Algorithmes, méthodes et modèles pour l'application des capteurs à ondes acoustiques de surface à la reconnaissance de signatures de composés chimiques

Olivier Hotel

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Chimie - Physique des Matériaux

présentée par

Olivier Hotel

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Algorithmes, Méthodes et Modèles pour l'Application des Capteurs à Ondes Acoustiques de Surface à la Reconnaissance de Signatures de Composés Chimiques

Thèse approuvée par un jury composé de :

HONEINE Paul	**Professeur, Univ. Rouen Normandie**	**Rapporteur**
JUTTEN Christian	**Professeur, Grenoble INP**	**Rapporteur**
CHECOURY Xavier	**Professeur, Univ. Paris Saclay**	**Examinateur**
SANGNIER Maxime	**Maître de conférences, UPMC**	**Examinateur**
DUVAL Laurent	**Ingénieur de recherche, IFP**	**Invité**
MER Christine	**Ingénieur de recherche, CEA-LIST**	**Invité**
SCORSONE Emmanuel	**Ingénieur de recherche, CEA-LIST**	**Invité**
POLI Jean-Philippe	**Ingénieur de recherche, CEA-LIST**	**Encadrant**
SAADA Samuel	**Ingénieur de recherche, CEA-LIST**	**Directeur de thèse**

Commissariat à l'Énergie Atomique et aux Énergies Alternatives

CEA-LIST - Laboratoire d'Intégration des Systèmes et des Technologies

***

Laboratoire pour l'Analyse de Données et l'Intelligence des Systèmes

Remerciements

Je remercie tout d'abord Samuel Saada, directeur du Laboratoire Capteurs Diamant du CEA-LIST, et Anthony Larue, directeur du Laboratoire pour l'Analyse des Données et l'Intelligence des Systèmes du CEA-LIST, de m'avoir permis de réaliser cette thèse dans leur laboratoire respectif et pour m'avoir permis de participer à plusieurs conférences.

Je tiens à remercier tout particulièrement Jean-Philippe Poli, Samuel Saada et Emmanuel Scorsone pour m'avoir fait confiance il y a 3 ans. Je les remercie pour leurs conseils sans lesquels ce travail n'aurait pu aboutir, pour leur patience et surtout pour les multiples relectures de ce manuscrit et des articles issus des travaux décrits dans ce document.

Je remercie Mr Honeine et Mr Jutten d'avoir accepté d'être rapporteurs de cette thèse ainsi que Mr Checoury et Mr Sangnier pour leur participation en tant qu'examinateur de ce travail.

De nombreuses personnes au sein du LCD et du LADIS m'ont permis de mener à bien cette thèse. Je remercie encore une fois Emmanuel Scorsone pour son expertise, Christine Mer pour la réalisation des capteurs SAW et pour ses précieux conseils lors de la conception et de la mise en œuvre des plans d'expériences et Nicolas Tranchant pour les manipulations des bouteilles de gaz. Je remercie également Aurélien Mayoue pour son expertise concernant les techniques multiparamétriques de reconnaissance d'odeurs, ainsi que Robin Delgado, Adrien Sourrisseau et Sébastien Klasa pour avoir réalisé les logiciels qui m'ont permis de réaliser les différentes expériences.

Glossaire

Liste des sigles

ANSI	American National Standarts Institute
CEA	Commissariat à l'Énergie Atomique et aux Énergies Alternatives
KIT	Karlsruhe Institute of Technology
IEEE	Institute of Electrical and Electronics Engineers
IMT	Institute of Microstructure Technology
LADIS	Laboratoire pour l'Analyse des Données et l'Intelligence des Systèmes
LCD	Laboratoire Capteurs Diamant
LIST	Laboratoire d'Intégration des Systèmes et des Technologies

Liste des molécules

C_7H_8	Toluène
$C_3H_9O_3P$	Méthylphosphonate de Diméthyle
$C_7H_7NO_2$	4-Nitrotoluène
CH_3OH	Méthanol
H_2S	Sulfure d'Hydrogène
NH_3	Ammoniac
SO_2	Dioxyde de Soufre

