HAL
open science

Semaine d'Etude Mathématiques et Entreprises 6 : Analyse et filtrage temps-fréquence de "bursts" ultrasonores : identification, classification, séparation

Meitner Cadena, Charles Guyon, Romain Hug, Ester Mariucci, Antonin Monteil, Thomas Oberlin

To cite this version:

Meitner Cadena, Charles Guyon, Romain Hug, Ester Mariucci, Antonin Monteil, et al.. Semaine d'Etude Mathématiques et Entreprises 6 : Analyse et filtrage temps-fréquence de "bursts" ultrasonores : identification, classification, séparation. 2013. hal-00933225

HAL Id: hal-00933225

https://hal.science/hal-00933225v1

Preprint submitted on 20 Jan 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Semaine d'Etude Mathématiques et Entreprises 6

Grenoble du 10 au 14 juin 2013

Analyse et filtrage temps-fréquence de "bursts" ultrasonores : identification, classification, séparation

Meitner CADENA[a] Charles GUYON[b]
Romain HUG[c] Ester MARIUCCI[c]
Antonin MONTEIL[d] Thomas OBERLIN[c]

[a] *Essec Business School & UPMC*
[b] *Laboratoire MIA, Université de La Rochelle*
[c] *Laboratoire Jean Kuntzmann, Université Joseph Fourier, Grenoble*
[d] *Université Paris-Sud*

Sujet proposé par :

Correspondant : Laurent DUVAL

Remerciements

Nous tenons à remercier vivement tous les organisateurs de la sixième Semaine d'Étude Maths-Entreprises, à Grenoble. En particulier, AMIES pour le support à l'initiative et l'équipe MaiMoSiNE pour avoir assuré l'animation scientifique, en les personnes de Marianne Clausel et Emmanuel Maître. Notre remerciement s'adresse aussi à la Maison Jean Kuntzmann pour nous avoir mis à disposition les locaux.

Nous remercions tout particulièrement IFP Énergies Nouvelles, en la personne de Laurent Duval, pour nous avoir proposé ce sujet. Laurent a été dès le début très clair dans l'exposition du problème, toujours disponible et à l'écoute. Finalement, nous avons apprécié cette collaboration et pour cela nous tenions à le remercier.

Résumé

Ce rapport est une présentation de nos résultats et de nos réflexions à propos du problème proposé par IFP Énergies Nouvelles pendant la sixième édition de la Semaine d'Étude Maths-Entreprises. Nous disposions d'enregistrements de bursts ultrasonores issus d'un problème physique de corrosion d'éprouvettes en métal. Le but était de donner une classification des signaux acoustiques visant à identifier les différentes typologies de corrosion. En Section 1 on trouve une présentation plus détaillée de la problématique enquêtée. Tous les approches considérées sont développées dans la Section 3, alors que dans la Section 2 on a passé en revue les outils mathématiques nécessaires.

Table des matières

1 Présentation du problème **3**
1.1 Problématiques 3
1.2 Les données 4
1.3 Méthodes utilisées 5

2 Outils mathématiques **5**
2.1 Quelques notions sur la théorie du signal 5
2.2 Temps fréquence 6
2.3 Filtrage 7
2.4 Classification 8

3 Méthodes utilisées **10**
3.1 Recherche des descripteurs pertinents 10
3.1.1 Analyse de la fréquence du pic maximal sur les données fréquentielles 10
3.1.2 Analyse de la fréquence et de l'amplitude du pic maximal sur les données fréquentielles 11
3.2 Approche sans descripteur (SVD) 13
3.3 Approche avec descripteur *a priori* 19
3.4 Descripteur basé sur le max 19
3.5 Distance de similarité entre spectres 27
3.5.1 Analyse de clusters basée sur moments 27
3.5.2 Regroupement hiérarchique et distance de Wasserstein . . 31

1 Présentation du problème

1.1 Problématiques

IFP Énergies Nouvelles est un organisme public de recherche, d'innovation et de formation dans les domaines de l'énergie, du transport et de l'environnement. Elle s'occupe, entre autres, des problèmes liés à l'extraction du pétrole dans la mer. Comme on peut le voir en figure 1, l'extraction du pétrole se fait à l'aide de tubes métalliques. Or, ces derniers subissent naturellement une corrosion qui doit être rigoureusement gardée sous contrôle. À ce propos, on peut remarquer qu'une étude qui date de l'année 2000 démontre que la corrosion est la cause du 15-20% des fuites d'hydrocarbures au large des côtes [6]. Une façon de traiter le problème de la corrosion passe par une analyse des données ultrasonores émises par des barres métalliques soumises à des conditions expérimentales contraignantes. On reproduit en laboratoire les conditions "naturelles" des tubes utilisés pour extraire le pétrole en gardant des barres métalliques sous pression et immergées dans une solution acide. Les dégradations dues à ces traitements génèrent des formes d'ondes ultrasonores de différentes natures, évoluant au cours du temps et qui sont enregistrées par des capteurs placés aux extrémités des barres (voir figure 2). La question qui se pose est alors de

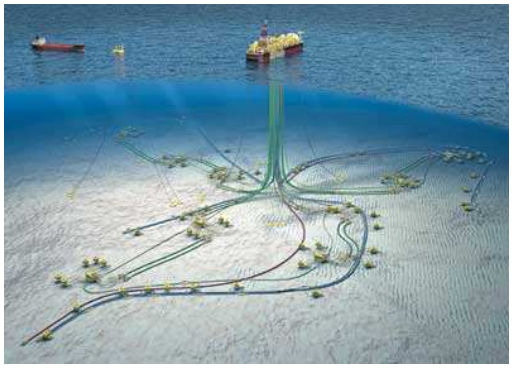


FIGURE 1 – Extraction du pétrole

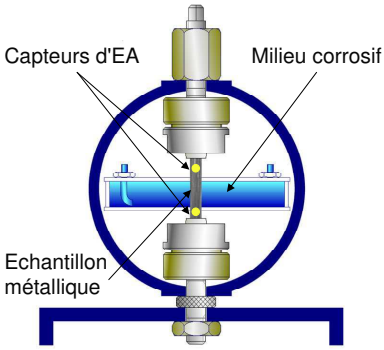


FIGURE 2 – Éprouvette métallique en laboratoire

déterminer comment, à partir des observations des ondes acoustiques, discerner les perturbations pertinentes. Dans ce cadre s'insère notre travail, qui vise à donner une classification des signaux acoustiques afin d'identifier les différentes typologies de corrosion.

1.2 Les données

Pour mener notre étude, nous disposions d'un échantillon d'environ 1500 signaux issus d'ondes ultrasonores amorties produites en laboratoire. Les données correspondent à 8 jours de prise de mesure et proviennent d'ondes dégradées par la présence d'un bruit. En figure 3, on retrouve la représentation typique d'un signal de l'échantillon. Comme ces signaux proviennent des réactions de corrosion de la barre métallique, ils interviennent ponctuellement à des instants aléatoires. Il a donc été choisi de lancer l'enregistrement automatiquement à chaque fois que l'amplitude du signal dépasse un certain seuil déterminé à l'avance. Il convient de remarquer qu'il y a nécessairement une évolution de la forme des signaux au cours du temps : En effet la forme de la barre, et donc ses propriétés acoustiques, sont modifiées par sa corrosion. En particulier l'étude du diagramme amplitude-fréquence réalisée plus bas révèle une augmentation de l'amplitude des signaux. On peut aussi remarquer que les signaux sont de plus en plus bruités par la dégradation de la barre.

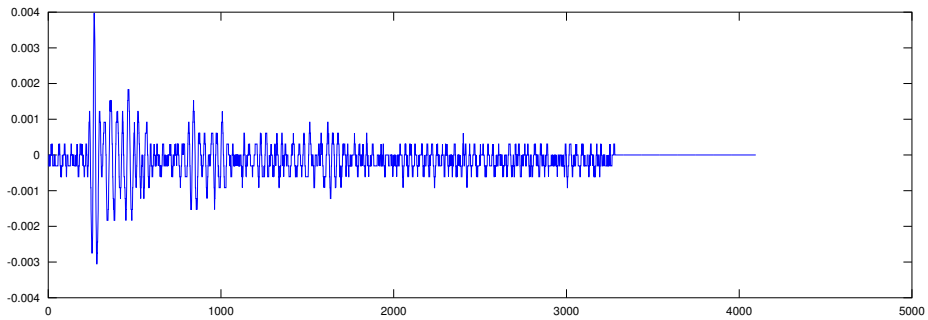


FIGURE 3 – Un signal des données

1.3 Méthodes utilisées

2 Outils mathématiques

2.1 Quelques notions sur la théorie du signal

Une façon naturelle de connaître un signal est d'observer son allure en fonction du temps : c'est la représentation temporelle, donnée par un oscillogramme. Or, plus souvent, on préfère donner une représentation fréquentielle plutôt que temporelle du signal, car elle va avoir des meilleures propriétés mathématiques. Une représentation fréquentielle passe par la notion de spectre d'un signal, c'est-à-dire, par la représentation en fonction de la fréquence des amplitudes des différentes composantes présentes dans le signal. La décomposition en série de Fourier est l'outil permettant de calculer le spectre d'un signal périodique. Plus précisément, en 1822, le mathématicien Joseph Fourier a montré que tout signal périodique de fréquence f_1 peut être décomposé en somme de signaux sinusoïdaux de fréquences f_n, toutes multiples de f_1 (appelée *fréquence fondamentale*). Ces signaux sinusoïdaux de fréquence $f_n = nf_1,\ n \in \mathbb{N}^*$, sont appelés *harmoniques*. Autrement dit, si $x(t)$ est un signal périodique de période T, on peut l'écrire de la façon suivante :

$$x(t) = x_0 + x_1 \sin(2\pi f_1 t + \varphi_1) + \cdots + x_n \sin(2\pi n f_1 t + \varphi_n), \tag{1}$$

avec

$$\begin{aligned}
x_0 &= \text{valeur moyenne du signal,}\\
x_1 &= \text{amplitude de la fréquence fondamentale,}\\
x_2 &= \text{amplitude de l'harmonique 2,}\\
&\vdots\\
x_n &= \text{amplitude de l'harmonique } n.
\end{aligned}$$

Cette décomposition peut aussi s'écrire de la façon suivante :

$$x(t) = x_0 + A_1 \cos(2\pi f_1 t) + B_1 \sin(2\pi f_1 t) + \cdots + A_n \cos(2\pi n f_1 t) + B_n \sin(2\pi n f_1 t), \tag{2}$$

avec

$$x_0 = \frac{1}{T}\int_0^T x(t)dt, \quad A_n = \frac{2}{T}\int_0^T x(t)\cos(2\pi n f_1 t)dt,$$
$$B_n = \frac{2}{T}\int_0^T x(t)\sin(2\pi n f_1 t)dt.$$

Comme il est bien connu, (1) et (2) sont équivalentes et on a :

$$x_n^2 = A_n^2 + B_n^2 \quad \text{et} \quad \tan(\varphi_n) = \frac{B_n}{A_n}.$$

Une fois que la décomposition d'un signal est calculée, on trace le spectre représentant les amplitudes x_i en fonction de la fréquence.

