Analyse d'homogénéité sur le cercle unité de $\mathbb{R}^2$

Alexandre Pizzut, Virgile Galle, Louis Alain

20/01/2013

Table des matières

Introduction De nombreux sujets d'astrophysique nécessitent une approche statistique pour être correctement appréhendés. Les échantillons étudiés par le biais de ces statistiques sont très majoritairement directionnels. Il faut donc travailler en accord avec ces données, c'est à dire ici sur la sphère unité. Or, bien qu'il existe déjà des techniques qui sont utilisables dans ce cas de figure, il s'agit d'une branche des statistiques assez jeune et dans laquelle il reste encore beaucoup à faire. En particulier, l'étude de certains rayonnements serait grandement facilitée par de nouveaux outils mathématiques pensés précisément pour les résoudre.

C'est dans ce cadre-ci que s'inscrit ce projet innovant de recherche en mathématiques, dont l'objectif est de mettre au point de nouvelles méthodes capables d'effectuer des tests d'homogénéité sur le cercle. Il s'agit d'une étape préalable à la mise au point de tests similaires sur la sphère, qui seraient ainsi directement applicables à des problèmes d'astrophysique.

Chapitre 1

Motivation physique du projet

Les explications de ce paragraphe proviennent essentiellement de l'article [1]

1.1 Les liens entre astrophysique et statistiques

L'astrophysique est à priori un monde régi par les lois de Newton et la théorie de la relativité d'Einstein, ce qui nous renvoie l'image d'un domaine de la physique déterministe dont on connait parfaitement les règles. L'œil profane ne voit donc pas ce que les statistiques pourraient avoir à faire avec cette branche de le physique où les phénomènes aléatoires semblent au premier abord peu présents.

Pourtant, c'est bel et bien le cas. Certaines théories concernant des objets quantiques reçus par la Terre requièrent une analyse statistique pour pouvoir être confirmées. C'est le cas de phénomènes tels que le Fond Diffus Cosmologique (CMB en anglais) et les rayons d'ultra-hautes énergies. Il s'agit en fait d'étudier par des méthodes statistiques des évènements tels que les rayonnements évoqués ci-avant pour en déterminer les caractéristiques, ou même de pouvoir les comparer à d'autres objets de l'univers pour établir des corrélations.

1.1.1 Le fond diffus cosmologique ou rayonnement fossile

Il s'agit d'un rayonnement cosmique très faible en énergie qui aurait été émis au moment de la naissance de l'univers, et que l'on continue à recevoir aujourd'hui, venant de très loin et voyageant vers nous depuis cette époque. La distribution directionnelle de ce rayonnement est donc une donnée précieuse pour les astrophysiciens, car elle les renseigne sur l'état de l'univers tel qu'il aurait été quelques centaines de milliers d'années après le Big Bang.

Cependant, la tâche est loin d'être simple. L'une des principales difficultés est de compenser les erreurs de mesures inévitables liées à l'imprécision des appareils de mesures. Une autre vient du rayonnement de notre galaxie que nous recevons dans une partie du ciel : il est si intense qu'il rend impossible la détection du rayonnement fossile dans cette zone, et une partie des données n'est de ce fait pas accessible. Il est donc nécessaire de développer des méthodes statistiques sur la sphère unité pour traiter ces données directionnelles difficilement exploitables à l'état brut et vérifier diverses hypothèses (on pourrait penser à l'uniformité de ce rayonnement, par exemple). Les différents terrains de recherche concernés

1.1.2 Les rayons de ultra-hautes énergies

Ces rayons sont des particules chargées telles que les électrons ou les protons qui percutent l'atmosphère terrestre à des énergies allant jusqu'à 10^{20}eV, ce qui correspond pour de si petits objets à des énergies cinétiques tout à fait colossales. Leur interaction avec notre atmosphère entraine des réactions en chaine qui aboutissent à l'émission de plusieurs milliards de particules que l'on peut détecter à la surface. On peut alors déduire de la réception de ces particules secondaires la provenance et l'énergie du rayon cosmique initial. Ces données sont précieuses, car elles sont la clé pour comprendre l'origine de ce rayonnement qui reste aujourd'hui inconnue.

De nombreuses théories sur ce le phénomène responsable de ces rayons sont émises par les astrophysiciens. Il y a tout d'abord celle de l'isotropie de ce rayonnement, ou dans une moindre mesure celle d'une corrélation avec la quantité de matière présente dans l'univers dans une direction particulière. Une autre hypothèse plausible est celle de la présence de sites d'émission lointains de ces particules tels que les Noyaux Actifs de Galaxie, qui correspondent donc depuis notre point de vue terrestre à des sources discrètes. Il apparait alors que les astrophysiciens auraient besoin de connaitre la distribution de ces rayons pour affirmer ou réfuter une de ces théories, ou même envisage une combinaison des deux, ce qui n'est pas exclu.

Mais plusieurs éléments rendent difficile la production de statistiques qui permettraient d'apporter des réponses. Tout d'abord, les instruments de mesure utilisés à l'heure actuelle ont une sensibilité directionnelle qui ne permet pas de détecter tous les rayons émis. En outre, l'extrême rareté de ces phénomènes est un obstacle majeur : la probabilité d'apparition de ces rayons décroit très rapidement avec leur énergie. Ainsi seuls 69 rayons d'une énergie supérieure à 10^{20}eV ont été détectés en 20 ans, ce qui est un véritable problème en ce qui concerne la convergence des statistiques. On pourrait alors choisir de se baser sur des rayons de moindre énergie mais que l'on observerait à fréquence plus élevée pour contourner cette difficulté... Cependant, plus ces particules qui sont chargées ont une énergie basse, plus elles sont sensibles aux différents champs magnétiques cosmiques qu'elles traversent, ce qui veut dire que leurs trajectoires sont malheureusement plus susceptibles d'être déviées. On aurait alors une importante perte d'information sur la provenance réelle des rayons qui ne pourrait être contournée, et fausserait inévitablement les résultats statistiques obtenus. Il y a donc là matière à réflexion pour les statisticiens.

1.2 Conséquences sur la recherche en statistiques

Ces problèmes physiques nouveaux appellent l'élaboration de méthodes statistiques innovantes capables de les traiter de manière efficace qui sont donc adaptées aux caractéristiques des données.

1.2.1 Les principales contraintes liées aux problèmes traités

Le contexte physique impose certaines règles au jeu de statistiques à employer. En premier lieu, on pense à celles liées à l'espace dans lequel vivent les données : on a des informations directionnelles qui nous imposent donc de produire des statistiques sur la sphère unité de $\mathbb{R}^3$. De nombreuses choses possibles en dimension 1 deviennent alors plus complexes ou impossible. Par exemple, il n'existe pas de bases d'ondelettes dans cet espace, rendant moins pratique leur utilisation. Il n'existe pas non plus de relation d'ordre évidente sur la sphère, ce qui rend certains tests qui fonctionnent bien sur la droite inapplicables dans notre cas.

Ensuite, les données elles-mêmes sont altérées, de manière plus ou moins importante. Il existe bien entendu un bruit de mesure, dont il est nécessaire d'atténuer les effets dans les statistiques employées, pour converger vers les résultats qu'on aurait obtenu avec le vrai signal. Diverses techniques permettent de traiter ces artéfacts de mesure. Mais un problème plus gênant est le fait que certaines données ne sont pas observées car il nous est aujourd'hui physiquement impossible d'y accéder. Il faut alors pour compenser ces données manquantes formuler des hypothèses sur ce qu'elles donneraient, ou choisir de restreindre son étude à un domaine où aucune observation n'est masquée ou non détectée.

Enfin, le faible nombre d'observations des phénomènes étudiés est un obstacle majeur en ce qui concerne la convergence des statistiques utilisées, qui sont beaucoup plus fiables quand elles sont utilisées sur des échantillons de taille importante. Ici, typiquement, moins d'une centaine de données est un nombre très inférieur à celui dont on aimerait disposer pour s'assurer d'avoir des résultats fiables.

1.2.2 Les différentes approches statistiques suggérées

Les problèmes astrophysiques évoqués ci-avant sont tous en lien direct avec la distribution directionnelle des phénomènes observés. Cependant on peut envisager plusieurs manières d'appréhender ces distributions selon l'hypothèse qu'il nous importe de vérifier. On distingue trois grands types d'approches statistiques que l'on pourrait chercher à développer sur ces jeux de données. Le grand point commun de ces approches est le fait qu'elles sont non paramétriques, afin d'être plus générales et avoir de meilleurs résultats de convergence dans la plupart des cas.

Une première possibilité est de construire un estimateur de la distribution des phénomènes astrophysiques que l'on étudie. Cette distribution empirique pourrait alors être utilisée comme terreau pour l'émergence de nouvelles théories, ou être ensuite injectée dans des tests statistiques.

Une deuxième idée serait d'effectuer des tests statistiques d'isotropie, car cette notion est récurrente quand il s'agit d'étudier un phénomène directionnel.

Une dernière alternative consisterait à développer des tests d'homogénéité, c'est-à-dire des tests qui permettraient de décider si deux échantillons distincts ont la même distribution, sans pour autant les connaitre au préalable. On pourrait par exemple vérifier à travers ce test si les rayons de ultra-hautes énergies sont corrélés aux noyaux actifs de galaxie en comparant leurs échantillons respectifs.

1.3 Notre projet

1.3.1 Choix du sujet

Ayant l'ambition de traiter un problème inédit sur lequel peu de recherche avait été faite pour pouvoir apporter notre pierre à l'édifice, nous avons choisi de nous intéresser aux tests d'homogénéité. En effet il existe très peu de littérature sur ce sujet, et le peu qui existe concerne des tests en une dimension. L'idée de notre projet est donc de développer des tests applicables à la sphère.

Néanmoins, passer à deux dimensions est un défi qui semble hors de notre portée dans le temps qui nous est imparti. Nous avons donc convenu avec notre encadrant de nous intéresser à des tests d'homogénéité sur le cercle,

ce qui permettrait de rester à une dimension tout en travaillant sur un ensemble périodique. Les particularités induites par cette périodicité se révèleraient précieuses lors du passage à des tests sur la sphère, qui présente elle aussi cette caractéristique.

1.3.2 Objectifs du projet

Pour développer ce projet, nous nous sommes fixés différents objectifs :
- Effectuer un état de l'art à travers cette étude documentaire qui nous permettra d'entamer le projet avec les connaissances nécessaires.
- Développer nos propres méthodes sur le plan théorique
- Les comparer en les implémentant sous le logiciel statistique R
- Publier nos résultats sous la forme d'un article de recherche rédigé en anglais

Ces jalons nous permettront d'être efficace et de laisser une trace de notre travail.

