

NOTES DE STATISTIQUE ET D'INFORMATIQUE

2011/2

INTRODUCTION AUX APPLICATIONS
DES ONDELETTES

C. CHARLES

Université de Liège – Gembloux Agro-Bio Tech
*Unité de Statistique, Informatique et Mathématique
appliquées à la bioingénierie*
GEMBLOUX
(Belgique)

INTRODUCTION AUX APPLICATIONS DES ONDELETTES

C. CHARLES *

RÉSUMÉ

Cette note technique fait suite à une note antérieure. Son objectif est de faire prendre conscience au lecteur de la large étendue d'applications des ondelettes. Elle se découpe en deux parties. La première illustre la théorie des ondelettes au moyen d'applications tournées vers la statistique. La deuxième se tourne vers les applications en traitement du signal et de l'image.

SUMMARY

The objective of this note is to make known a large number of applications of the wavelet theory. This note is divided in two parts. The first one briefly presents a few examples that illustrate the use of wavelets in statistics. The second one deals with applications in signal and image processing.

1. APPLICATIONS EN STATISTIQUE

Dans ce paragraphe, nous présentons deux exemples qui illustrent l'utilisation des ondelettes en statistique non-paramétrique. Il s'agit d'estimation de densité de probabilité et de diagramme de régression. Pour l'estimation de densité, nous proposons une nouvelle méthode non-paramétrique. Ceci veut dire que nous proposons une forme de lissage d'histogramme et non une équation exprimant la densité en fonction de sa variable. De même pour la régression, nous proposons une nouvelle méthode non-paramétrique. Ceci veut dire que nous proposons une nouvelle méthode pour retracer l'évolution de la moyenne de la variable en fonction de sa variable explicative et non pour trouver l'équation exprimant la variable en fonction de sa variable explicative. Dans chaque cas, nous

*Chargée de cours à l'Université de Liège, Gembloux Agro-Bio Tech (Unité de Statistique, Informatique et Mathématiques appliquées à la bioingénierie)

commençons par expliquer le contexte statistique du problème. Nous décrivons ensuite une solution basée sur les ondelettes. Ses avantages et inconvénients sont détaillés.

1.1. Utilisation des ondelettes pour l'estimation de densité

Une grande partie de la littérature consacrée à la statistique non-paramétrique concerne l'estimation de densité. Des aperçus sont donnés dans SILVERMAN, 1986 et IZENMAN, 1991. Toutes les méthodes proposées ont leurs propres avantages et inconvénients. Par exemple, la méthode du noyau bénéficie de l'héritage de toutes les propriétés de continuité et de différentiabilité du noyau, mais pose le problème du choix du paramètre de lissage. Espérant dépasser ce genre d'inconvénients et désirant tirer parti des nombreuses propriétés des ondelettes, des chercheurs, parmi lesquels Pinheiro et Vidakovic (PINHEIRO, VIDAKOVIC, 1997), ont adapté les estimateurs de densité par série orthogonale de Cencov (CENCOV, 1962). Dans cette section, nous présentons les estimateurs de densité par série orthogonale de Cencov et l'adaptation de celui-ci avec les ondelettes par Pinheiro et Vidakovic.

L'idée de Cencov est simple. La densité inconnue f de carré intégrable peut être représentée comme un développement en série orthogonale convergente

$$f(x) = \sum_{j \in J} a_j \psi_j(x), \quad (1)$$

où $\{\psi_j, j \in J\}$ est une base orthonormée de fonctions dans $L_2(D)$, $D \subset \mathbb{R}$ et J est un ensemble approprié d'indices. Par exemple, la base orthonormée peut être la base de Fourier ou une base d'ondelettes. A partir de l'équation 1, les coefficients a_j peuvent être exprimés comme

$$a_j = \sum_{i \in J} a_i \int \psi_i(x) \psi_j(x) dx = \int f(x) \psi_j(x) dx = E(\psi_j(X)). \quad (2)$$

Soit $\underline{X} = (X_1, X_2, \dots, X_n)$ un échantillon de la distribution inconnue f . Il semble naturel d'estimer a_j par

$$\hat{a}_j = \frac{1}{n} \sum_{i=1}^n \psi_j(X_i) \quad (3)$$

et $f(x)$ par

$$\hat{f}(x) = \sum_{j \in J} \hat{a}_j \psi_j(x). \quad (4)$$

Cependant, cet estimateur pourrait ne pas être bien défini. En fait,

$$\begin{aligned}
\hat{a}_j &= \frac{1}{n} \sum_{i=1}^n \psi_j(X_i) \\
&= \frac{1}{n} \sum_{i=1}^n \int \psi_j(x) \delta(x - X_i) dx \\
&= \int \left\{ \frac{1}{n} \sum_{i=1}^n \delta(x - X_i) \right\} \psi_j(x) dx \\
&= \int g(x) \psi_j(x) dx
\end{aligned}$$

où $g(x)$ est la fonction de probabilité empirique et δ est la fonction de Dirac. Comme \hat{a}_j et a_j sont identiques pour la fonction de probabilité empirique, ce qui suit est vrai pour tout échantillon \underline{X} :

$$\sum_{j \in J} \hat{a}_j \psi_j(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - X_i). \quad (5)$$

Cet estimateur a une variance infinie et n'est pas consistant au sens ISE (Integrated Square Error). La pratique standard est alors de sélectionner un nombre fini de coefficients empiriques \hat{a}_j et de les seuiller de manière appropriée.

Une base d'ondelettes est souvent choisie pour $\{\psi_j, j \in J\}$ à cause de sa localisation en temps et en fréquence qui permet d'obtenir un estimateur puissant. Dans ce cas-ci, $\{\psi_j\}_{j \in J} = \{\phi_{j_0, k}, \psi_{j, k}\}_{j \geq j_0, k \in \mathbb{Z}}$ et $\{a_j\}_{j \in J} = \{c_{j_0, k}, d_{j, k}\}_{j \geq j_0, k \in \mathbb{Z}}$. Nous estimons

$$\hat{c}_{j, k} = \frac{1}{n} \sum_{i=1}^n \phi_{j, k}(X_i)$$

pour $j = j_0, k \in \mathbb{Z}$ et

$$\hat{d}_{j, k} = \frac{1}{n} \sum_{i=1}^n \psi_{j, k}(X_i)$$

pour $j \geq j_0, k \in \mathbb{Z}$. La sélection de j_0 dépend de l'ondelette mère et de la régularité de la densité. Le lecteur peut se référer à PINHEIRO et VIDAKOVIC, 1997 pour plus d'explications. Ensuite, il est nécessaire de seuiller certains de ces coefficients comme spécifié plus haut. Kolaczyk (KOLACZYK, 1994) propose le seuil suivant $\lambda = \log(n)/\sqrt{n}$ pour les niveaux jusque $j_1 = \lfloor \log_2 n - 1 \rfloor$. Donoho (DONOHO, 1996) suggère de prendre le niveau $j_1 = \lfloor \log_2 n - \log_2(\log n) \rfloor$ où n est la taille de l'échantillon. De façon similaire, Delyon et Juditsky (DELYON et JUDITSKY, 1993) recommandent $j_1 = \lfloor \log_2 2n - \log_2(\ln n) \rfloor$. Ceci veut dire que, si nous utilisons un seuillage fort, nous obtenons certains nouveaux coefficients

$$\hat{d}_{j, k} = \begin{cases} \hat{d}_{j, k} & \text{si } |\hat{d}_{j, k}| \geq \lambda, \\ 0 & \text{sinon,} \end{cases}$$

$\forall j_0 \leq j \leq j_1, \forall k$, et si nous utilisons un seuillage doux, nous obtenons certains nouveaux coefficients

$$\hat{d}_{j, k} = \begin{cases} \hat{d}_{j, k} - \lambda & \text{si } \hat{d}_{j, k} \geq \lambda, \\ \hat{d}_{j, k} + \lambda & \text{si } \hat{d}_{j, k} \leq -\lambda, \\ 0 & \text{sinon,} \end{cases}$$

$\forall j_0 \leq j \leq j_1, \forall k.$

Pinheiro et Vidakovic (PINHEIRO et VIDAKOVIC, 1997) ont amélioré cet estimateur par ondelettes. Plutôt qu'estimer directement la densité inconnue f , ils estiment sa racine carrée : \sqrt{f} . Deux arguments justifient ce choix. Premièrement, comme beaucoup de méthodes standards d'estimation de densité, celle développée ci-dessus pourrait permettre d'obtenir un estimateur de densité avec des valeurs négatives. Une façon de contourner le problème est de tronquer l'estimateur en mettant à zéro les valeurs négatives et en normalisant la troncature. Une autre façon est d'estimer une transformation de f , telle que $\log f$ ou \sqrt{f} . Deuxièmement, en plus de la positivité, l'estimateur de densité doit avoir une intégrale de un. Le fait d'estimer \sqrt{f} va satisfaire cette contrainte grâce au fait que $1 = \int f = \langle \sqrt{f}, \sqrt{f} \rangle = \|\sqrt{f}\|^2$. En effet, si $\sqrt{f} = \sum_{j \in J} a_j \psi_j(x)$, alors $\|a_j\|_{l_2}^2 = \|\sqrt{f}\|^2 = 1$ par l'identité de Parseval. Ainsi, normaliser les coefficients revient à rendre un estimateur bona fide. Techniquement, Vidakovic calcule les coefficients de \sqrt{f} avec

$$\hat{c}_{j,k} = \frac{1}{n} \sum_{i=1}^n \frac{\phi_{j,k}(X_i)}{\sqrt{\hat{f}_n(X_i)}}$$

et

$$\hat{d}_{j,k} = \frac{1}{n} \sum_{i=1}^n \frac{\psi_{j,k}(X_i)}{\sqrt{\hat{f}_n(X_i)}}$$

pour un certain premier estimateur de la densité inconnue, \hat{f}_n . Le calcul de $\hat{c}_{j,k}$ ($\hat{d}_{j,k}$ resp.) est motivé de la façon suivante :

$$c_{j,k} = \langle \phi_{j,k}, \sqrt{f} \rangle = \int \phi_{j,k} \sqrt{f} = \int \frac{\phi_{j,k}}{\sqrt{f}} f$$

$$(d_{j,k} = \langle \psi_{j,k}, \sqrt{f} \rangle = \int \psi_{j,k} \sqrt{f} = \int \frac{\psi_{j,k}}{\sqrt{f}} f).$$

Le plus simple premier estimateur est l'histogramme. Pinheiro et Vidakovic suggèrent de choisir j_1 comme l'argument minimum de $E(j) = \sum_k \hat{d}_{j,k}^2$ et de seuiller les coefficients comme suit :

$$\hat{d}_{j,k} = I(\hat{d}_{j,k}^2 > \kappa \bar{\hat{d}}^2) \hat{d}_{j,k}$$

où $\bar{\hat{d}}^2$ est la moyenne de $\hat{d}_{j,k}^2$ et $\kappa \in \mathbb{R}$. Souvent, $\kappa = 0.5$ est choisi. Après le seuillage, les coefficients restants sont normalisés. L'identité de Parseval assure alors que l'estimateur est une densité bona fide.

En conclusion, l'algorithme général basé sur les ondelettes pour l'estimation d'une densité univariée est une méthode assez récente. L'estimateur est simple, adaptatif en localisation (choix de j_1) et régularité (choix de l'ondelette mère), et efficace. Nous avons vu que Pinheiro et Vidakovic ont à partir de là développé leur propre estimateur par ondelettes non négatif dont l'intégrale vaut 1. Ils ont montré que, sur de nombreux exemples, leur estimateur

est meilleur que celui des noyaux. Dans MULLER ET VIDA KOVIC, 1998, les auteurs ont montré que le rapport du MISE de l'estimateur par ondelettes sur le MISE de l'estimateur par noyaux est strictement inférieur à 1 sur de nombreux exemples. De nombreux autres travaux (ADRIAN, 2008, RENAUD, 1999, TRIBOULEY, 2008 pour n'en citer que quelques-uns) montrent que les estimateurs de densité par ondelettes ont une erreur minimale et sont avantageux du point de vue calculatoire. Ils offrent une meilleure estimation quand la densité recherchée n'est pas régulière. L'estimation de densité par ondelettes est proposée sur MatLab. Un minimum d'informations est disponible à l'adresse suivante : <http://www.mathworks.com/help/toolbox/wavelet/ug/f8-95760.html#f8-40906> (visité le 26/01/11).

