Université Aix-Marseille 2018-2019
Master Mathématiques appliquées et statistiques 2ème année

Représentation parcimonieuse des signaux

TP : Parcimonie et bases d'ondelettes.
Signaux parcimonieux dans une base d'ondelettes-Application au débruitage d'un signal

Le débruitage est l'opération qui consiste à restaurer des données qui ont subi une perturbation aléatoire lors de leur acquisition ou lors d'une transmission, ou à un autre moment de la chaine de traitement. Par exemple si on effectue un scanner à rayons X (une personne est placée dans un compartiment fermé et est soumise à un balayage de rayons X), ou une mesure à l'aide d'un microscope les détecteurs ne comptent jamais exactement le nombre de photons qui les frappent, et les mesures sont donc entachées d'une erreur qui est en général modélisée par un processus aléatoire.

Quand on doit débruiter des données on a donc affaire à des observations y d'un signal d'origine x qui a subi une perturbation. On supposera dans tout ce travail que l'on connait (ou en tous cas sait modéliser) cette pertubation, et en particulier son intensité à l'aide d'un paramètre α. Cela nous donne ainsi

$$y = \mathcal{D}_\alpha(x) \tag{1}$$

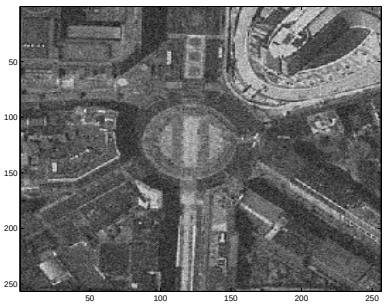
où $\mathcal{D}_\alpha$ modélise notre perturbation. Dans tout ce que nous ferons nos données sont supposées être des vecteurs de $\mathbb{R}^N$.

Un cas classique est le cas du bruit additif gaussien i.i.d, qui est utilisé pour modéliser les bruits dûs aux dispositifs électroniques dans les appareils de mesure, le bruit d'agitation thermique et aussi les perturbations lors de transmissions (dans des canaux terrestres, ou par satellite)... Le modèle (1) s'écrit alors dans ce cas particulier

$$y = x + b \tag{2}$$

avec $b = (b_i)_{i=0,\dots N-1}$ réalisation de $B = (B_i)_{i=0,\dots N-1}$ vecteur aléatoire composé de N variables aléatoires gaussiennes indépendantes centrées de variance $\alpha = \sigma^2$. On dit que B est un bruit blanc gaussien centré de variance σ^2.

Image originale



$\sigma = 20$

Un autre cas classique est celui de données perturbées par un bruit de Poisson (c'est comme on l'a dit un bruit qui est utilisé pour décrire le comptage de photons sur un détecteur). On a alors

$$y = \mathcal{P}(\alpha x) \tag{3}$$

où $\mathcal{P}(\alpha x) = (\mathcal{P}(\alpha x_0), \dots, \mathcal{P}(\alpha x_{N-1})$ et $\mathcal{P}(\alpha x_i), i = 0, \dots N-1$ est la réalisation de N variables aléatoires indépendantes qui suivent chacune une loi de Poisson de paramètre αx_i. En prenant α très proche de zéro on a ainsi la situation où les données proviennent d'un signal de très faible intensité.

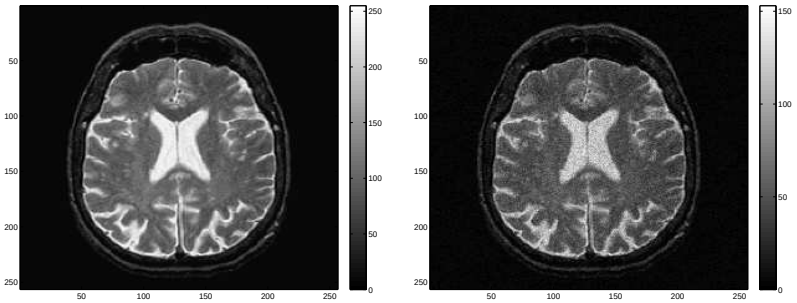
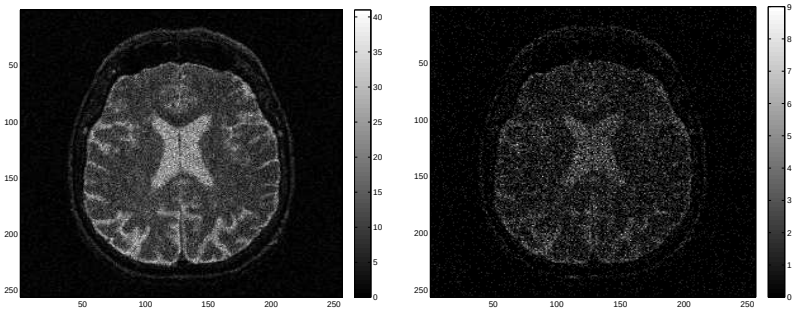


