

Différentes méthodes de détection de nouveautés/anomalies/outliers permettent d'étudier et de classifier des ensembles de données dont le label atypique est très faiblement représenté. Ces méthodes passent toutes par la caractérisation des données normales, le calcul d'une mesure de nouveauté et la comparaison à un seuil. Ces techniques peuvent être répertoriées en 5 classes distinctes [90] :

- les approches probabilistes,
- les approches basées sur les distances,
- les approches basées sur la reconstruction de la donnée,
- les approches basées sur la caractérisation des limites des données normales,
- les approches basées sur la théorie de l'information.

Le choix d'une méthode par rapport à une autre est effectué selon les caractéristiques des données comme la dimension ou la structure du domaine des données normales.

### 2.2.1 Les approches probabilistes

Ces approches estiment la distribution sous-jacente des données normales, définissant ainsi un modèle de normalité. La distribution des données nominales n'est généralement pas connue. Elle peut alors être estimée à partir de mélanges de gaussiennes [15]. Il s'agit là d'une méthode paramétrique dont les paramètres principaux sont le nombre de gaussiennes du mélange et leurs caractéristiques (moyenne et covariance). Chaque gaussienne correspond à un comportement normal. Les paramètres du modèle sont estimés à partir du maximum de vraisemblance via l'algorithme Expectation-Maximisation (EM). Une seconde approche pour estimer la distribution sous-jacente est l'estimation de la densité par noyau [49]. Cette méthode non-paramétrique consiste en l'application d'un noyau probabiliste au niveau de chaque donnée d'apprentissage normale. Une explication plus détaillée de cette approche est donnée dans la section 5.3. Le principal paramètre de cette approche est l'échelle du noyau considérée. Si elle est trop importante, la distribution est fortement lissée. Si elle est trop faible, la distribution est trop fortement liée aux données d'apprentissage (surapprentissage) et ne généralise pas le modèle de normalité.

La nouveauté est détectée par la comparaison de la donnée test à la distribution du modèle de normalité. Dans un cas unimodal, c'est-à-dire lorsque la distribution possède un unique mode, les nouveautés se trouvent au niveau des queues de distribution et sont donc détectées en fixant un seuil sur cette distribution. La théorie des valeurs extrêmes [37] est une des méthodes permettant de fixer un seuil de détection, elle permet de modéliser la distribution du maximum d'un échantillon de taille fixe. D'après le théorème de Fisher-Tippett [48], dépendant de la loi des données, la distribution des valeurs extrêmes correspond à une loi de Gumbel, Fréchet ou Weibull. La caractérisation de ce maximum permet alors de définir le seuil de détection. Dans un cadre multi-modal, les nouveautés ne sont plus caractérisées uniquement par les extrêmes de la distribution, mais aussi par les points de faibles densités entre les différents modes. Le seuil de détection peut alors être défini par une valeur de densité en dessous de laquelle les points sont considérés comme improbables donc nouveaux [32].

Ces approches permettent de définir la normalité à travers la distribution des données et de détecter la nouveauté à partir d'un seuil sur la densité des données. Ces approches sont dépendantes du nombre de données utilisées pour estimer la distribution. Cependant, en grande dimension l'apprentissage de la densité demande un nombre très important de données pour caractériser la distribution au niveau de toutes les régions de l'espace. Ainsi plus la dimension est grande, plus le nombre de données nécessaires pour caractériser tout l'espace est important. Il s'agit du "fléau de la dimension". Ces approches ne sont donc pas adaptées aux problématiques en grande dimension.

### 2.2.2 Les approches basées sur les distances

Elles consistent en la mesure de la distance entre une donnée et ses homologues normales. Plus la distance est grande, plus cette donnée peut être considérée comme nouvelle. Il existe deux grandes catégories d'approches utilisant les distances. La première considère les plus proches voisins des données tests comme les  $k$  plus proches voisins ( $k$ -NN) [49]. Elle permet de classifier les données comme nouvelles si elles se trouvent éloignées des données normales. Cette méthode est paramétrée par le nombre  $k$  de voisins à prendre en compte et la distance considérée. Le temps de calcul est généralement long pour de grosses bases de données. Une seconde méthode basée sur les voisins est le Local Outlier Factor (LOF) [17]. Elle compare la densité de voisins autour des données étudiées dans un certain rayon de voisinage (constituant le principal paramètre de l'approche) avec cette même densité calculée pour leurs voisins. Une densité de voisins plus faible signifie une isolation de la donnée et donc son caractère nouveau. La seconde catégorie d'approches basées sur les distances correspond au clustering des données dont la méthode la plus répandue est le  $k$ -means [49],  $k$  correspondant cette fois-ci au nombre de clusters. La détection de nouveautés s'effectue dans ce cadre par rapport à la distance au plus proche cluster estimé sur les données normales. Il existe plusieurs choix de distances possibles avec des propriétés différentes constituant un paramètre supplémentaire de ces approches.

En grande dimension, la notion de distance est mal établie, ces approches ne sont donc pas efficaces. De plus, la mise en place de ces approches nécessitent un très grand nombre de données pour caractériser correctement l'espace des données normales et pour éviter les fausses détections.

### 2.2.3 Les approches basées sur la reconstruction des données

Ces approches définissent un modèle caractérisant la normalité au sein des données à partir duquel il est possible de donner une estimation normale de la donnée test. La détection s'effectue alors sur le résidu de la reconstruction. Dans le cas d'une donnée normale, la reconstruction de cette dernière est proche de celle-ci et donc entraîne des résidus de reconstruction faibles. Pour une donnée contenant des nouveautés ou des anomalies, les résidus de la reconstruction sont plus importants. Pour ce type d'approche, l'élément déterminant est le nouvel espace de représentation, caractérisant la normalité des données et dans lequel les données sont représen-

tées [13]. Cet espace correspond à des dictionnaires pouvant être appris sur les données ou à des réseaux de neurones [68]. Les méthodes de représentation par dictionnaires décomposent la donnée dans un système linéaire avec des contraintes. Dans la suite nous donnons une explication plus détaillée de la représentation par dictionnaire. L'ACP [49] est la méthode la plus classique de représentation des données dans un dictionnaire permettant de maximiser la variance des données normales. La reconstruction est effectuée à partir des premières composantes principales modélisant la variabilité normale des données. Dans le cadre de la détection de nouveautés, il est également possible de projeter les données sur les dernières composantes principales n'ayant que peu d'information normale et détectant ainsi des outliers. Le principal paramètre de ce type d'approche est la nouvelle dimension dans laquelle les données sont représentées. Dépendant de la nouvelle dimension, ce type d'approche ne nécessite pas nécessairement un très grand nombre de données.