Liste des abréviations

4-NT	4-Nitrotoluene
	4-Nitrotoluène
AE	Autoencoder
	Autoencodeur
AR	Auto-Regressive
	Auto-Régressif
ARMA	Auto-Regressive Moving Average
	Auto-Régressif et Moyenne Glissante
BT	Bagged Trees
	Ensemble d'Arbres de Décision
DA	Discriminant Analysis
	Analyse Discriminante
DMMP	Dimethyl Methylphosphonate
	Méthylphosphonate de Diméthyle
DT	Decision Tree
	Arbre de décision
ES	Evolution Strategy
	Stratégie d'Évolution
GF	Generating Function
	Fonction génératrice
IDT	Interdigited Tease
	Transducteur Interdigité
KNN	k Nearest Neighbors
	k Plus Proches Voisins
KR	Kernel Regression
	Régression à Noyaux
LDA	Linear Discriminant Analysis
	Analyse Discriminante Linéaire
LMNN	Large Margin Nearest Neighbors
	Plus Proches Voisins à Vastes Marges
MA	Moving Average
	Moyenne Glissante
MLP	MultiLayer Perceptron
	Perceptron MultiCouches
MSE	Mean Squared Error
	Erreur Quadratique Moyenne

NN	Neural Network Réseau de Neurones
PCA	Principal Component Analysis Analyse en Composantes Principales
PLS	Partial Least Squares Moindres Carrés Partiels
PNN	Probabilistic Neural Network Réseau de Neurones Probabiliste
PPM	Part Per Million Partie Par Million
PSO	Particle Swarm Optimization Optimisation par Essaim Particulaire
RBF	Radial Basis Function Réseau à Fonctions Radiales
ROC	Region Of Convergence Région de Convergence
RT	Rise Time Temps de Montée
SA	Simulated Annealing Recuit Simulé
SAW	Surface Acoustic Wave Onde Acoustique de Surface
SIMO	Single Input Multiple Output Entrée Simple Multiple Sorties
SNR	Signal to Noise Ratio Rapport Signal sur Bruit
SVM	Support Vector Machine Machine à Vecteur de Support

Notations

$\mathbf{A_{n,m}}$	Matrice $\mathbf{A}$ de taille $n \times m$
$\mathbf{A_{i,j}}$	Élément de la $i^{ième}$ ligne et de la $j^{ième}$ colonne de la matrice $\mathbf{A}$
$\mathbf{A_{i,\_}}$	$i^{ième}$ ligne de la matrice $\mathbf{A}$
$\mathbf{A_{\_,i}}$	$i^{ième}$ colonne de la matrice $\mathbf{A}$
$\mathbf{A^{-1}}$	Inverse de la matrice $\mathbf{A}$
$\mathbf{A}^{\dagger}$	Pseudo-inverse Moore-Penrose de la matrice $\mathbf{A}$
$\mathbf{A^T}$	Transposée de la matrice $\mathbf{A}$
$\mathbf{A} \otimes \mathbf{B}$	Produit de Kronecker des matrices $\mathbf{A}$ et $\mathbf{B}$
$\mathbf{A} \odot \mathbf{B}$	Produit d'Hadamard des matrices $\mathbf{A}$ et $\mathbf{B}$
$\mathbf{I_n}$	Matrice identité de taille n
$\mathbf{0_{n,m}}$	Matrice de taille $n \times m$ remplie de 0
$\mathbf{1_{n,m}}$	Matrice de taille $n \times m$ remplie de 1
$a \cdot b$	Produit scalaire des vecteurs a et b
$\|\|a\|\|$	Norme euclidienne du vecteur a
$a[n]$	$n^{ième}$ terme de la séquence a
$a * b$	Produit de convolution des séquences a et b
$G(a, x)$	Fonction génératrice associée à la séquence a
S_a	Somme cumulée de la séquence a
S_a^k	Somme cumulée d'ordre k de la séquence a
$P^d[.]$	Polynôme de degré d
$\mathcal{P}_k(E)$	Ensemble des sous-ensembles de E de cardinal k
$\mathcal{B}_i$	$i^{ième}$ nombre de Bernoulli
$\mathcal{N}(\mu,\, \sigma^2)$	Loi normale de moyenne μ et d'écart-type σ
$\mathcal{U}(a, b)$	Loi uniforme entre a et b