Chapitre 2

Formalisation mathématiques et état de l'art

2.1 Formalisation du problème

Tests non paramétriques

On parle de tests non paramétriques lorsque l'on ne prend aucun paramètre en compte pour trouver le résultat du test. En ce sens que, pare exemple, n ne considère pas que les lois que l'on considère dépendent d'un paramètre et que le test revient à comparer ce paramètre. Les test non paramétriques sont donc des tests bien plus généralisables que les test paramétriques car la plupart des phénomènes physiques ne sont pas paramétrables si facilement que ça ou alors au prix de la perte d'information par modélisation (qui existe aussi en non paramétrique mais qui est plus faible).

Tous les tests que nous allons faire dans cette article sont non paramétriques. Cette partie du rapport se fonde sur les articles [2] et [3].

Formalisation

Dans le cadre donné par le chapitre 1, il faut donc formaliser le problème.

Soient deux échantillons $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ aléatoires de 2 variables X et Y dont on ne connait pas la loi. L'échantillon $(X_1, X_2, ..., X_n)$ correspond pratiquement aux 69 rayons de ultra-hautes énergies détectés à ce jour. L'échantillon $(Y_1, Y_2, ..., Y_m)$ va correspondre a un échantillon de données auxquelles on a accès, par exemple la position des étoiles, les nébuleuses, les "noyau de galaxies", mais aussi une autre famille de rayons, etc... Il est possible de connaitre la loi de l'échantillon $(Y_1, Y_2, ..., Y_m)$ suivant les paramètres que l'on prend.

Tests d'homogénéité

Notre problème est de savoir si les deux échantillons suivent la même loi. Mais pour ceci nous ne sommes pas obligé de connaitre cette loi. Le problème se pose donc ainsi :

$$H_0 : P_X = P_Y$$

$$H_1 : P_X \neq P_Y$$

De plus, nous avons vu que les variables sont normalement dans $\mathbb{R}^3$. Néanmoins, avant de passer de la droite à un ensemble à trois dimensions, nous allons nous pencher sur le problème du cercle. Nos variables aléatoires X et Y ont donc des réalisations dans le cercle unité. Il sera donc possible, sous la condition de définir une origine, d'identifier chaque réalisation bijectivement à un nombre dans l'intervalle $[0, 2\pi[$ correspondant à leur angle modulo 2π.

Ce type de test est appelé test d'homogénéité. Toute la suite de cet article se penchera sur ce type de test avec des variables à réalisation dans le cercle unité de $\mathbb{R}^2$.

2.2 État de l'art

Pour cet état de l'art, nous allons énumérer les résultats connus sur lesquels nos résultats se sont appuyés. Nous citerons les résultats pour des variables aléatoires de $\mathbb{R}$.

2.2.1 Test de Wilcoxon ou Test du rang

On a donc deux échantillons $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$, deux réalisations de taille n et m de deux variables aléatoires X et Y à valeur dans $\mathbb{R}$.

Construisons les vecteurs $X = (X_1, X_2, ..., X_n)$ de taille n, $Y = (Y_1, Y_2, ..., Y_m)$ de taille m et $Z = (X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m$ de taille n+m.

On rappelle la définition du rang d'un élément :

Pour le vecteur X, soit $i \in [1, n]$, on pose le rang du $i^{ème}$élément dans le vecteur X ainsi : $R_X(i) = 1+\sum_{j \neq i} 1_{\{X_j \leq X_i\}}$.

Pratiquement, cela revient à reclasser le vecteur X par ordre croissant et de regarder quelle place occupe le $i^{ème}$élément.

On voit tout de suite que se pose le problème de valeurs égales dans le vecteur. Nous considèrerons dans toute la suite de l'article que tout les Z_k sont différents (ce qui est presque certain dès que le variables X et Y sont à densité.

La statistique du test de Wilcoxon s'écrit alors :

$$W_n = \sum_{i=1}^{n} R_Z(i)$$

Que fait cette statistique ? Pour la comprendre plaçons nous dans l'hypothèse H_0.

Dans ce cas, les X_i et les Y_j sont répartis de la même manières. Ainsi on peut supposer qu'en reclassant le vecteur Z, les X_i soient répartis de manière homogène et c'est ce que vérifie la statistique W_n.

Voici les résultats que l'on a sous l'hypothèse H_0 :

$$E[W_n] = \frac{n(n+m+1)}{2}$$

$$Var(W_n) = \frac{nm(n+m+1)}{12}$$

$$\frac{W_n - E[W_n]}{\sqrt{Var(W_n)}} \underset{n\to+\infty}{\to} \mathcal{N}(0,1)$$

Ces trois résultats vont nous être très utiles dans l'explication de nos solutions.

2.2.2 Test de Kolmogorov-Smirnov

Soit deux échantillons $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$, deux réalisations de taille n et m de deux variables aléatoires X et Y à valeur dans $\mathbb{R}$.

Le test d'homogénéité s'écrit aussi :

$$H_0 : F_X = F_Y$$

$$H_1 : F_X \neq F_Y$$

où génériquement F_Z correspond à la fonction de répartition d'une variable aléatoire Z.

Principe Le test de Kolmogorov Smirnov vise à détecter toute forme de différence entre les distributions. Il repose sur l'écart maximum entre les fonctions de répartition empiriques.

Statistique du test Pour un test d'homogénéité de deux échantillons, on pose la statistique du test :

$$KS = \underset{x \in \mathbb{R}}{sup} \mid F_n^X(x) - F_n^Y(x) \mid$$

On rappelle que la fonction de répartition empirique d'un échantillon $(X_1, X_2, ..., X_n)$ est $F_n^X(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leq x\}}$

Il s'agit de mesurer l'écart vertical maximal entre les fonctions de répartition.

Le test de Kolmogorov-Smirnov est consistant pour toute hypothèse alternative c'est à dire la probabilité de rejeter H_0 tend

vers 1 lorsque $n, m \to +\infty$. Néanmoins, comme tous les tests omnibus c.P-DG. censés détecter toute forme de différence, il est peu puissant, avec un risque de deuxième espèce élevé. Il conclut un peu trop souvent à la compatibilité des données avec l'hypothèse nulle alors que l'hypothèse alternative est vraie. On dit qu'il est trop conservatif.

Région critique Pour le test d'homogénéité, nous rejetons l'hypothèse nulle lorsque l'écart maximum mesuré est anormalement élevé.

La région critique du test au risque de première espèce α s'écrit :

$$R.C. : KS \geq k_\alpha(n, m)$$

où k est donné dans la table des valeurs critiques de Kolmogorov-Smirnov.

Approximation pour les grands échantillons. Il en existe plusieurs, plus ou moins performantes. Attention, la convergence est lente, la précision est correcte lorsque n et m prennent des valeurs suffisamment élevées. Nous nous pencherons sur une approximation classique dans le cas général et dans le cas où n et m sont proportionnels.

Cette approximation est certainement la plus précise. Mais elle est assez complexe à calculer. Nous procédons en deux temps. Nous créons la statistique transformée :

$$KS' = \sqrt{\frac{n+m}{nm}} \times KS$$

Pour un test d'homogénéité la distribution de loi asymptotique de Kolmogorov-Smirnov (KS') sous l'hypothèse H_0associée s'écrit :

$$P\left(KS' \geq d\right) = 2\sum_{j=1}^{+\infty} (-1)^{j+1}\, exp\left(-2j^2d^2\right)$$

Cette formule nous permet d'obtenir directement la probabilité critique du test d'homogénéité. On voit bien que la loi de KS' ne dépend plus de la distribution de la loi des $(X_1, X_2, ..., X_n)$.

Lorsque les effectifs sont équilibrés, l'expression de la loi de Kolmogorov-Smirnov est simplifiée. Il s'écrit alors :

$$P\left(KS' \geq d\right) = \frac{\left(p!\right)^2}{(p-k)!\,(mp+k)!}$$

où k est tout entier positif tel que $d = \frac{k}{p}$.

2.2.3 Tests d'homogénéité et ondelettes

Cette partie va se baser principalement sur l'article [2]. Elle s'appuie sur des résultats sur les ondelettes et les estimateurs d'ondelettes qui sont développés dans l'Annexe A.

Principe Les tests d'homogénéité qu'il s'agit de traiter dans cette partie se fondent sur une estimation non paramétrique par ondelettes des densités des lois des échantillons indépendants $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ $\hat{f_X}$ et $\hat{f_Y}$. Explicitons l'origine de ces tests.

Une nouvelle formulation du test Dans l'article [2], tout découle d'une nouvelle formulation de l'hypothèse H_1. Si l'on appelle f la densité de $(X_1, X_2, ..., X_n)$ et g celle de$(Y_1, Y_2, ..., Y_m)$, toutes deux d'une classe de régularité R, on cherche en fait à déterminer si elles sont suffisamment éloignées, c'est-à-dire si la différence des densités appartient à un espace $S_{n,m}$, en se basant sur une distance l entre les densités.

Concrètement, on a :

$$H_1 : \ (f,g) \in S_{s,m}(C) \cap R \Leftrightarrow \{f, g \ densités, \ l(f-g) \geq Cr_{n,m}\}$$

où C est une constante positive et $r_{n,m}$permet de mesurer le taux de séparation et tend évidemment vers 0 quand n et m tendent vers l'infini.

Ici, on effectuera des tests dits *plug-in*, car on ne se sert bien évidemment pas directement des densités f et g auxquelles on n'a pas accès et que l'on ne cherche pas à déterminer mais d'estimateurs que l'on injecte dans la distance l.

Un test multiple à différentes échelles On construit des statistiques à différentes échelles j, $T_j = l(\hat{f}_{n,j} - \hat{g}_{m,j})$, qui sont calculées à partir des projetés des densités empiriques sur les espaces V_j. Le paramètre j vit dans une ensemble J tel que $j^0 \leq j \leq j^\infty$, où j^0 et j^∞sont choisis pour que le test soit optimal.

On définit alors le test à l'échelle j :

$$D_j = \Big\{ 0 \ si \ |T_j| \leq t_{n,m,j} \quad 1 \ sinon$$

où $t_{n,m,j}$est calibré par rapport à j et $r_{n,m}$.

On se sert alors du test

$$D = \underset{j^0 \leq j \leq j^\infty}{max} D_j$$

et on rejette l'hypothèse nulle si D vaut 1 et on l'accepte si D est nul. Il s'agit d'un test multiple, c'est à dire que l'on rejette H_0 dès qu'un des tests D_j la rejette. Cela revient à dire que pour accepter H_0 et ainsi décider que les échantillons ont même loi, il faut que les densités empiriques semblent être égales à chacune des échelles d'ondelette, en partant de l'échelle j^0 jusqu'à un certain seuil j^∞.

La principale difficulté de la mise en place de ce test est la détermination des paramètres r, t, et bien entendu de la distance l.