1.2. Utilisation des ondelettes pour la régression

Dans cette section, nous considérons uniquement les familles d'ondelettes qui forment une base orthonormée. Nous expliquons comment utiliser une base d'ondelettes pour construire un estimateur non paramétrique pour une fonction de régression m dans le modèle

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \dots, n. \quad (6)$$

Si on considère le modèle fixe, les X_i sont non aléatoires et les erreurs sont des variables normales indépendantes et identiquement distribuées $\epsilon_i \approx N(0, \sigma_\epsilon^2)$. Si on considère le modèle aléatoire, les (X_i, Y_i) sont des variables aléatoires indépendantes et identiquement distribuées (X, Y) avec $m(x) = E(Y|X = x)$ et $\epsilon_i = Y_i - m(X_i)$.

1.2.1. Modèle fixe

L'objectif est de construire un estimateur non paramétrique pour une fonction de régression $m \in L_2([0, 1])$ dans le modèle

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \dots, n, \quad n = 2^J, \quad J \in \mathbb{N}, \quad (7)$$

où $x_i = \frac{i}{n}$ et les erreurs ϵ_i sont des VA iid $\epsilon_i \approx N(0, \sigma_\epsilon^2)$. Notons que lorsque n n'est pas dyadique ou $x_i \neq \frac{i}{n}$ ou encore les erreurs ne sont pas des VA normales iid, la méthode proposée ci-dessous nécessite des adaptations non vues dans cette note.

L'idée générale est celle de Cencov qui considère que toute fonction m peut être décomposée dans une base à l'aide de coefficients :

$$m(x) = \sum_{j \in J} a_j \psi_j(x)$$

où $\{\psi_j, j \in I\}$ est une base orthonormée de fonctions dans $L_2(D), D \subset [0, 1]$, I est un ensemble approprié d'indices et $a_j = \int m(x) \psi_j(x) dx$. Pour $m \in L^2$,

$\sum_j a_j^2 < \infty$ entraîne que $m(x)$ est bien approximée en prenant seulement un petit nombre N de a_j .

Un estimateur par ondelettes consiste à choisir une base $\{\psi_j, j \in I\}$ d'ondelettes. Cet estimateur peut être linéaire ou non linéaire. L'estimateur par ondelettes linéaire procède par projection des données sur un espace de niveau plus grossier en prenant les N premiers coefficients. Cet estimateur est du type des noyaux. Une autre possibilité pour estimer m est de détecter quels coefficients relatifs aux détails contiennent l'information importante au sujet de la fonction m et de mettre à zéro les autres coefficients. Ceci donne lieu à un estimateur non linéaire. En pratique, il consiste à prendre les N plus grands coefficients en valeur absolue.

Estimateur par ondelettes linéaire

Supposons que nous avons des données $(x_i, Y_i)_{i=1}^n$ provenant du modèle défini à l'équation 7 et une base orthonormée d'ondelettes générée par une ondelette mère ψ et une ondelette père ϕ . L'estimateur linéaire procède en choisissant un niveau j_1 et représente une estimation de la projection de m dans l'espace $V_{j_1} \subset L_2(\mathbb{R})$ (pour plus d'explications, voir DELOUILLE, 2002 pour l'analyse multirésolution) :

$$\hat{m}(x) = \sum_{k=0}^{2^{j_0}-1} \hat{c}_{j_0,k} \phi_{j_0,k}(x) + \sum_{j=j_0}^{j_1-1} \sum_{k=0}^{2^j-1} \hat{d}_{j,k} \phi_{j,k}(x), \quad (8)$$

avec j_0 le niveau le plus grossier de la décomposition et $\hat{c}_{j_0,k} = c_{j_0,k}^Y$ et $\hat{d}_{j,k} = d_{j,k}^Y$. Le niveau j_1 joue un rôle de paramètre de régularisation : une petite valeur de j_1 veut dire que beaucoup de coefficients relatifs aux détails seront mis de côté, et ceci pourrait donc trop lisser. D'un autre côté, si j_1 est trop grand, trop de coefficients seront gardés et certaines bosses artificielles resteront probablement dans l'estimation de $m(x)$.

Grâce à l'orthogonalité de la transformée en ondelettes et l'égalité de Parseval, le risque L_2 (MISE ou Mean Integrated Square Error) de l'estimateur linéaire est égal au risque l_2 de ses coefficients d'ondelettes :

$$\begin{aligned} MISE &= E\|\hat{m}-m\|_{L_2}^2 = \sum_k E[\hat{c}_{j_0,k}-c_{j_0,k}^0]^2 + \sum_{j=j_0}^{j_1-1} \sum_k E[\hat{d}_{j_0,k}-d_{j_0,k}^0]^2 + \sum_{j=j_1}^{\infty} \sum_k (d_{j,k}^0)^2 \\ &= S_1 + S_2 + S_3, \end{aligned} \quad (9)$$

où

$$s_{j_0,k}^0 := \langle m, \phi_{jk} \rangle \quad \text{et} \quad d_{j_0,k}^0 := \langle m, \psi_{jk} \rangle \quad (10)$$

sont appelés coefficients théoriques dans le contexte de la régression. Le terme $S_1 + S_2$ constitue le biais stochastique tandis que S_3 est le biais déterministe. Le niveau j_1 optimal est tel que les deux biais sont de même ordre de grandeur.

En pratique, des méthodes de validation croisée sont souvent utilisées pour déterminer le niveau optimal.

Estimateur par ondelettes non linéaire

Étant donné le modèle de régression de l'équation 7, nous pouvons décomposer les coefficients relatifs aux détails \hat{d}_{jk}^Y des Y_i comme

$$\hat{d}_{jk}^Y = d_{jk} + \rho_{jk}, \quad (11)$$

où d_{jk} sont les coefficients relatifs aux détails des $m(x_i)$ et ρ_{jk} ceux associés aux ϵ_i . Si la fonction $m(x)$ permet une représentation en ondelettes creuse, seulement un petit nombre de coefficients relatifs aux détails d_{jk} contribueront au signal et seront non négligeables. Cependant, chaque coefficient empirique \hat{d}_{jk}^Y a une contribution non nulle provenant de la partie bruitée ρ_{jk} .

Supposons que le niveau de bruit n'est pas trop haut, de telle façon que le signal peut être distingué du bruit. Alors, par la propriété de parcimonie de l'ondelette, seuls les plus grands coefficients relatifs aux détails pourraient être inclus dans l'estimateur par ondelettes. Par conséquent, quand on veut estimer une fonction inconnue, on inclut seulement les coefficients qui sont plus grands en valeur absolue qu'un certain seuil. En réalité, on applique le seuillage fort. Maintenant, puisque chaque coefficient empirique consiste à la fois en une partie du signal et une partie du bruit, il serait peut-être souhaitable de seuiller tous les coefficients. C'est le seuillage doux.

La régression se fait donc en trois étapes :

1. appliquer la transformée en ondelettes aux observations $\{Y_i\}$ amenant donc $\hat{c}_{j_0}^Y$ et \hat{d}_j^Y pour $j = j_0, \dots, J - 1$
2. manipuler les coefficients relatifs aux détails au-dessus du niveau j_0 par un seuillage
3. inverser la transformée en ondelettes et produire une estimation de m .

Le choix de j_0 est souvent de 2 ou 3 en pratique, bien qu'une détermination par validation croisée est possible. La sélection du seuil est très importante. De nombreuses méthodes de sélection de seuil ont été développées. Le seuil universel est

$$t_{univ} = \sigma_d \sqrt{2 \log n}$$

où σ_d^2 est la variance des coefficients d'ondelettes empiriques.

Donoho et Jonhstone ont démontré les propriétés de convergence de cette méthode qui obtient de meilleurs résultats que l'estimateur par noyaux ou par splines pour des signaux singuliers. De plus sa complexité algorithmique est de $O(n \log n)$ contrairement à $O(n^2)$ pour les estimateurs par noyaux ou splines. En illustration, nous avons les figures 1 et 2 (HUANG, 2003).

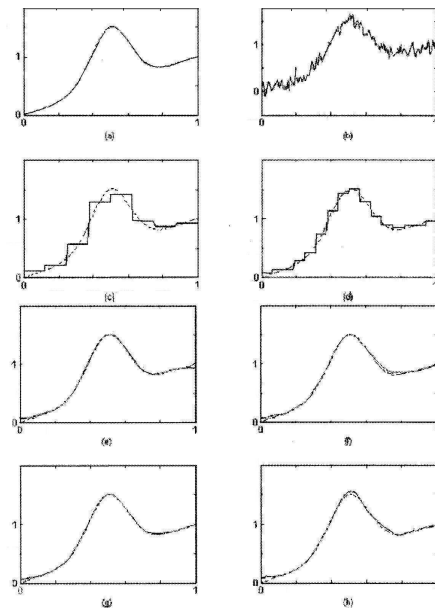


Figure 1. Régression faite à partir de 128 points : courbe, courbe bruitée, régressogramme 8, régressogramme 16, estimateur par série de cosinus, estimateur par noyaux, estimateur par spline, estimateur par ondelettes. (D'après HUANG, 2003).

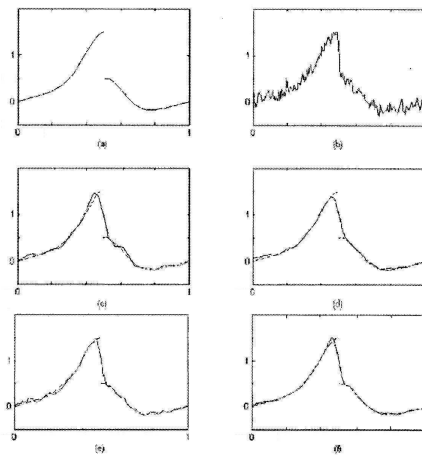


Figure 2. Régression faite à partir de 128 points : courbe, courbe bruitée, estimateur par série de cosinus, estimateur par noyaux, estimateur par spline, estimateur par ondelettes. (D'après HUANG, 2003).

1.2.2. Modèle aléatoire

Le modèle est le suivant :

$$Y_i = m(X_i) + \epsilon_i \quad i = 1, \dots, n,$$

avec (X_i, Y_i) variables aléatoires iid (X, Y) , $m(x) = E(Y|X = x)$ et ϵ_i variable aléatoire $N(0, \sigma_i^2)$.

L'estimateur à noyaux de Nadaraya-Watson est basé sur le fait que

$$m(x) = E(Y|X = x) = \int y f_{Y|X}(y|x) dx = \frac{\int y f_{XY}(x, y) dy}{f_X(x)}$$

et sur les densités estimées par la méthode des noyaux :

$$\hat{f}_{XY}(x, y) = \frac{1}{nh^2} \sum_i K_h(x - X_i) K_h(y - Y_i)$$

$$\hat{f}_X(x) = \frac{1}{nh} \sum_i K_h(x - X_i)$$

avec h paramètre de lissage. Par les propriétés du noyau K , on a

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n Y_i K_h(X_i - x)}{\sum_{i=1}^n K_h(X_i - x)}.$$

Cet estimateur à noyaux de Nadaraya-Watson peut être amélioré avec les ondelettes. Soit ϕ une fonction d'échelle à support compact. Soit K (noyau) défini comme

$$K(u, v) = \sum_k \phi(u - k) \phi(v - k).$$

On a

$$\hat{m}_j(x) = \frac{\sum_{i=1}^n Y_i K(2^j x, 2^j X_i)}{\sum_{i=1}^n K(2^j x, 2^j X_i)}.$$

On peut montrer que

$$\hat{m}_j(x) = \frac{\sum_k (\frac{1}{n} \sum_{i=1}^n Y_i \phi_{jk}(X_i)) \phi_{jk}(x)}{\sum_k (\frac{1}{n} \sum_{i=1}^n \phi_{jk}(X_i)) \phi_{jk}(x)}$$

qui n'est rien d'autre que l'estimateur de Nadaraya-Watson où les densités ont été estimées par la méthode des ondelettes. Une illustration se trouve en figure 3. De nouveau, on peut avoir un estimateur linéaire ou non-linéaire avec ou sans seuillage. Les avantages de cette méthode découlent des avantages de l'estimation de densité par ondelettes. La régression par ondelettes (fixe ou aléatoire) est proposée sur MatLab. Un minimum d'informations est disponible à l'adresse suivante : <http://www.mathworks.com/help/toolbox/wavelet/gs/f4-1021504.html> (visité le 26/01/11).