Table des matières

Remerciements **i**

Glossaire **v**

Notations **vii**

Introduction générale **2**

I Contexte de l'étude et positionnement des travaux 3

1 Multicapteurs à ondes acoustiques de surface pour l'identification de signatures chimiques 7

1.1 Capteur SAW et fonctionnalisation diamant 7

1.1.1 Principe de fonctionnement des capteurs SAW pour l'identification de signatures chimiques 7

1.1.2 Couche sensible en diamant fonctionnalisé 12

1.1.3 Approche multicapteurs 12

1.2 Protocoles expérimentaux 13

1.2.1 Présentation du dispositif utilisé 13

1.2.2 Environnement de laboratoire : toxiques chimiques . . . 13

1.2.3 Environnement partiellement contrôlé : capsules de café commerciales 15

1.2.4 Environnement non contrôlé : DMMP et 4-NT en sac à dos 17

1.2.5 Répétabilité des protocoles expérimentaux 18

2 Approches multiparamétriques pour l'identification de signatures chimiques 21

2.1 Extraction des descripteurs 21

2.1.1 Descripteurs directement issus de la réponse des capteurs 21

2.1.2 Descripteurs issus de la modélisation de la réponse des capteurs 22

2.2 Changement de représentation et réduction de dimension 23
2.2.1 Analyse en composantes principales 23
2.2.2 Autoencodeurs 24
2.2.3 Réduction de dimension 24
2.3 Apprentissage supervisé pour l'identification de signatures chimiques 25
2.3.1 Classifieurs linéaires 26
2.3.2 Approches connexionnistes 29
2.3.3 Approches à base de voisinages 35
2.3.4 Approches à base de règles 37
2.3.5 Approches basées sur la modélisation des signaux 38

3 Positionnement par rapport à l'état de l'art 41
3.1 Évaluation et analyse des performances des techniques de l'état de l'art 41
3.1.1 Protocole d'évaluation 41
3.1.2 Analyse des résultats du *benchmark* 47
3.2 Proposition d'une nouvelle approche pour l'identification de signatures chimiques 49

II Identification de signatures chimiques 53

4 Estimation des paramètres des contributions massique et viscoélastique 57
4.1 Justification du modèle utilisé 57
4.2 Formulation du problème d'optimisation 59
4.3 Comparaison de métaheuristiques pour la résolution du problème 64

5 Application à l'identification de signatures chimiques et à l'estimation de leur concentration 71
5.1 Application à l'identification de composés chimiques 71
5.1.1 Impact des performances du processus d'optimisation sur le taux de classification 73
5.1.2 Fusion des descripteurs 75
5.2 Application à l'estimation du profil de concentration 79
5.2.1 Déconvolution 79
5.2.2 Techniques de régression non paramétrique 85

6 Méthode de sélection des fonctionnalisations des capteurs 89
6.1 Définition d'un critère de séparabilité 89
6.2 Proposition d'un algorithme glouton pour la sélection de capteurs 93

6.2.1 Formulation du problème sous la forme d'un problème d'optimisation . . . 93
6.2.2 Algorithme glouton . . . 93
6.2.3 Analyse de l'algorithme . . . 95
6.3 Résultats expérimentaux . . . 96

