INSA – GM 5 Année 2018-2019
ANTOUN Cyril
ROUFFIAC Jean-Eudes

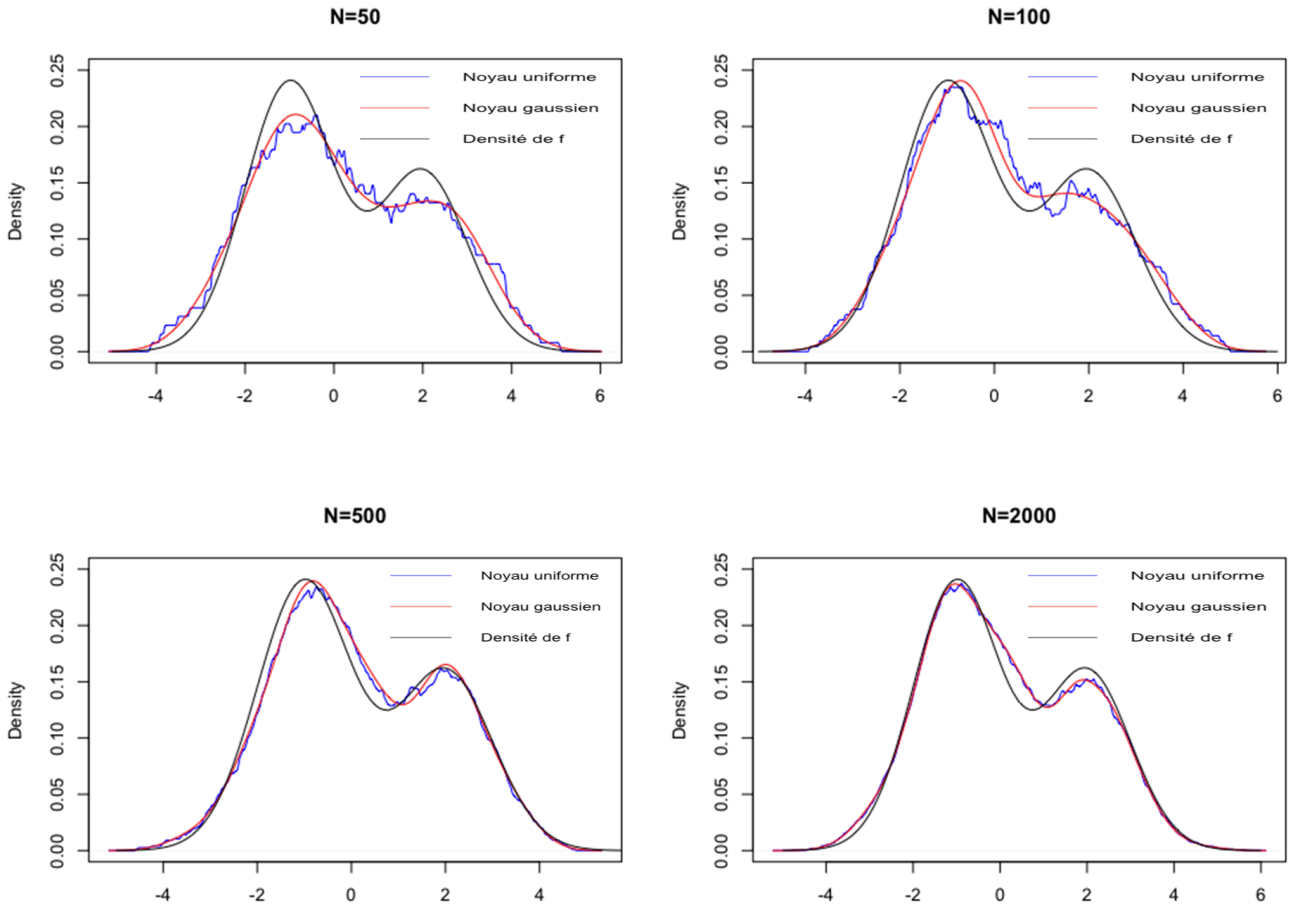
Régression Non Linéaire avec Applications sous R