Différentes distances possibles On a parlé précédemment de distance l sans l'expliciter. C'est qu'il en existe plusieurs, qui présentent différentes caractéristiques. L'article [2] en traite quatre :

- la distance en un point x_0 :

$$l(f-g) = (f-g)(x_0)$$

la distance sur un intervalle A :

$$l(f-g) = \int_A (f-g)$$

la distance dans L^2 :

$$l(f-g) = \int (f-g)^2$$

la distance dans L^∞ :

$$l(f-g) = \underset{x}{sup}(f-g)(x)$$

Nous ne nous intéresserons qu'aux deux dernières distances, qui sont les deux plus judicieuses dans la plupart des cas, et donnerons l'allure de T_j, r et t pour chacun de ces deux tests.

Statistique du test L^2 Ici, on a

$$T_j(L^2) = \frac{1}{(n \wedge m)((n \wedge m) - 1)} \sum_{i_1=1}^{n \wedge m} \sum_{\substack{i_2=1 \\ i_1 \neq i_2}}^{n \wedge m} U_{i_1 i_2}$$

avec

$$U_{i_1 i_2} = \sum_k (\phi_{jk}(X_{i_1}) - \phi_{jk}(Y_{i_1}))(\phi_{jk}(X_{i_2}) - \phi_{jk}(Y_{i_2}))$$

Ce test est critiquable car il ne fait pas intervenir l'intégralité des données dans le cas le plus général où $n \neq m$, et c'est pourquoi nous en proposerons plus au chapitre 3 un autre dont il s'inspire.

Statistique du test L^∞ Ici, la statistique de test est définie comme suit :

$$T_j(L^\infty) = \max_k |\hat{\alpha}_{j,k}(X_1, X_2, ..., X_n) - \hat{\alpha}_{j,k}(Y_1, Y_2, ..., Y_m)|$$

Détermination de J, r et t Nous nous basons ici sur le théorème 3 de l'article [2].

Supposons que les régularités des densités f et g sont exprimées en terme d'appartenance aux espaces de Besov, i.e. on suppose que R est de la forme :

$$R(p) = \{ f \in B_{s_f p\infty},\ g \in B_{s_g p\infty}\},\ p = 2 \text{ ou } \infty$$

Alors, en posant $N = \frac{nm}{n+m}$, les résultats suivants permettent d'obtenir un risque de première espèce α :

On choisit $J = [j^0, j^\infty]$ tel que :

$$2^{j^0} = log(N),\ 2^{j^\infty} = \begin{cases} \frac{N}{log(N)} & \text{si } p = \infty \\ \frac{N^2}{log^3(N)} & \text{si } p = 2 \end{cases}$$

L'article ne nous donne alors pas directement un seuil permettant de calibrer les différents tests T_i pour un risque de première espèce α. Néanmoins, il nous permet de normaliser chacun des estimateurs de pseudo-distances utilisés dans les tests aux différentes échelles grâce à un coefficient $c_{n,m,j}$ pour pouvoir ensuite effectuer un test multiple de manière simple, en prenant le maximum des T_i'ainsi normalisés et en les comparant à un unique seuil $t'_{n,m}$. Ce seuil nous demeure inaccessible tout de même inaccessible de manière théorique, car dépendant de la nature des lois des échantillons sur lesquels sont effectués les tests. On pourra envisager de les estimer pour certains types de lois par méthode de Monte-Carlo.

Voici les facteurs de normalisation :

$$c_{n,m,j} = \begin{cases} \left(\frac{2^j log(log(N))}{N^2} \right)^{\frac{1}{2}} & \text{si } p = 2 \\ \left(\frac{j + log_2(Llog(N))}{N} \right)^{\frac{1}{2}} & \text{si } p = \infty \end{cases}$$

où L est le support des échantillons.

Chapitre 3

Les solutions proposées

Dans cette partie, nous nous replaçons dans le cadre posé dans la partie 2.1.

Introduction Dans cette partie, nous allons voir plusieurs types de test pouvant permettre de résoudre notre problème posé auparavant. Cette partie sera composée de trois grands axes. Les deux premiers seront consacrés à l'utilisation des test de Wilcoxon et de Kolmogorov-Smirnov sur le cercle. Le troisième se penchera sur l'utilisation des ondelettes qui est motivée par "l'échec des 2 premières méthodes" ou du moins leur non optimalité.

3.1 Première méthode non réalisable avec Wilcoxon et Kolmogorov-Smirnov

Le principal problème lorsque l'on pose ces test sur le cercle est la détermination d'une origine qui pourrait enlever la symétrie de notre problème ou créer des effets de bords non désirés. C'est pourquoi nous allons faire varier notre origine ce qui va nous donner une collection de tests. La première méthode consiste à tourner de point d'observation en point d'observation. Mais comme nous allons le voir cette méthode pose le problème d'avoir un nombre de tests égal au nombre d'observations de notre échantillon. C'est pourquoi nous nous pencherons dans un deuxième temps sur une méthode expérimentale de choix d'un certain nombre de tests pour chaque pair d'échantillon. La nécessité de faire tourner l'origine mais seulement un certain nombre de fois donné nous conduira à seulement faire deux tests pour ces méthodes.

3.1.1 La définition des angles par rotation de l'origine point sur les point d'observation

La définition d'une origine est cruciale pour les 2 premières méthodes que nous allons exposer. Il nous faut donc poser clairement ce phénomène assez simple.

Nous allons réaliser une collection de n+m tests du rang sur le cercle en changeant l'origine. Le changement d'origine se fait de point d'observation en point d'observation. En effet, s'il on prend deux origines entre 2 points conjoints alors les 2 tests vont donner exactement le même résultat. On va donc faire $N = n + m$ tests.

Pour chaque point du vecteur $Z = (X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m)$, nous allons définir une série d'angles.

Donc pour chaque Z_i, on définit le vecteur $(\omega_{Z_i}^{(1)}, \omega_{Z_i}^{(2)}, ..., \omega_{Z_i}^{(n+m)})$ de taille n+m où $\omega_{Z_i}^{(j)}$ correspond à l'angle du point Z_i calculé par rapport à la $j^{ème}$-origine.

On peut pour plus de commodité définir la matrice carrée Ω de taille n+m telle que $\Omega_{ij} = \omega_{Z_i}^{(j)}$.

Prenons une origine "absolue" (cette détermination est totalement arbitraire) en prenant un point quelconque sur le cercle. On calcule alors les angles de chacun des Z_i que l'on note $\omega_{Z_i}^{(0)}$. On pose alors la première origine sur l'observation ayant l'angle le plus petit ; soit le Z_k tel que $k = \underset{i \in [1,n+m]}{min}(\omega_{Z_i}^{(0)})$.

Puis vient alors :

$$\forall i \in [1, n+m], \forall j \in [1, n+m], \omega_{Z_i}^{(j)} = \omega_{Z_i}^{(0)} - \omega_{Z_j}^{(0)}[2\pi]$$

On a ainsi défini la matrice Ω.

On remarque que tout élément de Ω est compris dans $[0, 2\pi]$.

3.1.2 Avec des tests de Wilcoxon

La méthode proposée à l'aide du Wilcoxon est la suivante.

Revenons à nos échantillons $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ répartis sur le cercle unité et posons le vecteur Z tel que $Z = (X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m)$ comme dans la partie 2.2.1.

On va réaliser n+m tests ce qui va nous donner accès à n+m statistiques de test (en faisant tourner l'origine).

On définit donc à l'aide du paragraphe 2.2.1, les statistiques suivantes :

$$\forall j \in [1, n+m], \forall i \in [1, n], R_Z^{(j)}(i) = 1 + \sum_{\begin{cases} k & = 1 \\ k & \neq i \end{cases}}^{n+m} 1_{\{\Omega_{kj} > \Omega_{ij}\}}$$

ce qui définit le rang d'un élément de X dans l'échantillon Z sur le cercle d'origine Z_j

Et on pose les statistiques de Wilcoxon associées :

$$\forall j \in [1, n+m], W_n^{(j)} = \sum_{i=1}^{n} R_Z^{(j)}(i)$$

D'après les résultats donnés en partie 2.2.1, on a donc $\forall j \in [1, n+m]$:

$$E[W_n^{(j)}] = \frac{n(n+m+1)}{2}$$

$$Var(W_n^{(j)}) = \frac{nm(n+m+1)}{12}$$

$$\frac{W_n^{(j)} - E[W_n^{(j)}]}{\sqrt{Var(W_n^{(j)})}} \underset{n\to+\infty}{\to} \mathcal{N}(0,1)$$

Mais il faut faire attention, on voit naturellement que les $W_n^{(j)}$ sont dépendants, ce qui va poser problème.

En effet, deux solutions sont alors possibles :

Soit de faire un test multiple, c'est à dire de prendre toutes ces statistiques, de voir si l'une d'entre elles est rejetée à un niveau α donné pour rejeter H_0. Mais cette méthode a tendance à beaucoup rejeter car dès qu'un cas n'est pas favorable on rejette tout le test ce qui est très contraignant.

Soit de faire un test "aggloméré", c'est à dire de créer une statistique de test faisant intervenir toutes les $W_n^{(j)}$ et dont on peut connaitre la loi ou la loi asymptotique pour en déduire un test.

Dans les deux cas, il y a un très gros problème. Nous avons un nombre de tests égal au nombre d'observations ne permettant pas de conclure sur la convergence de tous les tests en même temps. C'est pourquoi nous abandonnons ce chemin pour passer à une approche plus expérimentale.

3.1.3 Avec des tests de Kolmogorov-Smirnov

De la même manière, on va poser une collection de statistiques de tests en faisant varier l'origine. Seulement cette fois les statistiques seront basées sur les tests de Kolmogorov-Smirnov. Autre différence, nous réaliserons un test multiple sur ces statistiques.

En réalité, ayant une origine et des angles nous allons nous ramener au cas dans $\mathbb{R}$.

On a donc deux collections d'échantillons $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ répartis sur le cercle unité et posons le vecteur Z tel que $Z = (X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m)$ comme dans la partie 2.2.1. On pose aussi la matrice Ω.

Nous allons poser les fonctions de répartitions empiriques ayant pour origine la $j^{ème}$ origine où $j \in [1, n+m]$:

$$\forall \omega \in \mathbb{R}, F_X^{(j)}(\omega) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\Omega_{ij} \leq \omega\}}$$

et

$$\forall \omega \in \mathbb{R}, F_Y^{(j)}(\omega) = \frac{1}{m} \sum_{i=n+1}^{n+m} 1_{\{\Omega_{ij} \leq \omega\}}$$

On pose alors la statistique de Kolmogorov-Smirnov pour le cercle unité :

$$KS^{(j)} = \underset{\omega \in \mathbb{R}}{sup}(\mid F_X^{(j)}(\omega) - F_Y^{(j)}(\omega) \mid) = \underset{\omega \in [0,2\pi]}{sup}(\mid F_X^{(j)}(\omega) - F_Y^{(j)}(\omega) \mid)$$

Mais ici aussi le problème du nombre de statistiques de tests se pose, nous laissant alors la possibilité de faire comme en deuxième partie en gardant le même formalisme que celle-ci.