Image originale $\alpha = 0.5$



$\alpha = 0,1$ $\alpha = 0,01$

Nous allons nous intéresser au problème du débruitage dans le cas où le bruit est un bruit blanc gaussien centré et dans le cas où les signaux qui nous intéressent sont parcimonieux dans une base d'ondelettes : c'est à dire les signaux réguliers par morceaux.

1 Notations

— On note $\langle ., . \rangle$ le produit scalaire usuel sur $\mathbb{C}^N$ tel que $\langle x, y \rangle = \sum_{n=0}^{N-1} x_n \overline{y_n}$ pour $x = (x_i)_{i=0,\dots N-1} \in \mathbb{C}^N$ et $y = (y_i)_{i=0,\dots N-1} \in \mathbb{C}^N$. On note $|x|_2^2 = \sum_{n=0}^{N-1} |x_n|^2$ pour $x \in \mathbb{C}^N$.

— Si x est un signal dans $\mathbb{R}^N = \mathbb{R}^{2^n}$ et Φ une base d'ondelettes de $\mathbb{R}^N$ à n_0 niveaux de décomposition on note $D_k(x)$ les coefficients dits d'« approximation », ou coefficients d'« échelle »pour $k = 0, \dots 2^{n-n_0} - 1$ et $C_{j,k}(x)$ pour $j = n - n_0, \dots n-1$ et $k = 0, \dots 2^j - 1$ les coefficients en ondelettes de x ou coefficients de « détail ».

2 Parcimonie dans une base orthonormée

Dans cette partie on examine les propriétés de parcimonie des signaux réguliers par morceaux (non bruités) dans les bases d'ondelettes. Nous allons comparer les décompositions d'un signal régulier par morceaux dans différentes bases à l'aide d'histogrammes. Comme bases de décompositions d'un signal régulier nous prendrons la base canonique, la base de Fourier discrète orthonormalisée (voir proposition ci-dessous) et une base d'ondelettes, par exemple la base de Daubechies dont l'ondelette a 4 moments nuls, dite Daubechies 4.

Proposition 1

On considère la famille de vecteurs $\widetilde{\mathcal{E}} = \{\frac{1}{\sqrt{N}} e^\ell \in \mathbb{C}^N, \ell \in \mathbb{Z}\}$ *tels que pour* $\ell \in \mathbb{Z}$ *et pour* $n \in \{0, \dots N-1\}$ *la n-ième coordonnée du vecteur* e^ℓ *s'écrive*

$$e_n^\ell = e^{\frac{2i\pi\ell n}{N}}$$

La famille $\widetilde{\mathcal{E}}$ *est une base orthonormée de* $\mathbb{C}^N$.

2.1 Travaux pratiques

Exercice 1

1. *Charger le signal présent dans le fichier* `signal.txt`*. Normaliser le signal pour que ses valeurs soient dans* $[-1, 1]$ *mais que son allure soit conservée.*
2. *Calculer les coefficients en ondelettes du signal à l'aide de la librairie* `Pywavelet` *en prenant* $n_0 \leq 6$ *et l'ondelette* `'db4'`*.*
3. *Calculer les coefficients du signal dans la base* $\mathcal{E}'$*.*
4. *Effectuer l'histogramme des valeurs absolues des coefficients en ondelettes, du module des coefficients de Fourier, et également celui des valeurs du signal. Commenter.*

   Dans quelle base le signal est-il le plus parcimonieux ?

   On prendra soin d'avoir des histogrammes suffisamment précis pour pouvoir répondre à la question.

3 Débruitage par seuillage

3.1 Bruit blanc gaussien et coefficients en ondelettes

On étudie ici une méthode de débruitage par seuillage des coefficients en ondelettes.

