INSA – Génie Mathématique   Année 2018-2019

GM5 – Régression Non Linéaire avec Application sous R
Feuille TP 3
Estimation de densité et Tests d'ajustement

L'objectif de ce TP est d'étudier par simulations l'estimateur à noyau de la densité ainsi que les tests de Kolmogorov-Smirnov et Shapiro-Wilk.

Le compte-rendu ne devra pas excéder 8 pages et mettra l'accent sur les aspects statistiques.

Partie 1 Estimation de la densité

L'objet de cette première partie est d'étudier par simulations l'estimateur de Parzen-Rosenblatt. En `R`, la fonction `density` permet de construire cet estimateur à noyau de la densité. On peut spécifier le type de noyau (par exemple `kernel = 'rectangular'`), la valeur de la fenêtre (par exemple `bw = 1`), ...

1. Consulter l'aide en ligne de la fonction `density`.

```
> ? density
```

2. A l'aide des instructions suivantes :

```
n = 2000
u = runif(n)
x = rnorm(n)
Z = (u < 3/5) * (x-1) + (u> 3/5) * (x+2)
```

simuler 2000 réalisations d'une variable aléatoire Z de densité :

```
f = function(x){
    f = 3 * dnorm(x,-1,1) / 5 + 2 * dnorm(x,2,1) / 5
}
```

Cette densité est un mélange de deux densités gaussiennes $\mathcal{N}(-1,1)$ et $\mathcal{N}(2,1)$.

3. A l'aide de la fonction `seq`, créer un vecteur `t` constitué des réels compris entre -5 et 6 et équidistants de 0.1.

4. Définir une fenêtre graphique constituée de 4 figures, la première sera associée à l'estimateur de Parzen-Rozenblatt construit avec les $n = 50$ premières valeurs de `Z`, la deuxième avec les $n = 100$ premières, la troisième avec les $n = 500$ premières et la quatrième à l'ensemble des valeurs contenues dans `Z`, puis dans chaque figure

   (a) calculer les valeurs de l'estimateur de Parzen-Rozenblatt, construit avec un noyau uniforme, aux points du vecteur `t` et les enregistrer dans une liste de nom `est1`.

   (b) calculer les valeurs de l'estimateur de Parzen-Rozenblatt, construit avec un noyau gaussien, aux points du vecteur `t` et les enregistrer dans une liste de nom `est2`.

   (c) représenter sur la figure les deux estimateurs et la densité définie par la fonction `f`. On mettra les légendes adéquates.

5. Commenter les résultats obtenus.
6. Reprendre les questions 2 (on simulera seulement 500 valeurs), 3 et 4. Calculer les valeurs de l'estimateur de Parzen-Rozenblatt, construit avec les $n = 500$ valeurs de `Z`, un noyau gaussien et une fenêtre de la forme $s_Z\, n^{-\alpha}$ avec s_Z l'écart-type de Z et $\alpha \in]0, 1[$, aux points du vecteur `t`. On essaiera 8 valeurs différentes de α correctement choisies pour illustrer certains résultats théoriques du cours. Rassembler sur une même fenêtre graphique constituée de 4 figures, 4 estimateurs (1 estimateur par figure). Commenter les résultats obtenus.

Partie 2 Etude sur des données réelles

1. Récupérer les données de pollution à l'URL

   `http://lmi2.insa-rouen.fr/~bportier/Data/donpol.txt`
2. Pour chacune des variable (`ozon`, `temp` et `vent`), construire l'histogramme en fréquences et superposer l'estimation de la densité. On règlera la fenêtre de l'estimateur à noyau pour que l'allure de la densité soit satisfaisante.
3. Commenter les graphiques obtenus, que vous aurez préalablement agrémenté de légendes.
4. Le cas échéant, on pourra vérifier la normalité d'une variable à l'aide du test de Shapiro-Wilk.

Partie 3 Tests de Kolmogorov-Smirnov et Shapiro-Wilk

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

1. Simuler 200 échantillons de $n = 100$, puis $n = 500$ valeurs (on pourra aller jusqu'à 1000 si besoin) d'une variable aléatoire de loi
   (a) normale centrée réduite,
   (b) uniforme sur $[-2, 2]$,
   (c) de Student à 5 `ddl` et à 10 `ddl`,

   et pour chaque échantillon, tester au risque 5% l'hypothèse de normalité à l'aide du test de Shapiro-Wilk (fonction `shapiro.test`).

   Vous présenterez les résultats sous la forme d'un tableau indiquant lorsque vous êtes sous H_0 le niveau empirique du test et lorsque vous êtes sous H_1 la puissance empirique (le pourcentage de bonnes décisions).

   Commenter les résultats obtenus. Que peut-on dire notamment du niveau et de la puissance empirique du test sur ces exemples ?
2. Simuler 200 échantillons de $n = 100$, puis $n = 500$ valeurs d'une variable aléatoire de loi
   (a) normale centrée réduite,
   (b) uniforme sur $[-2, 2]$,
   (c) de Student à 5 `ddl` et à 10 `ddl`,

   et pour chaque échantillon, tester au risque 5%, à l'aide du test de Kolmogorov-Smirnov (fonction `ks.test`),

(a) l'hypothèse selon laquelle les données proviennent d'une loi normale centrée réduite ;

(b) l'hypothèse selon laquelle les données proviennent d'une loi uniforme sur $[-1, 1]$, sur $[-2, 2]$;

Vous présenterez les résultats sous la forme d'un tableau indiquant lorsque vous êtes sous H_0 le niveau empirique du test et lorsque vous êtes sous H_1 le pourcentage de bonnes décisions. Commenter les résultats obtenus. Que peut-on dire notamment du niveau et de la puissance empirique du test sur ces exemples ?

3. Quelles conclusions peut-on tirer sur le comportement des 2 tests ?