Pour les signaux complexes (audio, vidéo, etc...) le calcul du spectre est impossible, mais il est possible de le visualiser à l'aide d'un analyseur FFT. Le principe est le suivant.

On échantillonne le signal durant un temps T et on le convertit en un suite de N valeurs numériques $x_0 = x(0)$, $x_1 = x(T_e), \ldots, x_{N-1} = x((N-1)T_e)$. L'échantillonage lui-même se fait à une fréquence $f_e = \frac{1}{T_e}$ et la prise de N échantillons dure un temps T tel que $T = NT_e = \frac{N}{f_e}$. Plus précisément, la fréquence d'échantillonage f_e du signal est supposée vérifier le théorème d'échantillonage de Shannon, c'est-à-dire $f_e \geq 2B$, B étant la limite supérieure de l'étendu spectrale du signal (voir [9]). À partir de ces N échantillons, on peut calculer N points du spectre définis par leur abscisse $f(k)$ et leur ordonnée $S(k)$ en utilisant la transformée de Fourier discrète (TFD) définie par :

- fréquence : $f(k) = \frac{kf_e}{N}, \quad k = 0, \ldots, N-1,$
- amplitude : $S(k) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x(n)e^{\frac{-2n\pi ik}{N}}\right|.$

2.2 Temps fréquence

La représentation la plus simple utilise la transformée de Fourier à court terme (TFCT), obtenue en réalisant des transformées de Fourier "locales" grâce à une fenêtre glissante. Soit une fonction $g \in L^2(\mathbb{R})$, on définit la TFCT de $f \in L^2(\mathbb{R})$ au temps t et à la fréquence η par

$$V_f(\eta, t) = \int_{\mathbb{R}} f(\tau)g(\tau - t)^* e^{-2i\pi\eta\tau}\, d\tau. \tag{3}$$

La formule de Plancherel permet de réécrire cette définition dans le domaine des fréquences :

$$V_s(\eta, t) = \int_{\mathbb{R}} f(\xi)\hat{g}(\xi - \eta)^* e^{2i\pi\xi t}\, d\xi.$$

Pour obtenir une représentation bien localisée en temps et en fréquence, on doit utiliser une fenêtre g régulière et concentrée. En pratique, on utilisera des fenêtres réelles et paires. La fonction f est facilement reconstruite à partir de sa TFCT par

$$f(\tau) = \frac{1}{\|g\|^2} \iint_{\mathbb{R}^2} V_f(\eta, t) g(\tau - t) e^{2i\pi\eta(\tau - t)} \, dt d\eta, \tag{4}$$

le spectrogramme d'un signal, c'est-à-dire $|V_f(\eta, t)|^2$ dans le plan (η, t) est appelé le plan temps-fréquence.

2.3 Filtrage

Un filtre S est un système sélectif en fréquence : il transforme un signal d'entrée $x(k)$ en un signal de sortie $y(k)$ de façon à ce que uniquement les composantes comprises dans un certain intervalle de fréquence soient transmises. Ils existent plusieurs types de filtres, comme par exemple les filtres adaptés, les filtres de Widrow où encore le filtrage homomorphique (voir [3]). Dans ce contexte on considérera le filtre de Savitzky-Golay (1964) et un lissage employant l'algorithme *super-smoother* de Jerome Friedman (1984), voir [7] pour le premier et [4] pour le deuxième.

Nous allons ici expliquer brièvement le fonctionnement du filtre de Savitzky-Golay, car il s'avérera être celui mieux adapté à nos exigences. Il s'agit d'un filtre basé sur l'idée suivante :

- On considère $2M+1$ échantillons parmi les N données, centrés en un point $x(n_0)$;
- On fixe un degré $d < M$ et on cherche le polynôme de degré d et de coefficients a_k qui minimise l'erreur quadratique moyenne :

$$\mathcal{E}_d = \sum_{n=n_0-M}^{n_0+M} \left(\sum_{k=0}^{d} a_k n^k - x(n) \right)^2.$$

- On définit la sortie par $y(n_0) = a_0$, et on répète le procédé pour chaque point n_0 entre 0 et $N-1$ (proche de 0 et $N-1$ il faut utiliser des intervalles non symétriques).

Le point délicat est le choix des paramètres M et d. Le paramètre d est le plus souvent choisi égal à 2 ou 3, tandis que le paramètre M peut prendre des valeurs bien plus grandes. Si, comme dans notre cas, on s'intéresse à des pics, un choix naturel est de prendre $2M + 1$ égal à la largeur à mi-hauteur du pic le plus fin. En général, une valeur plus grande pour M donne une fonction plus lisse au risque de perdre des informations sur les pics les plus fins. En figure 4, on donne un exemple de lissage d'un spectre (représenté en haut dans un diagramme fréquence-amplitude) à l'aide du filtre de Savtzky-Golay (au centre) et de l'algorithme "super-smoother" (en bas). L'échantillon est constitué de N points, avec N de l'ordre de 2200 et le filtrage à été effectué avec les paramètres $d = 3$ et $M = 20$. Pour plus de propriétés du filtre de Savitzky-Golay et notamment pour une implémentation convenable, on fait référence à [8].

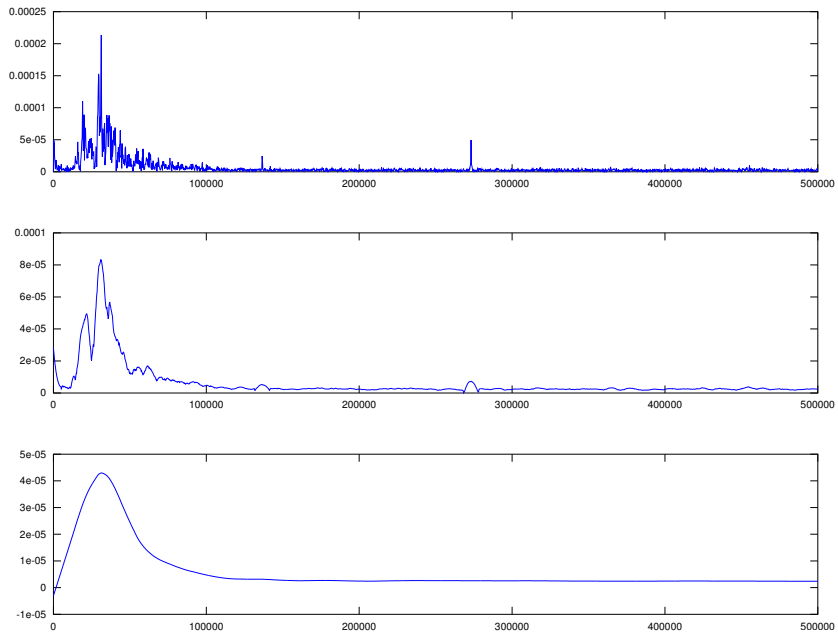


FIGURE 4 – Lissage d'un spectre avec Savitzky-Golay et "super-smoother". Paramètres : $N \sim 2200$, $d = 3$, $M = 20$.

2.4 Classification

Étant donné un ensemble $X = \{x_1, x_2, ...x_n\}$ de n objets dans un espace métrique et un nombre m de classes fixé à l'avance, le but du problème de classification est de séparer X en une partition $C = \{c_1, c_2, ...c_m\}$ suivant les «nuages de points» observés. Ce nuage se présente bien sûr avec des zones de plus ou moins fortes densités.

Algorithme des k plus proches voisins

Un premier algorithme très simple consiste à débuter avec un point (ou m points) et à classer chaque nouveau point dans la classe la plus représentée chez ses k plus proches voisins (il ne s'agit donc pas de trouver les classes, ce qui est un vrai problème d'apprentissage, mais de classer de nouveaux candidats).

Une façon quantitative d'évaluer la classification est calculer le *coefficient de corrélation intra-classes* ρ^2. Formellement, ρ^2 est définie en étant le rapport entre la variance intra-classes et la variance inter-classes :

$$\rho^2 = \frac{\sigma_{\bar{y}}^2}{\sigma_y^2}, \quad \sigma_{\bar{y}}^2 = \frac{\sum_{x=1}^{k} n_x(\bar{y}_x - \bar{y})^2}{n}, \quad \sigma_{\bar{y}}^2 = \frac{\sum_i (y_i - \bar{y})^2}{n},$$

où les données y_i, $i = 1, \ldots, n$ sont divisées en k groupes paramétrés par $x = 1, \ldots, k$, $\bar{y}$ est la moyenne de toutes les données y_i et $\bar{y}_x$ la moyenne du groupe x, qui a pour cardinal n_x si bien que $n = \sum n_x$.

Regroupement hiérarchique (voir [5])

On choisit une métrique d sur l'ensemble X des individus.
Le but est de classer les individus selon différentes classes avec pour critère de dissimilarité la métrique d : Deux individus d'une même classe doivent être proches pour la métrique d et inversement.

- Avantage : N'utilisant pas de paramètre, on ne perd pas d'information sur le signal : on ne fait pas de réduction de dimension.
- Inconvénients : Temps de calcul... Il dépend de la métrique d. Il y a éventuellement trop d'informations : La métrique choisie peut contenir beaucoup d'informations non discriminantes.
  La métrique d doit être robuste par rapport aux perturbations (bruits) dues par exemples aux capteurs utilisés lors des mesures.

Principe de l'algorithme : À la première étape chaque individu forme une classe à lui seul. A chaque étape de l'algorithme, on fusionne les deux classes les plus proches au sens d'une dissimilarité Δ basé sur la métrique d.
On continue le processus jusqu'à obtenir le nombre de classes voulu.

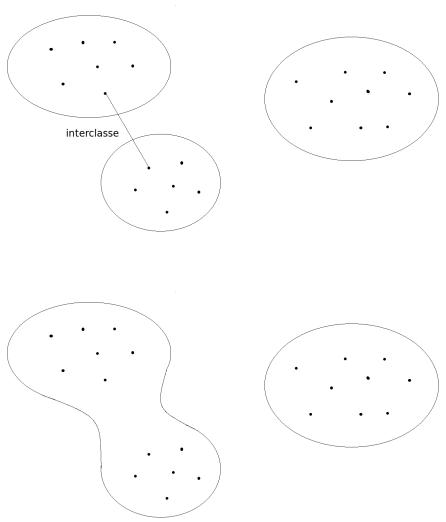


FIGURE 5 – Distance inter-classe / Fusion entre deux classes

Plus précisement, à l'étape n on dispose d'une classification $(C_i)_{i \in I_n}$ (une partition de X). Ensuite, on peut choisir de définir la mesure de dissimilarité inter-classe comme suit :

$$\Delta(C_1, C_2) = \underset{x \in C_1,\ y \in C_2}{min} d(x, y) \tag{5}$$

$$\Delta(C_1, C_2) = \underset{x \in C_1,\ y \in C_2}{max} d(x, y) \tag{6}$$

ou bien

$$\Delta(C_1, C_2) = \underset{x \in C_1,\ y \in C_2}{moyenne}\, d(x, y) \tag{7}$$

Une première solution consiste à utiliser la distance L^2 sur le spectre. Cette solution peut conduire pour certaines données à une classification peu pertinente : Par exemple une erreur de mesure d'un capteur entraînant une modulation du signal relevé (et donc un déphasage sur le spectre) suffit pour rendre la distance L^2 inadéquate. En effet, un déphasage (de l'ordre d'une période) induit une "forte" distance L^2 entre 2 signaux bruités : deux signaux visuellement proches ne le sont pas pour la norme L^2.