3.2 La deuxième méthode expérimentale avec Wilcoxon et Kolmogorov-Smirnov

Comme nous l'avons vu dans le paragraphe précédent, le nombre de tests effectués était égal à la taille de l'échantillon. Cela causait un grand nombre de problèmes pour réaliser les tests combinant toutes les statistiques intermédiaires de test. Nous allons donc procéder de la même manière mais en choisissant empiriquement pour chaque test le nombre optimal K de tests à réaliser. Mais après réflexion, il semble qu'il suffit de deux tests espacés de $\frac{\pi}{2}$ pour avoir un test n'ayant pas de problème dans le choix de l'origine.

3.2.1 Avec des tests de Wilcoxon

La méthode est la même que dans le paragraphe précédent sauf que l'on ne fait que deux tests. L'un en prenant une origine arbitraire n'importe où sur le cercle puis un faisant tourner de $\frac{\pi}{2}$.

La seule différence c'est que nous faisons un test multiple. Et dans ce cas, il nous faudra bien calibrer un niveau que l'on appellera élémentaire pour chaque test pour avoir un niveau global de α.

Pour montrer l'efficacité de ce test. Nous allons faire des simulations numériques avec des lois plus ou moins simples et plus ou moins critiques dans la partie simulation.

3.2.2 Avec un test de Kolmogorov-Smirnov

Sans aucun changement d'origine.

3.2.3 Avec des tests combinant les deux méthodes

Pour éviter de faire un changement d'origine, on peut aussi penser à faire un test multiple de deux tests en une unique origine prise arbitrairement sur le cercle, un de Wilcoxon et un des Kolmogorov-Smirnov. Cette méthode va avoir l'avantage de mesurer non seulement l'écart de médiane mais aussi l'écart des fonctions de répartitions. Cette méthode nous permet de nous ramener tout simplement à un problème d'homogénéité sur la droite entre $[0, 2\pi]$. Nous ferons aussi des simulations pour comparer l'efficacité de cette méthode par rapport aux deux précédentes.

3.3 Un test basé sur les ondelettes

On constate que les tests évoqués précédemment ont le défaut d'être incroyablement dépendants, si l'on n'en effectue qu'un seul, du choix de l'origine à partir de laquelle ils sont exécutés. Il est alors nécessaire de développer un artifice si l'on veut gommer les artéfacts liés à ce problème, tels que ceux détaillés ci-dessus. C'est ce qui motive l'élaboration d'un test à partir des ondelettes. Ces dernières présentent une certaine périodicité, et l'on est en droit de penser qu'elles sont plus adaptées à l'implémentation d'un test sur des données invariantes par rotation de 2π, à condition d'aller suffisamment loin dans les échelles.

3.3.1 Définition des ondelettes sur le cercle

Avant d'aller plus avant dans l'utilisation d'une méthode sur le cercle, il convient de définir les ondelettes dont on se servira sur le cercle.

Il y a tout d'abord l'ondelette mère, constante :

$$\phi_0(x) = \frac{1}{\sqrt{2\pi}}$$

Ondelette de Haar à la position k et à l'échelle j :

$$\forall k \in [0, 2^j - 1],\ \psi_{j,k}(x) = \frac{\sqrt{2}^{j-1}}{\sqrt{\pi}}\Big(1_{[2^{1-j}\pi k, 2^{-j}\pi(2k+1)[}(x) - 1_{[2^{-j}\pi(2k+1), 2^{1-j}\pi(k+1)[}(x)\Big)$$

On peut également définir ces ondelettes comme suit, à partir d'une ondelette ψ^0 comme suit :

$$\forall (j,k)\, o\grave{u}\, k \in [0, 2^j - 1],\ \psi_{j,k}(x) = 2^{\frac{j}{2}}\psi^0(2^j x - k),\ o\grave{u}\, \psi^0(x) = \frac{1}{\sqrt{2\pi}}(1_{[0,\pi[} - 1_{[\pi,2\pi[})(x)$$

3.3.2 Tests choisis

On effectuera le test présenté dans la version 2.2.3 qui concerne la pseudo-distance L^2. Néanmoins, peu convaincus de la pertinence de ne pas utiliser une partie de l'échantillon de plus grande taille pour se ramener à deux échantillons de même taille, nous effectuerons également un test adapté à des échantillons de taille différentes. Ce deuxième test est développé en annexe B.

3.3.3 Un grand défaut de la décomposition en ondelettes

Bien que les ondelettes présentent des avantages certains qui nous amènent à les utiliser, il n'est pas certain que les tests dont elles sont à la base soient insensibles au choix de l'origine si l'on ne va pas assez loin dans les échelles. Certaines lois, comme les lois uniformes ne poseront pas de problème pour le test simple, mais on peut imaginer que deux gaussiennes à faibles variance et de moyenne très proches ne soient pas aisément détectées par le test classique.

3.3.4 Un nouveau test ondelettes

Un bon moyen de gommer l'effet des possibles artéfacts liés à ce problème est de fonder les tests sur des distances moyennées. En faisant varier l'origine de l'échantillon modulo $2\pi(k-1)$ fois de $\frac{2^{1-j}\pi}{k}$ avec j l'échelle maximale à laquelle on va aller à partir de la première origine, et en utilisant la moyenne arithmétique de ces distances, on construit un test moins sensible aux défauts évoqués ci dessus.

Chapitre 4

Simulation

Introduction Cette partie va nous permettre de comparer les puissances des tests que nous avons proposés dans la partie précédente grâce à des courbes ROC . Pour chaque test, nous donnerons le code utilisé pour le logiciel R, ainsi que les choix que nous avons faits pour réaliser ces tests.

Nous allons tester l'homogénéité de deux échantillons pris suivant les lois suivantes qui varient suivant un paramètre ε.

Les $(X_1, X_2, ..., X_n)$ sont tels que $X \sim \mathcal{U}([0, 2\pi])$ et $(Y_1, Y_2, ..., Y_m)$ échantillon de Y sont tels que on a :

$$Y \sim (1 - \varepsilon)U_1 + \varepsilon U_2$$

où $U_1 \sim \mathcal{U}([0, 2\pi])$ et $U_2 \sim \mathcal{U}([\frac{5\pi}{6}, \frac{7\pi}{6}])$

Il s'agit en fait de perturber la loi avec laquelle on souhaite tester l'homogénéité par un paramètre ε que l'on fait varier. Bien entendu, plus epsilon est grand plus la distinction est simple.

Mais il nous faut préciser la notation $Y \sim (1 - \varepsilon)U_1 + \varepsilon U_2$. Elle signifie en fait, qu'il faut tirer une variable aléatoire A binomiale de paramètre m et ε : $A \sim \mathcal{B}(m, \varepsilon)$

On fait ensuite un tirage de taille m-A pour la loi U_1 et de taille A pour la loiU_2, que l'on remet aléatoirement dans un vecteur de taille m.

Dans tout nos tests on fera varier epsilon ainsi

$$\varepsilon = 0.1, 0.2, 0.3, 0.4, 0.5$$

4.1 La réalisation de courbes ROC

La fonction pour tracer une courbe ROC dépend du test que l'on va réaliser. Il dépend donc des variables aléatoires que le test fournit "en sortie". Voilà les méthodes que nous avons dues utiliser pour les tests suivants.

4.1.1 Pour un test simple donnant la réalisation en vecteur d'une variable aléatoire réelle positive

On la nomme ROCpos. Elle prend en argument deux vecteurs de même taille simulant des variables aléatoires par Monte Carlo. L'un sous l'hypothèse nulle, l'autre sous H_1 donnée. Elle demande aussi un vecteur NIV correspondant aux risques de première espèce qui seront en abscisse de la courbe ROC. L'inconvénient est que R n'aime pas avoir de return (plot) qui permettrait de retourner un graphe directement. De ce fait, la fonction retourne un vecteur puissance correspondant au vecteur niveau donné en argument. Il faut donc rajouter un plot après la fonction pour avoir le graphe ROC. Attention : La fonction ne prend en argument que des vecteurs triés préalablement et positifs.

4.1.2 Pour un test simple donnant la réalisation en vecteur d'une variable aléatoire réelle symétrique par rapport à 0.

Nous l'avons nommée ROC. Le principe de cette fonction est de se ramener au cas de ROCpos en passant simplement à la valeur absolue. Cette fonction nous sera utile pour le cas d'un test unique de Wilcoxon (quand ce test est bien sur centré).

4.1.3 Pour un test double donnant deux réalisations en vecteur de deux variables aléatoires réelles venant d'un test de Wilcoxon

Elle s'inspire aussi de la première avec la différence que cette fois, la fonction demande deux vecteurs issus des hypothèses H_0 et H_1 pour les deux tests.

On choisit de prendre la norme $\| \, . \, \|_\infty$ pour se doter d'une distance dans $\mathbb{R}^2$. Ainsi en prenant le maximum des deux échantillons sous H_0 et le maximum des deux sous H_1, il suffit de ré-appliquer la fonction ROC. Cette fonction se nomme ROCmult.

4.1.4 Pour un test double donnant deux réalisations en vecteur de deux variables aléatoires réelles venant de deux tests différents

Cette fonction sera utilisée dans le cas où l'on va combiner les tests de Wilcoxon et Kolmogorov-Smirnov. Cette fonction est bien plus compliquée que la précédente. En effet, comme les deux tests ne sont pas calibrés de la même manière, on ne peut pas prendre le maximum des deux aussi simplement.

Ce que nous allons donc faire est la chose suivante.

La fonction prend en argument :
- Deux variables non positives symétriques par rapport à 0 issues d'un premier test sous les hypothèses H_0 et H_1, on les nomme X_0 et X_1 , que l'on va prendre en valeur absolue.
- Deux variables positives issues d'un deuxième test sous les hypothèses H_0 et H_1, on les nomme Y_0 et Y_1.
- Une suite de niveau élémentaire allant de 0 à 1 notée α.

Nous allons faire une boucle sur les éléments de alpha.

Nous allons nous donner un niveau élémentaire $\frac{\alpha}{2}$ pour chacun des tests ce qui va nous donner accès à des quantiles q_X en utilisant X_0 et q_Y avec Y_0 . En calculant le nombre de points tels que $\mid X_0 \mid > q_X$ ou $\mid Y_0 \mid > q_Y$, ce nombre divisé par la longueur d'une des variables X_0 ou Y_0 (ce sont les mêmes) nous donne accès à la véritable valeur du niveau du test multiple qui est nécessairement inférieur à α car

$$P((\mid X_0 \mid > q_X) \cap (\mid Y_0 \mid > q_Y)) \leq P(\mid X_0 \mid > q_X) + P(\mid Y_0 \mid > q_Y) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$$

Remarque : il est plus facile de trouver le nombre de points tels que $\mid X_0 \mid < q_X$ et $\mid Y_0 \mid < q_Y$ ce qu'on a utilisé dans le code

Puis on calcule la puissance en trouvant le nombre de points tels que $\mid X_1 \mid > q_X$ ou $\mid Y_1 \mid > q_Y$ divisé par la longueur d'une des variables X_1 ou Y_1 .