Les réseaux de neurones [68] ont été particulièrement utilisés pour des tâches de classification supervisée mais également de détection de nouveautés. Les Replicator Neural Network (RNN) [53] et les autoencodeurs [109] sont des réseaux de neurones apprenant des sous-espaces permettant la reconstruction de la donnée d'entrée. Les réseaux sont donc appris sur des données normales afin de caractériser cet espace de normalité, la détection de nouveautés s'effectue également sur les résidus de la reconstruction. Les principaux paramètres de ces approches sont le pas du gradient mais également la structure et la composition des couches du réseau de neurones. La construction de ces réseaux nécessitent un très grand nombre de données pour éviter le surapprentissage.

Ces approches fonctionnent sur des données en grande dimension telles que des images. Les espaces dans lesquelles les données sont projetées permettent d'apprendre la structure et le comportement normal des données. Les paramètres de ces différentes approches correspondent aux paramètres des espaces dans lesquels les données sont projetées comme la nouvelle dimension ou la structure du réseau.

#### 2.2.4 Les approches basées sur la caractérisation des limites des données normales

Ces approches caractérisent les bordures de la normalité dans l'espace des données. Dans ce sens, elles ne tiennent pas compte de l'ensemble des données mais uniquement de celles se trouvant proches de ces bordures. La détection de nouveautés correspond alors à la comparaison de la donnée aux limites du domaine normal. Ces méthodes sont apparentées aux support vector machines (SVM) [49]. Un one-class SVM [99] permet de trouver un hyperplan dans un espace transformé des données d'apprentissage normales définissant de vastes marges par rapport à l'origine. Les principaux paramètres de ces approches sont le nombre de points définissant les bordures et un pourcentage de points pouvant être mal classifiés pour éviter le surapprentissage.

Le support vector data description (SVDD) [107] définit un domaine caractérisé par un centre et un rayon autour des données normales minimisant le volume de cette hypersphère. Cette

caractérisation s'effectue également dans un espace transformé. La nouveauté est détectée par l'appartenance ou la non-appartenance de la donnée à l'hypersphère. Les principaux paramètres sont similaires à ceux du one-class SVM.

Il existe d'autres approches caractérisant les bordures de l'espace de normalité non liées au SVM comme les one-class random forests [38]. Ces approches sont également intéressantes sur des données en grande dimension car seuls les points aux bordures sont étudiés. Elles requièrent tout de même un grand nombre de points pour la calibration du modèle.

### 2.2.5 Les approches basées sur la théorie de l'information

Ces approches proviennent de l'intuition que la présence d'une donnée nouvelle ou atypique au sein d'un ensemble de données normales entraîne un changement de la quantité d'information estimée sur toutes les données. La mesure de l'information [15] permet alors de détecter des nouveautés en calculant l'entropie (possible mesure de l'information) pour des sous-ensembles des données. Lorsque l'entropie diminue, cela signifie que la donnée retirée est différente du reste [56].

Ces approches de théorie de l'information ne nécessitent aucun a priori sur les données, seule la mesure de l'information des données est considérée. Cependant il est indispensable de définir la mesure considérée pour l'information. De plus la détection s'effectue uniquement lorsque plusieurs observations sont atypiques dans le jeu de données. Dans le cas contraire, le retrait d'une seule observation peut ne pas être suffisant pour entraîner une modification significative de la mesure d'information. De plus, ces mesures passent souvent par la définition d'une densité de probabilité. Ce type d'approche n'est donc pas adapté en grande dimension.

## 2.3 La détection de nouveautés appliquée aux données vibratoires

### 2.3.1 Application aux données temporelles et fréquentielles

Les idées de détection de nouveautés et d'anomalies sont courantes en traitement du signal. Les signaux récupérés sur des systèmes industriels complexes sont en grande majorité des données considérées comme normales. La complexité des systèmes rend impossible de considérer toutes les signatures atypiques possibles. Une pratique standard du traitement du signal consiste à calculer le spectre à partir des données vibratoires temporelles (lorsque ces dernières sont stationnaires) et de comparer le pics fréquentiels obtenus avec ceux issus de comportements normaux. Il s'agit bien d'une approche de détection de nouveautés où les irrégularités sont caractérisées par l'apparition de pics inconnus. La mesure de nouveauté peut alors être obtenue par un calcul résiduel ou par une matrice de correspondance des fréquences d'apparition des différents pics. Les signaux temporels ou fréquentiels sur lesquels les méthodes de détection de nouveautés sont appliquées

sont généralement stationnaires.

Dans [55], une approche de type peak-over-threshold<sup>1</sup>(POT) (approche liée aux valeurs extrêmes) est étudiée pour la détection de défauts de roulements à partir de mesures vibratoires sur des périodogrammes. Il s'agit de la transformée de Fourier discrète appliquée à des signaux temporels en régime stationnaire. Les maximums des périodogrammes sans signature atypique sont calculés pour chaque fréquence, ces derniers sont comparés à d'autres périodogrammes de la base d'apprentissage. Les points supérieurs aux maximums calculés sur les périodogrammes sont modélisés par une loi de Pareto sur laquelle le seuil de détection est déterminé.

Une approche de type one-class SVM est utilisée dans [23] pour la détection d'anomalies sur des données vibratoires temporelles. Différents indicateurs du signal temporel sont calculés et projetés dans un espace de dimension réduite sur lequel le one-class SVM est appris.

Une comparaison de différentes approches de détection de nouveautés a été réalisée dans [111] à partir de différents indicateurs classiques des signaux temporels et en ordre. Les différentes approches comparées sont l'ACP, l'estimation de densité par noyau gaussien, les k-means, les k plus proches voisins, les one-class SVM et les autoencodeurs. Les résultats montrent un comportement varié des différentes méthodes suivant le type d'anomalie présent dans les données.

Ces méthodes sont établies dans le domaine temporel ou angulaire du signal en supposant la stationnarité de ce dernier. Notre problématique consiste à étudier des représentations temps-fréquence de ces signaux permettant de tenir compte du caractère non-stationnaire.