Rapport de TP n°3

Titre : *Estimation de densité et Tests d'ajustement*

Partie 1

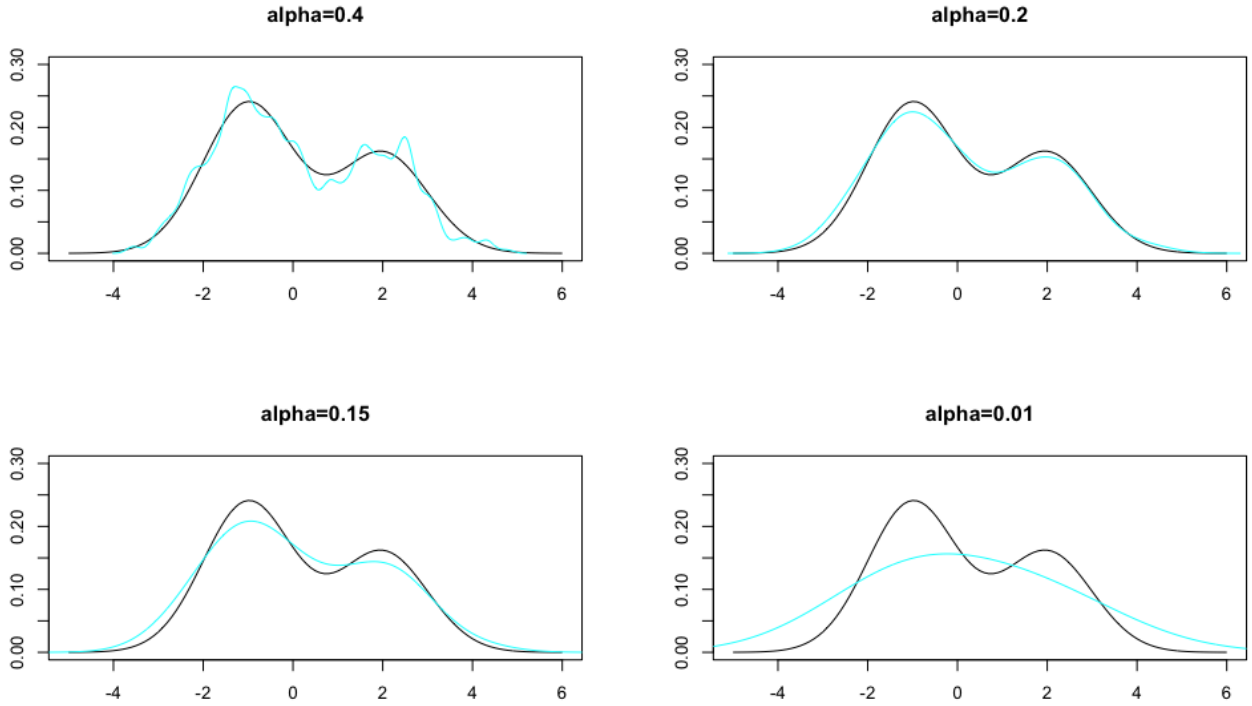
Dans cette première partie, on veut étudier par simulations l'estimateur de Parzen-Rosenblatt. On cherche d'abord à comprendre l'influence du type de noyau et du nombre de réalisations utilisés sur cet estimateur. Ainsi nous travaillons avec les noyaux rectangulaire et gaussien. Nous choisissons quatre valeurs N du nombre de réalisations : 50, 100, 500 et 2000. Les résultats sont les suivants :



Premièrement, on remarque que plus le nombre N de réalisations est élevé, plus nos deux densités estimées se rapprochent de la densité définie par la fonction f (mélange de deux densités gaussiennes $\mathcal{N}(1,1)$ et $\mathcal{N}(2,1)$). Ensuite, les graphiques montrent qu'avec N=50 et N=100 (donc pour de petits échantillons), le noyau gaussien offre une courbe d'estimation de la densité plus lisse que celle obtenue avec le noyau uniforme. Mais dès N=500, on ne voit plus la différence. Par conséquent, on déduit que le noyau gaussien et le noyau

uniforme sont appropriés pour calculer notre esimateur de la densité à partir de N=500. De plus, après plusieurs tests sour R, on peut dire que l'estimateur de Parzen-Rosenblatt donne une bonne estimation de la dénsité définie par la fonction f dès N=500 réalisations.