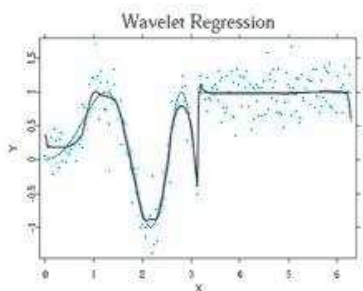


Figure 3. Régression par ondelettes $n = 256$.

2. APPLICATIONS EN TRAITEMENT DU SIGNAL ET DE L'IMAGE

Dans ce paragraphe, nous présentons quelques exemples qui illustrent l'utilisation des ondelettes dans le traitement du signal et de l'image. Il s'agit de débruitage, d'analyse en ondelettes croisée et de cohérence par ondelettes, de la déconvolution et de la compression. Dans chaque cas, nous commençons par expliquer le problème. Nous décrivons ensuite une solution basée sur les ondelettes. Les avantages et inconvénients sont mis en avant.

2.1. Utilisation des ondelettes pour le débruitage

L'estimation de signaux dans du bruit est un grand exemple de l'efficacité des ondelettes. En fait, dans un bruit de fond de conversations anglaises, il est facile de suivre une discussion en français. De même, l'estimation d'un signal mêlé à du bruit peut s'optimiser en trouvant une représentation qui sépare le signal du bruit (MALLAT, 1988). Par leur localisation en temps et en fréquence, les ondelettes permettent une discrimination efficace du signal et du bruit.

En sciences expérimentales, les signaux sont habituellement bruités. Il est souvent, mais pas toujours, raisonnable de considérer que le bruit est normalement distribué. Dans des techniques impliquant un processus de comptage, les données sont modélisées par une distribution de Poisson. Pour couvrir aussi bien le cas normal que le cas de Poisson, nous décrivons la méthode de seuillage par ondelettes de Donoho et Johnstone (DONOHO et JOHNSTONE, 1994) pour un bruit normalement distribué et l'algorithme de Kolaczyk pour un bruit de Poisson (KOLACZYK, 1996). Il est clair que ce dernier est basé sur le premier. Notons que cette section est relative à des signaux unidimensionnels mais que les résultats sont aisément généralisables à des signaux multidimensionnels.

Notons également qu'il n'y a pas de différences fondamentales entre le filtrage d'un bruit et la régression. La démarche est identique. Historiquement, les ondelettes ont d'abord été utilisées dans le débruitage avant d'être accaparées par les statisticiens pour la régression.

2.1.1. Filtrage d'un bruit normalement distribué : Donoho et Johnstone

Modèle

On suppose que les observations $(X_i)_{i=0}^{n-1}$ ($n = 2^J; J > 0$) peuvent être modélisées comme la somme d'un signal à estimer $(\lambda_i)_{i=0}^{n-1}$ et d'un bruit blanc normalement distribué $(W_i)_{i=0}^{n-1}$ de variance σ^2 . Nous avons donc :

$$X_i = \lambda_i + W_i \quad i = 0 \dots n - 1. \quad (12)$$

De nombreuses approches ont été suggérées pour filtrer un bruit normalement distribué. La plupart d'entre elles sont basées sur la régularisation par spline, l'estimation par noyau, le développement en série de Fourier, pour n'en citer que quelques-unes. Plus particulièrement, la dernière décennie du siècle passé a été le témoin de l'émergence d'une méthode puissante basée sur les ondelettes. L'approche habituelle de Donoho est de développer les données bruitées en série d'ondelettes, d'extraire les coefficients d'ondelettes significatifs par seuillage et ensuite d'utiliser l'inverse de la transformée en ondelettes sur les coefficients débruités. Le succès de cette approche est principalement basé sur d'importantes propriétés d'optimalité de cet estimateur par ondelettes, sur la représentation parcimonieuse des séries d'ondelettes pour une large gamme de fonctions et sur sa rapidité.

Meilleur estimateur

En décomposant $X = (X_i)_{i=0}^{n-1}$ dans une base d'ondelettes orthonormée $\{\psi_{j,k}\}_{j,k}$, nous trouvons les coefficients d'ondelettes :

$$\langle X, \psi_{j,k} \rangle = \langle \lambda, \psi_{j,k} \rangle + \langle W, \psi_{j,k} \rangle \quad (13)$$

où les $\langle W, \psi_{j,k} \rangle$ sont des variables aléatoires normales indépendantes et identiquement distribuées possédant une moyenne nulle et une variance σ^2 . Dans MALLAT, 1988, Mallat explique que le filtre de Wiener généralisé donne l'estimateur suivant pour λ :

$$\hat{\lambda} = \sum_{j,k} \langle X, \psi_{j,k} \rangle \theta_{j,k} \psi_{j,k} \quad (14)$$

où $\theta_{j,k} = \frac{|\langle \lambda, \psi_{j,k} \rangle|^2}{|\langle \lambda, \psi_{j,k} \rangle|^2 + \sigma^2}$ minimise un critère de performance global : le "Mean Integrated Squared Error (MISE)", $E(\|\lambda - \hat{\lambda}\|_2^2)$. Notons ϵ_a la valeur minimum du MISE. Cette méthode est théorique. Elle ne peut pas être implémentée car il est impossible de calculer $\langle \lambda, \psi_{j,k} \rangle$. Notons que le rôle du facteur $\theta_{j,k} (< 1)$

consiste à seuiller les coefficients $\langle X, \psi_{j,k} \rangle$ dans le but d'enlever le bruit de ces coefficients et par conséquent d'enlever le bruit de X .

Estimateur simple

L'estimation de $\theta_{j,k}$ peut être simplifiée en restreignant ses valeurs à 0 ou 1. Si le MISE est minimisé sous cette contrainte, il peut être montré que

$$\theta_{j,k} = \begin{cases} 1 & \text{si } |\langle \lambda, \psi_{j,k} \rangle|^2 > \sigma^2 \\ 0 & \text{si } |\langle \lambda, \psi_{j,k} \rangle|^2 \leq \sigma^2. \end{cases} \quad (15)$$

Malheureusement, cet estimateur-ci requiert le calcul de $\langle \lambda, \psi_{j,k} \rangle$ et n'est dès lors pas implémentable. Ce seuillage est appelé le seuillage simple et son erreur, ϵ_{simple} , satisfait :

$$\epsilon_{simple} \geq \epsilon_a \geq \epsilon_{simple}/2. \quad (16)$$

Estimateur par seuillage de Donoho-Johnstone

Afin d'obtenir un estimateur de λ implémentable, Donoho et Johnstone ont suggéré un estimateur par seuillage fort où $\theta_{j,k}$ prend les valeurs suivantes :

$$\theta_{j,k} = \begin{cases} 1 & \text{si } |\langle X, \psi_{j,k} \rangle| > T \\ 0 & \text{si } |\langle X, \psi_{j,k} \rangle| \leq T \end{cases} \quad (17)$$

où le seuil T est égal à $\sqrt{2 \log(n)} \sigma$.

Cette dernière valeur a été choisie pour T sur base de l'argument suivant. Sous l'hypothèse nulle $H_0 : \lambda = 0$, l'équation 12 devient $X = W$, qui implique à son tour que $\langle W, \psi_{j,k} \rangle = \langle X, \psi_{j,k} \rangle$. Afin d'obtenir $\lambda = 0$, nous devons dès lors avoir $T > |\langle W, \psi_{j,k} \rangle|$. Cependant, sous l'hypothèse alternative $H_a : \lambda \neq 0$, nous devons être prudents afin d'éviter des valeurs de seuil trop grandes qui mettraient tous les coefficients à zéro, spécialement dans les cas où $\langle X, \psi_{j,k} \rangle \neq \langle W, \psi_{j,k} \rangle$. Donoho et Johnstone (DONOHO et JOHNSTONE, 1994) ont suggéré une valeur de seuil ayant une grande probabilité d'être juste au-dessus de la valeur maximum prise par $|\langle W, \psi_{j,k} \rangle|$. Choisir $T = \sqrt{2 \log(n)} \sigma$ permet de satisfaire cette contrainte puisque :

$$\lim_{n \rightarrow +\infty} P\left(T - \frac{\sigma \log(\log(n))}{\log(n)} \leq \max_{j,k} |\langle W, \psi_{j,k} \rangle| \leq T\right) = 1, \quad (18)$$

et

$$\lim_{n \rightarrow \infty} \frac{\log(\log(n))}{\log(n)} = 0. \quad (19)$$

Donoho et Johnstone ont aussi montré que le MISE de leur estimateur par seuillage est simplement relié à l'estimateur simple présenté ci-dessus :

$$\epsilon_{thresholding} = E(\|\lambda - \hat{\lambda}_{thresholding}\|_2^2) \leq (2 \log(n) + 1)(\sigma^2 + \epsilon_{simple}). \quad (20)$$

L'estimateur par seuillage de Donoho-Johnstone donne lieu à un algorithme de filtrage en trois étapes :

1. la décomposition des observations dans une base d'ondelettes,
2. le T -seuillage de tous les coefficients,
3. l'application de l'inverse de la transformée par ondelettes sur les coefficients seuillés.

En plus de T , d'autres paramètres doivent être choisis : l'échelle grossière et la famille d'ondelettes. Johnstone propose $[\log_2(n) - \log_2(\log_{10}(n))]$ pour le choix de l'échelle grossière L (dans DONOHO, 1996) ; Juditsky suggère $[\log_2(n) - \log_2(\ln(n))]$ (dans JUDITSKY, 1994). Un choix éclairé de la famille d'ondelettes est de prendre une famille donnant un nombre maximum de coefficients d'ondelettes $\langle \lambda, \psi_{j,k} \rangle$ proches de zéro. De cette manière, le signal λ est concentré sur un petit nombre de grands coefficients et le signal n'est pas confondu avec le bruit qui est uniformément répandu sur tous les coefficients d'ondelettes. Le seuillage est plus efficace. Afin d'obtenir un nombre maximum de petits coefficients d'ondelettes, il faut être attentif à trois critères dans le choix de l'ondelette.

1. Il est préférable de choisir une ondelette qui possède beaucoup de moments nuls. Par définition, une ondelette avec m moments nuls est orthogonale aux polynômes de degré $m - 1$. En fait, on peut démontrer que si le signal est régulier et si l'ondelette choisie possède assez de moments nuls, alors les coefficients d'ondelettes seront petits pour les fines échelles.
2. Il est préférable de choisir une ondelette avec un petit support. En fait, si le signal contient une singularité en t_o et si t_o est dans le support de $\psi_{j,k}$, le coefficient d'ondelette correspondant sera grand. La taille du support de l'ondelette et le nombre de moments nuls sont a priori indépendants. Cependant, nous pouvons prouver que les contraintes qui sont imposées sur les ondelettes orthogonales impliquent que si l'ondelette possède p moments nuls, son support sera au moins de taille $2p - 1$. Par conséquent, il existe un compromis entre le nombre de moments nuls et la taille du support de l'ondelette. Si le signal possède des singularités isolées mais est régulier entre les singularités, il est préférable de choisir une ondelette avec un grand nombre de moments nuls afin de produire un grand nombre de petits coefficients. Si le nombre de singularités augmente, il est préférable de diminuer la taille du support et, par conséquent, de réduire le nombre de moments nuls.
3. Finalement, dans le choix de l'ondelette, nous devons aussi faire attention à la régularité de l'ondelette. Celle-ci a principalement une influence esthétique sur l'erreur introduite durant le seuillage des coefficients d'ondelettes. Une erreur régulière est toujours moins visible qu'une erreur irrégulière. C'est pourquoi l'ondelette de Haar est rarement utilisée avec le seuillage de Donoho-Jonhstone.