Ainsi on obtient un vecteur contenant les niveaux ainsi que les puissances du test multiple. Cette fonction est appelée ROCmult2.

4.2 Les méthodes avec Wilcoxon et Kolmogorov-Smirnov

Tout d'abord, il est important de se munir des deux tests sur le cercle. Le test de Kolmogorov-Smirnov comme on l'a mentionné auparavant est déjà contenu dans la bibliothèque de R sous le nom de ks.test.

Le test de Wilcoxon l'est aussi mais nous l'avons ré implémenté sous une forme la plus vectorielle possible.

4.2.1 Méthode 1 : Un test simple de Wilcoxon

On voit que ce test n'est vraiment pas puissant tout seul pour détecter des inhomogénéités même importantes. En fait, ceci est du au fait que dans notre simulation, la perturbation c'est à dire la loi uniforme sur $[\frac{5\pi}{6}, \frac{7\pi}{6}]$ est centrée en π comme la loi de X. Or le test de Wilcoxon détecte les écarts entre les médianes ce qui justifie notre motivation de changer d'origine.

4.2.2 Méthode 2 : Avec un test multiple de Wilcoxon avec variation de l'origine

On remarque bien que quand ε diminue la puissance du test diminue amplement car il est plus dur de discerner les X des Y. Ces courbes toutes seules non sont pas très utiles et doivent être comparées à d'autres courbes ROC.

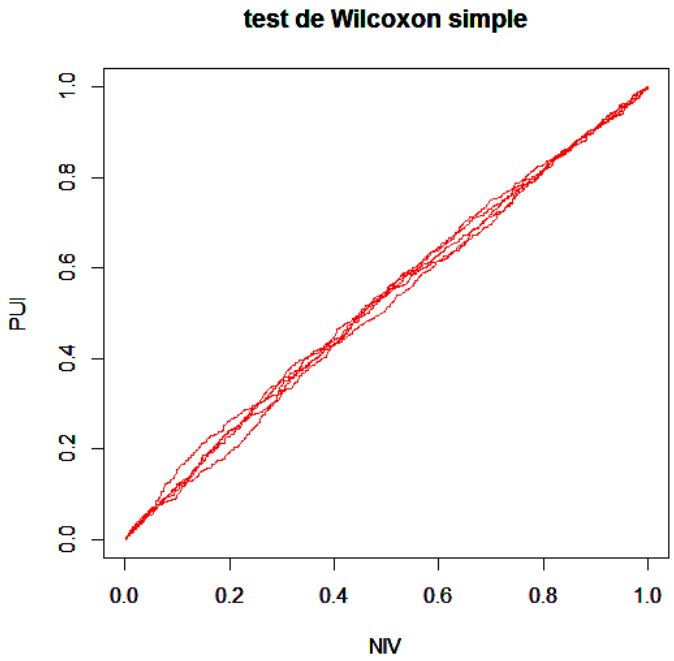

FIGURE 4.2.1 –

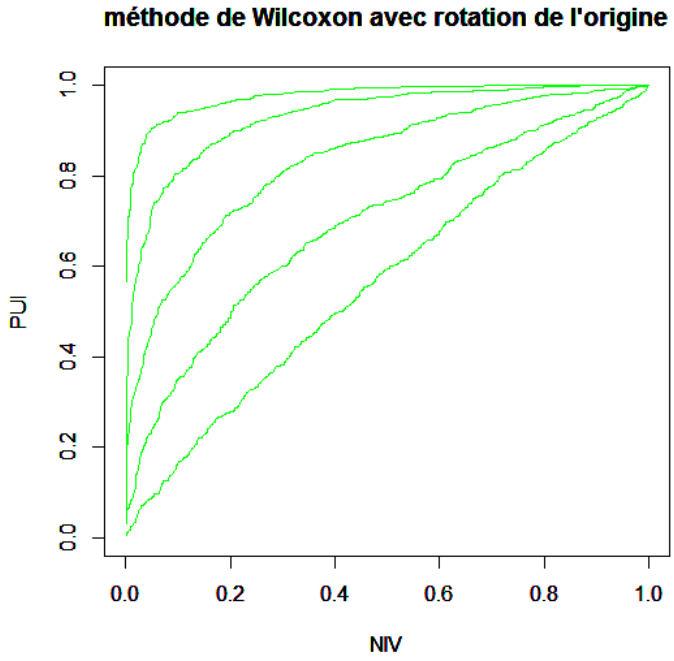

FIGURE 4.2.2 –

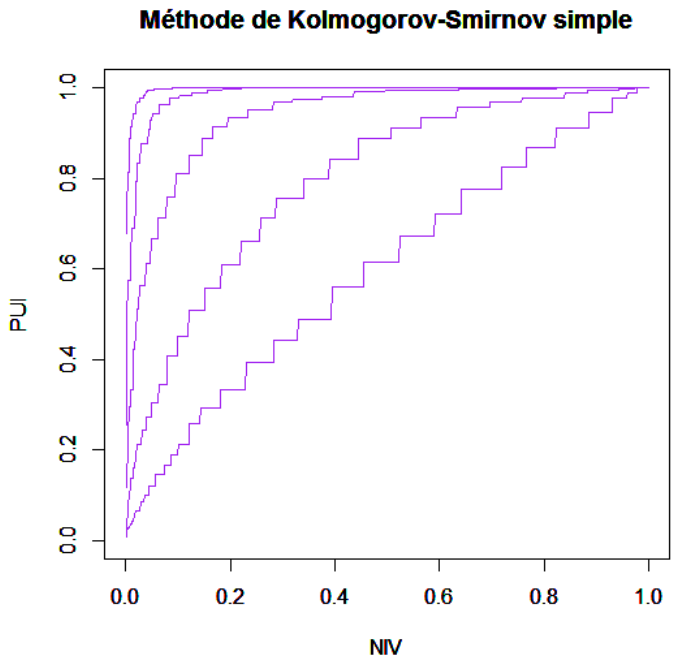

FIGURE 4.2.3 –

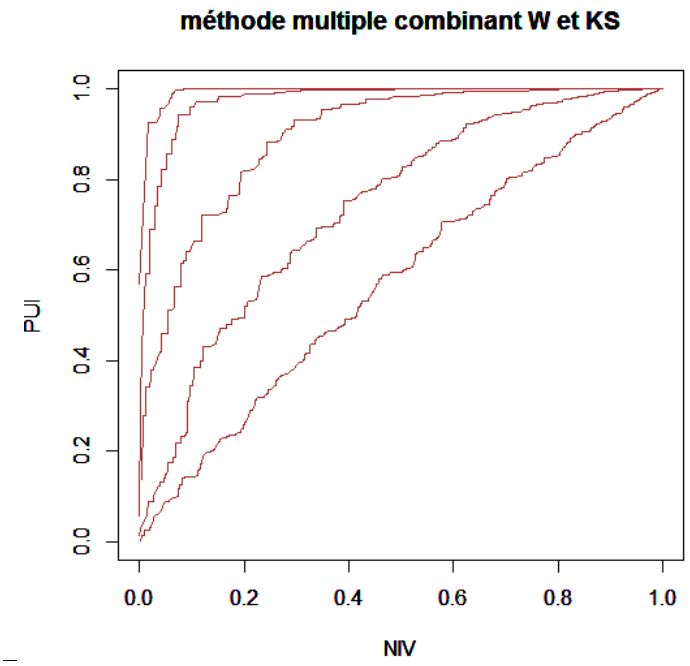

FIGURE 4.2.4 –

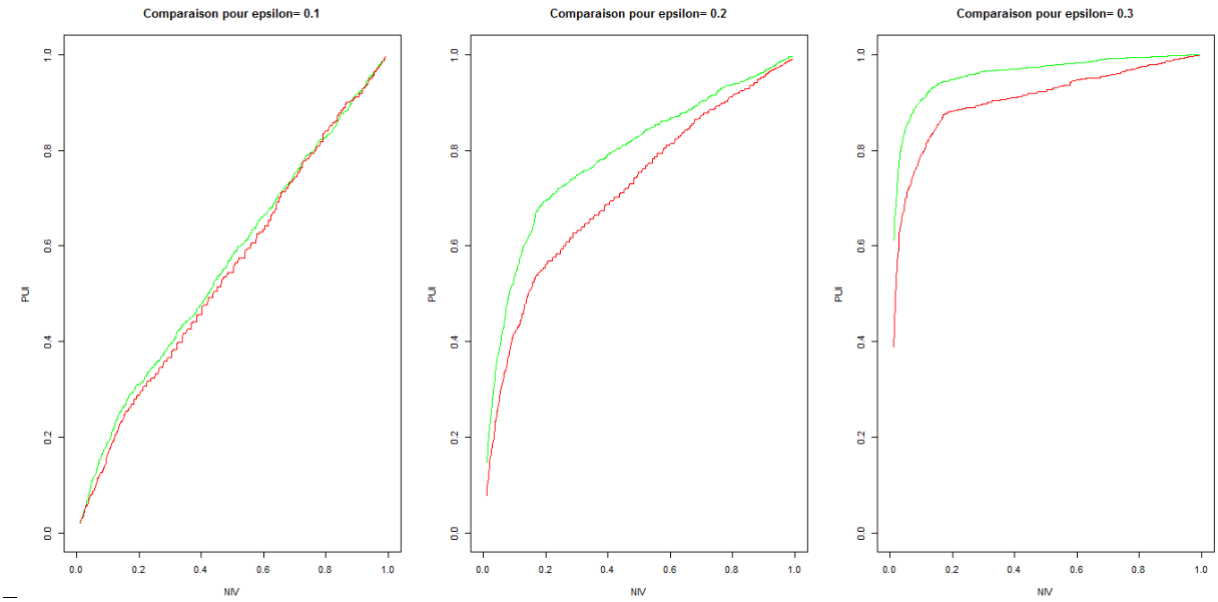

FIGURE 4.3.1 –

4.2.3 Méthode 3 : Avec un test simple de Kolmogorov-Smirnov

4.2.4 Méthode 4 : Avec des tests combinant Wilcoxon et Kolmogorov-Smirnov sans changement d'origine

4.3 Méthode 5 :Un test basé sur les ondelettes

L'enjeu ici est de développer un test qui se fonde sur une décomposition en ondelettes de Haar. Cependant on a pour ce faire plusieurs possibilités : préfèrera-t-on un test qui ne travaille qu'avec des échantillons de même taille, quitte à délaisser une partie du plus grand, ou préfèrera-t-on se servir de l'intégralité des données ? Vaut-il mieux effectuer un test avec une origine arbitraire ou se servir d'une distance moyennée calculée en faisant varier l'origine ? Nous allons pour répondre à ces questions et se servir d'un test ondelettes optimal comparer des courbes ROC empiriques de chacun de ces tests.