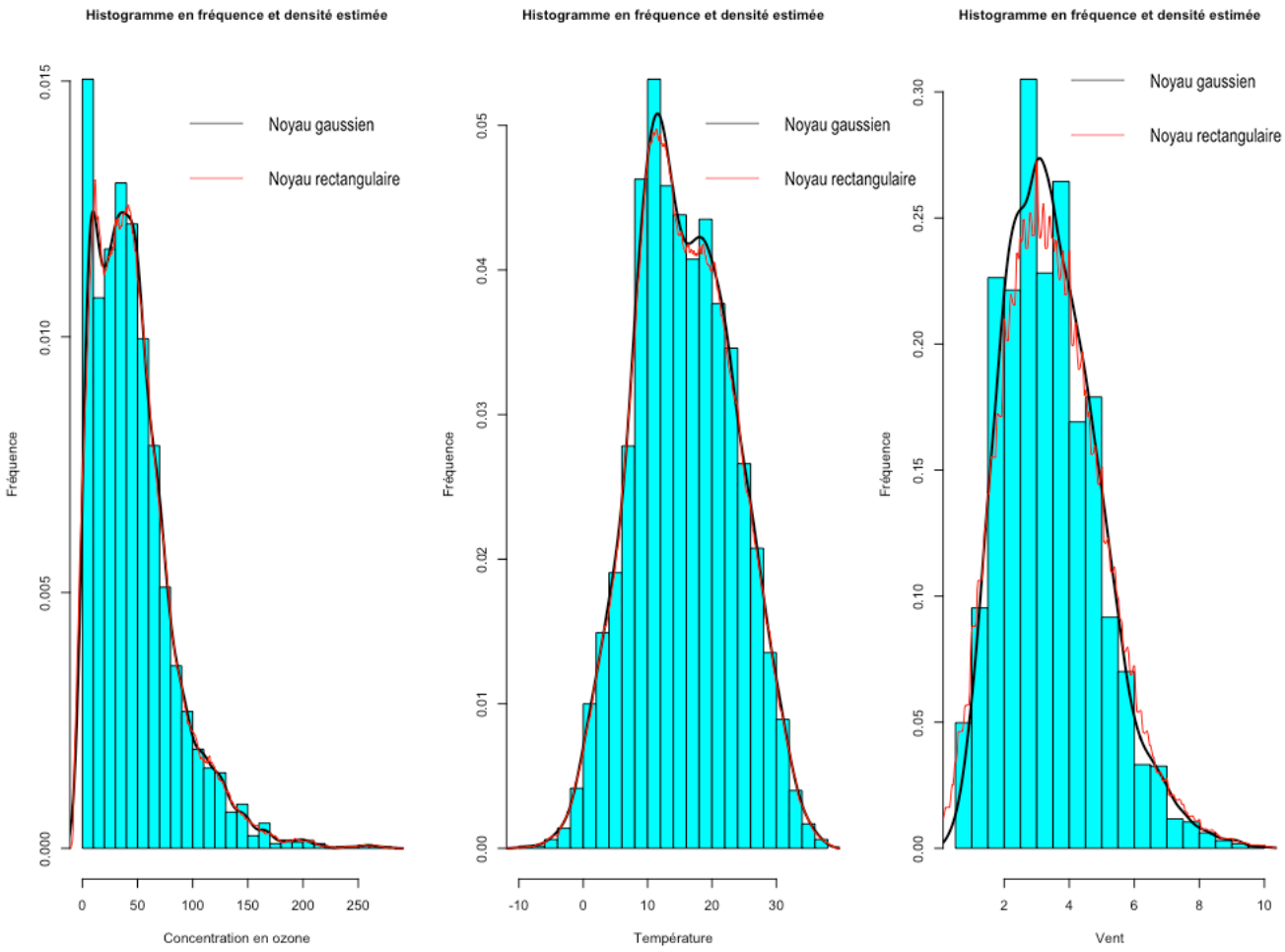
Les résultats précédents ont été obtenus grâce à une fenêtre optimale pré-calculée par R. On veut maintenant étudier l'influence de cette fenêtre sur l'estimateur de Parzen-Rosenblatt. La figure suivante représente les estimations avec un noyau gaussien et un nombre de réalisations fixé à 500. On prendra une fenêtre de la forme $s_z n^{-\alpha}$ avec s_z l'écart type de Z et nous ferons varier les α entre 0 et 1.