III Vers l'identification des composés d'un mélange de composés chimiques 101

7 Problématiques liées à l'identification des mélanges de composés chimiques 105
7.1 Typologie des problèmes et des approches pour l'identification de mélanges . . . 105
7.1.1 Exhaustivité de la base d'apprentissage . . . 105
7.1.2 Connaissance *a priori* du nombre de composés du mélange 106
7.1.3 Problèmes et approches pour l'identification des mélanges 106
7.1.4 Résultats expérimentaux . . . 107
7.2 Proposition d'approches phénoménologiques pour la modélisation de la réponse des SAW à un mélange . . . 109
7.2.1 Proposition de modèles empiriques . . . 109
7.2.2 Validation expérimentale des modèles . . . 109

8 Estimation du nombre de composés dans un mélange et application à l'identification des mélanges 115
8.1 Estimation du nombre de constituants d'un mélange . . . 115
8.1.1 Extension du modèle de type somme pondérée . . . 115
8.1.2 Formulation et résolution d'un problème de régression linéaire . . . 117
8.2 Résultats expérimentaux . . . 120
8.3 Application à l'identification de mélanges . . . 122

Conclusion générale 128

A Processus linéaires temps invariant et réponse impulsionnelle 129
A.1 Processus linéaires temps invariant . . . 129
A.2 Réponse impulsionnelle . . . 130
A.3 Réponse impulsionnelle et processus linéaires temps invariant . . . 130

B Définition et propriétés des fonctions génératices 131
B.1 Fonctions génératrices . . . 131
B.2 Produit de convolution . . . 131
B.3 Approximation numérique . . . 132

C Métaheuristiques d'optimisation **135**
C.1 Recuit simulé 135
C.2 Stratégie d'évolution $\lambda + \mu$ 136
C.3 Optimisation par essaim particulaire 136
C.4 Optimisation sous contraintes 137
C.5 Critères d'arrêt 138
C.6 Optimisation de l'implémentation 138

D Linéarisation de la réponse des SAW à un mélange **139**

Bibliographie **147**

Introduction générale

L'identification de signatures chimiques de composés volatils est une problématique actuelle majeure dans de nombreux domaines tels que l'industrie agroalimentaire, l'environnement, la médecine et la défense. Récemment, des systèmes multicapteurs offrant un grand intérêt en termes de rapidité, sélectivité, robustesse et portabilité, permettant de répondre à ces problématiques, se sont développés. Parmi les technologies existantes, la technologie utilisant des capteurs à ondes acoustiques de surface, du fait de sa grande sensibilité et de sa rapidité de détection, est très prometteuse et fait l'objet de nombreuses recherches. Ces dernières portent principalement sur la conception et le développement de nouvelles couches sensibles. Si le développement de polymères constitue la majorité des travaux, l'utilisation de couches sensibles en diamant est une approche originale et innovante initiée au Laboratoire Capteurs Diamant (LCD). Le diamant étant un matériau particulièrement stable, les couches ainsi obtenues permettent de fabriquer des capteurs très robustes. La fabrication de tels capteurs, avec différentes terminaisons de surface et affinités chimiques, permet d'envisager leur utilisation dans un réseau de capteurs afin de créer un système sélectif grâce à des approches multiparamétriques.

Les objectifs de cette thèse sont multiples et se placent dans le cadre d'une collaboration entre le LCD et le Laboratoire pour l'Analyse des Données et l'Intelligence des Systèmes (LADIS) de l'institut LIST. Un des premiers objectifs visé est de développer des méthodes multiparamétriques permettant d'identifier des signatures chimiques de composés volatils. Le second objectif vise à estimer la concentration de ces derniers. Enfin, le troisième objectif se focalise sur la caractérisation des performances des algorithmes développés à partir de mesures expérimentales sur des toxiques chimiques, des capsules de café, des gaz toxiques et des explosifs. Ce manuscrit s'articule autour de trois parties.