Nous supposons que nous avons à notre disposition des observations y d'un signal d'origine x telles que

$$y = x + b$$

où b est une réalisation d'un bruit blanc gaussien noté $B = (B_i)_{i=0,\dots N-1}$. Nous noterons σ^2 la variance de B_i.

Remarquons d'abord que si nous préférons travailler avec la décomposition en ondelettes du signal sur n_0 niveaux plutôt qu'avec le signal lui-même cela nous donne

$$D_k(y) = D_k(x) + D_k(b) \text{ pour tout } k = 0, \dots 2^{n-n_0} - 1 \tag{4}$$

et

$$C_{j,k}(y) = C_{j,k}(x) + C_{j,k}(b) \text{ pour tout } j = n - n_0, \dots n \text{ et pour tout } k = 0, \dots 2^{n-n_0} - 1 \tag{5}$$

Les variables $D_k(B)$ pour $k = 0, \dots 2^{n-n_0} - 1$, $C_{j,k}(B)$ pour $j = n - n_0, \dots n - 1$ et pour tout $k = 0, \dots 2^{n-n_0} - 1$ sont des variables gaussiennes indépendantes centrées de même variance σ^2 que les variables aléatoires qui constituent B.

En effet nous avons le résultat suivant

Proposition 2

Soit $\{\Phi^0, \dots \Phi^{N-1}\}$ *une base orthonormée de* $\mathbb{R}^N$*.*

Si $B = (B_i)_{i=0,\dots N-1}$ *est un vecteur aléatoire constitué de* N *variables aléatoires gaussiennes centrées i.i.d de variance* σ^2 *alors* $(\langle B, \Phi^k \rangle)_{k=0,\dots N-1}$ *est un vecteur aléatoire gaussien de même loi que* B*.*

Ainsi (3) est équivalent à (4) et (5). Débruiter y revient à débruiter les $C_{j,k}(y)$ qui ont subi une perturbation qui a les mêmes propriétés statistiques que celle subie par le signal d'origine. L'usage est en effet de ne pas toucher aux $D_k(y)$ avec l'hypothèse que si le bruit n'est pas trop fort, vu leur grande amplitude, la perturbation ne les aura que légèrement perturbés.

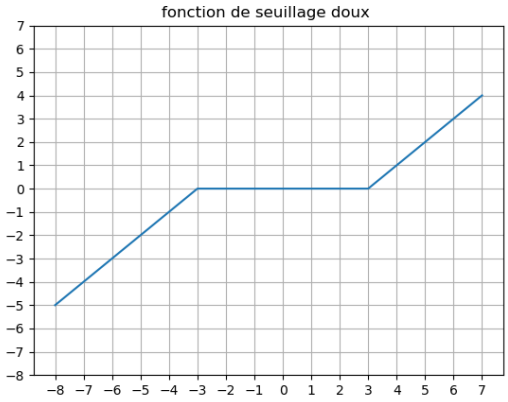


FIGURE 1 – On trace la courbe de la fonction de seuillage doux $\beta = s_T(\alpha)$. Pour quelle valeur de T est-elle ici représentée ?

3.2 Travaux pratiques

Exercice 2 (*Bruitage d'un signal*)

Ecrivez une fonction `addwhitenoise.py` *qui prend en entrée un signal* `x` *et un écart type* σ *et ajoute au signal une réalisation d'un bruit blanc gaussien centré de variance* σ^2. *N'oubliez pas de la commenter pour indiquer ce qu'elle fait.*

On pourra utiliser la bibliothèque `numpy.random` *adéquatement.*

Test : *Tester cette fonction sur différents signaux (par exemple un signal constant par morceaux, et le signal du fichier signal.txt) et pour différentes valeurs de* σ. *Visualiser à chaque fois le signal original, le signal bruité, les coefficients en ondelettes du signal original et ceux du signal bruité.*

Nous allons chercher à débruiter le signal en seuillant les coefficients, c'est à dire en ne gardant que les coefficients au-dessus d'un certain seuil.