Nous terminons cette section relative au filtrage d'un bruit normalement distribué avec quelques remarques. Notons premièrement qu'un seul seuil est proposé pour tous les coefficients d'ondelettes. Notons également que nous ne mentionnons ici que l'estimateur par seuillage fort de Donoho-Johnstone. Il existe l'estimateur par seuillage doux de Donoho-Johnstone (DONOHO, 1996). La propriété générale du seuillage doux est qu'il assure avec une grande probabilité que

l'estimateur est au moins aussi régulier que le signal à estimer. Cependant, il produit souvent une plus grande erreur quadratique que le seuillage fort. Finalement, notons aussi que le seuil T dépend de la variance σ qui est souvent inconnue. Mallat (MALLAT, 1988) propose de l'estimer par $\hat{\sigma} = \frac{1}{0.6745} \text{Median}(|\langle X, \psi_{1,j} \rangle|)_{0 \leq j < n/2}$. La figure 4 montre le seuillage d'une fonction bruitée utilisant les seuillages doux et fort de Donoho-Johnstone. Cet algorithme peut être généralisé à plusieurs dimensions comme illustré dans la figure 5. Le succès de cette technique par rapport aux autres est lié à sa meilleure complexité algorithmique et à ses meilleurs résultats principalement lorsque le signal à estimer est singulier.

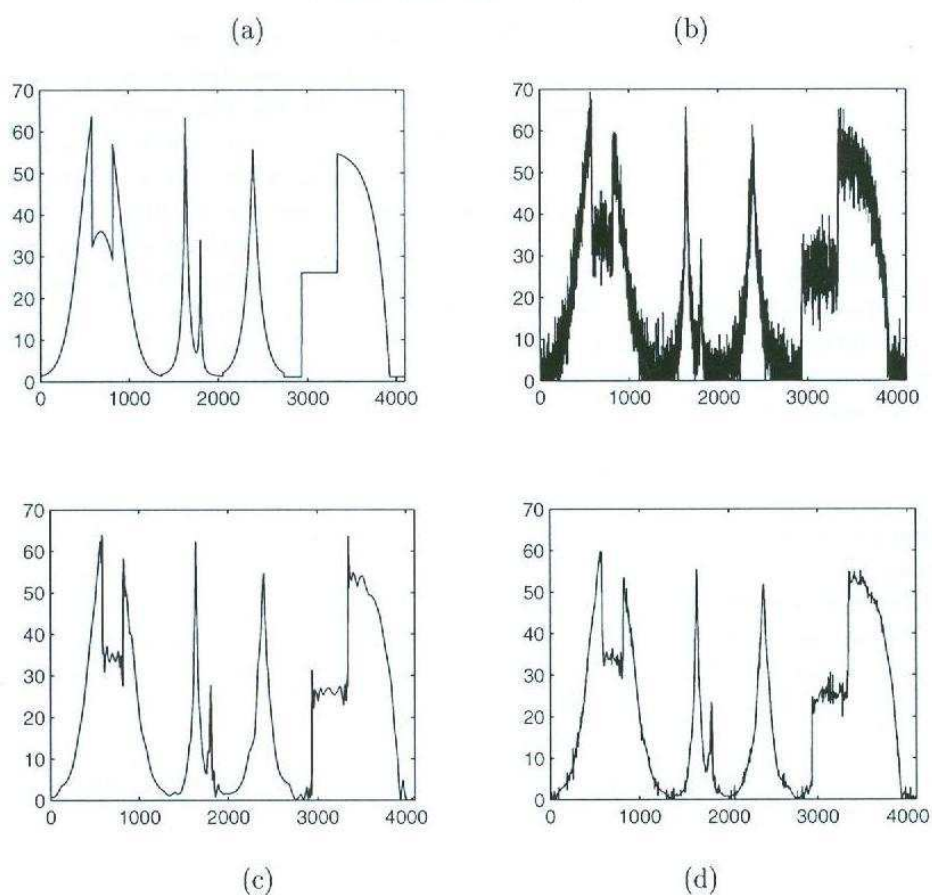
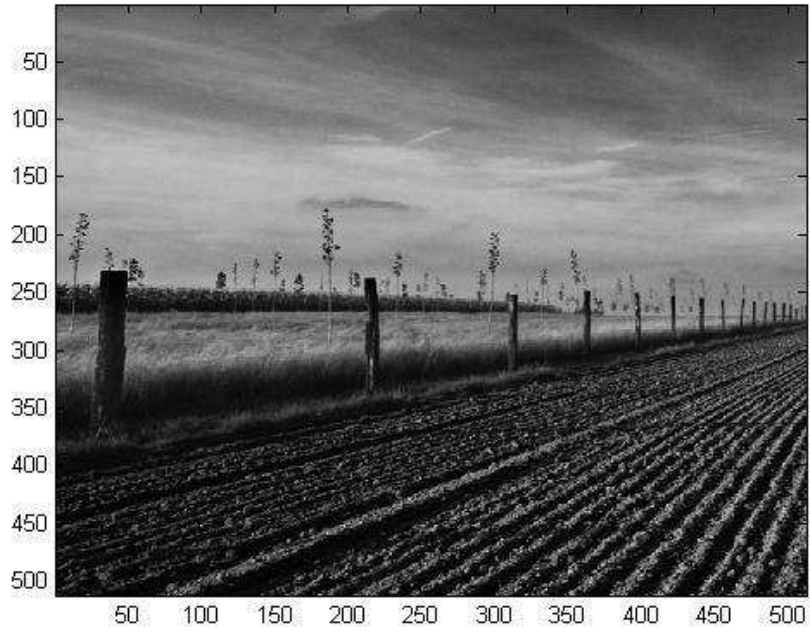
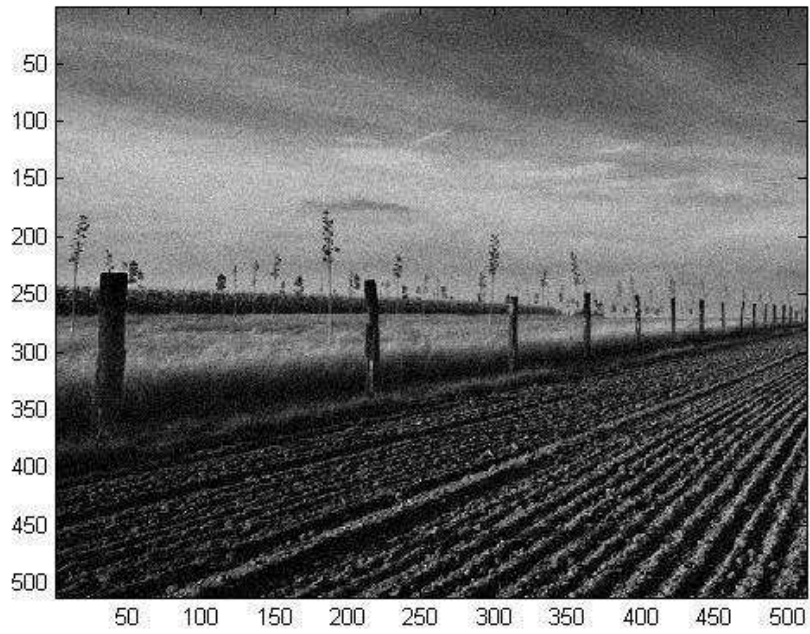


Figure 4. (a) : signal original ; (b) : signal bruité ; (c) : estimation avec un seuillage fort ; (d) : estimation avec un seuillage doux.

(a)



(b)



(c)

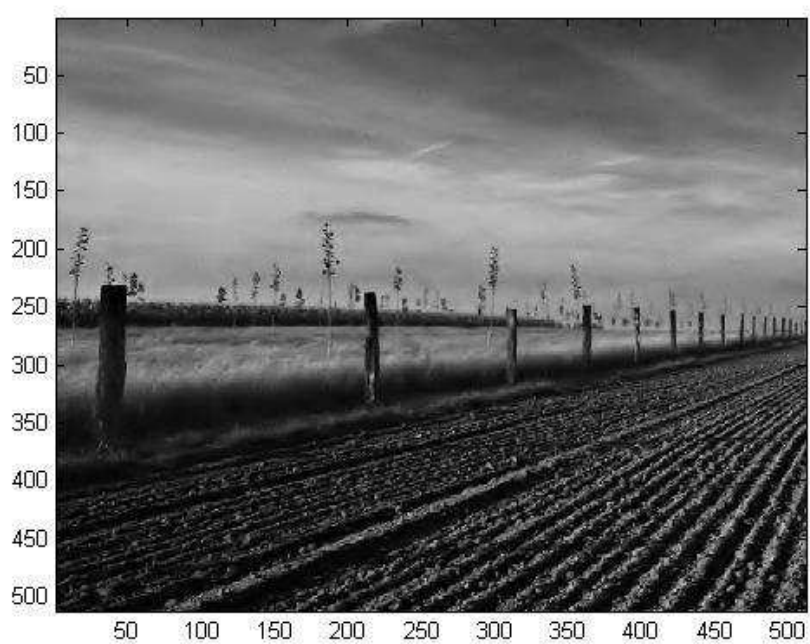


Figure 5. (a) : signal original; (b) : signal bruité; (c) : estimation par ondelettes avec un seuillage fort.

2.1.2. Filtrage du bruit de Poisson : Anscombe et Kolaczyk

Il est souvent raisonnable de considérer le bruit de signaux expérimentaux comme étant normalement distribué. Cependant, dans des techniques où la détection implique un processus de comptage (comme en spectroscopie : les données correspondent au *nombre* d'électrons détectés comme une fonction de leur perte d'énergie), les données sont mieux modélisées par une distribution de Poisson. Malgré son grand nombre d'applications, peu d'études ont été réalisées sur le filtrage du bruit de Poisson. Ceci est probablement dû à la difficulté d'étudier des signaux avec une variance non constante.

Modèle

Nous considérons un processus de Poisson (p.p.) non-homogène sur $[a, a + np]$ ($a \in \mathbb{R}, n \in \mathbb{N}, p \in \mathbb{R}$) :

$$N_t \equiv N(a, t] \sim Po(\Lambda((a, t])) \quad (21)$$

où $\Lambda((a, t]) = \int_a^t \lambda(s) ds \quad \forall t \in [a, a + np]$ et λ est l'intensité du p.p. Considérons que ce processus est observé à intervalles de taille p . Ces observations peuvent être considérées comme un ensemble de comptage cumulatif : $N_a, N_{a+p}, \dots, N_{a+np}$. Nous notons alors $X_i = N_{a+(i+1)p} - N_{a+ip}$ ($i = 0, \dots, n-1$). Celles-ci sont des variables aléatoires de Poisson indépendantes : $X_i \sim Po(\lambda_i)$ où $\lambda_i = \int_{a+ip}^{a+(i+1)p} \lambda(t) dt$ doit être estimé. Ceci explique pourquoi le comptage de particules suit habituellement une distribution de Poisson. La distribution de Poisson ne possède qu'un seul paramètre, λ , et est notée $Po(\lambda)$.