3 Méthodes utilisées

3.1 Recherche des descripteurs pertinents

Une première approche à notre problème a été de rechercher, directement, des descripteurs pertinents. Notre analyse a été surtout "exploratoire", ne disposant pas d'accès aux apparats expérimentals mais uniquement à un jeu de données. Ce qui suit est une reproduction des indicateurs qu'on a testé dans l'ordre chronologique de découverte.

3.1.1 Analyse de la fréquence du pic maximal sur les données fréquentielles

Le premier descripteur auquel on s'est intéressé a été la fréquence relative au pic d'amplitude maximale. En figure 6, sur la gauche, on trouve la classification via algorithme des 3 plus proches voisins en ayant normalisé les données par rapport à la fréquence. Pour évaluer cette classification, on a calculé la corrélation intra-classes, qui c'est avérée être égale à 6.94. Ensuite, on a cherché à comprendre si un lissage du spectre pouvait amener à une meilleure classification. Sur la droite dans la figure, on peut trouver la même classification faite sur le spectre lissé par convolution avec une gaussienne. En regardant l'indice de corrélation intra-classes, qui dans ce cas est 4.89, on s'aperçoit qu'il y a une légère amélioration.

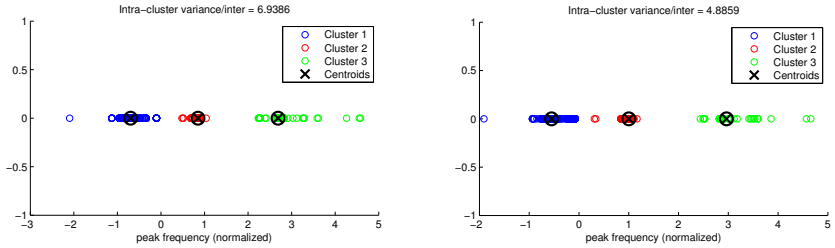


FIGURE 6 – Classification des fréquences d'amplitude maximale avant et après lissage

3.1.2 Analyse de la fréquence et de l'amplitude du pic maximal sur les données fréquentielles

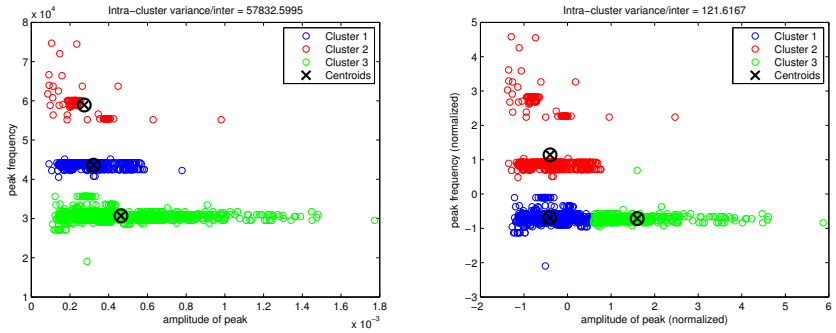


FIGURE 7 – Classification selon l'amplitude maximale et sa fréquence, avant et après normalisation

Ensuite, on a ajouté comme descripteur l'amplitude maximale du pic. Dans la figure 7 on peut trouver, sur la gauche, la classification via algorithme des 3 plus proches voisins sur les données non normalisées, alors que sur la droite, le même algorithme a été appliqué aux données normalisées en fréquence et en amplitude. Bien que l'ajout de l'amplitude comme descripteur apporte des informations, seul la fréquence du pic paraît réellement discriminant.

Pour améliorer les résultats obtenus on a essayé, dans un deuxième temps, de lisser le spectre de façon à réduire le bruit et à obtenir une classification plus satisfaisante. On a commencé par lisser le spectre avec la fonction supsmu de matlab (il s'agit de l'implémentation de la fonction super smoother de Friedman). Les résultats n'ont pas été du tout satisfaisants : dans la figure 8 on trouve le diagramme de l'amplitude du pic maximal, avec sa fréquence ; on y trouve aussi reporté l'index du signal (c'est-à-dire, l'instant du temps de l'observation). L'image du haut est le nuage des points correspondants aux données brutes et celle d'en bas après application du filtrage "super smoother" ; on s'aperçoit qu'il s'agit d'un lissage qui modifie trop l'allure du spectre, en perdant trop d'information. On s'est donc tourné vers un filtrage qui garde mieux la forme du spectre, en choisissant d'appliquer le filtre de Savitzky-Golay, avec paramètres $d = 3$ (le degré du polynôme) et $M = 20$ (la demi-largeur de l'intervalle d'échantillonage). Le résultat est reporté dans la figure du milieu ; dans la figure 9 on visualise la classification via l'algorithme des k plus proches voisins, en choisissant $k = 3$ dans la partie de gauche et $k = 4$ sur la droite.

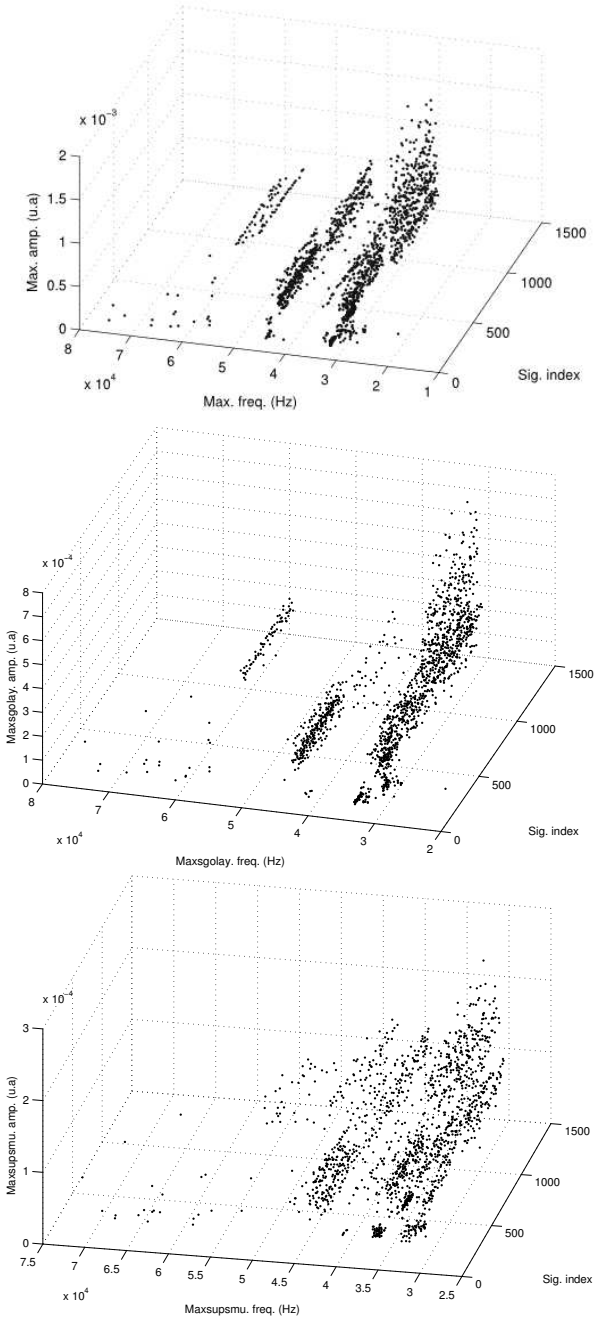


FIGURE 8 – Visualisation des données en amplitude maximale et sa fréquence, sans filtrage, avec filtrage de Savitzky-Golay et avec filtrage “super-smoother”

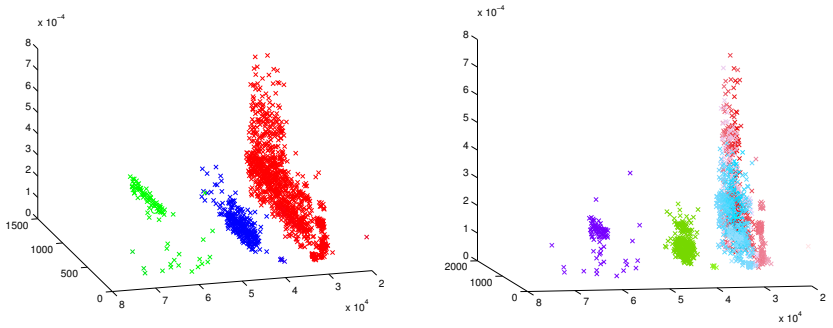


FIGURE 9 – Classification selon l'amplitude maximale et sa fréquence du signal filtré avec Savitzky-Golay en trois et quatre classes, respectivement

3.2 Approche sans descripteur (SVD)

En adoptant un point de vue matriciel, plaçons les données dans la matrice $A \in \mathbb{R}_{n \times m}$. Les vecteurs colonnes de $A_{:j}$ correspondent à chacun des $m = 1448$ enregistrements. Un enregistrement contient n=4096 valeurs (Fig. 10).

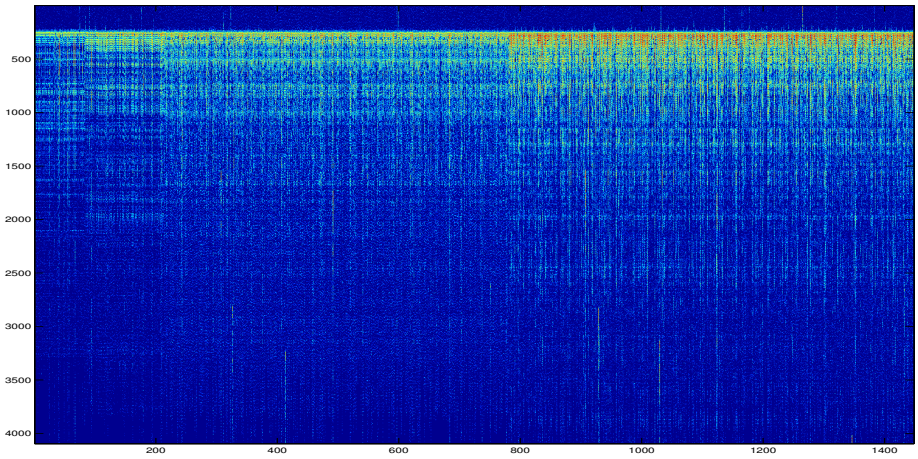


FIGURE 10 – Matrice A où chaque enregistrement est représenté en colonne.

Dans un premier temps, une simple réduction de dimension à l'« aveugle » est directement appliquée sur les données sans utiliser de descripteur spécifique et donc sans *a priori*. La dimensions de ces vecteurs descripteur sera de 2048.