On remarque qu'il est nécessaire de trouver un compromis entre le biais et la variance. En effet plus α est faible plus le biais devient grand et donc l'estimation s'applatit. Inversement, lorsqu'α augmente la variance aussi et donc l'estimation n'est plus lisse. Par conséquent, pour des valeurs trop extrêmes de α l'estimateur s'éloigne de la densité définie par la fonction f. Il est donc important de trouver un compromis biais-variance. On peut conclure que la fenêtre optimale se calcule pour un α de 0.2.

Partie 2

Dans cette deuxième partie, on cherche à construire une estimation de la densité pour la concentration en ozone, la température et le vent. Pour ce faire on détient 3252 mesures relevées pour ces trois variables. Ici on se sert de l'estimateur de Parzen-Rosenblatt avec les noyaux gaussien et rectangulaire.
La fênetre choisie est la valeur optimale calculée par R. On remarque cependant que l'estimation de la densité pour la variable vent et pour le noyau rectangulaire n'est pas satisfaisante car la courbe n'est pas lisse. Comme on vient de le voir dans la partie précédente, le choix de la fenêtre est important. On change alors la fenêtre en choisissant un alpha de 0.08.
Les résultats sont présentés dans la figure ci-dessous. On superpose à l'histogramme en fréquence l'estimateur à noyau de la densité.



Nous remarquons que pour les trois variables le noyau gaussien donne une estimation

de densité plus lisse que le noyau rectangulaire. Le noyau gaussien semble donc le plus adapté pour estimer la densité des trois variables. On peut aussi voir que l'estimation de la densité pour les variables ozone et vent présente une légère asymétrie contrairement aux estimations de la variable température qui semblent d'avantage réparties de manière symétrique. De plus, certaines courbes montrent deux maxima locaux. Clairement, les variables ozone et vent ne sont pas distribuées normalement. On se demande maintenant si la variable température suit une loi gaussienne. Pour ce faire, on teste alors la normalité pour la variable température à l'aide du test de Shapiro-Wilk. On trouve une p-value de $1.388.10^{-9}$ qui est donc inférieure à 5%. On rejette alors au risque 5% l'hypothèse nulle et on conclut que la variable température n'est pas gausienne.

Partie 3

L'objet de cette partie est d'étudier le niveau empirique et la puissance empirique du test de Shapiro-Wilk et de Kolmogorov-Smirnov.

Nous simulons 200 échantillons de N valeurs d'une variable aléatoire de loi :

- Une loi normale centrée réduite
- Une loi uniforme sur [-2 ;2]
- Une loi de Student à 5 ddl
- Une loi de Student à 10 ddl

Pour chaque échantillon, nous testons au risque 5% l'hypothèse de normalité à l'aide du test de Shapiro-Wilk.

Nous considérons les N données des échantillons comme étant les réalisations indépendantes d'une variable aléatoire X. Nous testons donc l'hypothèse nulle :

H_0 :"La loi de la variable X est gaussienne"

contre l'hypothèse alternative

H_1 :"La loi de la variable X n'est pas gaussienne"

Pour obtenir le niveau empirique du test nous calculons le pourcentage de rejet à tort sous H_0, et pour obtenir la puissance empirique nous calculons le pourcentage de bonnes décisions sous H_1. Les résultats sont présentés dans le tableau suivant :

Loi	Gausienne	Uniforme	Student 5 ddl	Student 10 ddl
Hypothèse	H_0	H_1		
N=100	6.5%	99.5%	45%	19%
N=500	2.5%	100%	98%	78.5%
N = 1000	3%	100%	100%	92%