4.3.1 Faut-il garder ou non l'intégralité des échantillons ?

Nous avons d'abord fait un test qui n'utilisait qu'une partie du plus grand échantillon (courbe rouge) et un autre qui se sert de toutes les données (courbe verte) . Ces tests ont été faits pour des échantillons de tailles comparables avec $n = 100$ et $m = 200$, puis avec des tailles très différentes ($n = 50$, $m = 400$)

On constate alors que le test se servant de toutes les données est le plus puissant. C'est donc celui-là que l'on va retenir.

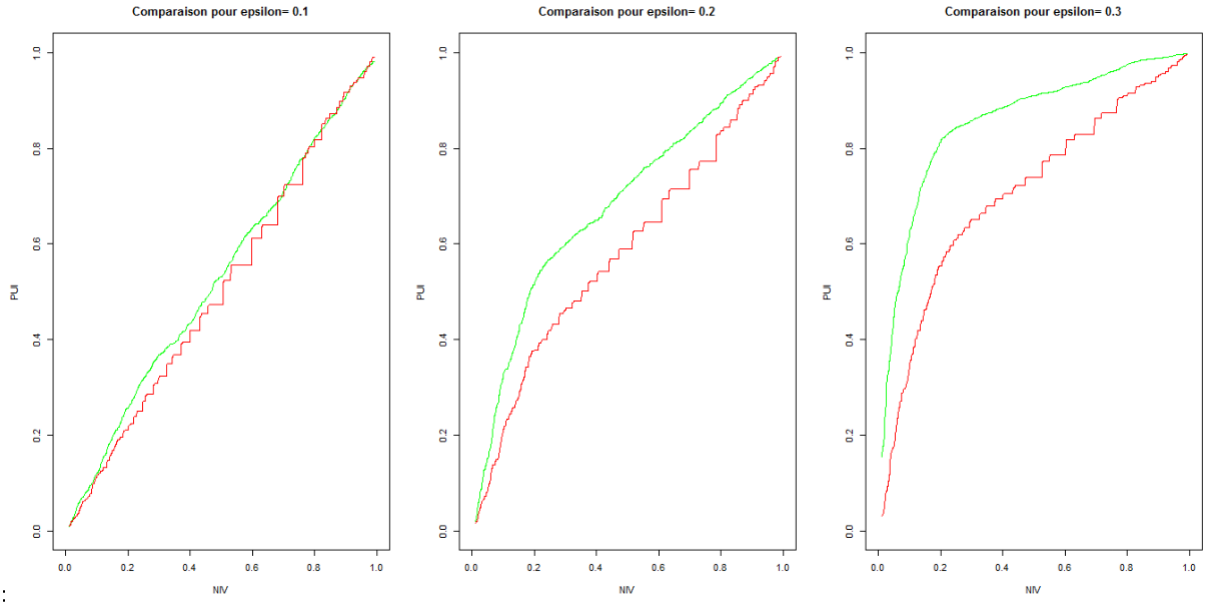

Figure 4.3.2 :

4.3.2 Un test simple ou moyenné ?

Ici, on se sert du test retenu précédemment de deux manières : d'abord de manière simple (courbe verte), puis en le faisant travailler avec des estimateurs moyennés calculés en ayant fait tourner l'origine comme expliqué en 3.3.4 (courbe rouge). Nous l'avons effectué avec des tailles d'échantillon valant$n = 100$ et $m = 200$ dans un premier temps, puis avec des tailles très différentes ($n = 100$, $m = 800$)

Il apparait alors que le test le plus puissant est le test simple, et qu'il n'est pas utile de faire varier l'origine dans le cadre des ondelettes pou ces lois. Le test ondelette utilisé par la suite sera donc le test simple faisant intervenir l'intégralité des deux échantillons.

Voici la méthode donc retenue pour les epsilons donnés en introduction.

4.4 Comparaison des méthodes

Nous allons comparer les méthodes pour les différents epsilons.

Pour ceci nous avons sauvé dans chaque simulation pour ne pas avoir à les refaire deux vecteurs. Les vecteurs COMPNIVMi et COMPPUIMi ou i est la $i^{ème}$méthode utilisée.

Chaque méthode a une couleur :
– La méthode 1 : Wilcoxon seul en rouge
– La méthode 2 : Wilcoxon double avec variation de l'origine en vert
– La méthode 3 : Kolmogorov-Smirnov seul en violet
– La méthode 4 : Wilcoxon et Kolmogorov-Smirnov combiné sans changement d'origine en marron
– La méthode 5 : Ondelettes en bleu
Voici les différentes courbes pour les ε :