Algorithme d'Anscombe

L'algorithme d'Anscombe transforme les données X_i en utilisant la transformation de Anscombe $Y_i = 2\sqrt{X_i} + 3/8$. Celle-ci rend les données quasi gaussiennes avec un niveau de bruit relativement constant de 1 (STARCK, MURTAGH, BIJAOU, 1998). Ensuite, il procède comme si les données possédaient réellement un bruit gaussien. La méthode de Donoho et Johnstone avec le seuil $T = \sqrt{2 \log(n)}$ est utilisée. Cet algorithme est critiqué pour son lissage excessif aux fines échelles et son lissage timide aux grandes échelles. Ceci est illustré à la figure 7.

Algorithme TIPSH

Kolaczyk a développé un algorithme visant à fournir une alternative à l'algorithme d'Anscombe, finement adaptée aux signaux plongés dans un bruit de Poisson, plus particulièrement les signaux correspondants aux explosions de rayons Gamma. Kolaczyk a proposé d'étendre la solution de Donoho-Johnstone au cas de Poisson. De façon semblable au filtrage d'un bruit normalement distribué (voir équation 14), nous savons que $\hat{\lambda}$, estimateur de λ , peut s'écrire :

$$\hat{\lambda} = \sum_{j,k} \langle X, \psi_{j,k} \rangle \theta_{j,k} \psi_{j,k}, \quad (22)$$

où $\{\psi_{j,k}\}_{j,k}$ représente l'ensemble des fonctions formant une base orthonormée d'ondelettes. Kolaczyk généralise l'estimateur par seuillage de Donoho-Johnstone en imposant

$$\theta_{j,k} = \begin{cases} 1 & \text{si } |\langle X, \psi_{j,k} \rangle| > T \\ 0 & \text{sinon.} \end{cases} \quad (23)$$

Kolaczyk trouve un seuil dépendant de l'échelle j qu'il note alors t_j :

$$t_j = 2^{-j/2} \{ \log(n_j) + \sqrt{\log^2(n_j) + 2\log(n_j)\lambda^*2^j} \} \quad (24)$$

où $n_j = 2^{J-j}$ et λ^* est un réel constant défini plus tard.

L'algorithme final procède selon les trois mêmes étapes que l'algorithme de Donoho-Johnstone :

1. la décomposition des observations dans une base d'ondelettes de Haar,
2. le t_j -seuillage à chaque échelle des coefficients relatifs aux détails
3. l'application de la transformée de Haar inverse aux coefficients seuillés.

Cependant, les estimateurs utilisant la base de Haar ont tendance à ressembler à une fonction en escalier. Ceci est dû à la nature de l'ondelette de Haar. Cela pourrait être un problème quand λ possède un certain degré de régularité. Kolaczyk suggère alors d'utiliser la transformée en ondelettes de Haar invariante par translation qui évite ce problème. L'algorithme résultant de Kolaczyk est appelée "Translation Invariant Poisson Smoothing using Haar Wavelets", ou TIPSH. Pour plus de détails sur celui-ci, le lecteur pourra se référer aux articles de Kolaczyk (KOLACZYK, 1996 et KOLACZYK, 1997).

Dans ses articles (KOLACZYK, 1996 et KOLACZYK, 1997), Kolaczyk a étudié les signaux d'explosions de rayons Gamma. Ce genre de signal est caractérisé par un fond relativement constant et par des pics abrupts occasionnels. Deux illustrations sont données en figure 6. Kolaczyk a choisi l'hypothèse nulle $\lambda = \lambda^*$ qui correspond au fond sans les pics. C'est toujours le cas avec des explosions de rayons Gamma. Il estime λ^* en prenant la moyenne d'au moins 60% des observations. Ceci est justifié par le fait que ces observations sont une partie du fond. C'est pourquoi elles sont des réalisations de $Po(\lambda^*)$. Kolaczyk a réalisé une simulation pour ses signaux et a déclaré que le MISE est minimisé en choisissant un paramètre L petit quand on utilise le seuillage fort et une valeur pour L moyenne quand on utilise le seuillage doux (dû au biais important dans ce cas). Les figures 8 et 7 présentent certains résultats obtenus avec les algorithmes TIPSH et Anscombe.

Ces deux figures nous permettent de comparer l'algorithme TIPSH et l'algorithme standard basé sur la transformation d'Anscombe. Nous voyons que l'approche standard lisse trop le signal aux fines échelles et pas assez aux échelles grossières. Kolaczyk affirme que son algorithme TIPSH a une plus petite erreur que l'erreur obtenue avec la transformation d'Anscombe, et que le biais est particulièrement réduit. Ces deux arguments rendent TIPSH particulièrement

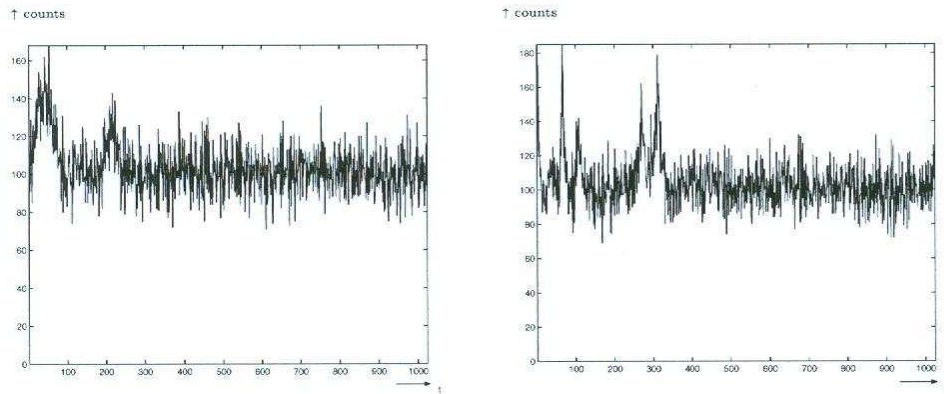


Figure 6. Deux explosions de rayons Gamma (simulation). D'après CHARLES, 2003.

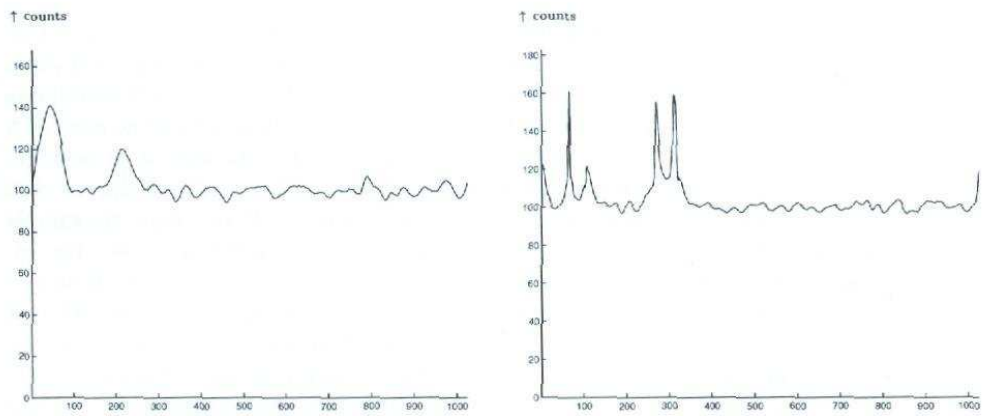


Figure 7. Estimation des fonctions d'intensité correspondant à la figure 6 au moyen de l'algorithme d'Anscombe. D'après CHARLES, 2003.

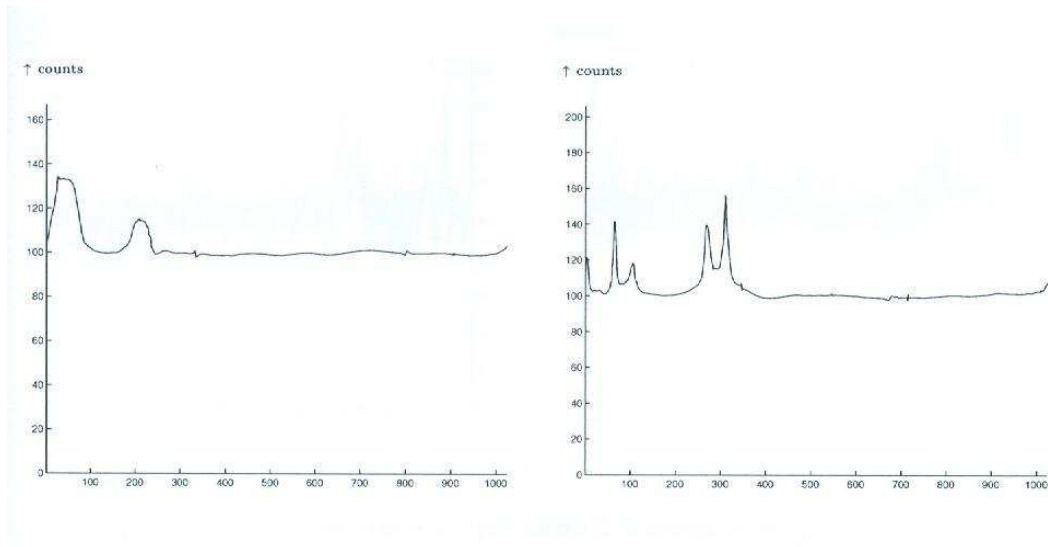


Figure 8. Estimation des fonctions d'intensité correspondant à la figure 6 au moyen de l'algorithme TIPSH. D'après CHARLES, 2003.

séduisant.

2.2. Analyse en ondelettes croisée et cohérence par ondelettes

L'analyse en ondelettes croisée et la cohérence par ondelettes sont deux méthodes permettant de découvrir des liens dans le domaine temps-fréquence entre deux signaux.

La transformée en ondelettes continue d'un signal discret $\mathbf{x} = x(t_n)_{1 \leq n \leq N}$ avec $t_n = n\delta t$ est l'ensemble de ses coefficients d'ondelettes. Ceux relatifs aux détails valent

$$W^{\mathbf{X}}(n, s) = \frac{1}{\sqrt{\frac{s}{\delta t}}} \sum_{i=0}^{N-1} x(t_i) \psi^*\left(\frac{i-n}{\frac{s}{\delta t}}\right)$$

où * indique le complexe conjugué lorsqu'on travaille avec une ondelette complexe. Dans l'analyse en ondelettes croisée ou la cohérence par ondelettes, la transformée en ondelettes continue est souvent utilisée avec l'ondelette de Morlet (qui est complexe).

2.2.1. Spectre de puissance en ondelettes

Le spectre de puissance en ondelettes d'un signal x est défini comme

$$P^X(n, s) = |W^X(n, s)|^2.$$

Il permet de quantifier l'importance de la variabilité du signal expliquée par l'ondelette à chaque pas de temps et à chaque échelle.

S. Jenouvrier de l'équipe "Ecologie des oiseaux et des mammifères marins" (CNRS) (JENOUVRIER, 2004) a étudié, au moyen du spectre de puissance en ondelettes, les périodicités des variations de la taille de la population de trois espèces d'oiseaux marins (Figure 9). Le spectre de puissance se retrouve dans la figure 10.

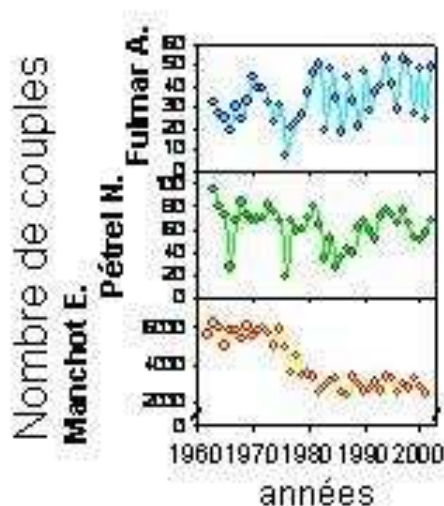


Figure 9. Taille de la population de trois espèces d'oiseaux marins. D'après JENOUVRIER, 2004.