La première partie décrit dans un premier temps le contexte de l'étude, la description du principe de fonctionnement des capteurs à ondes acoustiques de surface, les mécanismes de transduction associés et présente les dispositifs, les protocoles expérimentaux et les bases de données utilisées tout au long

de cette étude. Dans un second temps, nous présentons un état de l'art des algorithmes et des méthodes proposés dans la littérature pour identifier des signatures chimiques. Cette partie présente successivement les différents descripteurs proposés et décrit les techniques de normalisation, de changement de représentation ainsi que les algorithmes d'apprentissage supervisé utilisés dans le cadre de l'identification de signatures chimiques. Cette partie se termine par une comparaison expérimentale des techniques de l'état de l'art permettant de mettre en évidence leurs limitations et motivant de ce fait les travaux entrepris dans les parties suivantes.

La seconde partie présente une méthode permettant d'estimer les paramètres clés des mécanismes de transduction en les considérant comme des solutions d'un problème d'optimisation et compare, dans le cadre des problèmes d'identification de signatures chimiques, les performances que ceux-ci permettent d'obtenir avec celles de l'état de l'art. Une méthode permettant de fusionner ces paramètres avec les descripteurs proposés dans la littérature est également détaillée. Nous décrivons ensuite une méthode permettant d'estimer la concentration d'un composé chimique. Pour finir, nous abordons le problème de la sélection des couches sensibles des capteurs en introduisant un critère de séparabilité de signatures chimiques ainsi qu'un algorithme glouton permettant de sélectionner les couches sensibles les plus appropriées pour une application donnée.

La troisième partie est dédiée aux problèmes d'identification des composés d'un mélange. Elle décrit d'abord la typologie des problèmes d'identification ainsi que les approches permettant de résoudre ces derniers. Puis, des modèles phénoménologiques sont proposés pour représenter la réponse des capteurs à ondes acoustiques de surface exposés à un mélange et expérimentalement validés. Enfin, la description d'une méthode permettant d'estimer le nombre de constituants d'un mélange termine cette troisième partie.

Première partie

Contexte de l'étude et positionnement des travaux

Introduction

L'objectif de cette partie est de présenter le contexte de cette thèse et de décrire le positionnement des travaux effectués par rapport à l'état de l'art.

Le premier chapitre décrit les capteurs chimiques à ondes acoustiques de surface et l'approche multicapteurs mise en œuvre pour développer des systèmes de type nez électronique dédiés à la reconnaissance de signatures chimiques. L'approche multicapteurs nécessite des couches sensibles ayant des affinités chimiques différentes. La fonctionnalisation des capteurs via une couche de nanodiamants ayant des terminaisons de surface différentes permet d'obtenir une grande variété d'affinités chimiques. Les protocoles expérimentaux présentés dans la seconde partie, associés aux expériences réalisées, viseront à obtenir plusieurs jeux de données. Ces derniers permettront, dans la suite de cette étude, d'évaluer de manière quantitative les performances des approches décrites.

Le second chapitre présente les principales approches multiparamétriques proposées dans la littérature qui peuvent être appliquées à l'identification de signatures chimiques. Il fait successivement un état de l'art des descripteurs proposés, des techniques de changement de représentation et de réduction de dimension ainsi que des algorithmes d'apprentissage supervisé proposés pour l'identification de signatures chimiques. Une comparaison quantitative des différentes approches de l'état de l'art sur les bases de données termine ce second chapitre.

Les résultats du *benchmark* sont analysés et commentés dans le troisième chapitre. Ce dernier se termine par la description de l'approche proposée dans les parties suivantes et par ses motivations.

Chapitre 1

Multicapteurs à ondes acoustiques de surface pour l'identification de signatures chimiques

Introduction

Les capteurs à ondes acoustiques de surface, appelés capteurs SAW (*Surface Acoustic Wave*) sont basés sur une technologie qui fait intervenir différents mécanismes de transduction qui sont ici présentés. Nous détaillerons ensuite le dispositif expérimental et les bases de données utilisés tout au long de ce manuscrit.