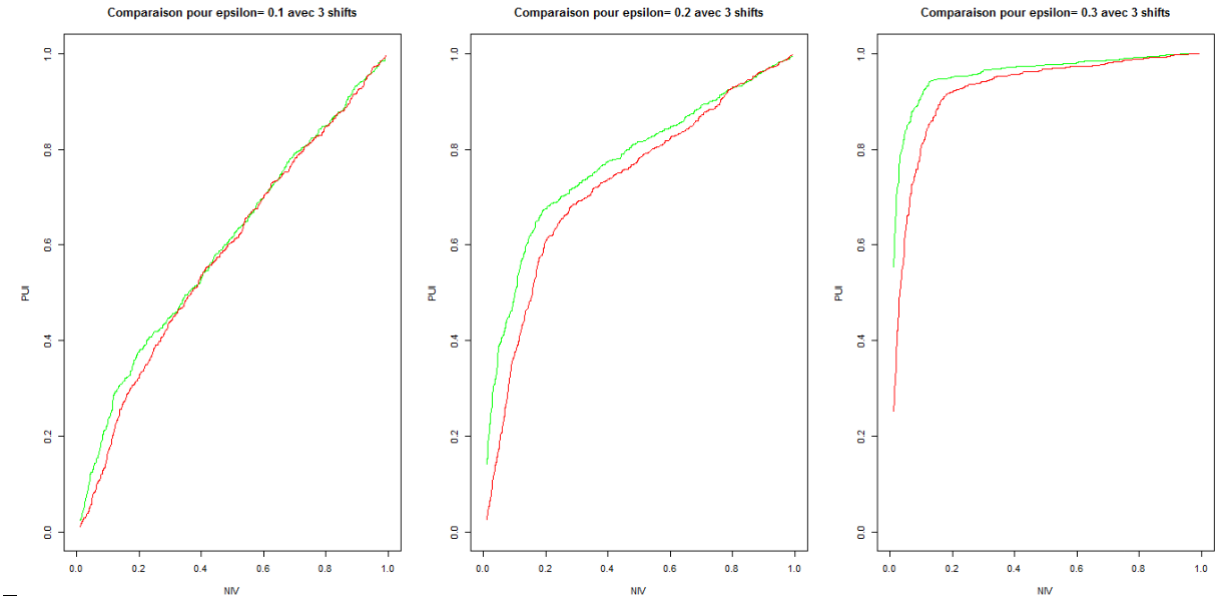

FIGURE 4.3.2 –

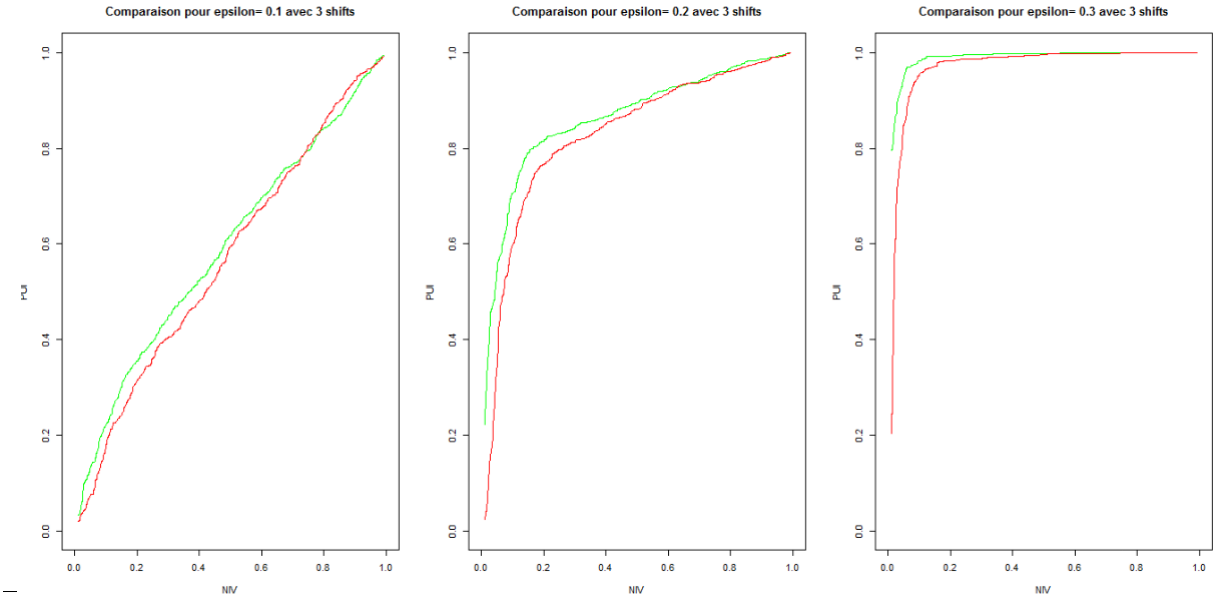

FIGURE 4.3.3 –

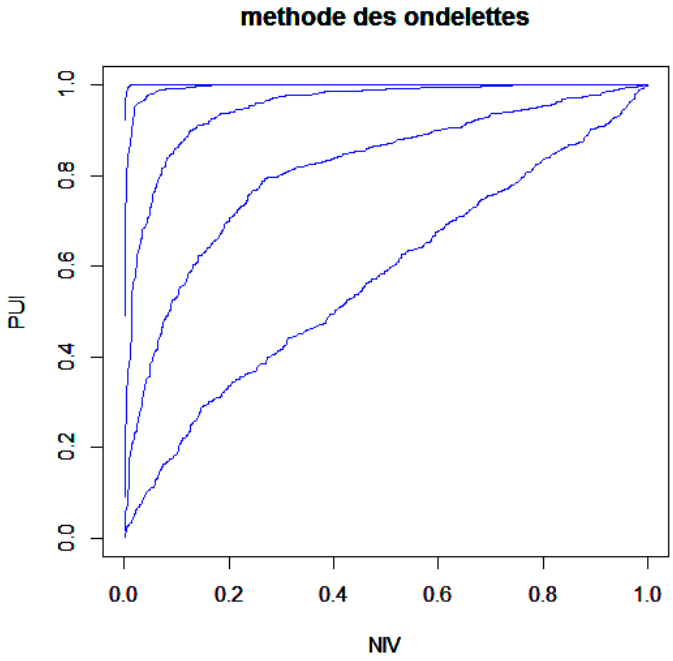

FIGURE 4.3.4 –

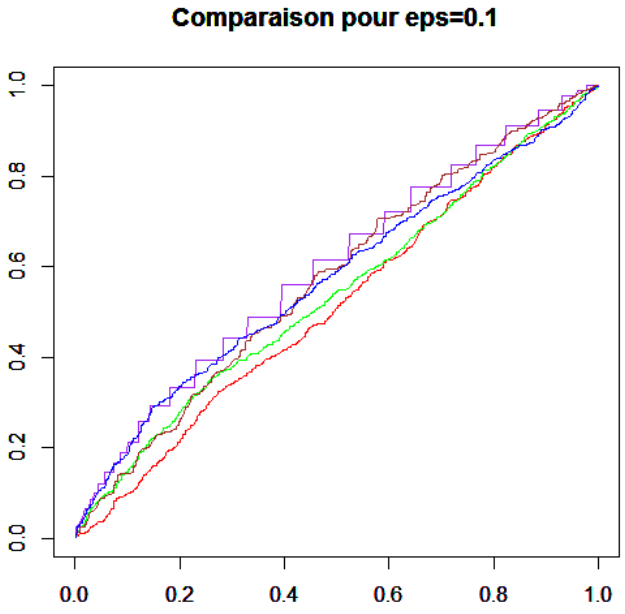

FIGURE 4.4.1 –

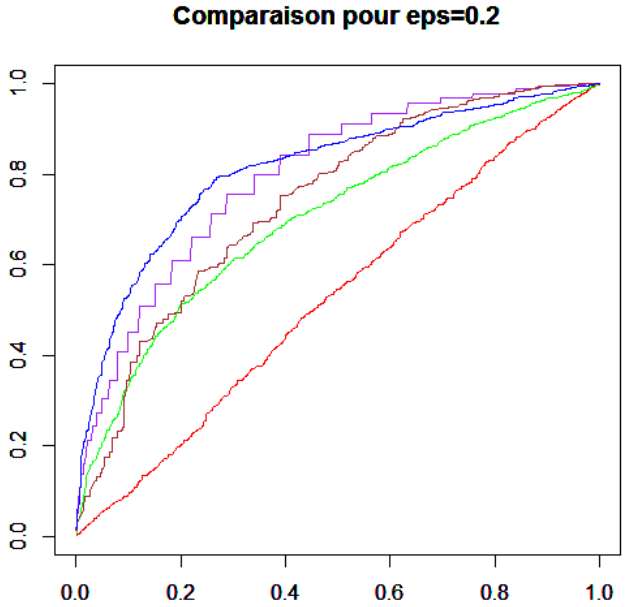

FIGURE 4.4.2 –

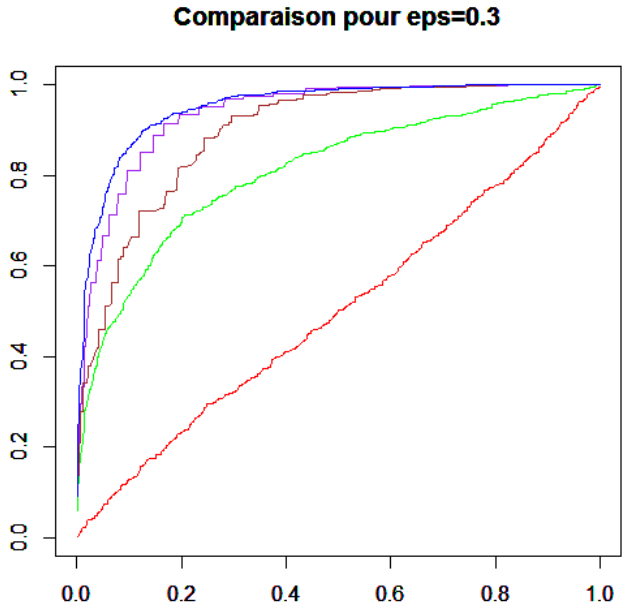

FIGURE 4.4.3 –

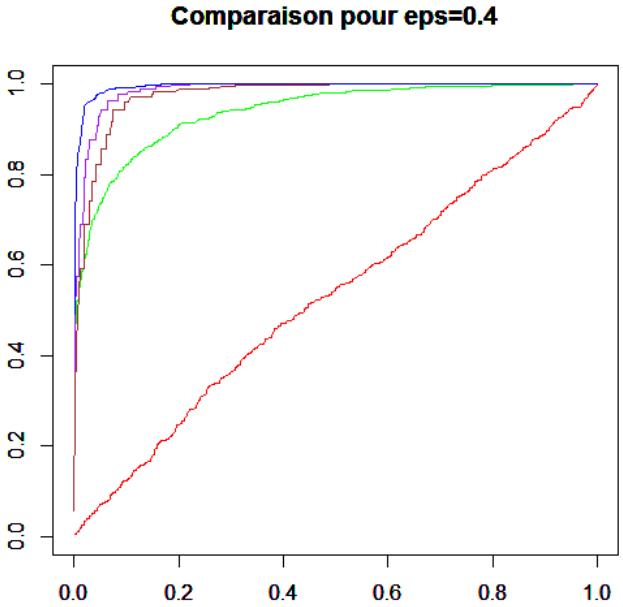

FIGURE 4.4.4 –

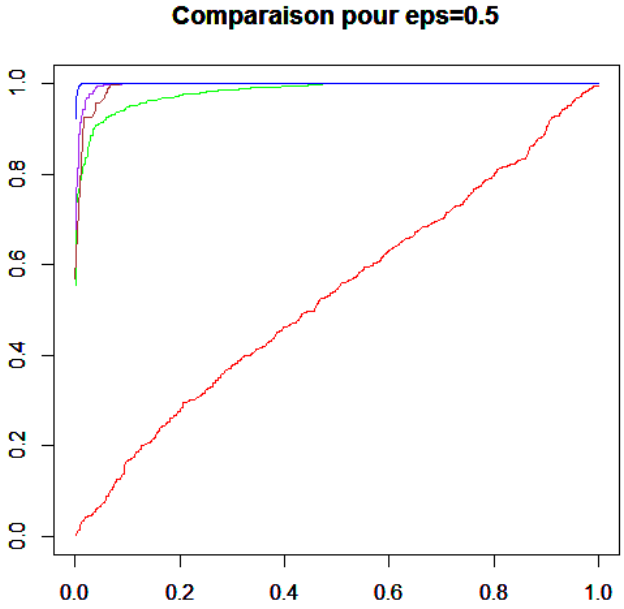

FIGURE 4.4.5 –

Conclusion

Il apparait donc que le meilleur des tests pour les lois que nous avons étudiées, est celui des ondelettes sélectionné dans la partie 4.3, c'est à dire le test faisant intervenir l'intégralité des données sans variation d'origine.

Nénamoins, il semble que le choix du meilleur dépende fortement des lois que l'on simule. C'est pourquoi, il faudrait faire différents types de simulation (par exemple des lois localisées) pour juger pertinemment si une méthode est meilleure que les autres.

Une grande amélioration apportée par ce rapport est le choix de prendre en compte toutes les données expérimentales pour construire le test ondelettes, ce qui n'était pas fait par Butucea-Tribouley.

Bibliographie

[1] Gilles Fay, Jacques Delabrouille, Gerard Kerkyacharian and Dominique Picard, TESTING THE ISOTROPY OF HIGH ENERGY COSMIC RAYS USING SPHERICAL NEEDLETS, 23 June 2012, Ecole Centrale Paris , CNRS and Universite Paris Diderot, arXiv :1107.5658v2

[2] P. Baldi, G. Kerkyacharian, D. Marinucci and D. Picard, ASYMPTOTICS FOR SPHERICAL NEEDLETS, 2009, Università di Roma Tor Vergata, LPMA Université de Paris X, Università di Roma Tor Vergata and LPMA Université de Paris 7, DOI : 10.1214/08-AOS601

[3] David L. Donoho, Iain M. Johnstone, Gerard Kerkyacharian, Dominique Picard, DENSITY ESTIMATION BY WAVELET THRESHOLDING, April 1993, Stanford University, Universite de Picardie,Universite de Paris VII

[4] C. Butucea & K. Tribouley, Nonparametric homogeneity tests, 2004, Université Paris X

Annexes

Annexe A : Ondelettes, Principes et Résultats

Motivation d'utilisation des ondelettes

Tout comme on décompose les fonctions de $L^2(\mathbb{R})$ sur la base des fonctions trigonométriques, les ondelettes sont le meilleur moyen pour décomposer une fonction sur une base de fonctions orthonormées. Néanmoins, elles présentent un avantage de taille : Alors que les coefficients dans la décomposition de Fourier décroissent avec une puissance dépendante de la régularité de la fonction que l'on veut décomposer, les coefficients d'ondelettes peuvent décroitre même si la fonction admet des discontinuités sur un ensemble de mesure nulle pour la mesure de Lebesgue. En d'autres termes, la décomposition en ondelettes est la plus appropriée quand on étudie des fonctions de $L^2(\mathbb{R})$ pouvant présenter des discontinuités brusques sur certains points, ce qui est notre cas.

Dans cette partie, les théorèmes s'appuieront sur l'article [2]

Ondelettes : une base de $L^2(\mathbb{R})$

Nous citerons d'abord quelques théorèmes généraux sur les ondelettes avant de les construire.

A noter que tous les théorèmes que nous allons développer sont pour des fonctions $f : \mathbb{R} \to \mathbb{R}$, notre but sera de les appliquer dans le cadre de fonction $g : \mathcal{C} \to \mathbb{R}$où$\mathcal{C}$ est le cercle unité dans $\mathbb{R}^2$.

On sait qu'il existe une fonction $\phi : \mathbb{R} \to \mathbb{R}$telle que :

(1) $\{\phi(x-k), k \in \mathbb{Z}\}$est une famille orthonormale de $L^2(\mathbb{R})$. Soit V_0le sous-espace engendré

(2) $\forall j \in \mathbb{Z}$, $V_j \subset V_{j+1}$où on pose V_jl'espace engendré par $\{\phi_{jk}, k \in \mathbb{Z}\}$et $\phi_{jk}(x) = 2^{\frac{j}{2}}\phi\left(2^j x - k\right)$.