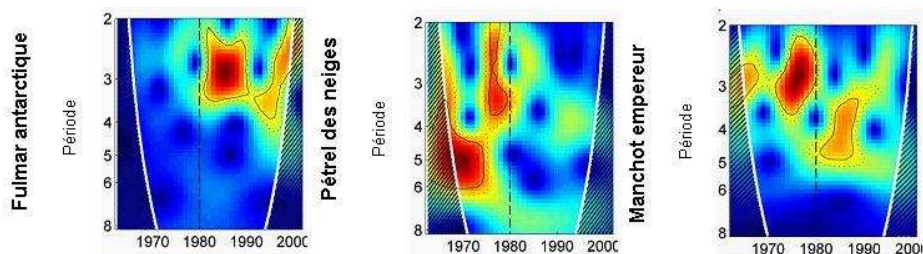


Figure 10. Spectres de puissance correspondant à la figure 9. D'après JENOUVRIER, 2004.

Habituellement, les scientifiques utilisent un code de couleur. Les couleurs orangées signifient que la série temporelle étudiée présente des fortes variations dans les périodicités correspondantes. A l'inverse, les couleurs bleutées reflètent de faibles variations de la série temporelle dans les périodes considérées. Afin de quantifier si les périodicités mises en évidence dans les régions orangées sont significatives, et non pas obtenues par le simple fait du hasard, on utilise des méthodes de ré-échantillonnage (méthodes bootstrap). Les périodicités significatives sont indiquées par une ligne noire pour le seuil 5 pourcent (c'est à dire que l'on a 5 pourcent de risque que la périodicité observée soit le fait du hasard) et par une ligne noire pointillée pour le seuil 10 pourcent. Ce sont ces lignes qui sont importantes à visualiser et que l'on remarque sur la figure 10.

Les périodicités des variations du nombre de couples de manchot empereur, de pétrel des neiges et de fulmar antarctique, sont similaires (autour de 3 et 5 ans) et ne sont pas constantes au cours du temps. L'analyse d'ondelettes met en évidence des changements brusques de périodicité autour de 1980. Ces résultats suggèrent un changement de régime de l'environnement à la fin des années 1970. Pendant ce changement de régime, des fortes anomalies chaudes auraient profondément affecté les populations d'oiseaux marins. Un mécanisme probable peut être une modification de la périodicité du cycle de la glace qui entraînerait une diminution importante des proies consommées par ces oiseaux, et par des effets en cascade dans la chaîne alimentaire, des changements brusques des populations de prédateurs supérieurs.

Considérons un autre exemple tiré de GRINSTED, MOORE, JEVREJEVA, 2004. Les images ont été reproduites via le package MatLab disponible à

www.pol.ac.uk/home/research/waveletcoherence.

Nous désirons examiner le possible lien entre l'étendue de la glace dans la mer Baltique et l'échange de masse atmosphérique entre l'Arctique et le Nord Atlantique. Pour cela, nous disposons de deux mesures : l'oscillation arctique (AO) qui caractérise l'échange de masse atmosphérique décrit ci-dessus et l'étendue de glace maximum annuel (BMI). Ces deux séries sont illustrées en figure 11.

Leur spectre de puissance respectif se trouve dans la figure 12. Les contours noirs désignent le seuil de 5 pourcent et le cône d'influence où les effets de bords peuvent se faire sentir est estompé. Il y a des caractéristiques communes dans les deux spectres de puissance comme le pic significatif dans la période 5 ans autour de 1940. Les deux séries ont aussi une haute puissance dans la période 2-7 ans dans l'intervalle de temps 1860-1900, bien que pour l'AO la puissance ne dépasse pas le niveau de 5 pourcent. Il y a également une haute puissance dans la période 8-16 ans dans les années 1950-2000. Cependant, la similitude entre ces deux schémas est assez faible et il est difficile de dire si ce qu'il y a en commun résulte d'une simple coïncidence. Ceci justifie l'intérêt de l'analyse en ondelettes croisée développée ci-dessous.

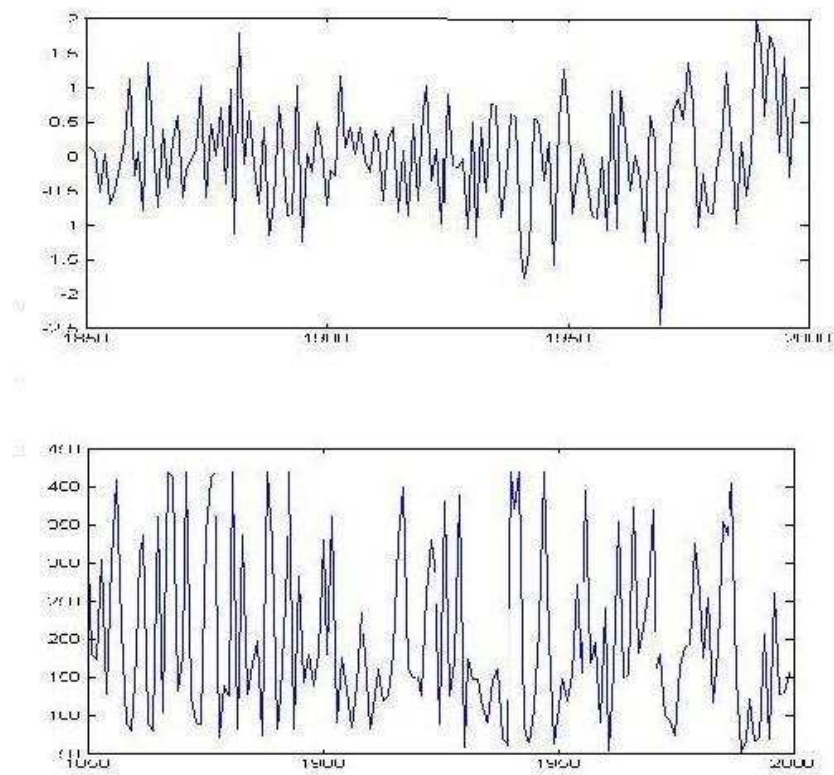


Figure 11. Séries temporelles AO (au-dessus) et BMI (en-dessous).

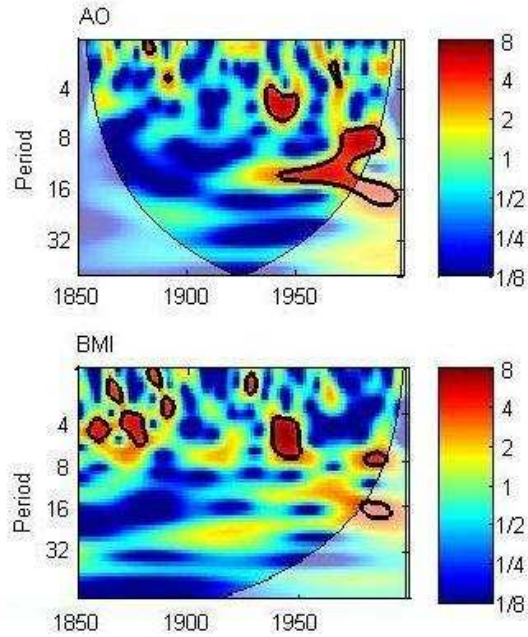


Figure 12. Spectres de puissance des séries temporelles AO et BMI.

2.2.2. Analyse en ondelettes croisée

L'analyse en ondelettes croisée est utilisée pour analyser le lien entre deux signaux à partir d'un spectre de puissance *commun*. L'analyse en ondelettes croisée de deux signaux $\mathbf{x} = x(t_n)$ et $\mathbf{y} = y(t_n)$ est définie par

$$W^{\mathbf{XY}}(n, s) = W^{\mathbf{X}}(n, s)W^{\mathbf{Y}*}(n, s)$$

où $W^{\mathbf{Y}*}(n, s)$ est le complexe conjugué de $W^{\mathbf{Y}}(n, s)$. La puissance en ondelettes croisée est dès lors définie par $|W^{\mathbf{XY}}(n, s)|$. Il faut cependant faire attention qu'un coefficient de la puissance en ondelettes croisée peut être élevé car le spectre de puissance en ondelettes des deux signaux est élevé ou que le spectre de puissance en ondelettes d'un seul signal est très élevé.

La différence de phase entre les deux signaux est définie par

$$\phi(n, s) = \arctan\left(\frac{I(S(\frac{W^{\mathbf{XY}}(n, s)}{s}))}{R(S(\frac{W^{\mathbf{XY}}(n, s)}{s}))}\right).$$

où S est un opérateur de lissage ressemblant à l'ondelette mère.

L'analyse en ondelettes croisée de l'AO et du BMI est visible dans la figure 13. La phase est représentée par des flèches (orientée à droite : en phase, orientée

à gauche : antiphase). Nous voyons que les caractéristiques communes trouvées par le spectre de puissance se retrouvent ici comme étant significatives. Pour conclure à une relation de cause à effet, il faut que les phénomènes enregistrés soient en phase ou en antiphase. Cela nous rassure donc de voir que tous les secteurs avec une puissance commune significative sont en antiphase. Ainsi, BMI pour une grande partie reflète AO. En dehors des régions à puissance significative, nous avons beaucoup d'antiphase. Nous spéculons donc qu'il y a un lien plus fort entre l'AO et le BMI que ce que nous montre l'analyse en ondelettes croisée.

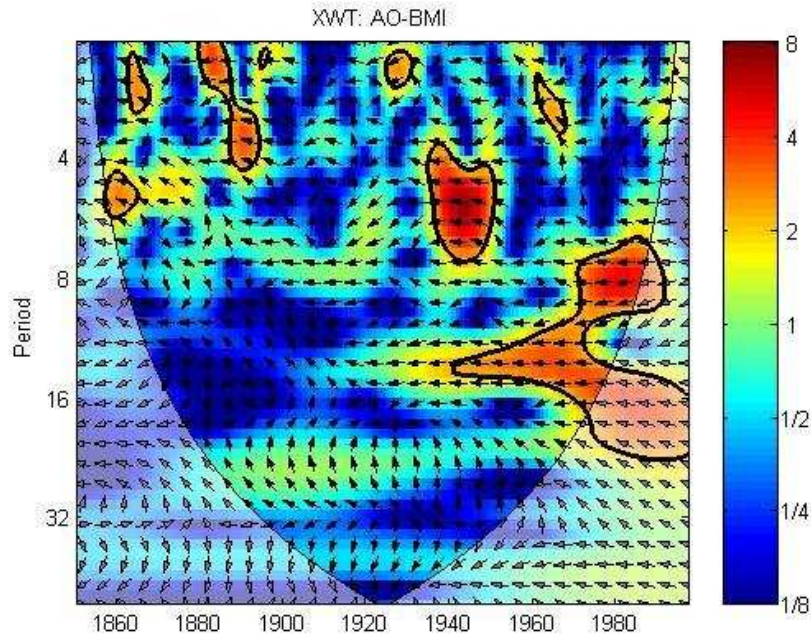


Figure 13. Analyse en ondelettes croisée des séries AO et BMI.

2.2.3. Analyse de cohérence par ondelettes

L'analyse de cohérence par ondelettes a le même but que celui de la puissance en ondelettes croisée. Elle trouve des régions dans l'espace temps-fréquence où les deux signaux covarient mais n'ont pas nécessairement une haute puissance commune. La cohérence par ondelettes de deux signaux est décrite par

$$R^2(n, s) = \frac{|S(\frac{W^{XY}(n, s)}{s})|^2}{S(\frac{W^X(n, s)}{s})S(\frac{W^Y(n, s)}{s})}$$

où S est un opérateur de lissage ressemblant à l'ondelette mère. Cet opérateur est important sinon $R^2(n, s) = 1$. $R^2(n, s)$ donne une valeur comprise entre 0 et

1 qui fournit une information sur la relation entre les deux signaux. Des résultats différents pour l'analyse en ondelettes croisée et pour l'analyse de cohérence permet d'identifier des régions dans l'espace temps-fréquence avec des probables petites puissances. Les régions de puissance commune petite sont des régions avec une petite analyse croisée mais avec une haute cohérence.

La racine carrée de la cohérence de l'AO et du BMI est montrée en figure 14. En comparaison avec l'analyse croisée, une plus large région est considérée comme significative et toutes ses régions montrent une antiphase. Les oscillations de l'AO sont manifestées dans le BMI de 2-20 ans, suggérant que le BMI reflète l'AO.