Pré-traitement

Un changement d'espace et une normalisation sont effectués. Empiriquement, il semble plus pertinent d'avoir une interprétation et d'effectuer des traitements dans le domaine spectral. De plus, le fait de travailler sur le module du

spectre garantit une l'invariance en translation temporelle. Soit B, la matrice contenant le log-module du demi-spectre de A (Fig. 11, plot de gauche).

$$F_{kj} = e^{i2\pi \frac{(k-\frac{1}{2})(j-\frac{1}{2})}{n}}, \quad k = 1 \dots n, \quad j = 1 \dots n$$

$$B_{kj} = \log\left(\epsilon + |FA|_{kj}\right), \quad n_2 = \frac{n}{2}, \qquad k = 1 \dots n_2, \quad j = 1 \dots m, \quad \epsilon = 10^{-15},$$

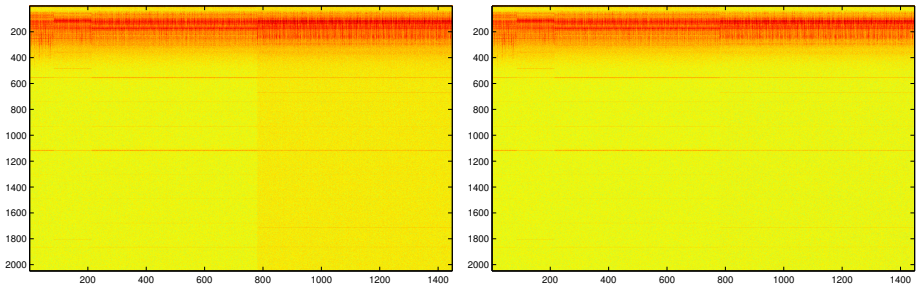


FIGURE 11 – La matrice $B \in \mathbb{R}_{n_2 \times m}$ et la matrice C (*i.e.* B normalisée)

On constate un offset sur la partie droite de l'image, il semble donc judicieux de normaliser/centrer chaque vecteur colonne qui sera stocké dans C (Fig. 11, plot de droite),

$$C_{kj} = \frac{B_{kj}}{\frac{1}{n_2}\sum_{k=1}^{n_2} B_{kj}} - 1, \quad \text{donc} \sum_{k=1}^{n_2} C_{kj} = 0_j, \quad k = 1 \dots n_2, \quad j = 1 \dots m$$

Enfin, isoler la bande de fréquence la plus pertinente. Celle-ci semble être située principalement entre les indices 30 et 335 : $X = C_{30:335,:}$ (Fig 12).

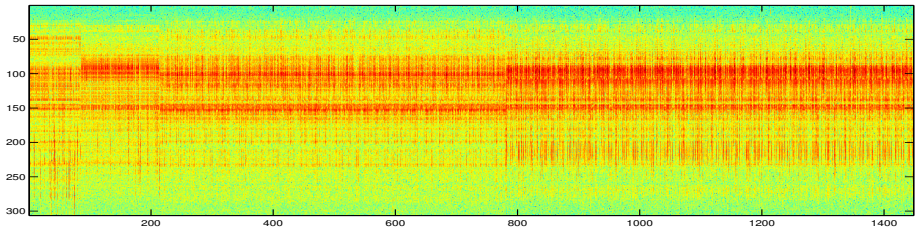


FIGURE 12 – Matrice X

Dès lors, nous ne travaillerons plus que sur la matrice X.

Singular Value Decomposition (SVD)

On peut directement appliquer une SVD sur X, c'est-à-dire, la décomposition matricielle qui vérifie :

$$X = UDV, \quad D = \text{diag}, \quad U'U = V'V = I$$

et qui répond au problème de minimisation en cascade suivant,

$$\left|\begin{array}{lll} X^{(0)} = X, & \underset{U_{:,1},D_{1,1},V_{1,:}}{\text{argmin}} & ||X^{(0)} - U_{:,1}D_{1,1}V_{1,:}||_F \\ X^{(1)} = X^{(0)} - U_{:,1}D_{1,1}V_{1,:}, & \underset{U_{:,2},D_{2,2},V_{2,:}}{\text{argmin}} & ||X^{(1)} - U_{:,2}D_{2,2}V_{2,:}||_F \\ X^{(2)} = X^{(1)} - U_{:,2}D_{2,2}V_{2,:}, & \underset{U_{:,3},D_{3,3},V_{3,:}}{\text{argmin}} & ||X^{(2)} - U_{:,3}D_{3,3}V_{3,:}||_F \\ \dots & \dots & \end{array}\right.$$

Ainsi, les valeurs singulières sont ordonnées : $D_{1,1} \geq D_{2,2} \geq D_{3,3} \geq \cdots \geq D_{m,m}$ Comme le montre la Fig. 13, on peut reconstruire successivement le signal X avec une somme de matrices de rang 1.

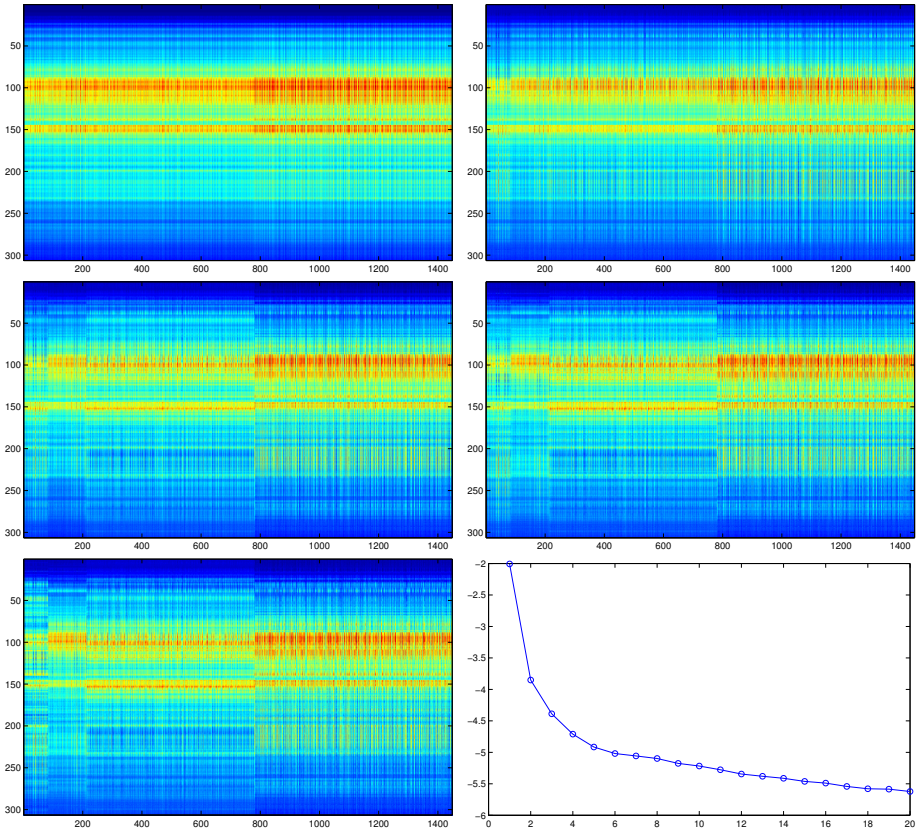


FIGURE 13 – Différentes reconstructions de la matrice X avec 1,2,3,4 et 5 composantes principales. Le dernier plot correspond au log des valeurs singulières.

À l'observation du plot du log des valeurs singulières, il est raisonnable de penser que l'essentiel de l'information (du moins le nécessaire discriminatif), est porté par environ 6 composantes principales.

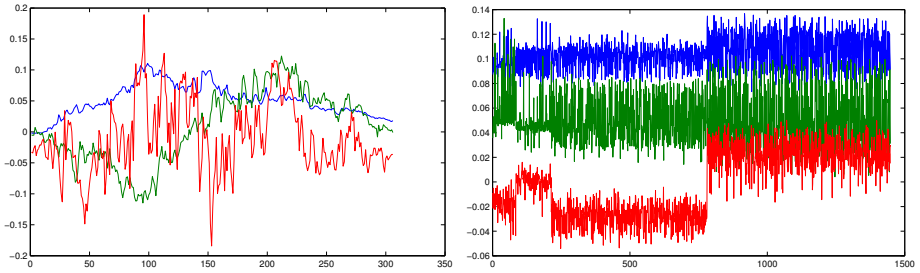


FIGURE 14 – Les 3 premières composantes principales $U_{:,1:3}$ (à gauche) ansi que les poids associés $V_{1:3,:}$ (à droite) . Rouge=1, Vert=2 et Bleu=3 ème composante.

À en juger $V_{1:3,:}$ (Fig. 14), il semble exister 4 classes, visiblement très corrélées temporellement (c'est à dire, dans l'ordre chronologique des enregistrements). On peut supposer que $X_r = V_{1:6,:}$ est suffisant pour être discriminatif. La Fig. 15 illustre les différentes réponse du k-means si on l'applique avec plusieurs choix de nombre de classes (k).

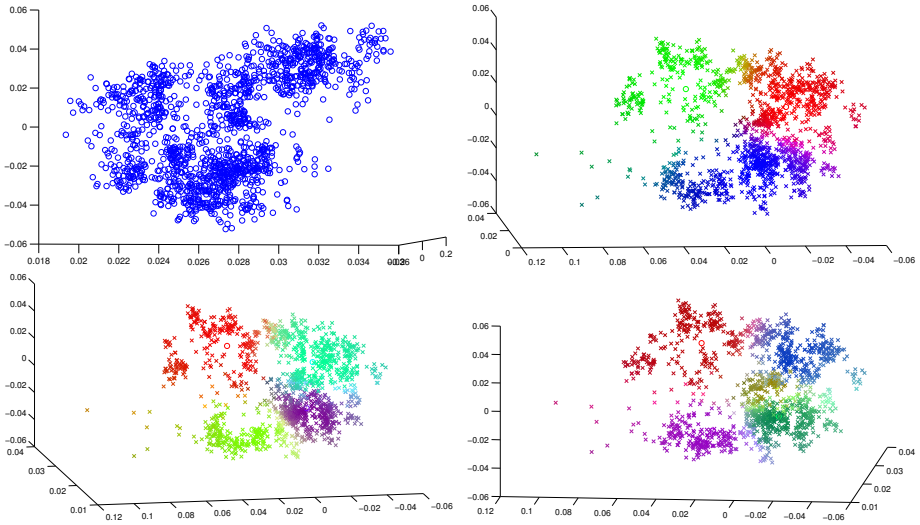


FIGURE 15 – La matrice X_r visualisée dans les 3 premières directions principales. Les trois plot en couleur sont les clustering associés pour 3, 4 et 5 classes.

Kernel MDS

Une réduction de dimension non-linéaire peut être appliquée pour visualiser les données. Un MDS (Multi-Dimensionnal Scaling) est très proche d'une PCA à la différence où l'on dispose en entrée, non pas des coordonnées cartésiennes des objets, mais de leurs distances pair à pair (stocké dans une matrice symétrique à diagonale nulle). Un kernel MDS repose essentiellement sur les formules de

passage entre matrice de similarité/corrélation et matrice de distance, ainsi que l'introduction de la non-linéarité faite par l'ajout intermédiaire d'une fonction ϕ.