(3) il existe $A > 0$ tel que $supp(\phi) = [-A; A]$

Dans ce cas, on a $V_{j+1} = V_j \oplus W_j$en posant W_jl'orthogonal de $V_j$dans V_{j+1}

On sait qu'il existe alors $\psi : \mathbb{R} \to \mathbb{R}$telle que :

(a) $\{\psi(x-k), k \in \mathbb{Z}\}$ est une base orthonormée de W_0

(b) $\{\psi_{jk}, k \in \mathbb{Z}, j \in \mathbb{Z}\}$ est une base orthonormée de $L_2(\mathbb{R})$ où $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi\left(2^j x - k\right)$

(c) ψ a la même régularité que ϕ

Théorème 1 : En prenant les notations ci-dessus, on a :

$$\forall j_0 \in \mathbb{Z}, L^2(\mathbb{R}) = V_{j_0} \oplus (\underset{j \geq j_0}{\oplus} W_j)$$

i.e.

$$\forall f \in L^2(\mathbb{R}), \forall x \in \mathbb{R}, f(x) = \sum_{k \in \mathbb{Z}} \alpha_{j_0 k} \phi_{j_0 k}(x) + \sum_{j \geq j_0} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}(x)$$

où $\forall j \geq j_0, \forall k \in \mathbb{Z}, \alpha_{j_0 k} =< f, \phi_{j_0 k} >_{L^2}$ et $\beta_{jk} =< f, \psi_{jk} >_{L^2}$

Nous avons donc une décomposition pour n'importe quelle fonction de $L^2(\mathbb{R})$. Ce qui va maintenant nous intéresser est bien sur le comportement des $(\alpha_{j_0 k})_{k \in \mathbb{Z}}$ et des $(\beta_{jk})_{j \in \mathbb{Z}, k \in \mathbb{Z}}$ en fonction de f.

Ondelettes de Haar : A partir du résultat du théorème 1, en prenant $\psi_{jk}(x) = 2^{\frac{j}{2}} \psi\left(2^j x - k\right)$ et $\forall x \in \mathbb{R}, \psi(x) = 1_{[0;\frac{1}{2}[}(x) - 1_{[\frac{1}{2};1]}(x)$, on définit les ondelettes de Haar.

Remarque : on remontre ici que $L^2(\mathbb{R})$ est séparable car on a trouvé une famille orthormée, dense et dénombrable de fonctions de $L^2(\mathbb{R})$

Espaces de Besov

Nous allons donner 2 définitions équivalentes des espaces de Besov, dont l'une en terme de coefficient d'ondelettes.

On fixe $(s, p, q) \in \mathbb{N}^3$tels que $s > 0$, $1 \leq p \leq +\infty$ et $1 \leq p < +\infty$et on pose B_{spq}l'espace de Besov caractérisé par ces 3 entiers

$1^{ère}$**caractérisation** On pose E_j projecteur sur V_j et $D_j = E_{j+1} - E_j$ le projecteur complémentaire à E_j dans V_{j+1}

on dit que

$$f \in B_{spq}$$

si et seulement si

$$\| E_0(f) \|_{L^p} + (\sum_{j \geq 0} (2^{js} \| D_j(f) \|_{L^p})^q)^{\frac{1}{q}} < +\infty$$

en utilisant le résultat précédent, on sait que il existe ϕ, ψ, $(\alpha_{0k})_{k \in \mathbb{Z}}$ et $(\beta_{jk})_{j \in \mathbb{Z}, k \in \mathbb{Z}}$ tels que $\forall x \in \mathbb{R}, f(x) = \sum_{k \in \mathbb{Z}} \alpha_{0k} \phi_{0k}(x) + \sum_{j \geq 0} \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}(x)$

Dans ce cas, $E_0(f) = \sum_{k \in \mathbb{Z}} \alpha_{0k} \phi_{0k}$ et $D_j(f) = \sum_{k \in \mathbb{Z}} \beta_{jk} \psi_{jk}$

D'où la deuxième caractérisation.

$2^{ème}$**caractérisation** on dit que

$$f \in B_{spq}$$

si et seulement si

$$\| \alpha_{0.} \|_{l^p} + (\sum_{j \geq 0} (2^{j(s+\frac{1}{2}-\frac{1}{p})} \| \beta_{j.} \|_{l^p})^q)^{\frac{1}{q}} < +\infty$$

Pour rappel, $\| u_{j.} \|_{l^p} = (\sum_{k \in \mathbb{Z}} u_{jk}^p)^{\frac{1}{p}}$

Soit f une fonction de $L^2(\mathbb{R})$, elle peut ne pas être "très régulière" mais appartenir à B_{spq}et donc avoir ses coefficients d'ondelettes $(\alpha_{j_0 k})_{k \in \mathbb{Z}}$ et $(\beta_{jk})_{j \in \mathbb{Z}, k \in \mathbb{Z}}$ qui décroissent (dans le sens où ils convergent pour la norme l^p)

Estimateurs d'ondelettes

On considère $(X_1, X_2, ..., X_n)$ un échantillon empirique d'une loi X de densité f

On sait d'après le théorème 1 que f est caractérisé par ces coefficients d'ondelettes $(\alpha_{j_0 k})_{k \in \mathbb{Z}}$ et $(\beta_{jk})_{j \in \mathbb{Z}, k \in \mathbb{Z}}$

Pour construire une estimation de la densité de f, on construit des estimateurs de ces coefficients. Cependant il faut aussi définir un j_0 adéquat pour cet échantillon.

On suppose donc que $j_0 = j_0(n) = J^*$

on prend donc :

$$\forall k \in \mathbb{Z}, \hat{\alpha_{J^* k}} = \frac{1}{n} \sum_{i=1}^{n} \phi_{J^* k}(X_i)$$

et

$$\forall j \geq J^*, \forall k \in \mathbb{Z}, \hat{\beta_{jk}} = \frac{1}{n} \sum_{i=1}^{n} \psi_{jk}(X_i)$$

et dans ce cas, on pose alors logiquement :

$$\forall x \in \mathbb{R}, \hat{f_X}(x) = \sum_{k \in \mathbb{Z}} \hat{\alpha_{J^* k}} \phi_{J^* k}(x) + \sum_{j \geq J^*} \sum_{k \in \mathbb{Z}} \hat{\beta_{jk}} \psi_{jk}(x)$$

On remarque que $\hat{\alpha_{J^* k}} = \frac{1}{n} \sum_{i=1}^{n} \phi_{J^* k}(X_i) \underset{n \to \infty}{\to} < \phi_{J^* k}, f_X > = \alpha_{J^* k}$ et de la même manière $\hat{\beta_{jk}} = \frac{1}{n} \sum_{i=1}^{n} \psi_{jk}(X_i) \underset{n \to \infty}{\to} < \psi_{jk}, f_X > = \beta_{jk}$

Par Perceval, on a alors :

$$\| \hat{f_X} - f_X \|_{L^2}^2 = \sum_{j \leq J^*} \sum_{k \in \mathbb{Z}} (\hat{\beta_{jk}} - \beta_{jk})^2$$

Remarque : le J^*optimal est donné par la relation $J^* \underset{n \to \infty}{\sim} log_2(\frac{n}{log(n)})$

Seuillage Le seuillage est l'opération qui consiste à s'affranchir de coefficients d'ondelettes trop petits que l'on considère "inexistant" dans la décomposition de f.

Deux opérations de seuillage sont possibles :

le seuillage doux (soft thresholding) et le seuilage dur (hard thresholding).

Pour ceci, on se munit de fonction $\delta_s$et $\delta_h$définie par

$$\delta_s : \begin{cases} \mathbb{R} \times \mathbb{R}^+ & \to \mathbb{R} \\ (x, \lambda) & \mapsto signe(x)(\mid x \mid -\lambda)_+ \end{cases} \text{ et } \delta_h : \begin{cases} \mathbb{R} \times \mathbb{R}^+ & \to \mathbb{R} \\ (x, \lambda) & \mapsto x1\{\mid x \mid > \lambda\} \end{cases}$$

Et ainsi on pose un nouvel estimateur. Si on prend $\delta = (\delta_s, \delta_h)$,

on pose alors

$$\forall k \in \mathbb{Z}, \alpha_{\tilde{J^*}k} = \delta(\alpha_{\hat{J^*}k}, \lambda_{J^*})$$

et

$$\forall j \geq J^*, \forall k \in \mathbb{Z}, \beta_{jk}^{\tilde{}} = \delta(\beta_{jk}^{\hat{}}, \lambda_j)$$

et ainsi

$$\forall x \in \mathbb{R}, \tilde{f_X}(x) = \sum_{k \in \mathbb{Z}} \alpha_{\tilde{J^*}k} \phi_{J^*k}(x) + \sum_{j \geq J^*} \sum_{k \in \mathbb{Z}} \beta_{jk}^{\tilde{}} \psi_{jk}(x)$$

Remarque : on peut prendre pour simplifier $\lambda_j = \lambda$ mais l'estimateur est plus optimum avec $\lambda_j \underset{n \to \infty}{\sim} \sqrt{\frac{j}{n}}$

Ces deux seuillages permettent d'annuler les coefficients s'ils sont inférieurs à λ, la différence entre les deux étant que le seuillage doux diminue tous les coefficients de λ(s'ils sont positifs et les augmente de λ s'ils sont négatifs) donnant une faible variance dans le risque quadratique alors que le seuillage dur laisse intacte les autres coefficients donnant un biais plus faible à l'estimateur.

Des résultats existent sur les vitesses de convergence au sens minimax de $\hat{f_X}$ et $\tilde{f_X}$ mais sont complexes.

Annexe B

Le principe de ce test est de former à l'échelle j un estimateurT_j de la norme L^2 du projeté de la différence des densités associées aux échantillons $(X_1, X_2, ..., X_n)$ et $(Y_1, Y_2, ..., Y_m)$ sur l'espace V_j. Ainsi, on cherche à estimer :

$$\|P_j(f_X - f_Y)\| = \sum_{k \in [0, 2\pi]} (\alpha_{j,k} - \beta_{j,k})^2$$

La clé est donc de trouver un estimateur de $(\alpha_{j,k} - \beta_{j,k})^2 = \alpha_{j,k}^2 - 2\alpha_{j,k}\beta_{j,k} + \beta_{j,k}^2$.

On se sert alors des estimateurs

$$\hat{\alpha^2}_{j,k} = \frac{1}{n(n-1)} \sum_{i \neq j} \phi(X_i)\phi(X_j) = \frac{1}{n(n-1)} \Big((\sum_i \phi(X_i))^2 - \sum_i \phi^2(X_i) \Big)$$

$$\hat{\beta^2}_{j,k} = \frac{1}{m(m-1)} \sum_{i \neq j} \phi(Y_i)\phi(Y_j) = \frac{1}{m(m-1)} \Big((\sum_i \phi(Y_i))^2 - \sum_i \phi^2(Y_i) \Big)$$

$$\hat{\alpha_{j,k}\beta_{j,k}} = \frac{1}{nm} \sum_{i,j} \phi(X_i)\phi(Y_j)$$

On se sert alors de ces estimateurs pour construire la pseudo-distance T_i et on procède comme dans le test à tailles d'échantillon égales décrit en 2.2.3.