### 2.3.2 Application sur les harmoniques du signal

Les études citées ci-dessus sont effectuées sur des signaux en régime stationnaire sur lesquels la transformée de Fourier peut être appliquée. Cependant dans les signaux non-issus de banc d'essai, la stationnarité est rarement vérifiée. Des méthodes classiques comme la transformée de Fourier ou des indicateurs sur les signaux temporels ne sont plus pertinents. Il est tout de même possible de construire des représentations temps-fréquence afin d'étudier ces signaux à partir de la STFT, des transformées en ondelettes,... Cela permet de prendre en compte l'aspect fréquentiel du signal (donc les éléments périodiques), mais aussi la non-stationnarité à travers l'aspect temporel. Il est également possible d'étudier juste quelques ordres ou harmoniques spécifiques du signal, c'est sur les premières harmoniques qu'apparaissent certains défauts comme les problèmes d'équilibrage.

Dans [31, 34], des méthodes de détection de nouveautés basées sur une estimation de la distribution des valeurs extrêmes, à partir d'un mélange de gaussiennes, ont été étudiées pour la détection d'anomalies de moteurs d'avions dans un cadre multivarié et multimodal. Les valeurs extrêmes dans ce cadre sont définies comme les valeurs les moins probables, c'est-à-dire celles dont les densités de probabilité sont les plus faibles. Les valeurs extrêmes sont calculées sur banc

---

1. méthode des excès

d'essai [31] pour les premiers ordres des spectrogrammes vibratoires normaux de chaque arbre du moteur, et en vol [34] à partir des points de plus fortes intensités.

Une approche de détection de nouveautés à partir du one-class SVM a été développée dans [54]. Les intensités vibratoires d'ordres prédéfinis des spectrogrammes normaux sont récupérées afin d'apprendre le domaine des données saines.

Notre étude consiste à détecter toute trace des signatures vibratoires inusuelles sur les spectrogrammes. Il est donc indispensable de définir des modèles sur les spectrogrammes et non pas sur quelques ordres de ce dernier.

### 2.3.3 Application aux spectrogrammes

Un masque normal des spectrogrammes [52] a été présenté dans le chapitre 1 permettant la suppression de l'information normale des spectrogrammes et de comptabiliser les éléments restants. Cette approche s'apparente à des méthodes de détection d'anomalies basées sur une distance correspondant à la concordance de l'information vibratoire présente entre les spectrogrammes normaux et le spectrogramme étudié.

Une approche similaire a été développée dans [33]. Les spectrogrammes ont été subdivisées en sous-zones sur lesquelles une normalisation est effectuée en modélisant le bruit par une loi Gamma afin de l'homogénéiser. Cette normalisation est effectuée intra-zone et entraîne des artefacts sur les bords des sous-zones définies. La distribution des valeurs extrêmes est estimée pour chaque sous-zone. Le nombre de points dépassant le seuil de détection défini à partir de la distribution estimée est comptabilisé pour chaque sous-zone des spectrogrammes normaux. L'approche est réitérée pour chaque spectrogramme et le nombre de points dépassant les seuils de détections sont comparés. Une anomalie est déclarée lorsqu'une des zones du spectrogramme étudié possède significativement plus de points détectés que ses homologues des spectrogrammes normaux.

Dans [63], une référence (définie par la moyenne et la variance) est apprise pour tous les points des spectrogrammes en ordre à partir de données normales. Les spectrogrammes tests sont alors comparés à cette référence pour déterminer les points anormaux. Afin que les points soient considérés comme anormaux, il est nécessaire qu'ils définissent une région continue avec un nombre de points et une surface suffisante. Pour que le spectrogramme soit déclaré comme anormal, la surface totale doit être supérieure à un seuil. Il est également possible de récupérer les raies le long de certains ordres particuliers en collectant les voisinages des maxima locaux au niveau des points détectés [64].

Notre travail est complémentaire à ces différentes approches. Nous cherchons à détecter finement toutes signatures atypiques sur les spectrogrammes à partir de notre base de données construite, c'est-à-dire déterminer l'ensemble des points composant les signatures inusuelles.

## 2.4 Caractérisation de la base de données de spectrogrammes construite

### 2.4.1 Répartition des données en sous-ensembles

La figure 2.1 présentait le besoin de plusieurs bases de données disjointes pour mettre en place et tester nos modèles :

- une base d'**apprentissage** comportant uniquement des données normales sans signature inusuelle pour mettre en place le modèle de normalité ( $\Omega_{App}$ ) ;
- une base de **validation** contenant des données avec et sans signatures inusuelles pour calibrer le seuil de détection et les paramètres optimaux du modèle par cross-validation ( $\Omega_{Val}$ ) ;
- une base de **test** pour présenter les résultats et évaluer les performances de nos approches, ( $\Omega_{Test}$ ).

Les données de la base d'apprentissage et des données normales de validation forment un même ensemble dans lequel nous sélectionnons aléatoirement la base d'apprentissage pour définir le modèle et les données normales de la base de validation. Les données atypiques de la base de validation restent identiques, nous pouvons en sélectionner un sous-ensemble pour nos études. La base de test reste la même pour toutes les différentes approches afin de pouvoir les comparer.

Nous avons présenté (section 1.4) la base de données ( $\mathcal{B}_0$ ) construite et indexée à partir de l'extraction automatique des zones atypiques sur les données textuelles d'annotations manuelles des experts. Nous avons également présenté nos propres annotations de quelques points en différentes classes ( $\mathcal{B}_1$ ) sur quelques spectrogrammes (section 1.5.3) afin d'obtenir une vérité terrain plus fine que sur l'échelle d'un patch. Toutes les données se trouvant dans  $\Omega_{Test}$  correspondent à des spectrogrammes annotés sur quelques points (donc  $\Omega_{Test} \subsetneq \mathcal{B}_1$ ). Nous souhaitons nous servir d'eux pour donner des résultats numériques. Les spectrogrammes annotés ponctuellement mais ne faisant pas partie de  $\Omega_{Test}$  font partie de la base de validation  $\Omega_{Val}$  afin de calibrer les métaparamètres des différentes approches comme les seuils de détection. Les moteurs de  $\Omega_{Test}$  restent les mêmes tout au long de l'étude.