La méthode peut s'appliquer aussi bien sur les données pleines X, ou sur les données réduite X_r (de la section précédente). On commence par créer Y, la matrice des individus centrés,

$$Y = \begin{cases} X' - u_m \text{mean}(X') \\ X'_r - u_m \text{mean}(X'_r) \end{cases} \quad , \; u_m = 1 \text{ vecteur colonne de taille } m \text{ contenant des } 1$$

Puis Z, la matrice de covariance de taille $m \times m$,

$$Z = YY'.$$

M est la matrice de distance pair à pair, où $M_{ii} = 0$ et $M_{ij} = M_{ji}$,

$$M = \text{diag}(Z)u'_m + u_m \text{diag}(Z)' - 2Z.$$

Remarque : La matrice de covariance des données centrées et la matrice de distance **contiennent exactement les même informations**. Il est possible de retrouver la matrice de covariance Z grace à la relation dite de *Torgerson*,

$$Z = -\frac{1}{2}\left(I - \frac{uu'}{u'u}\right) M \left(I - \frac{uu'}{u'u}\right).$$

Ainsi, si les données étaient normalisées ($\sum_j Y_{kj}^2 = 1_k$), alors $|Z_{kj}| \leq +1$ [1]. L'idée du MDS est de partir de la matrice de distance M, puis d'appliquer un noyau non-linéaire ϕ tel que $\begin{cases} \phi(0) = 1 \\ \phi(\infty) = 0 \end{cases}$ et de revenir à une matrice type covariance. La stratégie générale du MDS est schématisée Fig.16 et les résultats sont présentés Fig.17.

1. $Z_{kj} = +1$ signifie que les données sont totalement corrélé, 0 décoréllé, et -1 anticorrélé.

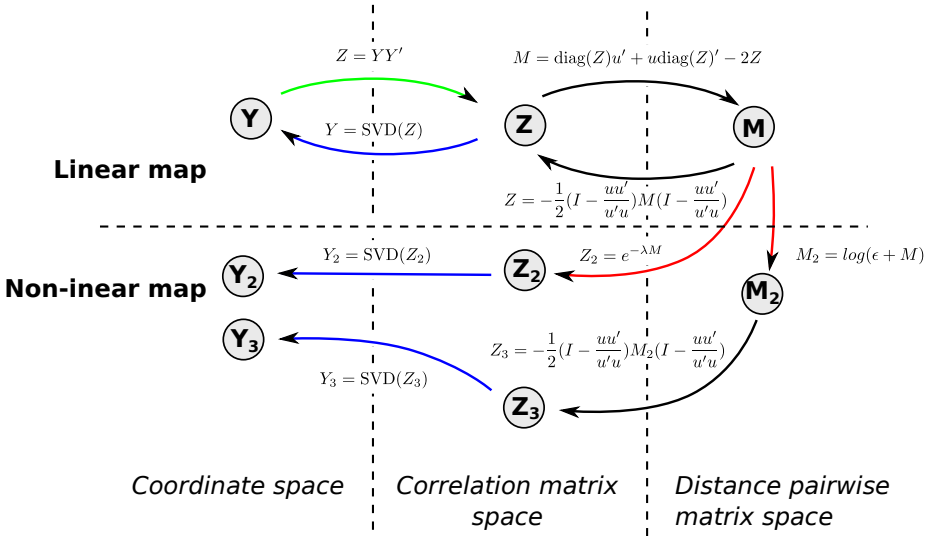


FIGURE 16 – Diagramme résumant les différentes aspects du MDS. Les flèches noires, vertes, bleues et rouges sont respectivement des maps linéaires, quadratiques, d'ordre $\frac{1}{2}$ et non-linéaires.

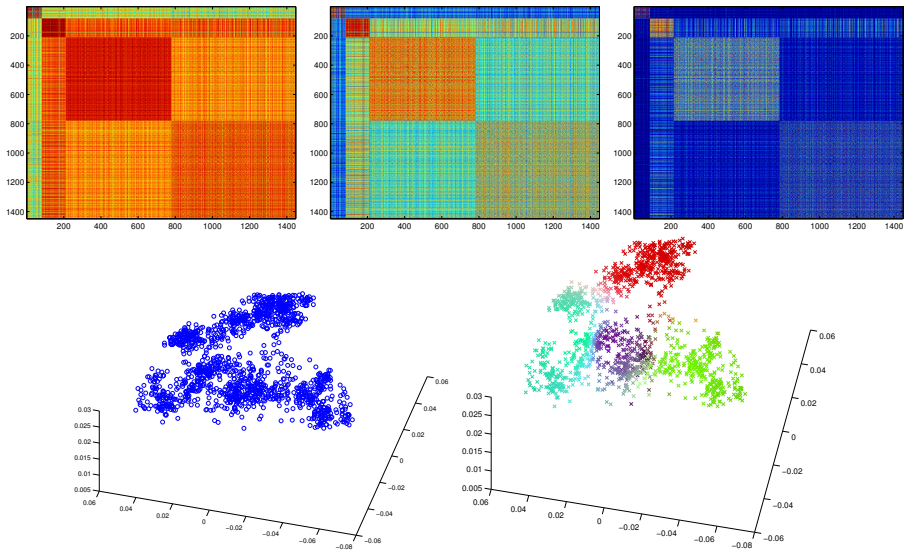


FIGURE 17 – **Première ligne**, plusieurs matrices de distance issues de différent MDS : linéaire, non-linéaire TYPE 1 et TYPE 2. Le choix du paramètre λ est fixé tel que $\mathsf{mean}(\phi(M)) = \frac{1}{\sqrt{d}}$, avec $d = 6$ dans notre cas (dimension du descripteur). **Deuxième ligne**, résultats du MDS offrant le visuel le plus satisfaisant (non-linéaire TYPE 1).

3.3 Approche avec descripteur *a priori*

Si on considère le signal comme une densité de probabilité, un choix de descripteur peut correspondre aux indicateurs statistique usuels (mean, med, max, etc...). Nous aborderons 2 types de descripteurs. Un basé sur une caractérisation par quantiles et un autre basé sur l'idée que la position du max est l'information discriminative entre les 4 groupes identifiés.

Descripteur basé sur la médiane

Les quantiles permettent de caractériser une distribution et peuvent offrir un potentiel discriminatif important. Si nous prenons les 8 octiles de chaque signaux, cela nous donne une signature à 8 dimensions. Le nombre de quantile choisi est limité à 2^k puisque l'algorithme que nous utilisons est basé sur une itération de médianes. Pour des raisons arbitraires, nous fixerons k à 3.

La Fig. 18 représente les résultats de ce choix de descripteur, après réduction de dimensions du type de celle décrite à la section précédente.

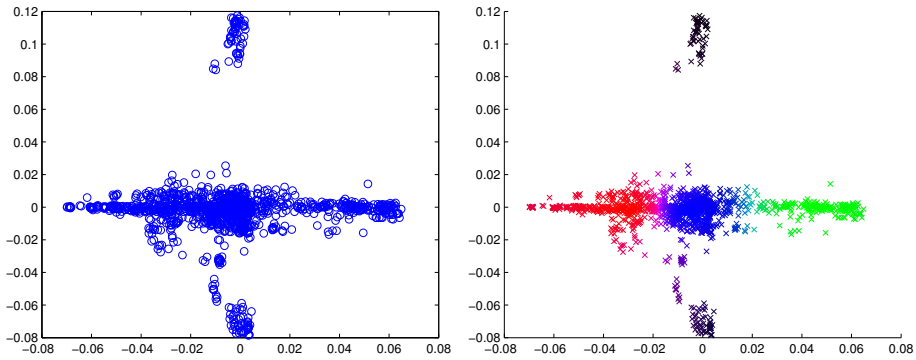


FIGURE 18 – Méthode basé sur les octiles, réduction de dimension et visualisation, puis clustering.

3.4 Descripteur basé sur le max

Reprenons la matrice X, puis observons 2 signaux appartenant à 2 classes différentes (Fig. 19).

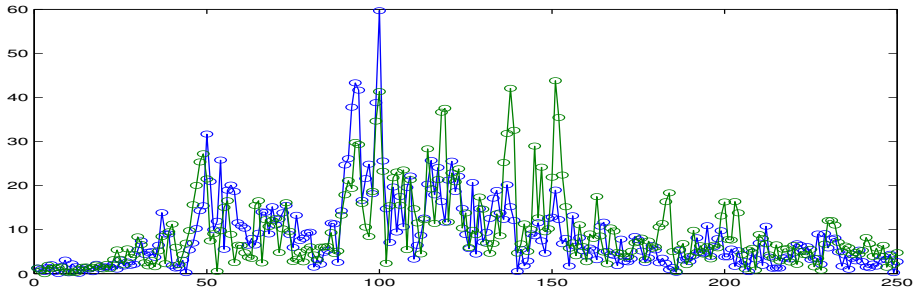


FIGURE 19 – Deux signaux appartenant à 2 classes différentes.

Le maximum de la courbe bleu est à la position 100, tandis que le maximum de la courbe verte est à 153. Afin d'être plus robuste que seulement ne retenir que l'index du max comme valeur discriminative, il est intéressant d'ajouter d'autres éléments. Regardons par exemple, les positions des 12 plus grandes valeurs de chacunes des courbes (Fig. 20, plot de gauche).

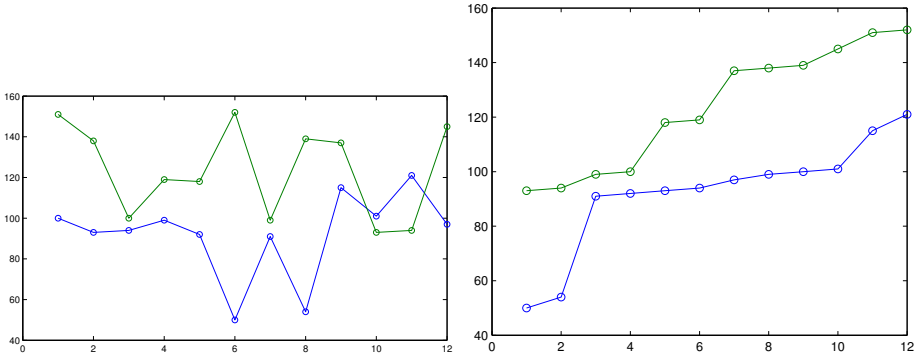


FIGURE 20 – A gauche : positions des 12 plus grandes valeurs de 2 signaux (bleu et vert) appartenant à 2 classes distinctes. A droite : Positions triées.