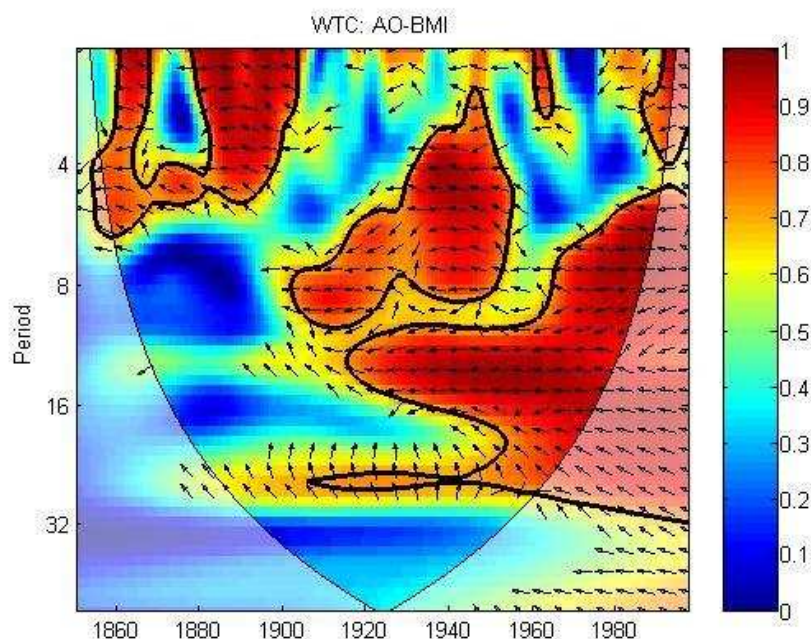


Figure 14. Cohérence de l'AO et du BMI.

2.3. Utilisation d'ondelettes pour la déconvolution

Dans beaucoup de techniques spectroscopiques, les spectres peuvent être modélisés comme la convolution bruitée d'une fonction instrumentale avec un vrai signal à estimer. En principe, cette estimation du vrai signal pourrait être trouvée à travers la déconvolution.

Une approche bien connue pour la déconvolution est basée sur la transformée de Fourier MALLAT, 1988. Le théorème de convolution affirme que si le

signal enregistré $f = g * h$ (g convolué par h), alors $F = GH$ où F , G et H sont les transformées de Fourier de f , g et h . Si h est le signal à retrouver et g la fonction instrumentale, ceci permet que h soit la transformée de Fourier inverse de (F/G) . Cette technique ne fonctionne vraiment pas bien quand G a de très petites valeurs, comme montré dans la figure 15. Dans cet exemple, les très petites valeurs mises à zéro par l'ordinateur ont été mises à la plus petite valeur de l'ordinateur différente de zéro par notre algorithme. De plus, cela fonctionne vraiment mal sur des signaux expérimentaux car cela agit souvent comme un amplificateur de bruit et dès lors cela requiert un filtrage drastique comme vu dans la figure 16.

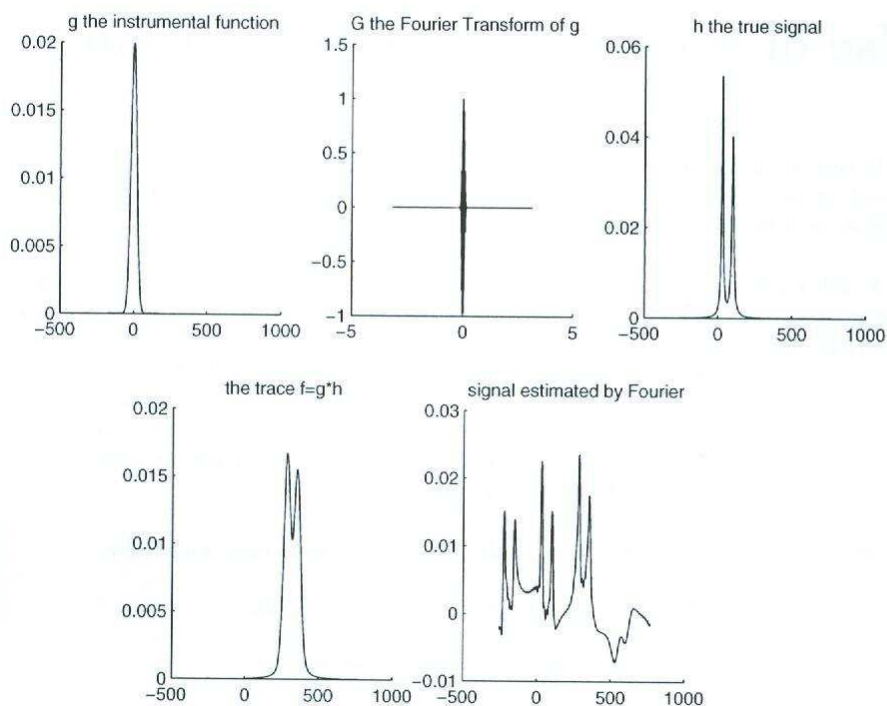


Figure 15. Modèle : $f = g * h$. Mauvaise déconvolution par Fourier due aux très petites valeurs de G (la transformée de Fourier de g). Le vrai signal h n'est pas retrouvé. D'après CHARLES, 2003.

Dans le domaine des ondelettes, nous avons une formule similaire. Dans le cas continu, nous obtenons

$$Wf(u, s) = (g * Wh(., s))(u); \quad Lf(u, s) = (g * Lh(., s))(u).$$

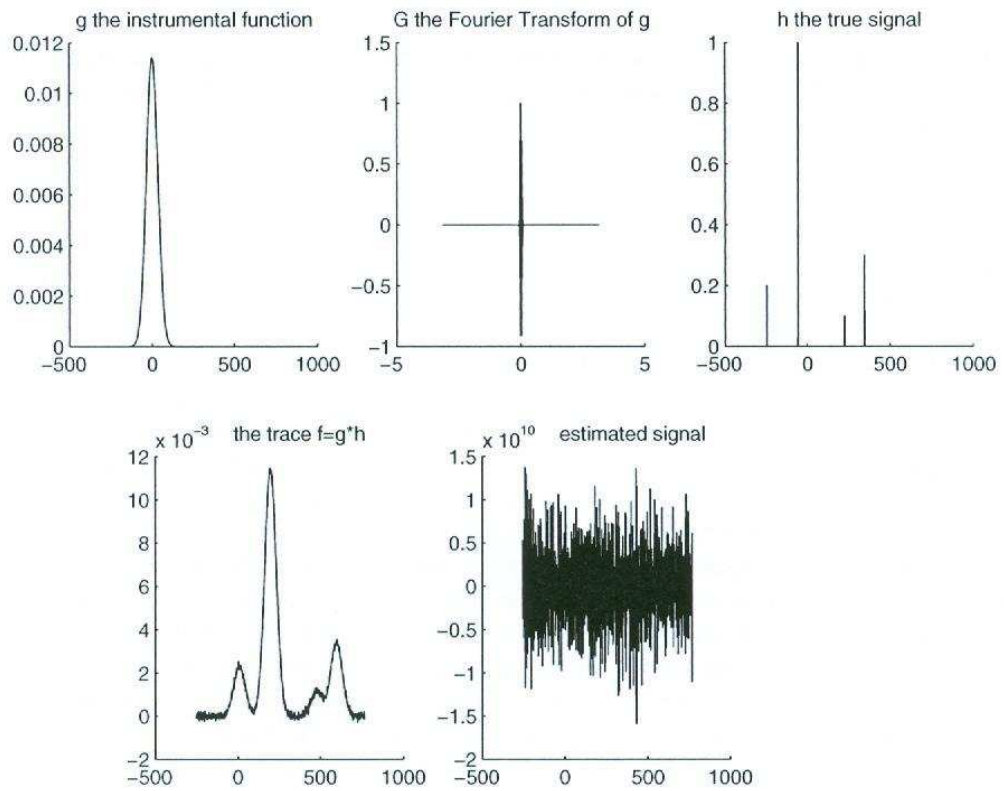


Figure 16. Modèle : $f = g * h + W$ où W est le bruit. Mauvaise déconvolution par Fourier due aux petites valeurs de G (la transformée de Fourier de g) et dû au bruit. Le vrai signal h n'est pas retrouvé. D'après CHARLES, 2003.

Dans le cas discret, utilisant une transformation en ondelettes non-orthogonales, nous avons

$$d_{(j,k)}^f = (g * d_{(j,\cdot)}^h)(k); \quad c_{(j,k)}^f = (g * c_{(j,\cdot)}^h)(k). \quad (25)$$

Malheureusement, nous pouvons voir que, contrairement à la transformée de Fourier où le produit de convolution se transforme en un simple produit, le produit de convolution est préservé par la transformée en ondelettes.

Cependant, un résultat théorique existe dans le cas d'une convolution par une gaussienne. Nous supposons que $f(t) = (f_0 * g_\sigma)(t)$ où g_σ est une gaussienne de variance σ^2 :

$$g_\sigma(t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{t^2}{2\sigma^2}}.$$

Soit $\psi = (-1)^n \theta^{(n)}$ avec $\theta(t) = \lambda e^{-\frac{t^2}{2\sigma^2}}$. Nous avons

$$Wf(u, s) = \left(\frac{s}{s_0}\right)^{n+\frac{1}{2}} Wf_0(u, s_0) \quad \text{avec } s_0 = \sqrt{s^2 + \frac{\sigma^2}{\beta^2}}.$$

Ces formules permettent de retrouver les coefficients d'ondelettes de f_0 et dès lors de retrouver f_0 en utilisant la transformée en ondelettes inverse.

Mais les ondelettes interviennent principalement dans des procédés itératifs de déconvolution. Une solution simple au problème de déconvolution serait d'optimiser les moindres carrés : $\min_h \|f - g * h\|^2$. Cependant, ce problème est mal posé au sens de Hadamard : la solution n'est pas unique et n'est pas stable. De faibles variations du signal observé f vont entraîner de fortes variations du signal restauré h . Une procédure assez standard pour éviter ces instabilités ou pour "régulariser" le problème est de modifier la fonction à minimiser de façon à ce qu'elle contienne l'erreur mais aussi une certaine connaissance a priori de ce que pourrait être la solution. Si, par exemple, on sait que h doit avoir une petite α -norme, la fonction à minimiser devient

$$\|f - g * h\|^2 + \mu \|h\|_\alpha^2$$

où μ est une constante positive appelée paramètre de régularisation. La solution basée sur les ondelettes est de minimiser

$$\|f - g * h\|^2 + \sum_{j,k} \mu_{(j,k)} \|d_{j,k}\|^p$$

où $\{d_{j,k}\}_{j,k}$ représente l'ensemble des coefficients en ondelettes de la fonction h . Si on sait que la fonction h possède une représentation en ondelettes creuse dans une certaine base orthonormée d'ondelettes, alors $\sum_{j,k} \mu_{(j,k)} \|d_{j,k}\|^p$ sera petit. Le problème de déconvolution est donc transformé en un problème d'optimisation qui est résolu de façon itérative. La figure 17 illustre ce principe. Pour plus d'information, voir DAUBECHIES, DEFRISE, DEMOL, 2004.

