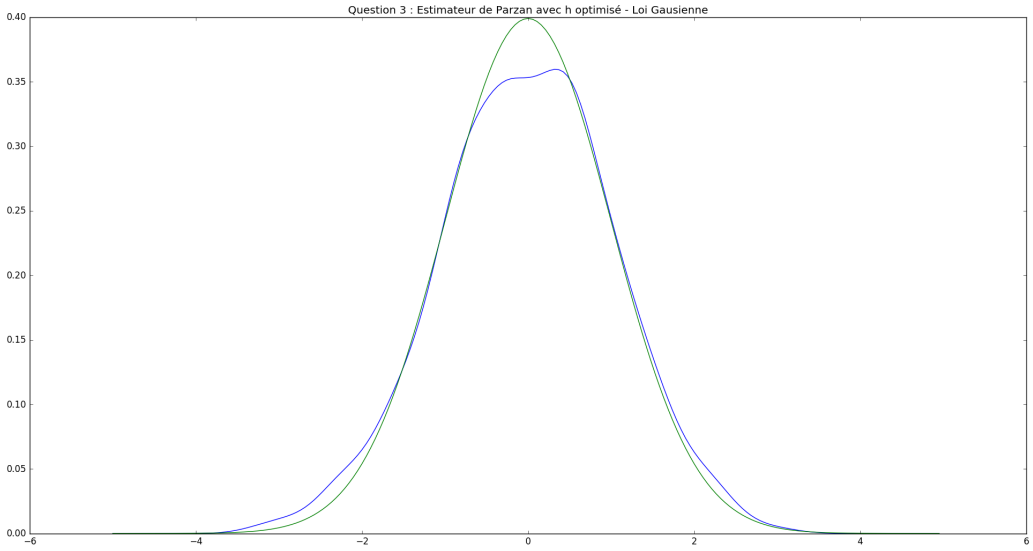


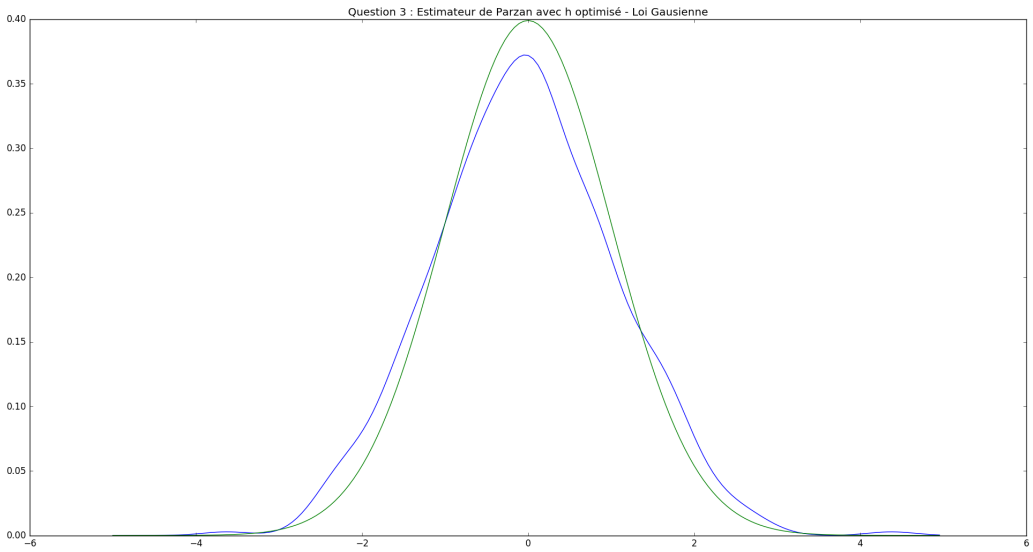


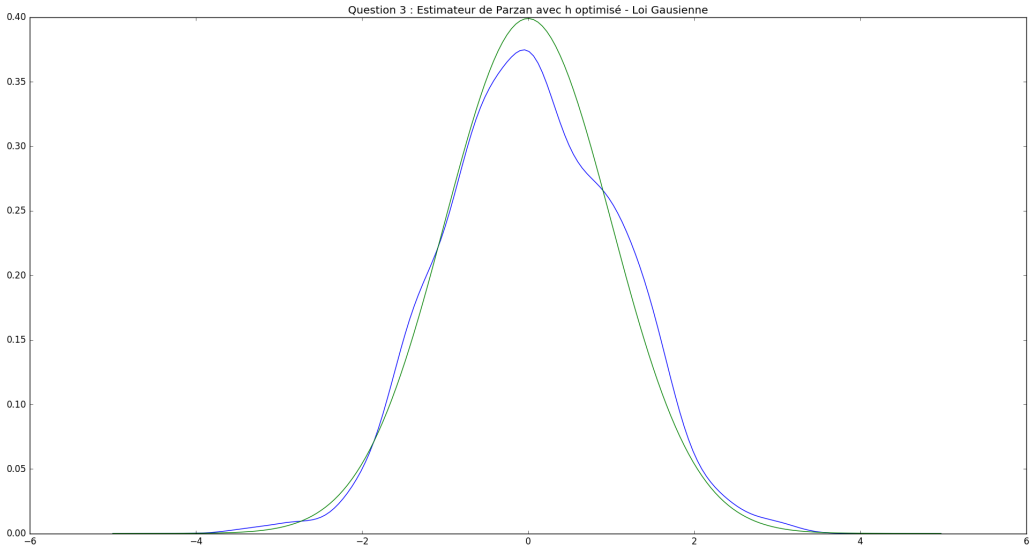
Question 3 :
On affiche 5 fois le graphe avec optimisation de h ; on constate que l'approximation est bonne (meilleure que dans 1. et 2. avec le même n), mais présente quelques variations.




Question 3 : Estimateur de Parzan avec h optimisé - Loi Gausienne
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6


Question 3 : Estimateur de Parzan avec h optimisé - Loi Gausienne
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−6
−4
−2
0
2
4
6





Question 4 :
On constate effictivement que l'estimateur converge vers f lorsque n --> l'infini

Question 4 : Comparaison d'estimateurs de Parzan avec h optimisé