La répartition des données au niveau de chaque patch commence par une classification de ces derniers suivant qu'ils sont normaux ou atypiques. Chaque patch  $\{Z_{\mathcal{K}_j}\}_{j=1, \dots, \text{card}(\mathcal{K})}$  est comparé à la base de données en vérifiant l'intersection entre le patch du spectrogramme étudié et les zones atypiques extraites de ce même spectrogramme. Les patchs dont la surface d'intersection est supérieure à un seuil sont considérés comme atypiques et sont envoyés dans la base de validation  $\Omega_{Val}^j$  du patch  $j$  correspondant. Les patchs dont la surface d'intersection est inférieure à un seuil sont envoyés aléatoirement dans la base d'apprentissage  $\Omega_{App}^j$  ou dans la base de validation  $\Omega_{Val}^j$  du patch  $j$  correspondant. Le choix d'un seuil non nul de la surface d'intersection est dû à la récupération d'information normale lors de l'extraction des zones atypiques. Les données d'apprentissage et de validation sont donc différentes selon le patch étudié tandis que les données

de test restent les mêmes sur tous les patchs. Ce processus de répartition des données est détaillé dans l'algorithme 5. Pour les études ponctuelles (point à point) des spectrogrammes, nous utilisons la base  $\mathcal{B}_1$  des annotations ponctuelles. La base de test reste la même (contenant déjà les annotations ponctuelles) et la base de validation correspond aux spectrogrammes de  $\mathcal{B}_1$  non utilisés dans la base de test. La base d'apprentissage reste sélectionnée de la même manière.

---

**Algorithme 5 :** Répartition des données en base d'apprentissage et base de validation

---

**Données :** Base de données  $\Omega$ , base de données  $\mathcal{B}_1$  des spectrogrammes annotés ponctuellement, la subdivision  $\mathcal{K}^{128}$ , la surface minimale d'intersection  $S_{min}$ , le pourcentage de patchs normaux sélectionnés pour la base d'apprentissage  $\tau_{App}$ , les données test  $\Omega_{Test}$

**Résultat :** Les données réparties  $\Omega_{normal}^j, \Omega_{ano}^j, \Omega_{App}^j, \Omega_{Val}^j$

Initialisation :  $\forall j \quad \Omega_{normal}^j = \emptyset, \quad \Omega_{ano}^j = \emptyset, \quad \Omega_{App}^j = \emptyset, \quad \Omega_{Val}^j = \mathcal{B}_1 \setminus \Omega_{Test}$  ;

**pour**  $j$  in  $1, \dots, \text{card}(\mathcal{K}^{128})$  **faire**

**pour** chaque moteur  $i$  de  $\Omega$  **faire**

**si**  $\exists k \text{ zone}_{ano}^i(k) : Surface(Z_j, \text{zone}_{ano}^i(k)) > S_{min}$  **alors**

| Ajout de  $i$  à  $\Omega_{ano}^j$

**sinon**

| Ajout de  $i$  à  $\Omega_{normal}^j$

**fin**

**fin**

Sélection aléatoire parmi  $\Omega_{normal}^j \setminus \Omega_{Test}$  de  $\tau_{App}$  moteurs pour  $\Omega_{App}^j$

$\Omega_{Val}^j = \Omega_{normal}^j \setminus \{\Omega_{Test} \cup \Omega_{App}^j\} \cup \Omega_{ano}^j \setminus \Omega_{Test}$

**fin**

**retourner**  $\Omega_{normal}^j, \Omega_{ano}^j, \Omega_{App}^j, \Omega_{Val}^j, j \in \{1, \dots, \text{card}(\mathcal{K}^{128})\}$

---

$\text{zone}_{ano}^i(k)$  correspond aux zones atypiques présentes dans la base de données annoté (Figure 1.8).

La figure 2.2 présente la répartition entre spectrogrammes normaux et atypiques pour chaque patch parmi les  $n = 493$  moteurs. Chaque sous-rectangle (encadré en noir) correspond à un des patchs, sa position correspond à la position dans le spectrogramme. L'axe des  $N_2$  a donc été divisé en 3 intervalles et l'axe des fréquences en 18 intervalles à partir de la décomposition en patchs carrés de taille 128 pixels. Le coin inférieur droit du sous-rectangle contient le taux de patchs normaux dans la base de données, et le coin supérieur gauche le taux de patchs atypiques. Comme nous l'avions énoncé dans le chapitre précédent (Figure 1.9), les données ne sont pas équilibrées. La grande majorité des patchs possèdent très peu de données atypiques. Les approches de type one-class/détection d'anomalies, permettant de caractériser le comportement normal des patchs, sont donc adaptées à notre problématique. Deux patchs présentent un nombre de données atypiques supérieur à celui des données normales. Il s'agit des patchs les plus complexes où de nombreux types de signatures inusuelles apparaissent. Utiliser des approches supervisées pour caractériser les données atypiques pourrait paraître plus efficace sur ces patchs. Cependant la parcimonie des signatures inusuelles au sein des patchs et la grande variabilité de

ces dernières rendent cette analyse délicate. Caractériser la normalité reste donc plus pertinent. Nous cherchons également à mettre en place des méthodes non dépendantes de la subdivision en patchs établie. Il est donc indispensable que les approches étudiées fonctionnent sur l'intégralité des patchs sans tenir compte de la répartition des moteurs dans ces derniers.