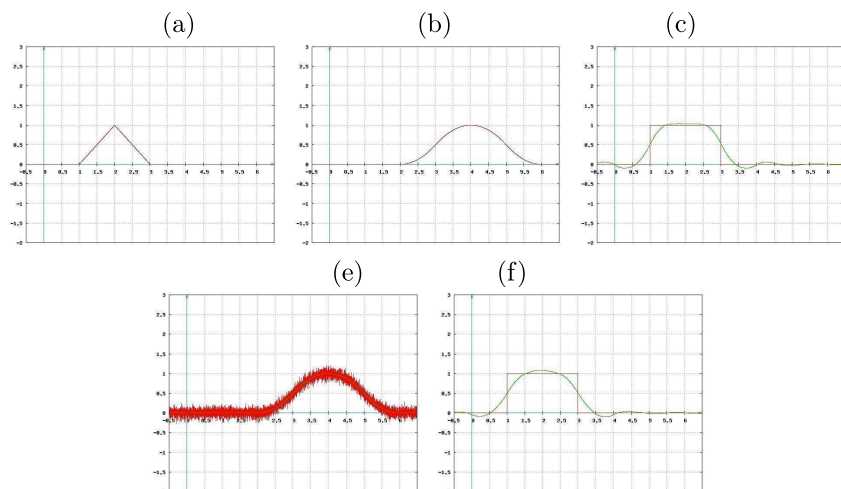


Figure 17. (a) g ; (b) f ; (c) h et h estimé; (e) f bruité; (f) h et h estimé pour f bruité.

2.4. Utilisation des ondelettes pour la compression

Réduire un litre de jus de fruit à quelques grammes de poudre concentrée est typiquement une compression avec perte. Le goût de la boisson reconstituée est semblable au goût du jus de fruit initial mais a souvent perdu de sa finesse. Dans la compression de signaux ou d'images, nous sommes confrontés au même compromis entre qualité et compression. Les applications principales concernent le stockage de données et la transmission à travers des canaux à bande passante limitée (MALLAT, 1988).

Un codeur par transformée décompose un signal dans une base orthogonale et quantifie les coefficients de décomposition, c'est-à-dire que ces coefficients sont remplacés par des valeurs proches mais appartenant à un ensemble déterminé. La quantification est donc l'étape dans laquelle on perd réellement de l'information mais qui fait gagner beaucoup de place. La distortion du signal reconstitué est minimisée par une optimisation de la quantification, de la base et de l'allocation de bits. Pour information, une image en niveaux de gris a typiquement 512^2 pixels, chacun codé sur 8 bits. Actuellement, les meilleurs algorithmes de compression d'images sont des codages par transformée, sur des bases de cosinus (jpeg standard) ou d'ondelettes (jpeg2000). L'efficacité de ces bases vient de leur capacité à construire des approximations non linéaires précises à l'aide de quelques vecteurs non nuls. Avec moins de 1 bit par pixel, on reconstruit des images visiblement parfaites. A 0.25 bit par pixel, l'image reste de bonne qualité. A titre d'exemple, la vitesse de compression d'une image $60 \times 480 \times 24$ sur un pentium est de 3s pour le format jpeg standard et de 1s pour le format jpg 2000.

En illustration, l'image 18 est compressée suivant le jpeg standard (figure 19) et suivant jpeg2000 (figure 20). Les deux images compressées ont la même taille (16k). On remarque sur la figure 19 des artefacts de pixels 8x8 un peu partout. Le résultat est également assez mauvais autour du texte. Le ratio moyen de compression pour un jpeg standard est de 1 :25 alors que pour un jpeg2000 il est de 1 :50.



Figure 18. Image originale (<http://www.fnordware.com/j2k/jp2samples.html>).



Figure 19. Image de la figure 18 compressée avec jpeg.



Figure 20. Image de la figure 18 compressée avec jpeg2000.

2.5. Autres utilisations des ondelettes

Les ondelettes ont de nombreuses autres applications qui n'ont pas été expliquées dans cette note. Elles peuvent par exemple aider dans la classification de signaux. Si le but est de classifier des signaux monodimensionnels selon des classes de signaux invariantes par translation, la classification peut se faire par arbre de décision où le dictionnaire de questions est :

$$Q_{j,\underline{d},\bar{d},\theta}(X) = 1$$

\Leftrightarrow il existe deux extrema locaux de la transformée à l'échelle j tels que

$$\left\{ \begin{array}{l} \underline{d} \leq |u_i - u_k| \leq \bar{d} \\ \min\{|\langle \psi_{j,u_i}, X \rangle|, |\langle \psi_{j,u_k}, X \rangle|\} \geq \theta. \end{array} \right.$$

Si le but est la segmentation d'une image, elle peut se faire en suivant les étapes :

1. extraction (au moyen d'ondelettes) de vecteurs de caractéristiques de l'image sur des fenêtres aléatoires,
2. itération du calcul des vecteurs jusqu'à stabilisation du réseau de neurones,
3. K-Means non supervisé sur les vecteurs.

Les ondelettes peuvent aussi être utiles dans la détection *automatique* de singularités d'un signal grâce aux lignes de modulus maxima. Elles peuvent être utiles dans la détection *automatique* des contours dans une image par une version multiéchelle de l'algorithme de Canny. Elles sont également utilisées avec efficacité dans le domaine des séries temporelles, de l'interpolation, etc.

3. EN GUISE DE CONCLUSION

Dans cette note technique, nous avons développé diverses applications de base de la transformée en ondelettes. La régression et l'estimation de la densité par ondelettes semblent être une bonne alternative à la méthode des noyaux. L'efficacité des ondelettes dans les méthodes de filtrage est incontestable, que ce soit pour un modèle gaussien ou un modèle poissonien. L'analyse cohérente par ondelettes est utilisée pour examiner le lien possible entre deux signaux. En déconvolution, les ondelettes sont efficaces lorsque le problème est transformé en un problème d'optimisation. Quant à la compression via les ondelettes, son efficacité est universellement reconnue car les ondelettes sont à la base du format jpeg2000. Les ondelettes sont donc un outil qui peut être utilisé dans un grand nombre de domaines différents. Son succès provient non seulement de son analyse temps-fréquence mais aussi de ses algorithmes rapides.

BIBLIOGRAPHIE

- ADRIAN P., ANAND R. [2008]. *Maximum Likelihood wavelet density estimation with application to image and shape matching*. IEEE Transaction on image processing, 17(4) :458-468.
- CENCOV N.N. [1962]. *Evaluation of an unknown distribution density from observations*. Doklady, 3 :1559-1562.
- CHARLES C. [2003]. *Some wavelet applications to signal and image processing*. PhD Thesis. FUNDP.
- CHARLES C., LECLERC G., PIREAUX J.-J. RASSON J.-P. [2003]. *Wavelets applications in surfaces sciences*. Surface and Interface analysis.
- COCQUEREZ J.P., PHILIPP S. [1995]. *Analyse d'images : filtrage et segmentation*. Masson.
- DAUBECHIES I., DEFRISE M., DE MOL C. [2004]. *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*. Communications on Pure and Applied Mathematics, vol. LVII, 1413-1457.
- DELOUILLE V. [2002]. *Nonparametric stochastic regression using design-adapted wavelets*. Thèse de doctorat, Université Catholique de Louvain.
- DONOHO D.L., COIFMAN R.R. [1995]. *Translation-invariant denoising*. Wavelets and Statistics, A. Antoniadis and G. Oppenheim, Springer-Verlag.
- DELYON B., JUDITSKY A. [1993]. *Wavelet estimators, global error measures : revisited*. Technical report, irisa-inria.
- DONOHO D.L., JOHNSTONE I.M. [1994]. *Ideal spatial adaptation via wavelet shrinkage*. Biometrika, 81 :425-455.
- DONOHO D.L., JOHNSTONE I.M., KERKYACHARIAN G., PICARD D. [1996]. *Density estimation by wavelet thresholding*. The annals of statistics, 24 :508-539.
- DONOHO D.L. [1996]. *Denoising via soft thresholding*. IEEE Trans. Inf. Theory, 41 :613-627.

- GRINSTED A., MOORE J.C., JEVREJEVA S. [2004]. *Application of the cross wavelet transform and wavelet coherence to geophysical time series*. *Nonlinear Processes in Geophysics*, 11 :561-566.
- HUANG W. [2003]. *Wavelet regression with an emphasis on singularity detection*. Master of Science. Texas.
- IZENMAN A. [1991]. *Recent developments in nonparametric density estimation*. *JASA* 86, 413 :205-224.
- JENOUVRIER S. [2004]. *Influence de la variabilité environnementale sur les stratégies démographiques des populations de prédateurs supérieurs : la communauté d'oiseaux marins en antarctique*. Thèse de doctorat. Université Paris 6.
- JUDITSKY A. [1994]. *Wavelet estimators : adapting to unknown smoothness*. IRISA Publication interne, 815.
- KOLACZYK E.D. [1994] *Wavelet methods for the inversion of certain homogeneous linear operators in presence of noisy data*. PhD thesis, Stanford University.
- KOLACZYK E.D. [1996] *Estimation of intensities of burst-like poisson processes using haar wavelets*. *Biometrika*, 46 :352-363.
- KOLACZYK E.D. [1997] *Non-parametric estimation of gamma-ray burst intensities using Haar wavelets*. *The Astrophysical Journal*, 483 :340-349.
- KOLACZYK E.D. [1998] *A method for wavelet shrinkage estimation of certain poisson intensity signals using corrected thresholds*. *Statistica Sinica*, :119-135. *Biometrika*, 46 :352-363.
- MALLAT S. [1988]. *A wavelet tour of signal processing*. Academic Press.
- MULLER P., VIDAKOVIC B. [1998]. *Bayesian Inference with wavelets : Density Estimation*. *Journal of Computational and Graphical Statistics*, vol. 7, 4 :456-468.
- NOWAK R.D., BARANIUK R.G. [1997]. *Wavelet-domain filtering for photon imaging systems*. *Proc. SPIE, Wavelet Applications in Signal and Image Processing V*, 3169 :55-66.
- NOWAK R.D., HELLMAN R., NOWAK D., BARANIUK R.G. [1996] *Wavelet-domain filtering for nuclear medicine imaging*. *Proc. SPIE Imaging Conf.*, pages 279-290.
- PINHEIRO A., VIDAKOVIC B. [1997]. *Estimating the square root of a density via compactly supported wavelets*. *Computational statistics and data analysis*, 25 :399-415.
- RENAUD O. [1999]. *Density estimation with wavelets : variability, invariance and discriminant power*. PhD Thesis. Ecole polytechnique fédérale de Lauzanne.
- SILVERMAN B.W. [1986]. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.
- STARCK J.L., MURTHAG F., BIJAOUI A. [1998]. *Image Processing and Data analysis. The multiscale approach*. Cambridge University Press.
- TIMMERMANN K.E., NOWAK R.D. [1999]. *Multiscale modeling and estimation of poisson processes with application to photon-limited imaging*. *IEEE Transactions Information Theory, Special Issue on Multiscale Signal Analysis and Its Applications*, 45 :846-862.

TRIBOULEY K. [2008]. *Practical estimation of multivariate densities using wavelet methods*. *Statistica Neerlandica*, 19(1) :41-62.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant de l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de l'Université de Liège/Gembloux Agro-Bio Tech et du Département Agriculture et milieu naturel, Unité Systèmes agraires, Territoire et Technologies de l'Information du Centre wallon de Recherches agronomiques.

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

*Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLoux (Belgique)
E-mail : sima.gembloux@ulg.ac.be*

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

<http://www.fsagx.ac.be/si/>

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

- PALM R. [2007]. Etude des séries chronologiques par les méthodes de lissage. *Notes Stat. Inform.* (Gembloux) 2007/1, 22 p.
- PALM R. [2007]. L'analyse des correspondances multiples : principes et application. *Notes Stat. Inform.* (Gembloux) 2007/2, 28 p.
- PALM R. [2008]. Détermination de la répétabilité et de la reproductibilité d'une méthode de mesure normalisée selon la norme ISO 5725-2. *Notes Stat. Inform.* (Gembloux) 2008/1, 22 p.
- CHARLES C., LECHARLIER L., RENAUD F. [2008]. Introduction à LATEX. *Notes Stat. Inform.* (Gembloux) 2008/2, 21 p.
- CHARLES C. [2008]. Introduction à OCTAVE. *Notes Stat. Inform.* (Gembloux) 2008/3, 19 p.
- PALM R., BROSTAUX Y. [2009]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 2009/1, 17 p.
- CHARLES C. [2011]. Introduction aux ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/1, 22 p.