Soit $T \in \mathbb{R}^+$ paramètre fixé. On définit la fonction dite de seuillage « doux » s_T définie sur $\mathbb{R}$ comme suit

$$s_T(\alpha) = \begin{cases} \alpha - T\text{sign}(\alpha) & \text{si } |\alpha| - T \geq 0 \\ 0 & \text{sinon} \end{cases}$$

avec $\text{sign}(\alpha) = 0$ si $\alpha = 0$ et $\text{sign}(\alpha) = \frac{\alpha}{|\alpha|}$ sinon (voir figure 1).

On pourrait choisir de faire ce qu'on appelle un seuillage dur c'est à dire de considérer l'opérateur $\tilde{s}_T$ *tel que*

$$\tilde{s}_T(\alpha) = \begin{cases} \alpha & \textit{si } |\alpha| - T \geq 0 \\ 0 & \textit{sinon} \end{cases}$$

Cet opérateur a le désavantage d'être très brutal (il est discontinu comme fonction de α*) : il suffit que le coefficient considéré prenne une valeur un tout petit peu inférieure à* T *pour être mis brutalement à zéro, alors que celui qui aura une valeur légèrement supérieur à* T *ne l'est pas. On lui préferera dans beaucoup de cas l'opérateur* s_T.

Exercice 3

1. *On voudrait programmer une fonction dont l'aide est la suivante*

```
computes the soft thresholding of all the coefficients of a vector x above
a given threshold T, i.e xT satisfies for all available indices k
```

```
xT(k)=0 if |x(k)|<=T
xT(k)=x(k)-T sgn(x(k)) if |x(k)|>T

xT=Softthreshold(x,T) computes the soft thresholded vector xT.
Input is vector x and the threshold parameter T
```

La fonction pywt.threshold de la bibliothèque PyWavelet avec l'option `'soft'` *correspond-elle à ce qu'on cherche ? Si oui illustrer son action sur l'exemple du signal* $x = (-K_0, -K_0+1, \dots 0, 1, 2, \dots K_0-1)$.
Si non programmer une fonction `Softthreshold.py` *qui effectue les opérations.*

2. *Comment l'utiliser pour restaurer les deux Diracs d'un signal qui s'écrit* $y = a_0\delta^{k_0} + a_1\delta^{k_1} + b$, *où* b *est la réalisation d'un bruit blanc gaussien d'écart-type* $\sigma = \frac{\max(|a_0|,|a_1|)}{20}$ *? Illustrer cette situation par une ou des simulations.*

On se propose donc de débruiter le signal en seuillant les coefficients en ondelettes et en conservant les coefficients d'échelle. Nous allons donc calculer une estimation $\tilde{x}$ du signal original x à l'aide d'un opérateur S_T tel que

$$\tilde{x} = S_T(y) = \sum_{k=0}^{2^{n-n_0}-1} D_k(y)\phi_k + \sum_{j=n-n_0}^{n-1}\sum_{k=0}^{2^j-1} s_T\left(C_{j,k}(y)\right)\psi_{j,k}$$

Exercice 4

1. *Programmer une fonction* `thresholdsignal.py` *qui calcule le signal* $\tilde{x}$ *à partir de la donnée d'une observation* y *bruitée, de la donnée d'un seuil* T, *d'une ondelette (par exemple* `'db4'`*) et d'un niveau de décomposition* n_0.

2. *Tester cette fonction pour différentes valeurs de* $n_0 = 1, \dots 6$ *et* $\sigma = 0.01$ *et pour deux ondelettes différentes (par exemple Daubechies 4 et l'ondelette de Haar) avec* $T = 2\sigma$. *Pour quels niveaux* n_0 *le débruitage semble-t-il le plus efficace ? Quelle en est la raison d'après vous ? On pourra faire le même genre de tests avec d'autres valeurs de* σ.

Pour évaluer la performance d'une méthode de restauration d'un signal on calcule souvent le SNR (≪ Signal to noise ratio ≫), ou rapport signal à bruit avec x le signal de référence et $\tilde{x}$ l'estimation calculée, supposement débruitée dans notre cas. La formule est la suivante

$$SNR = -10\log_{10}\left(\frac{\|x-\tilde{x}\|^2}{\|x\|^2}\right)$$

Le programme Python qu'on peut utiliser est le suivant

```
import numpy as np

def snr(xtilde,x):
    s = 10*np.log10(np.mean(x**2)/np.mean((x-xtilde)**2))
    return s
```

Exercice 5

Pour $\sigma = 0.05$ *fixée tracer la courbe du SNR en fonction de* T/σ. *Quel serait le seuil optimal selon le critère du SNR ? Cela correspond-il à votre perception ?*

Un choix couramment recommandé dans la littérature est $T = \sigma\sqrt{2\log_2(N)}$.