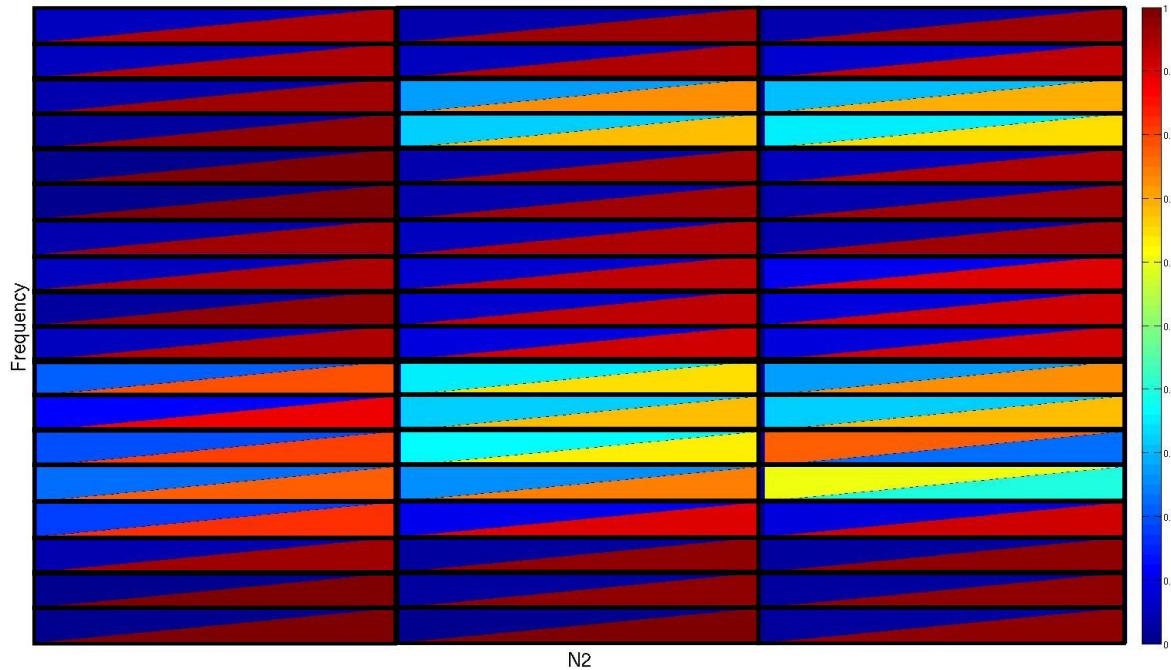


FIGURE 2.2 – Ratio de données normales et atypiques sur les différents patchs allant du bleu (faible proportion) au rouge (forte proportion). Chaque sous-rectangle correspond à un patch avec la partie inférieure droite correspondant au taux de données normales sur ce patch et la partie supérieure gauche aux données atypiques. La grande majorité des patchs possède peu de données atypiques, quelques patchs se distinguent avec une forte proportion de données possédant des signatures inusuelles.

#### 2.4.2 Visualisation des résultats

Dans la thèse, nous présentons des résultats visuels de détection des signatures inusuelles sur un patch spécifique contenant différents types de signatures atypiques. Il s'agit du patch sur lequel quelques points de certains spectrogrammes ont été annotés. Nous avons sélectionné 5 données de ce patch spécifique dans la base de test pour présenter visuellement les résultats, une donnée ne contenant pas de signatures inusuelles et les 4 autres possédant différents formes de signatures atypiques (Figure 2.3). Les approches développées ne sont pas spécifiques à un type de signatures mais caractérisent la normalité présente sur les patchs afin de détecter tout type de signatures inusuelles. Nous représentons sur la figure 2.3 les différents patchs utilisés pour la représentation visuelle des résultats dans ce manuscrit avec un encadrement des signatures inusuelles présentes. Le patch sans encadrement correspond au patch normal.

A partir de nos différentes approches, nous cherchons à détecter les points composant ces

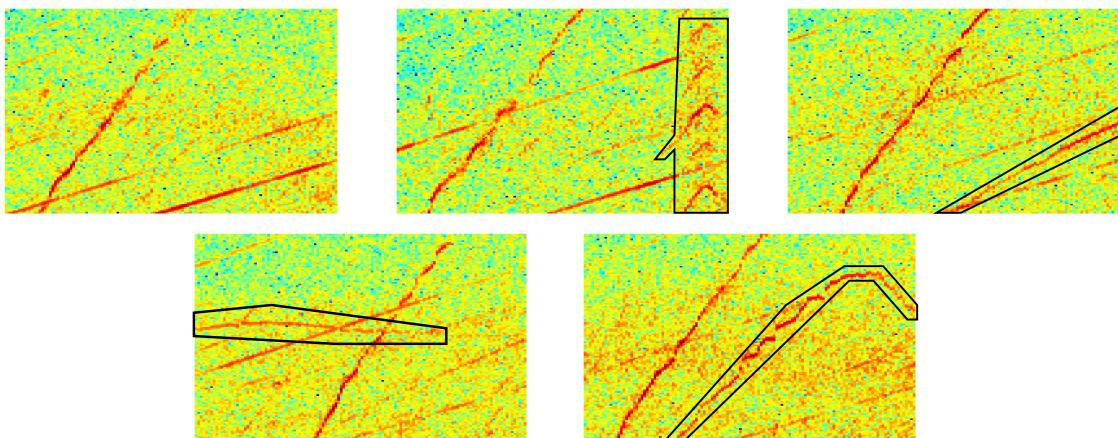


FIGURE 2.3 – Patches de la base de test utilisés pour présenter les résultats visuels des différentes approches. Le premier patch ne contient pas de signature inusuelle, les suivants possèdent tous différentes signatures inusuelles encadrées.

signatures atypiques et ainsi les mettre en évidence pour des détections visuelles. Pour cela, nous modélisons les comportements normaux de chaque patch dans leur globalité et ponctuellement.

## **Deuxième partie**

# **Les approches de représentation globale par dictionnaire**



## Introduction

Les approches par dictionnaires permettent de représenter les données en grande dimension dans un nouvel espace. Dans le cadre de la détection de nouveautés, ce dictionnaire doit permettre de caractériser le comportement normal de nos données, c'est-à-dire l'ensemble des raies présentes sur la très grande majorité des spectrogrammes. Cette approche est appliquée patch par patch de manière indépendante. Les dictionnaires sont définis sur chaque patch et n'ont aucune relation avec les dictionnaires définis sur d'autres patchs. Les patchs sont définis à partir de la subdivision  $\mathcal{K}^{128}$  dont chaque patch est de dimension  $128 \times 128$ . Afin de caractériser la normalité, ces dictionnaires sont calibrés sur les patchs normaux sans signature inusuelle. Nous nous servons donc des zones atypiques extraites de la base de données et de la classification des patchs établis afin de sélectionner les patchs d'apprentissage.

Nous cherchons à définir un espace, défini par les atomes du dictionnaire, caractérisant les éléments normaux sur les patchs afin d'y projeter nos données. Nous obtenons ainsi une représentation normale des données à partir de laquelle nous reconstruisons les patchs. Dans ces reconstructions issus du dictionnaire de normalité, les signatures atypiques présentes dans les données se retrouvent absentes ou réduites. Les reconstructions correspondent alors à des estimations normales des patchs. Nous utilisons des dictionnaires dont la reconstruction s'écrit comme une combinaison linéaire des atomes de ces derniers. Ce choix est volontaire et vient de notre interprétation des spectrogrammes comme une superposition de différentes raies pouvant être interprétées comme des sources. Les atomes du dictionnaire sont appris sur les données ou sélectionnés dans un dictionnaire plus large à partir d'un seuillage des coefficients de la décomposition. Nous avons étudié deux de ces dictionnaires :

- les curvelets [22] forment un dictionnaire non-adaptatif (du fait de la non modification des atomes en fonction des données) défini à partir de fonctions. Les atomes de ce dictionnaire s'apparentent à des raies sur de petites échelles et dans différentes orientations. Ce dictionnaire permet donc la caractérisation des signatures vibratoires en combinant différents atomes pour reconstruire la raie.
- La Non-Negative Matrix Factorization (NMF) [70], un dictionnaire adaptatif dont les atomes sont appris à partir des données. Les atomes de ce dictionnaire prennent en compte la structure globale du patch par la procédure d'apprentissage. Ce dictionnaire impose des combinaisons additives de ces atomes uniquement. Les signatures atypiques non présentes dans les données d'apprentissage ne sont pas bien caractérisées par le dictionnaire et ne peuvent pas être reconstruites.