1.1 Capteur SAW et fonctionnalisation diamant

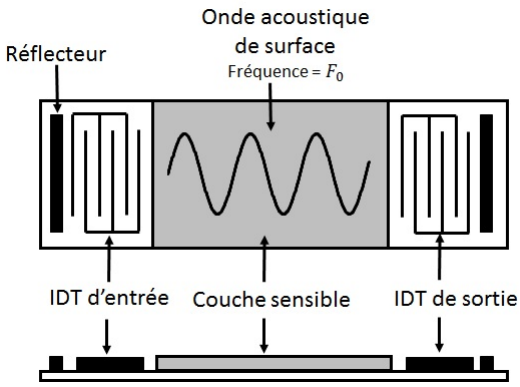
1.1.1 Principe de fonctionnement des capteurs SAW pour l'identification de signatures chimiques

De manière générale, un capteur chimique permet de transformer une information chimique en un signal mesurable. Ces capteurs sont le plus souvent composés d'une couche sensible qui transforme l'information chimique en une forme d'énergie mesurable et d'un transducteur permettant de convertir cette énergie en un signal mesurable [Janata, 2009]. Plusieurs technologies de capteurs permettent de convertir la présence de molécules sur la couche sensible en un tel signal. Ces technologies peuvent notamment exploiter des phénomènes de nature optique, électrochimique, électrique, mécanique, ou encore gravimétrique [Liu *et al.*, 2012].

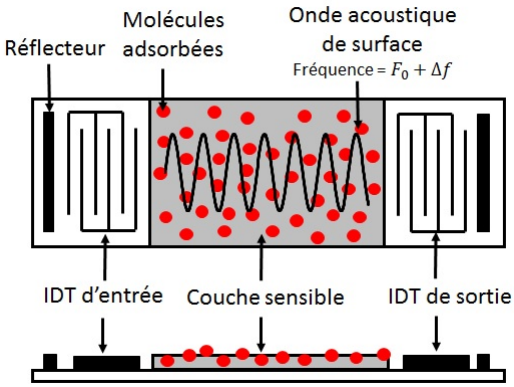
Dans le cadre de cette thèse, nous utilisons uniquement des capteurs à ondes acoustiques de surface (SAW). Ces capteurs sont des résonateurs constitués d'un matériau piézoélectrique sur lequel sont déposés deux transducteurs interdigitaux (*Interdigited Tease*, IDT) aux extrémités d'une cavité recouverte d'une couche sensible. Les IDT vont permettre de générer une onde acoustique de surface en convertissant l'énergie électrique en énergie mécanique. Le système est conçu de façon à ce que la cavité soit résonante. La fréquence de résonance de la cavité est particulièrement sensible aux modifications de surface. Ainsi, en fonction des interactions entre la couche sensible déposée sur la cavité et le milieu environnant, la fréquence de résonance mesurée évoluera. Les interactions physico-chimiques des molécules cibles avec cette couche sensible perturbent la propagation de l'onde acoustique à la surface du résonateur piézoélectrique et induisent de ce fait une variation de fréquence mesurable. Comme le mentionne Tard [Tard, 2013], les capteurs SAW ont de nombreux avantages : leurs procédés de fabrication sont, à l'heure actuelle, compatibles avec les technologies de salle blanche, les rendant de ce fait aisément fabricables. Ils offrent également des seuils de sensibilité extrêmement bas pour de nombreux composés chimiques [Afzal, 2011, Wen *et al.*, 2007]. Néanmoins, ils sont très sensibles à l'humidité et à la température [García et Aparicio, 2002].

La direction de l'onde de propagation peut être soit dans le plan de la surface du matériau piézoélectrique, dans le cas de capteurs SAW à onde de Love, soit dans un plan orthogonal à la surface comme c'est le cas pour les capteurs SAW à onde de Rayleigh ou de Lamb [Ballantine et Wohltjen, 1989] [Janata, 2009].