3.3 Justification de la stratégie pour débruiter

L'objectif est de débruiter les coefficients en ondelettes et nous pouvons chercher une modélisation de notre problème qui nous permet de justifier la stratégie concrète que nous avons choisie. Remarquons que nous cherchons à restaurer un signal parcimonieux dans une base d'ondelettes. Or le fait que le signal est parcimonieux dans une base d'ondelettes peut être vu comme le fait que ses coefficients sont la réalisation d'un processus aléatoire qui suit une loi de probabilité qui encourage la parcimonie. Un choix très classique est celui de la loi de Laplace.

Nous considérons ainsi que $C_{j,k}(y) = C_{j,k}(x) + C_{j,k}(b)$ correspond à la réalisation d'un processus aléatoire $V = (V_0, \dots V_{K-1})$ avec $V = U + W$ où $U = (U_0, \dots, U_{K-1})$ est un vecteur aléatoire avec les U_i qui sont i.i.d de loi de Laplace de paramètres $(0, L)$ (et dont les coefficients en ondelettes de x seraient des réalisations), et W un bruit blanc centré gaussien indépendant de U (dont les coefficients en ondelettes de b seraient des réalisations).

Nos observations seront notées $(v_0, \dots v_{K-1})$. Nous cherchons à estimer les u_i qui vont maximiser la vraisemblance de nos observations. En effet notons $f(u, v)$ la densité jointe du couple (U_i, V_i) et g_U la densité de U_i alors que g_W est la densité de W_i. Enfin g_V est la densité de V_i. Nous définissons la densité à posteriori

$$f_{V_i/U_i}(u, v) = \frac{f(u, v)}{g_U(u)}$$

Comme $f(u, v) = g_U(u)g_W(v - u)$ nous obtenons $f_{V_i/U_i}(u, v) = g_W(v - u)$.

De même nous posons $f_{U_i/V_i}(u, v) = \dfrac{f(u, v)}{g_V(v)} = \dfrac{g_U(u)g_W(v - u)}{g_V(v)}$.

Rappelons que dans notre problème de débruitage nous avons l'information des valeurs qu'a prises V. Nous estimons ainsi u_i en maximisant $u \mapsto f_{U_i/V_i}(u, v_i)$, c'est à dire en minimisant $u \mapsto -\ln\big(f_{U_i/V_i}(u, v_i)\big)$. Nous avons donc une estimée $\tilde{u}_i$ de u_i en calculant

$$\tilde{u}_i = \arg\min_{u\in\mathbb{R}} \left(\frac{|u - v_i|^2}{2\sigma^2} + \frac{1}{L}|u| \right) = \arg\min_{u\in\mathbb{R}} \left(\frac{|u - v_i|^2}{2} + \frac{\sigma^2}{L}|u| \right) \tag{6}$$

Posons $T = \frac{\sigma^2}{L}$. On a la propriété suivante.

Proposition 3

Soit $T > 0$, et $t_0 \in \mathbb{R}$ Soit $f : t \mapsto \frac{|t-t_0|^2}{2} + T|t|$ définie sur $\mathbb{R}$.
Le minimum de f est atteint pour $t = s_T(t_0)$.

Pour simplifier les notations nous posons $\gamma = (\gamma_m)_m$ avec $\gamma_m = C_{j,k}(y)$ en reindexant les coefficients en ondelettes pour les sommer avec $m = 0, \dots K - 1$ et $K = 2^n - 2^{n-n_0}$. Les coefficients en ondelettes de $\tilde{x} = S_T(y)$ sont solutions du problème

$$\text{Trouver le minimum en } t = (t_m) \text{ de } F(t) = \frac{|t - \gamma|_2^2}{2} + T|t|_1$$

où $|t|_1 = \sum_m |t_m|$

4 Compte-rendu du TP

Le compte-rendu est à rendre par groupe de deux maximum (avec mention sur les documents des membres du groupe) sur Ametice.