Ces dictionnaires ne sont généralement pas utilisés dans le cadre de la détection d'anomalies. Nous les avons donc adaptés pour qu'ils permettent de répondre à notre problématique. Dans cette partie, nous montrons que ces dictionnaires sont pertinents et complémentaires pour finement détecter les points atypiques sur les patchs.



# Chapitre 3

## Représentation par dictionnaire fixe - les curvelets

### 3.1 Introduction

#### 3.1.1 La représentation par dictionnaire

La représentation des données a un rôle primordial dans les performances des méthodes de machine learning [13]. Cette représentation doit avoir un sens aussi bien mathématique qu'explicatif vis-à-vis du problème étudié. Elle doit permettre de réaliser l'étude et de mettre en valeur les éléments pertinents sur les données au niveau de la représentation. Pour la détection dans un cadre supervisé, il est important que la représentation des données soit discriminante, tandis que pour la détection de nouveautés, il est important que la représentation généralise la normalité.

Dans cette partie, nous avons opté pour des dictionnaires reconstruisant les données de manière linéaire. Ce choix a été réalisé car nous considérons les spectrogrammes comme une superposition linéaire de différentes raies pouvant être considérées comme des sources. Les dictionnaires correspondent à un ensemble d'éléments sur lesquels les données sont représentées ; chaque élément du dictionnaire est un atome. Les dictionnaires étudiés permettent de représenter les données comme une combinaison linéaire des atomes de ce dictionnaire. Dans ce cadre, il existe deux moyens de représenter une donnée  $x \in \mathbb{R}^p$  à partir d'un dictionnaire  $\mathcal{D} = [d_1, \dots, d_r] \in \mathbb{R}^{p \times r}$  [98, 93] :

- la méthode d'analyse où la donnée  $x$  est représentée par sa décomposition  $C_{\mathcal{D}}^A$  dans le dictionnaire  $\mathcal{D}$ , c'est à dire par le produit scalaire de  $\mathcal{D}$  avec  $x$ ,

$$C_{\mathcal{D}}^A = \mathcal{D}^T x,$$

- la méthode de synthèse où la donnée  $x$  est considérée comme une combinaison linéaire des atomes de  $\mathcal{D}$  pondérés par la représentation  $C_{\mathcal{D}}^S$ , la décomposition est obtenue par des

méthodes d'optimisation.

$$x = \mathcal{D}C_{\mathcal{D}}^S$$

Ces deux approches sont identiques dans le cas où le dictionnaire  $\mathcal{D}$  est orthogonal. Lorsque  $r < p$ , la donnée est représentée dans un espace de plus petite dimension, on parle alors de réduction de dimension [35]. Lorsque  $r > p$  la représentation est effectuée dans un espace de plus grande dimension, il s'agit alors d'un dictionnaire sur-complet [4]. Ce second cas de figure est utilisé plus particulièrement pour obtenir des représentations parcimonieuses des données.

Les atomes des dictionnaires peuvent être définis de manière analytique par des fonctions, ou sont appris à partir des données. Un état de l'art de ces différentes approches est réalisé dans [97]. Les dictionnaires analytiques possèdent leurs atomes définis par des fonctions prédéfinies. Ce type de dictionnaire est non-adaptatif, car les atomes ne sont pas modifiés en fonction des données, mais possède des propriétés mathématiques intéressantes. Ces dictionnaires donnent une décomposition en général unique et rapide à calculer. Ils sont basés sur la représentation d'un signal complexe à partir de classes de fonctions mathématiques plus simples. Ils sont généralement sélectionnés car leurs atomes correspondent à la structure des données étudiées et sont susceptibles de donner une représentation parcimonieuse de ces dernières. Les dictionnaires *data-driven* apprennent directement leurs atomes sur les données. Les représentations acquises à partir de ces dictionnaires ne sont généralement pas uniques. Afin de rendre ces dictionnaires et les décompositions dans ces derniers plus robustes et d'éviter le surapprentissage, des contraintes sont ajoutées lors de l'apprentissage telles que la parcimonie, la positivité, l'invariance par translation ou par rotation... Ce type de dictionnaire s'est rapidement développé dans la fin du 20ème siècle et le début du 21ème grâce à son adaptabilité aux données et aux méthodes d'optimisation comme les « basis pursuit » [28].

Nous nous intéressons à ces approches de dictionnaire dans un cadre de la détection de nouveautés sur nos spectrogrammes. Nous définissons le modèle de normalité de nos données à partir de ces dictionnaires. Nous ne cherchons pas à extraire des caractéristiques sur les données pour la détection d'anomalies, mais à représenter les éléments normaux des données afin de pouvoir donner une reconstruction normale (sans signature atypique) de ces dernières. Nous définissons la représentation des données par la combinaison du dictionnaire et la décomposition des données dans ce dernier. Dans le cadre de la détection de nouveautés, nous définissons la représentation normale par la reconstruction des données à partir du dictionnaire caractérisant la normalité, il s'agit donc d'une estimation normale des données. Dans ce chapitre, nous étudions les dictionnaires non-adaptatifs à travers les curvelets [22], les dictionnaires adaptatifs sont étudiés dans le chapitre suivant.

### 3.1.2 Les dictionnaires non-adaptatifs

Ces dictionnaires permettent de représenter les données à partir d'ensemble de fonctions simples. L'exemple le plus connu de ce type de dictionnaire est la transformée de Fourier caractérisant les signaux à partir de fonctions sinusoïdales. Cependant, ne bénéficiant pas de fonction de localisation, cette transformée ne permet pas de caractériser efficacement des signaux discontinus. Des approches multi-échelles tenant compte de la localisation ont permis d'obtenir de meilleurs résultats de caractérisation pour certains types de données. Il s'agit de la transformée en ondelettes [80], fondée sur une famille de fonctions paramétrées par un facteur d'échelle et de position. Cette transformée est inversible et permet de décomposer efficacement des signaux discontinus sans aucun apriori avec peu de coefficients. Autrement dit, la représentation dans ce dictionnaire est généralement parcimonieuse. Dans le cas de la transformée de Fourier, les signaux discontinus influent sur toutes les fréquences et donc sont représentés par un grand nombre de coefficients. Il y a un choix à effectuer par rapport à la nature des fonctions utilisées pour les atomes du dictionnaire. Il existe différentes catégories d'ondelettes comme les ondelettes de Haar, de Daubechies, de Morlet,...[80]. Ces dictionnaires constituent généralement une base de  $L^2$ . La décomposition dans ces dictionnaires se calcule donc généralement à partir d'un produit scalaire entre les données et les atomes, donc par analyse.