Deux technologies de résonateur existent à l'heure actuelle. Les transducteurs SAW à lignes de retard sont constitués d'électrodes de type peignes interdigités (IDT) : la différence de polarisation entre deux peignes génère des contraintes mécaniques, et donc une onde au niveau des électrodes d'entrée tandis que le phénomène inverse se produit au niveau des électrodes de sortie [Voiculescu et Nordin, 2012]. La technologie de type résonateur ajoute des réflecteurs latéraux pour que l'onde soit maintenue dans une cavité réduisant ainsi les pertes du dispositif et améliorant ainsi son facteur de qualité [Lange *et al.*, 2008]. La figure 1.1 montre la structure et le principe de fonctionnement d'un capteur SAW de type résonateur : la couche sensible se trouve au milieu des peignes interdigités d'entrée et de sortie de telle sorte qu'elle soit sur le chemin de l'onde acoustique et les réflecteurs placés aux extrémités du capteur permettent de piéger l'onde au niveau de celle-ci.



(a) Capteur SAW en l'absence de molécules cibles



(b) Capteur SAW en présence de molécules cibles

FIGURE 1.1 – Principe de fonctionnement d'un capteur SAW de type résonateur à onde de Rayleigh.

Mécanismes de transduction

Dans cette section, nous décrivons les phénomènes physiques qui influent sur la fréquence de résonance. Nous limitons notre étude au cas des capteurs SAW de type résonateur à onde de Rayleigh. Les phénomènes en jeu étant bien connus de la communauté, cette section est un résumé des résultats établis dans [Ballantine *et al.*, 1997, Tard, 2013]. Ces travaux ont mis en évidence que la variation de fréquence de résonance des capteurs SAW est la superposition de trois phénomènes.

Contribution massique : cette contribution, due à l'augmentation de la masse de la couche sensible suite à l'adsorption de molécules, est décrite par

l'équation :

$$\Delta f = -\frac{C_m F_0^2}{\rho_S} \Delta m$$

où Δf est la variation de fréquence, F_0 est la fréquence de résonance, C_m est le facteur de sensibilité massique et ρ_S est la densité massique de surface. Deux éléments primordiaux sont mis en évidence par cette équation : le premier étant le fait que les capteurs SAW permettent de détecter tout type de molécules puisqu'elles ont toutes une masse ; le second étant le fait que la contribution massique induit un déplacement de fréquence négatif.

Contribution viscoélastique : cette contribution est due au stockage et à la dissipation de puissance dus à la déformation de la couche sensible sous l'effet de l'onde. Dans le cas des couches acoustiquement fines, dans lesquelles l'onde se déplace uniformément dans l'épaisseur de la couche, le déplacement en fréquence est donné par l'équation :

$$\Delta f = -2\pi h \Delta \left(C_1 \big(\rho - \frac{\mu}{\upsilon_0^2}\big) + C_2 \rho + C_3 \big(\rho - \frac{4\mu}{\upsilon_0^2} \frac{\lambda + \mu}{\lambda + 2\mu}\big) \right) F_0^2$$

où υ_0 est la vélocité acoustique, λ et μ sont les constantes de Lamé et où les C_i décrivent les vélocités de surface dans les trois directions. Les termes C_1 et C_2 étant généralement négligeables devant le troisième, la perturbation viscoélastique peut être définie par [Hietala *et al.*, 2001] :

$$\Delta f = \frac{2\pi h C_3}{\upsilon_0^2} \Delta \big(E \frac{4\upsilon - 5}{5\upsilon^2 + \upsilon - 4} \big) F_0^2$$

où E est le module d'Young et υ le coefficient de Poisson. Le module d'Young est une constante reliant l'allongement relatif d'un matériau à des forces de traction ou de compression, tandis que le coefficient de Poisson relie l'allongement relatif orthogonal aux forces auxquelles il est soumis. De plus, les auteurs de [Philip et Hess, 2003] ont montré que le coefficient de Poisson n'est que très peu modifié, si bien qu'il peut être considéré comme constant. Ainsi, seule la variation du module d'Young a un effet sur le déplacement en fréquence, et dans le cas d'une augmentation de celui-ci, le déplacement en fréquence est positif.