Une première méthode (I) consiste à utiliser ce vecteur de valeurs comme signature et d'appliquer ensuite une réduction de dimension du type de celle effectuée dans la section précédente (Fig. 21).

```
[tmp , idx] = sort(X, 'descend')
Y = idx(1 :12, :)
```

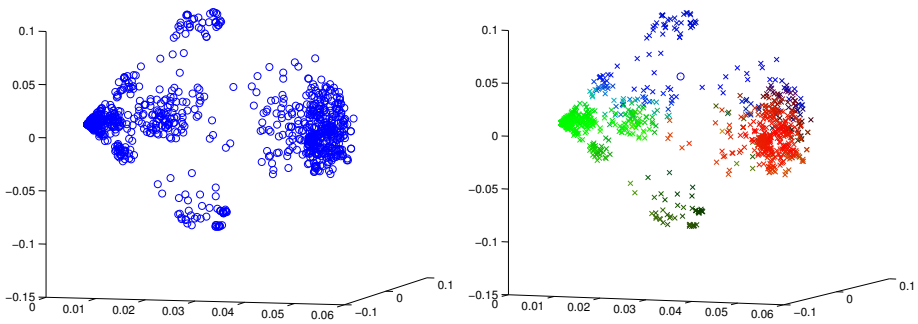


FIGURE 21 – Résultats obtenus par la méthode (I) après réduction de dimensions

Une autre méthode (II) consiste à trier à nouveau ces valeurs (Fig. 20, plot de droite).

$$\mathsf{Y2} = \mathsf{sort}(\mathsf{Y})$$

Y2 peut également servir de descripteur (à 12 dimensions). Celui-ci semble plus robuste par le fait que les positions des k-index sont triées. Si l'on tente une explication, on pourrait dire que ces valeurs traduisent une certaine **dispersion au voisinage du maximum**. Les résultats sont présentés après réduction dimensionnelle (Fig. 22).

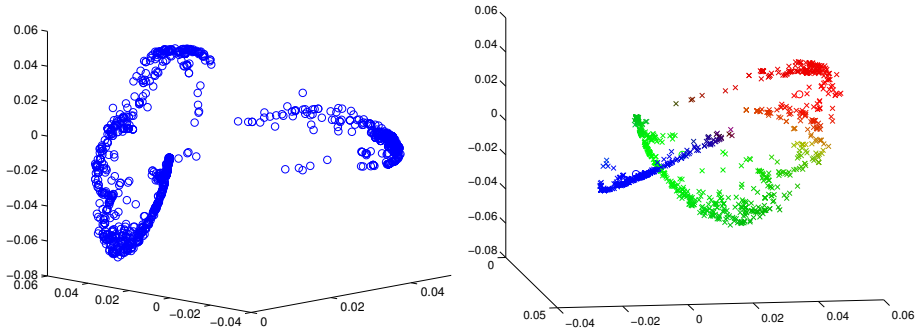


FIGURE 22 – Résultats obtenus par la méthode (II) après réduction de dimension

Enfin, en reprenant la méthode (II), il est plus judicieux d'extraire un descripteur à 2 paramètres dont les valeurs seraient les coefficients d'une régression linéaire de $Y2$ dans L_1 (méthode III). Cela correspond au problème de minimisation suivant,

$$W_{k1} = 1, \quad W_{k2} = k, \quad k = 1 \ldots 12, \quad \underset{Y}{\operatorname{argmin}} \ ||WY' - X||_1$$

Pour résoudre cela (numériquement), il suffit d'itérer pour chaque $k = 1 \ldots m$.

$$D_{(k)} = \operatorname{diag}\left(\frac{1}{\epsilon + |WY'_{k,:} - X_{:,k}|}\right)$$
$$Y_{k,:} = (W'D_{(k)}W)^{-1}W'D_{(k)}X_{:,k}$$

Remarque : Ces 2 paramêtres représentent la position du max, ainsi qu'un genre de dispersion. Un plot de ce descripteur à 2 dimensions (donc sans réduction de dimension) est présenté Fig. 23. Ce descripteur semble offrir un pouvoir discriminant plus pertinent que les méthodes « aveugles ».

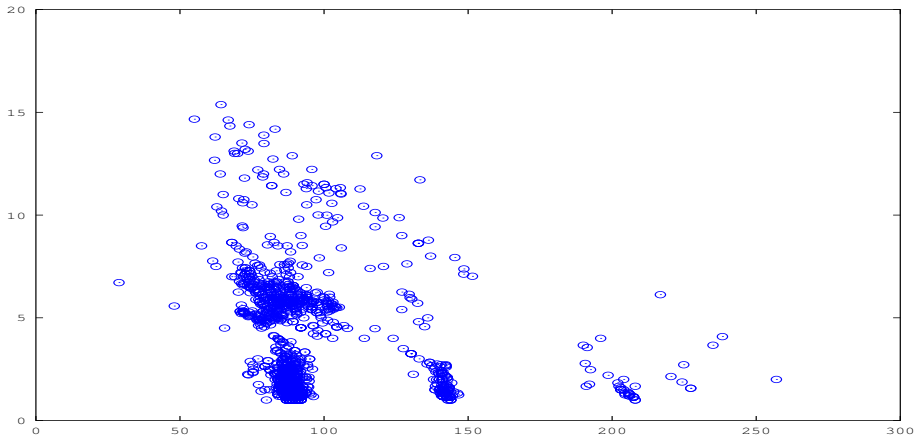


FIGURE 23 – Visualisation du descripteur à 2 dimensions de la méthodes (III)

Approche Temps fréquence

Les représentations temps-fréquence permettent de révéler le contenu fréquentiel d'un signal, tout en conservant une certaine localisation temporelle. La figure 24 représente le spectrogramme d'un signal étudié (le plan temps-fréquence). On observe que le signal est composé de trains d'ondes successifs (ou "bursts"), qui semblent situés à peu près à la même fréquence. Il semble naturel de se focaliser sur le premier burst, qui est d'amplitude maximale, les bursts suivant pouvant correspondre à des phénomènes de réflexion et ne portant pas d'information intrinsèque.

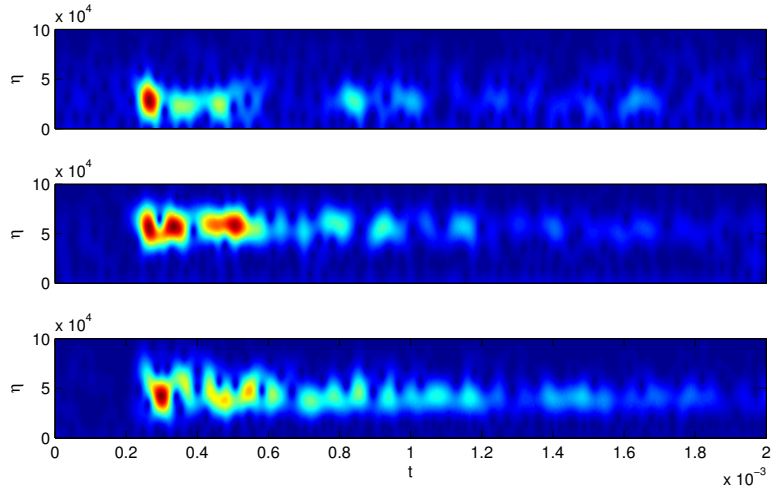


FIGURE 24 – Spectrogramme de trois signaux de la base de donnée.

Notre premier essai consiste à extraire pour chaque signal de la base de données le point du plan temps fréquence où le spectrogramme est maximal. La position fréquentielle de ce point ainsi que son amplitude sont des descripteurs intéressants. La classification correspondante, obtenue par l'algorithme k-means avec trois classes, est représentée dans la figure 25. Ce résultat est relativement décevant, puisque ces descripteurs ne semblent pas produire des classes bien distinctes.

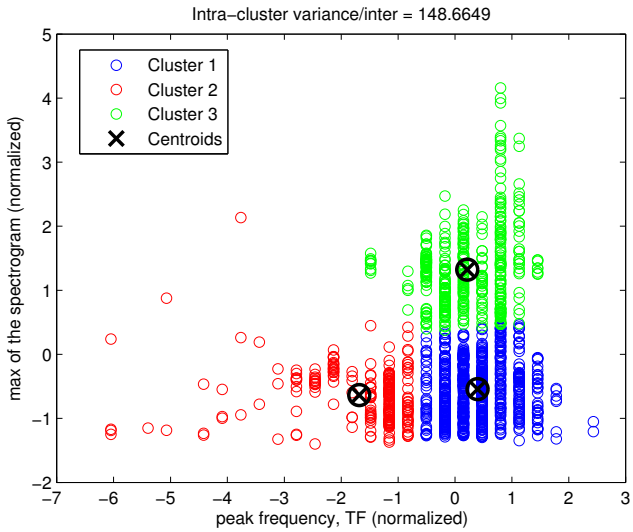


FIGURE 25 – Les signaux de la base de donnés représentés par les descripteurs temps-fréquence. Les couleurs correspondent aux trois classes obtenues par k-means.

Modélisation paramétrique

Malgré plusieurs tentatives, les approches temps-fréquence ne nous ont pas permis d'obtenir des classifications pertinentes. Ceci peut être dû à plusieurs facteurs, notamment au fait que le contenu fréquentiel semble stable au cours du temps. Cela nous a donc encouragé à envisager de nouvelles méthodes basées sur la transformée de Fourier. Une première idée consiste à modéliser les spectres des signaux, à partir d'une connaissance (ou d'une intuition) physique. Dans notre cas, il semble naturel de considérer nos signaux comme des ondes harmoniques, amorties exponentiellement. Notre modèle de signal dépendra ainsi de trois paramètres (α, β, ω) correspondant respectivement à l'amortissement, l'amplitude maximale et la fréquence des ondes :

$$f_{\alpha,\beta,\omega}(t) = \beta \cos(2\pi\omega t) e^{-\alpha t}, \ t > t_0. \tag{8}$$

Un rapide calcul permet d'obtenir le modèle des spectres correspondant :

$$|\hat{f}_{\alpha,\beta,\omega}(\nu)| = \frac{2\beta}{|\alpha + 2i\pi(\nu - \omega)|}. \tag{9}$$

Une régression paramétrique standard (en utilisant l'algorithme de Matlab avec les options par défaut) sur chaque signal permet de représenter les signaux par leurs paramètres, et semble fonctionner correctement. La figure 26 montre ainsi le spectre de deux signaux de la base, ainsi que leur spectre modélisé

correspondant. Le résultat semble correct, et relativement stable (on n'a pas besoin de débruiter les signaux auparavant).

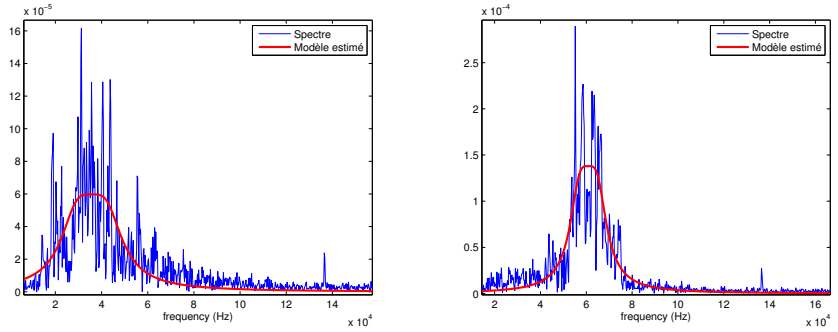


FIGURE 26 – Deux spectres de signaux de la base de données (en bleu), avec leur modélisation paramétrique (en rouge).