La décomposition en ondelettes est pertinente pour caractériser des singularités ponctuelles sur des signaux unidimensionnels ou multidimensionnels. Cependant, dans les dimensions supérieures, les singularités peuvent également correspondre à des hyperplans qui sont alors caractérisés par un grand nombre de coefficients. La transformée en ondelettes ne permet plus une représentation parcimonieuse en grande dimension. Bien qu'elle tienne compte de la position, la transformation en ondelettes classique ne prend pas en compte les orientations et donc la géométrie des singularités. Ceci entraîne la nécessité d'un grand nombre de coefficients afin de caractériser une discontinuité linéaire en 2 dimensions. De nouvelles approches basées sur les ondelettes et la représentation multi-échelle ont vu le jour pour pallier ce problème et principalement caractériser des singularités courbes dans des images, donc en dimension 2. Dans [42], un dictionnaire appelé les wedgelets est mis en place, il consiste en la division de l'image en 4 carrés dyadiques et la définition d'une droite séparatrice pour chacun de ces carrés. Cette droite permet de représenter la partie supérieure du carré par une valeur et la partie inférieure par une autre. Si la droite séparatrice ne permet pas une assez bonne caractérisation du carré dyadique, c'est-à-dire que ce carré est traversé par une forme non linéaire, ce dernier est à nouveau décomposer en 4 sous-carrés dyadiques et chacun d'entre eux est une nouvelle fois divisé par une droite séparatrice. Le processus est itéré jusqu'à avoir une bonne caractérisation des données. Le dictionnaire correspond alors à l'ensemble des carrés dyadiques et des droites séparatrices correspondantes. Il permet de caractériser efficacement des formes courbes en les décomposant en sous-formes linéaires à travers des carrés dyadiques d'échelle de plus en plus fine. Les ridgelets [21] permettent de caractériser des discontinuités linéaires, elles sont basées sur une application des ondelettes le long de droites, les curvelets [22] caractérisent les discontinuités courbes. Plusieurs autres approches permettent de caractériser des discontinuités sur des images comme les

contourlets [39], les bandelets [67]... Le scattering network [18] permet de représenter les données dans un réseau convolutionnel où chaque noeud correspond au module des convolutions successives de transformées d'ondelettes de Morlet. Cette représentation a l'avantage d'être invariante par translation et rotation. Une étape non-linéaire de seuillage est généralement ajoutée à ces représentations non-adaptatives afin de donner une représentation parcimonieuse car un nombre limité de coefficients est généralement suffisant pour caractériser et/ou reconstruire la donnée d'entrée.

Le dictionnaire des curvelets permet de caractériser les formes courbes, nous utilisons donc ce dictionnaire afin de caractériser les raies normales présentes sur les patchs sans signature atypique. La caractérisation des raies normales permet de définir un dictionnaire de normalité dans lequel nous pouvons projeter les patchs pour obtenir une reconstruction normale de ces derniers et étudier les résidus associés afin de vérifier la présence potentielle de signatures inusuelles sur le patch.

## 3.2 La transformée en curvelet

La transformée en curvelet [22] permet la caractérisation de singularités courbes. Elle est basée sur la transformée en ridgelet [21] et la caractérisation d'une forme courbe à partir d'une succession de formes linéaires à petite échelle (Figure 3.1).

### 3.2.1 La transformée en ridgelet

Les ridgelets ont été introduites dans [19] et sont définies sous la forme d'une composition de fonctions *ridges* avec une ondelette  $\psi$  (3.1). Elles sont caractérisées par un paramètre d'échelle  $a > 0$ , un paramètre de position  $b \in \mathbb{R}$  et un paramètre d'orientation  $\theta \in [0, 2\pi[$ . Ce nouveau paramètre diffère des ondelettes et permet la caractérisation de toutes formes linéaires sous différentes directions. On note  $\psi_{a,b,\theta}$  la ridgelet de paramètre  $(a, b, \theta)$ .

$$\psi_{a,b,\theta}(t = (t_1, t_2)) = a^{-\frac{1}{2}}\psi\left(\frac{t_1 \cos \theta + t_2 \sin \theta - b}{a}\right) \quad (3.1)$$

Cette fonction est constante le long des droites  $t_1 \cos \theta + t_2 \sin \theta = \text{constante}$  et se comporte comme une ondelette perpendiculairement. Dans le domaine des fonctions ridges, une singularité linéaire est définie par un point (la constante à laquelle est associée la droite la caractérisant). Dans cet espace, la singularité linéaire devient donc une singularité ponctuelle pouvant être caractérisée efficacement par les ondelettes avec peu de coefficients. Les ridgelets permettent donc une représentation efficace des formes linéaires.

La transformée en ridgelet  $\mathcal{R}_x$  d'un signal  $x$  correspond alors au produit scalaire entre la

fonction de ridgelet  $\psi_{a,b,\theta}$  et  $x$ .

$$\mathcal{R}_x(a, b, \theta) = \int \bar{\psi}_{a,b,\theta}(t)x(t)dt$$

$\bar{\psi}$  et  $\hat{\psi}$  sont respectivement le complexe conjugué et la transformée de Fourier de  $\psi$  satisfaisant la condition d'admissibilité :

$$\int \frac{|\hat{\psi}(\lambda)|^2}{\lambda} d\lambda < \infty$$

Cette transformée est inversible pour un signal  $x$  intégrable et de carré intégrable, respecte la relation de Parseval et se généralise dans toutes les dimensions.