Le compte-rendu du TP consiste en une archive contenant

- un fichier .pdf écrit en Latex ou un fichier Note_book contenant le texte (commentaires, démonstrations mathématiques) et les figures du compte-rendu
- les fichiers `.py` qui ont servi pendant la programmation, qui doivent être commentés : on doit dès les premières lignes et tout au cours du programme savoir ce qu'ils font et ce qu'ils calculent ou illustrent.

Si le texte est rendu en .pdf il est indispensable que parmi les fichiers `.py` on trouve pour chaque exercice un fichier dont le titre commence par `demo` et indexé par le numéro de l'exercice qu'il suffit d'executer pour illustrer informatiquement ce qui est indiqué dans le compte-rendu.

Dans tous les cas dans le compte-rendu on doit trouver

1. Des commentaires et autant que possible des explications sur les simulations présentées et les phénomènes constatés. On pourra choisir d'approfondir numériquement certains points particuliers qui vous paraissent pertinents et intéressants et détailler les conclusions qu'on peut tirer de ces expériences numériques.
2. Une étude mathématique partielle ou exhaustive. Elle peut s'appuyer par exemple sur la démonstration des propositions indiquées dans le sujet, ou encore de la partie 3.3, ou sur tout autre développement dans le cadre du sujet, débruitage par seuillage des coefficients en ondelettes.
3. **Pour aller plus loin :** voici quelques pistes (non exhaustives)
   — on peut examiner ce qui se passe si on seuille les coefficients de Fourier du signal bruité et analyser les phénomènes constatés.
   — on peut s'intéresser au cas du débruitage d'images par seuillage des coefficients en ondelettes

Grille de notation approximative

Ci-dessous quelques critères qui guideront la notation : quatre critères vont être pris en compte
- programmation : les codes ne doivent pas être buggés et comporter des commentaires qui expliquent et détaillent ce qu'ils font.
- commentaires des parties simulations : ils doivent aider à comprendre ce qui est simulé, et le point principal est leur clarté et leur précision. On attend aussi de chercher à comprendre ce que l'on fait et donc d'expliquer au lecteur le plus possible les raisons pour lesquelles on choisit de calculer telle ou telle quantité.
- explications théoriques : un ou plusieurs développements mathématiques qui vous intéressent ou vous semblent pertinents doivent être présentés, expliqués et articulés avec la partie simulation. Il est possible de choisir d'autres points que ceux qui sont suggérés par l'énoncé du TP et cela est a priori encouragé à l'avertissement près que le plagiat sera durement sanctionné.
- forme du document : le document doit être clair et correctement rédigé. Il doit comporter une introduction et un plan et être rédigé en français clair et autant que possible correct.

Pour vous donner une idée voilà à peu près ce qui vous attend selon le degré d'investissement dans le travail.
- Note entre 0 et 5/20 : les codes ne tournent pas en général. En particulier les fichiers `demoexo1.py, demoexo2.py, demoexo3.py,..., demoexo5.py` ne sont pas présents ou sont buggés. La rédaction est très succinte voir inexistante. La forme du document laisse à désirer et le français est incorrect ou parfois même incompréhensible.
- Note entre 5 et 8/20 : certains codes tournent, mais pas tous. En particulier un ou deux fichiers `demoexo...` ne tournent pas. Les commentaires sont généraux et peu précis et reprennent le sujet du TP avec presque aucun apport personnel ni détails sur ce qui est fait, ou alors les commentaires sont juste descriptifs sans aucun effort d'analyse.
- Note entre 8/20 et 10/20 : les codes tournent mais les commentaires restent peu précis ou alors purement descriptifs, et il y a peu de réflexion personnelle dans le compte-rendu. On ne comprend pas bien pourquoi tel ou tel calcul est produit.
- Note entre 10/20 et 15/20 : les codes tournent. Une partie jusqu'à toutes les parties mathématiques sont développées et détaillées. Les simulations sont expliquées clairement. Il y a des efforts de rédaction pour articuler les commentaires des simulations et de la partie mathématique.
- Note entre 15/20 et 20/20 : les codes tournent, les commentaires de la partie simulation et de la partie mathématique s'articulent et expliquent ce qui est fait. On va même dans certains cas jusqu'à aller plus loin et proposer d'explorer une piste qui n'est pas présente stricto senso dans le sujet.