Pour évaluer la pertinence de ces descripteurs, on représente dans la figure 27 chacun des signaux dans l'espace de paramètre (α, β, ω). Comme pour les autres decripteurs, on réalise également une classification, et l'on colorie chaque point selon la couleur de la classe auquel il appartient. La figure 27 représente le résultat des classifications en imposant 5 et 6 classes. Il est clair que ces descripteurs permettent de faire ressortir des structures, et donc de classifier assez correctement les signaux.

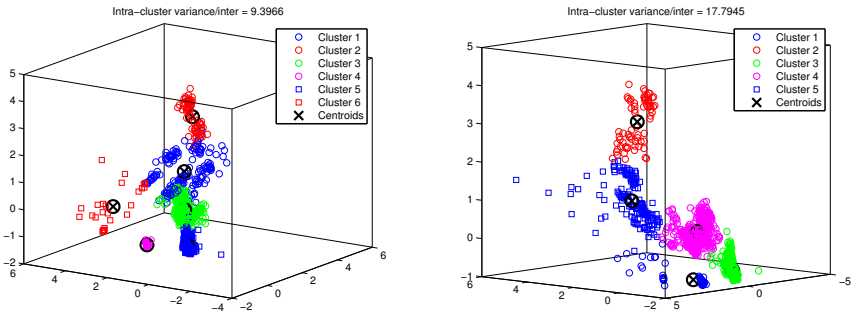


FIGURE 27 – Classifications correspondantes, avec l'algorithme k-means utilisant 6 et 5 classes.

Pour comparer qualitativement la classification obtenue, on peut représenter les signaux dans l'espace (α, β, ω), mais en les coloriant selon une autre classification. On vérifie ensuite "à l'oeil nu" si les couleurs correspondent aux structures visibles. La figure 28 montre une telle comparaison, en utilisant les

descripteurs paramétriques et les descripteurs basiques (position et amplitude du maximum du spectre).

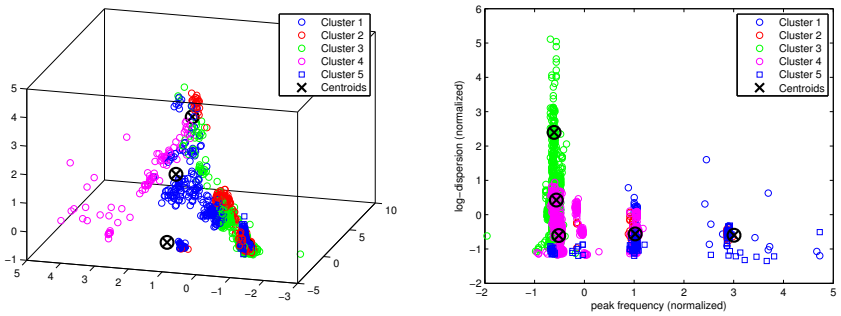


FIGURE 28 – Comparaison avec la classification basique. A gauche, les signaux sont représentés dans l'espace des descripteurs paramétriques, et coloriés selon la classification basique. A droite, on fait l'inverse. On utilise toujours le k-means, avec 5 classes.

Fréquence et amplitude instantanée

Nos signaux accoustiques étant des ondes modulées en amplitude et en fréquence, il est possible de s'intéresser à d'autres grandeurs spectrales que leur spectre ou leur spectrogramme. Pour cela, on commence par calculer le signal analytique $f_a = (\mathcal{I} + i\mathcal{H})f$, où $\mathcal{H}$ est la transformée de Hilbert. Cette opération consiste à enlever toutes les fréquences négatives, et on obtient ainsi un signal complexe s'écrivant $f_a(t) = a(t)e^{2i\pi\phi(t)}$. On peut alors définir la fréquence instantanée $\phi'(t)$ et l'amplitude instantanée $a(t)$ qui caractérisent le signal. La figure 29 trace ces grandeurs instantanées pour un signal de notre base de données. On observe à nouveau que la fréquence du signal semble quasiment constante au cours du temps, alors que son amplitude instantanée dessine des pics amortis, correspondants à chacun des bursts visualisés dans le plan temps-fréquence. Nous pouvons envisager plusieurs des cripteurs tirés de cette représentation, comme la dispersion des bursts, leur nombre, ou les variations moyennes des grandeurs. Nous n'avons malheureusement pas eu le temps de tester de telles approches.

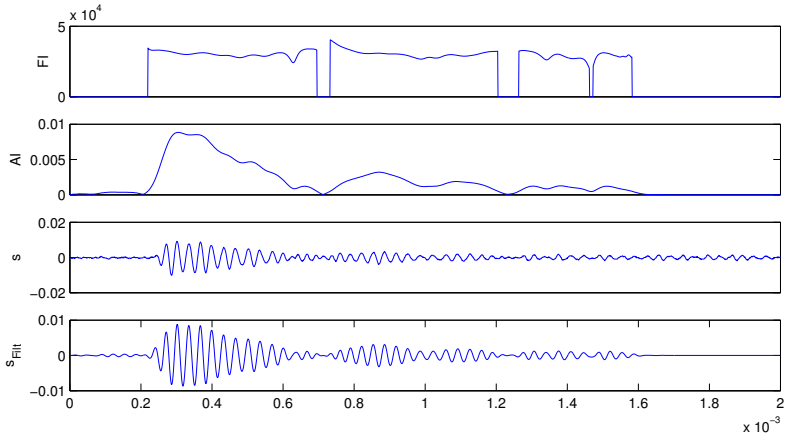


FIGURE 29 – Un signal acoustique ainsi que ses grandeurs instantanées. De bas en haut : le signal filtré, le signal original, son amplitude et sa fréquence instantanées.

3.5 Distance de similarité entre spectres

3.5.1 Analyse de clusters basée sur moments

Nous nous intéressons ici à l'emploi de la moyenne et de la variance, ou de la déviation standard, dans la construction de clusters. A ce propos, notre première remarque est que de tels indicateurs sont en relation avec les deux premiers moments de la variable aléatoire qu'on analyse. En effet, la moyenne correspond exactement au premier moment et la variance est obtenue à partir des deux premiers moments. Bref, on pourrait procéder à la construction de clusters en utilisant les deux premiers moments.

Par ailleurs, étant donné qu'une variable aléatoire dispose, en général, de plusieurs moments, on est tenté d'utiliser plusieurs moments pour notre analyse. Certaines variables ont des moments de tous les ordres, et donc tous peuvent potentiellement être utilisés. De plus, les moments peuvent correspondre à d'autres ordres que les entiers.Nous allons présenter notre stratégie d'analyse. Pour cela, on montrera dans un premier temps la procédure de calcul des moments. Afin d'évaluer la qualité des clusters formés, on appliquera une analyse de discrimination linéaire dont on peut déduire le pourcentage de cas mal classés, indicateur qui sera utilisé pour établir des pistes pour des analyses ultérieures.

Moments d'une variable aléatoire

Étant donné que par la suite nous exploiterons la notion de moment en considérant des ordres réels, on donne la définition suivante pour $r \in \mathbb{R}^+$. On rappelle qu'une variable aléatoire positive X de densité continue f admet un moment d'ordre r, $r \in \mathbb{R}$ si :

$$m^{(r)} = E(X^r) = \int_0^\infty x^r f(x)dx, \text{ avec } E(X^r) < +\infty,$$

(voir par exemple [5]).

Les cas classiques sont par exemple le cas $r = 1$, qui correspond à la moyenne, ou le cas $r = 2$ qui intervient dans la formulation variance : $\sigma^2 = m^{(2)} - (m^{(1)})^2$. Pour la suite nous nous intéresserons à une variable aléatoire de densité spectrale f, supposée connue seulement en certains points x_i. Par conséquent nous nous intéresserons non pas à m^r, mais à son estimateur $\hat{m}^{(r)}$:

$$\hat{m}^{(r)} = \frac{1}{n} \sum_{i=1}^{n} x_i^r f(x_i).$$

Construction de clusters

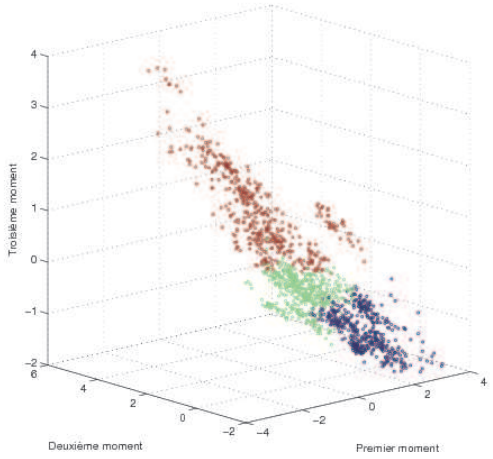
Avec les données disponibles, on peut calculer environ 1500 moments pour un ordre donné, soit un pour chaque série disponible.

Afin de procéder à la construction des différentes classes, nous utiliserons en premier lieu les expressions $\left(\hat{m}^{(r)}\right)^{1/r}$ pour comparer les différents moments avec $r = 0,5, 1, 0, 1, 5, 2, 0, 2, 5, 3, 0, 3, 5, 4, 0, 4, 5,$ et $5, 0$. Nous avons en outre normalisé ces résultats afin d'éviter d'avoir à traiter des variables trop grandes. À la série $\left(\left(\hat{m}_i^{(r)}\right)^{1/r}, i = 1, \ldots, s\right)$, nous préférerons donc utiliser la série $(y_{i,r})$ définie pour $i = 1, \ldots, s$ par :

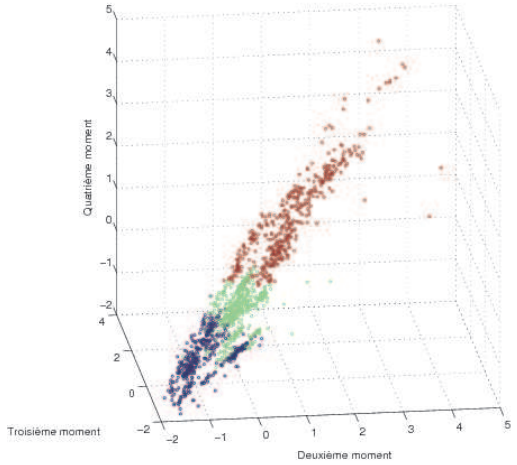
$$y_{i,r} = \left[\left(\hat{m}_i^{(r)}\right)^{1/r}\right]' = \frac{\left(\hat{m}_i^{(r)}\right)^{1/r} - \frac{1}{s}\sum_{j=1}^{s}\left(\hat{m}_j^{(r)}\right)^{1/r}}{\left[\frac{1}{s-1}\sum_{i=1}^{s}\left(\left(\hat{m}_i^{(r)}\right)^{1/r} - \frac{1}{s}\sum_{j=1}^{s}\left(\hat{m}_j^{(r)}\right)^{1/r}\right)^2\right]^{1/2}}.$$

La stratégie utilisée pour construire les classes, au nombre de 3, est le regroupement hiérarchique aussi connue sous le nom de Méthode de Ward ou méthode de minimisation de la variance intra-classe, (voir par exemple [1]).

Avec ces outils on obtient les premiers résultats représentés dans la Figure 30. On y remarque que les classes recherchées sont différenciées entre elles, mais pas complètement. Il reste certains sous-ensembles où les classes sont superposées. Un autre fait à souligner dans ces résultats est la linéarisation des classes lorsque les ordres des moments augmentent. En effet, les observations dans le premier graphique sont plus dispersés que dans le deuxième graphique.