Pour des échantillons de 100 données, le niveau empirique du test est de 6.5%, ce qui est supérieur au niveau théorique de 5%. On peut cependant remarquer que lorsque N vaut 500, le niveau empirique est de 2.5%, et ce niveau reste stable pour des échantillons de 1000 données.
Dans le cas des lois de Student, le test a tendance à donner de mauvaises conclusions lorsque N est petit car le pourcentage de bonnes décisions est faible. Mais lorsque la taille des échantillons augmente, ce pourcentage augmente et se rapproche de 100%. À noter que la puissance empirique est moins bonne quand le degré de liberté est élevé.
Dans le cas d'une loi uniforme, le test nous donne des puissances empiriques de 100%.
Pour résumer, le niveau empirique du test est bien inférieur à 5% quand la taille de

l'échantillon est assez grande, et la puissance empirique s'améliore lorsque le nombre de données disponibles est grand. On peut donc conclure que lorsque nous disposons d'un nombre d'échantillons de taille importante, le test de Shapiro-Wilk est fiable.

De la même façon, nous allons maintenant utiliser le test de Kolmogorov-Smirnov sur 200 échantillons de N données dans le but de tester s'ils sont tirés de populations suivants une certaine loi. Nous travaillons de nouveau avec les quatre mêmes lois que précédemment pour pouvoir examiner ce test.
Nous testons donc l'hypothèse nulle

$$H_0 :" F = F_0"$$

contre l'hypothèse alternative

$$H_1 :" F \neq F_0"$$

où F_0 est le fonction de répartition d'une loi donnée.
Prenons d'abord F_0 la fonction de répartition d'une loi normale centrée réduite.

Loi	Gausienne	Uniforme	Student 5 ddl	Student 10 ddl
Hypothèse	H_0	H_1		
N=100	5.5%	63%	6%	5.5%
N=500	3.5%	100%	33%	5.5%
N = 1000	7.5%	100%	45.5%	11%

Le niveau empirique du test varie entre 3.5% et 7.5% et semble donc indépendant de la taille de l'échantillon.
Dans le cas d'une loi uniforme, le test est performant quand N vaut 500 ou 1000 mais ne performe pas correctement pour N égale à 100 puisqu'on relève seulement 63% de bonnes décisions.
Dans le cas des lois de Student, le pourcentage de bonnes décisions augmente avec N mais reste inférieur à 50%. Nous avons testé avec un maximum de N=1000. On peut alors se demander à partir de quelle valeur de N le test nous permettra d'atteindre plus de 95% de bonnes décisions. Pour des échantillons tirés d'une population suivant une loi de Student à 5 ddl, on atteint 97% de bonnes décisions pour N égale à 2100. Pour 10 ddl, on atteint 97% de bonnes décisions pour N égale à 7800.

On choisit maintenant F_0 la fonction de répartition d'une loi uniforme sur [-1 ;1].

Loi	Gausienne	Uniforme	Student 5 ddl	Student 10 ddl
Hypothèse	H_1			
N=100	97.5%	100%	99.5%	99%
N=500	100%	100%	100%	100%
N = 1000	100%	100%	100%	100%

On remarque que toutes les puissances empiriques calculées sont excellentes. Le test fait moins de 5% d'erreur dans tous les cas. Et il n'y aucun faux positif pour N égale à 500 et 1000 avec les quatre lois étudiées.

Pour terminer, on se donne F_0 la fonction de répartition d'une loi uniforme sur [-2 ;2].

Loi	Gausienne	Uniforme	Student 5 ddl	Student 10 ddl
Hypothèse	H_1	H_0	H_1	
N=100	56.5%	4.5%	29.5%	41.5%
N=500	100%	4.5%	99.5%	100%
N = 1000	100%	4.5%	100%	100%

On remarque que quelque soit la taille de l'échantillon, le niveau empirique du test est de 4.5%.
Pour une valeur de N égale à 500 et 1000, les puissances empiriques sont toutes égales à 100% à 0.5% près. Cependant, le pourcentage de bonnes décisions ne dépasse pas 50% quand la taille de l'échantillon est faible (N=100).

On peut tirer comme conclusion que le test de Kolmogorov-Smirnov semble moins puissant que le test de Shapiro Wilk quand on veut tester le normalité des variables. Pour savoir si les données provivennent d'une loi uniforme, le test de Kolmogorov-Smirnov donne des résultats très satisfaisants dès lors que l'on prend un nombre d'échantillon assez grand.