La transformée, ainsi définie, permet d'obtenir une représentation des signaux continus. La version discrète de la transformée en ridgelets [21] correspond à une discréétisation spécifique  $(a_j, b_k, \theta_l)$  de ses paramètres afin que l'ensemble des fonctions de ridgelets  $\{\psi_{j,k,l}\}$  forme une frame. Inspirés par la transformée en ondelette, les paramètres d'échelle  $a$  et de position  $b$  sont échantillonnés de manière dyadique avec une dépendance de la discréétisation du paramètre de position par rapport à celle de l'échelle. La résolution du paramètre d'orientation  $\theta$  augmente mécaniquement avec le paramètre d'échelle et est donc dépendante de ce paramètre.

$$a_j = a_0 2^{-j} \quad b_{j,k} = 2\pi k 2^{-j} \quad \theta_{j,l} = 2\pi l 2^{-j}$$

L'ensemble des fonctions

$$\{\psi_{j,k,l}(t) = 2^{\frac{j}{2}} \psi(2^j(t_1 \cos(2\pi l 2^{-j}) + t_2 \sin(2\pi l 2^{-j}) - 2\pi k 2^{-j}))\}_{j,k,l} \quad (3.2)$$

forme une frame permettant la décomposition des signaux. La transformée en ridgelet [21] discrète s'effectue par produit scalaire entre le signal et l'ensemble des fonctions de ridgelets définies dans (3.2).

La transformée en curvelet se calcule à partir des ridgelets orthonormales  $\rho_\lambda$  définies dans [40] et paramétrées par  $\lambda = (j, k, i, l, \epsilon) \in \Lambda$ .  $j$  et  $k$  correspondent respectivement aux paramètres d'échelle et de position de la fonction ridge,  $i$  et  $l$  définissent les paramètres d'échelle et de position angulaire,  $\epsilon$  est un paramètre de genre lié aux paramètres d'échelle. Nous renvoyons à [40] pour plus de détails sur la construction de ces ridgelets particulières. Le dictionnaire défini à partir des ridgelets orthonormales constitue une "tight" frame permettant la conservation de l'énergie entre les signaux et leurs représentations à partir de ces ridgelets particulières.

### 3.2.2 Les ridgelets multi-échelles

Les ridgelets définies au dessus permettent d'analyser les données en 2 dimensions dans leur globalité. Les ridgelets multi-échelles [22] ont pour but de pouvoir analyser une donnée sous différentes échelles de cette donnée. Il est possible de considérer les ridgelets orthonormales comme

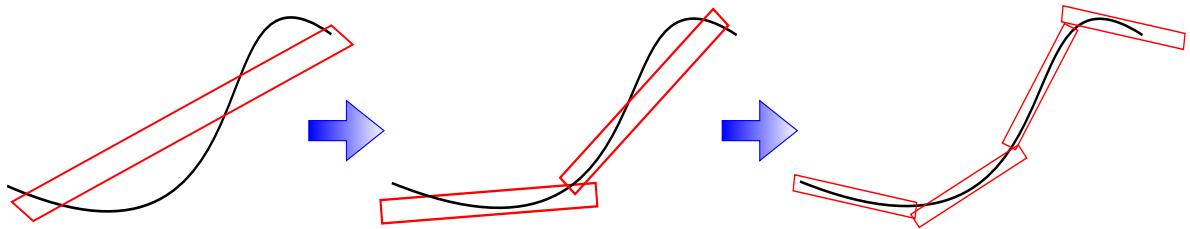


FIGURE 3.1 – Caractérisation d'une forme courbe à différentes échelles. La taille des éléments caractérisant la forme diminue avec l'échelle. La représentation de la forme courbe est améliorée en augmentant le nombre de formes linéaires à plus petites échelles.

des éléments linéaires de taille fixe et de largeur plus ou moins fine. Cependant, dépendant de la forme considérée, une certaine taille de ridgelets peut être plus pertinente qu'une autre (Figure 3.1). Il faut donc qu'elles puissent s'adapter aux différentes dimensions des formes étudiées. Pour cela, les ridgelets multi-échelles sont définies à partir d'un opérateur d'échelle  $T_Q$  qui transpose la ridgelet orthonormale de son espace de départ  $[0, 1]^2$  vers un carré dyadique  $Q = (s, k_1, k_2)$  définit sur  $[\frac{k_1}{2^s}, \frac{k_1+1}{2^s}] \times [\frac{k_2}{2^s}, \frac{k_2+1}{2^s}]$ ,  $s$  définit l'échelle de subdivision des données en carrés dyadiques et  $(k_1, k_2)$  la position du carré étudié. L'énergie de la ridgelet est également répartie de manière lisse sur les points voisins du carré dyadique par une fonction de fenêtrage  $\omega$  ayant comme propriété  $\sum_{k_1, k_2} \omega(x_1 - k_1, x_2 - k_2) \equiv 1$ .

Les ridgelets multi-échelles  $\{\psi_\mu, \mu = (\lambda, Q) \in (\Lambda, \mathcal{Q}_s)\}$  (3.3) correspondent à une application des ridgelets orthonormales paramétrées par  $\lambda$  sur des extractions de carrés dyadiques à différentes échelles  $s$  paramétrés par  $Q \in \mathcal{Q}_s$  avec  $\mathcal{Q}_s$  l'ensemble des carrés dyadiques de taille  $2^{-s}$ . L'application des ridgelets orthogonales sur différentes tailles d'images permet de considérer différentes tailles pour les ridgelets (correspondant à la taille de l'image) tout en étudiant différentes échelles et orientations des ridgelets.

$$\psi_{\mu_s} = 2^s T_Q(\omega \cdot \rho_\lambda), \quad \mu_s = (\lambda, Q) \in (\Lambda, \mathcal{Q}_s) = \mathcal{M}_s \quad (3.3)$$

Il faut distinguer l'échelle  $j$  des ridgelets et l'échelle  $s$  des multi-ridgelets qui correspond à la subdivision en carrés dyadiques sur lesquels les ridgelets définies à plusieurs échelles  $a_j$  sont appliquées.

### 3.2.3 La construction de la transformée en curvelet

Les curvelets sont utilisées pour caractériser des formes courbes sur les données. Elles sont issues de la transformée en ridgelet appliquée à différentes échelles  $s$  de la donnée. Les différentes échelles caractérisent l'idée que chaque courbe peut être approximée par une série de formes linéaires sur de petites échelles (Figure 3.1). En diminuant l'échelle, les formes linéaires caractérisent bien mieux les courbures en prenant de plus petites tailles, mais nécessairement leur nombre augmente. Le meilleur moyen de caractériser les données  $x$  est donc d'appliquer les ridgelets orthonormales sur des échelles  $s$  de plus en plus fines. Il s'agit de caractériser  $x$  à partir