(a) Moments 1, 2, et 3



(b) Moments 2, 3, et 4

FIGURE 30 – Identification de 3 clusters en utilisant des moments

Afin d'évaluer la qualité des classes trouvées, nous utiliserons l'analyse discriminante linéaire (ADL), voir par exemple [2], en prenant la variable objective comme mesure de dissimilarité. Pour finir nous analyserons le taux de cas mal classés par l'ADL.

Dans ce problème nous construisons donc trois classes par la méthode du regroupement hiérarchique. Tout ceci est bien sûr fait dans l'optique d'explorer un certain nombre de pistes à propos de l'apport potentiel de l'utilisation des

moments pour la classification de ce type de données.

La Table 1 montre quelques résultats qui nous permettent de dégager des pistes intéressantes sur l'emploi des moments pour la construction de classes. On observe tout d'abord que le pourcentage de cas mal classés est plus élevé si le nombre de variables utilisées diminue ; cependant, le neuvième essai nous montre le contraire par rapport aux essais 10, 11, et 12. On observe ensuite que l'augmentation du nombre de variables permet généralement de réduire le nombre de cas mal classés, mais on peut trouver des situations où ce comportement n'est pas vérifié : comparer par exemple les essais 2 et 3. Enfin, l'augmentation de l'ordre des moments ne coïncide pas toujours avec la diminution des cas mal classés : une telle diminution est constatée en comparant les essais qui commencent avec 0,5 et 1,0, mais non avec 2,0. Pour résumer, s'il y a clairement un ensemble de moments pour lesquels on peut réduire le taux de cas mal classés, ce fait n'est pas forcement lié à un grand nombre de moments, comme l'essai 9 le montre, ni même avec des ordres de moments plus bas ou plus élevés.

La meilleure combinaison est celle présentée dans l'essai 9. Cette combinaison montre que le problème d'une classification basée sur les moments revient à un problème de minimisation du taux de cas mal classés selon l'ADL (ce critère pourrait changer si on décidait de changer la mesure d'évaluation de la qualité des classes). Notons bien qu'on n'a analysé qu'un nombre assez limité d'ordres de moments, et donc qu'il serait pertinent d'étudier l'apport d'autres moments (par exemple 1,1, 1,2, etc...).

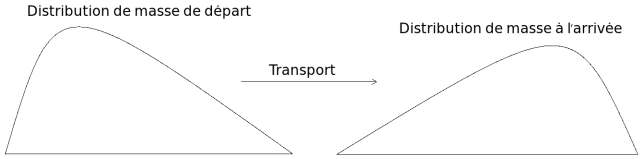
Essais	Variables (ordres des moments)	Cas mal classés (%)
1	0,5, 1,0, et 1,5	9,96
2	0,5, 1,0, 1,5, et 2,0	5,48
3	0,5, 1,0, 1,5, 2,0, et 2,5	6,18
4	0,5, 1,0, 1,5, 2,0, 2,5, et 3,0	6,04
5	1,0, 1,5, et 2,0	6,98
6	1,0, 1,5, 2,0, et 2,5	4,64
7	1,0, 1,5, 2,0, 2,5, et 3,0	4,66
8	1,0, 1,5, 2,0, 2,5, 3,0, et 3,5	3,51
9	1,5, 2,0, et 2,5	1,93
10	1,5, 2,0, 2,5, et 3,0	3,78
11	1,5, 2,0, 2,5, 3,0, et 3,5	2,83
12	1,5, 2,0, 2,5, 3,0, 3,5, et 4,0	2,27
13	2,0, 2,5, 3,0	3,41
14	2,0, 2,5, 3,0, et 3,5	2,88
15	2,0, 2,5, 3,0, 3,5, et 4,0	3,84
16	2,0, 2,5, 3,0, 3,5, 4,0, et 4,5	4,80

0.0 2.5 5.0 7.5 10.0

TABLE 1 – Evaluation du classement des clusters selon l'ADL

3.5.2 Regroupement hiérarchique et distance de Wasserstein

Comme expliqué précédemment et comme on a pu l'observer, le regroupement hiérarchique pour la norme L^2 n'a pas du tout porté ses fruits. Nous avons donc testé cette méthode avec une distance plus adaptée que la distance L^2 pour notre problème. Nous avons choisi d'utiliser une distance de Wasserstein entre spectres (en module). La théorie du transport optimal associée à cette distance est très utilisée en traitement d'image (distance et transport d'histogramme) bien qu'elle ne se soit pas encore développée en acoustique.
La distance de Wasserstein entre deux mesures positives μ et ν sur $\mathbb{R}$ de même masse correspond à l'énergie nécessaire pour transporter la distribution de masse μ sur la distribution de masse ν sachant que le coût pour le transport d'une masse m sur une distance d vaut $d^p \times m$. Ici $p \geq 1$. Si $p = 1$ le coût est proportionnel à la masse ainsi qu'à la distance. Ce problème peut être reformulé du point de vue eulerien par sa formulation cinétique : Le problème consiste alors à minimiser le travail total d'un déplacement conservatif (qui vérifie l'équation de conservation de la masse).



On note généralement $W_p(\mu, \nu)$ cette distance. Elle est donc définie de la manière suivante :

$$W_p(\mu, \nu)^p = \underset{T \in \mathcal{T}(\mu,\nu)}{inf} \left(\int_{\mathbb{R}} |T(x) - x|^p \, d\mu(x) \right)$$

où $\mathcal{T}(\mu, \nu) = \{T : \mathbb{R} \to \mathbb{R}$: "T envoie la distribution μ sur ν"$\}$=$\{T : \mathbb{R} \to \mathbb{R}$: $T_{\#}\mu = \nu\}$.
En particulier, dans cette définition, les densités μ et ν sont de même "masse". Il faudrait donc renormaliser les signaux. Cependant ceci implique une très forte dépendance du signal au bruit qui est un fort perturbateur de "masse", et qui donc, une fois les signaux renormalisés, provoque de fortes variations relatives des fréquences du spectre dominant.
Une solution à ce problème est d'autoriser la présence d'un terme source de masse, dans la formulation eulerienne, plutôt que de considérer un déplacement conservatif. De cette façon les fréquences du spectre dominant ne seront pas changées. On peut alors tester une classification hiérarchique pour la distance de Wasserstein.
Cette méthode a l'avantage (comparativement à la classification L^2) d'être robuste par rapport aux translations. En fait la distance de Wasserstein entre deux densités translatées d'une constante c est égale à $|c|$.

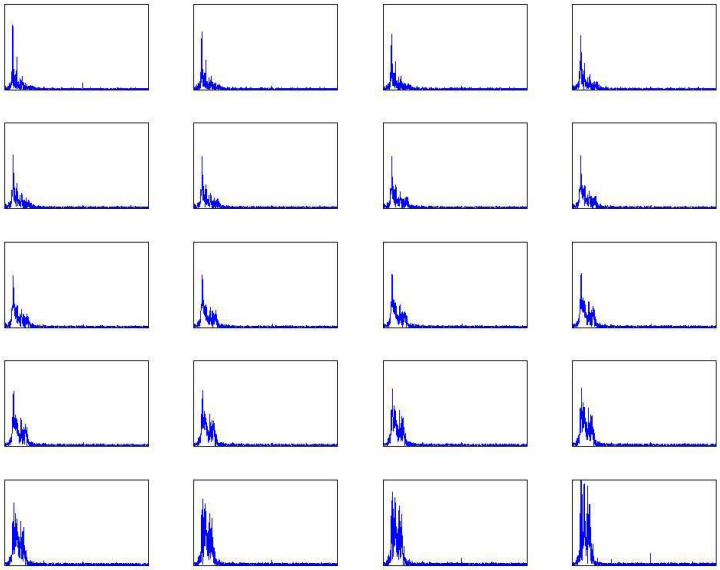


FIGURE 31 – Transport optimal avec terme source entre les spectres de deux signaux

Cependant cette méthode présente des inconvénients indéniables :

1. Cette approche est considérablement limitée par le temps de calcul assez long de la distance de Wasserstein entre deux signaux (Algorithmes de descente de gradient, de Benamou-Brenier, méthodes proximales, etc...). En particulier il n'est pas envisageable en quelques jours et sur un ordinateur de bureau de faire une classification hiérarchique en considérant l'intégralité des données fournies dans le cadre du projet.
2. Ce critère n'est probablement pas assez robuste (par rapport aux bruits engendrés notamment par la dégradation de la barre) : Une petite perturbation risque d'influencer fortement la distance de Wasserstein. Ce problème peut être atténué en lissant le spectre.

Pour ces raisons, même si le temps de calcul pour toutes les données n'est pas prohibitif (quelques jours sur un ordinateur de bureau), avec le peu de temps que nous avions, nous n'avons pu faire les calculs que sur un nombre limité de données. C'est pourquoi l'interprétation statistique de ces résultats ne peut pas être totalement pertinente.
En outre, cette méthode ne nous a pas permis d'observer avec certitude une corrélation avec la classification en Fréquence/Amplitude maximale : En utilisant les mesures de dissimilarités (5) et (6), on obtient des résultats qui semblent

confirmer la pertinence des trois classes obtenues avec la fréquence du pic mais ce n'est pas le cas pour (7). Voici les résulats partiels que nous avons obtenus qui sont, encore une fois, à nuancer avec le peu de données utilisées :

1. Pour une dissimilarité définie avec le min :
   dissimilarité intra-classes = 0.013 (minimum des distances entre 2 éléments d'une même classe).
   dissimilarité inter-classes = 0.025 (minimum des distances entre 2 éléments appartenant à 2 classes différentes).
2. Pour une dissimilarité définie avec le max :
   dissimilarité intra-classes = 1.0
   dissimilarité inter-classes = 1.4

Toutefois, une classification hiérarchique au sein de chacune des trois classes obtenue avec la fréquence du pic semble montrer l'existence de sous-classes ce qui nous a laissé penser que cette méthode devrait plutôt être utilisée pour affiner la classification déjà obtenue avec les descripteurs pic/amplitude max.

Références

[1] The cluster procedure. *SAS OnlineDoc : Version 8*, 1999.

[2] The discrim procedure. *SAS/STAT(R) 9.2 User's Guide*, 2009.

[3] Frédéric de Coulon. *Théorie et traitement des signaux*. Traité d'Électricité [Treatise on Electricity], VI. Georgi Publishing Co., St., 1984.

[4] J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Stanford University, 1984.

[5] Loève. Probability theory 1. *Springer-Verlag*, 1977.

[6] R. Patel and J. Rudlin. Analysis of erosion/corrosion incidents in offshore process plant and the implications for ndt. *Insight*, 42 (1) :17–21, 2000.

[7] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.*, 36 (8) :1627–1639, 1964.

[8] Ronald W. Schafer. What is a savitzky-golay filter ? [lecture notes]. *IEEE Signal Process. Mag.*, 28(4) :111–117, 2011.

[9] Claude E. Shannon. Communication in the presence of noise. *Proc. I.R.E.*, 37 :10–21, 1949.