Contribution électro-acoustique : cette contribution, due au couplage onde - porteurs de charges, est décrite par à l'équation :

$$\Delta f = -\frac{K^2}{2} \Delta \left(\frac{1}{1 + \frac{\upsilon_0 C_g}{\sigma_s}} \right) F_0$$

où σ_s est la conductivité de surface, C_q est la somme des permittivités diélectriques de l'aire et du substrat :

$$C_q = \epsilon_0 + \epsilon_q,$$

υ_0 est la vélocité acoustique de l'onde et K est le coefficient de couplage électromécanique. Pour les capteurs en quartz, tels que ceux utilisés dans le cadre de cette étude, $K = 0,11$ ce qui rend ce phénomène négligeable devant les deux autres. Cette hypothèse a d'ailleurs été confirmée expérimentalement lors des travaux menés par B. Tard [Tard, 2013]. Les mesures effectuées en mesurant la conductivité de surface ont montré que cette contribution impliquait une variation de fréquence de l'ordre du micro-hertz tandis que les deux autres impliquaient des variations de fréquence de plusieurs dizaines de hertz.

Modélisation de la réponse des capteurs SAW

Il a également été établi que ces perturbations peuvent être modélisées par des équations différentielles linéaires du premier ordre [Raj *et al.*, 2012] :

$$\begin{aligned}
\tau_m \frac{\partial F_m}{\partial t} + F_m &= K_m c \\
\tau_v \frac{\partial F_v}{\partial t} + F_v &= K_v c \\
\tau_e \frac{\partial F_e}{\partial t} + F_e &= K_e c
\end{aligned}$$

F_m, F_v et F_e étant respectivement les variations de fréquence dues aux contributions massique, viscoélastique et électro-acoustique et c étant le profil de concentration. Les coefficients τ_m, τ_v et τ_e sont les constantes de temps de ces équations ; elles caractérisent la rapidité de l'évolution du déplacement en fréquence dans le temps. Les coefficients K_m, K_v et K_e sont les gains de ces équations ; ces derniers caractérisent le régime stationnaire des solutions de ces équations. La variation totale de fréquence est alors donnée par

$$\Delta f = F_m + F_v + F_e$$

ou, si l'on néglige la contribution électro-acoustique

$$\Delta f = F_m + F_v.$$

Il faut toutefois remarquer que, dans la littérature, des modèles dans lesquels les constantes de temps sont des fonctions linéaires de la concentration ont été proposés [Manai, 2014].

1.1.2 Couche sensible en diamant fonctionnalisé

La couche sensible d'un capteur SAW est un élément clé pour développer une approche multicapteurs afin d'aborder la reconnaissance de signatures chimiques. Elle influence non seulement la sélectivité du capteur, c'est-à-dire sa capacité à ne répondre qu'à certaines molécules cibles, mais également sa sensibilité, c'est-à-dire sa capacité à répondre à des molécules cibles même si celles-ci sont à des concentrations très faibles.

Les capteurs SAW sont recouverts, le plus souvent, d'une couche sensible composée de polymères [Harsanyi, 1995, Grate et McGill, 1995, Grate, 2000]. De manière plus marginale, des couches sensibles à base de nanotubes de carbone [Penza *et al.*, 2007] ou de feuilles de graphène [Arsat *et al.*, 2009] ont également été proposées.

L'utilisation du diamant comme couche sensible est une approche originale développée et brevetée par le Laboratoire Capteurs Diamant du CEA-LIST qui sert de base aux travaux décrits dans cette thèse. Cette approche consiste à déposer sur la surface du capteur SAW une couche sensible constituée de nanocristaux de diamant. Après le dépôt sur la surface d'une solution contenant des nanocristaux de diamant, le solvant est retiré et les capteurs sont exposés à un plasma constitué d'hydrogène et de méthane pendant une courte durée pour fixer les nanocristaux sur la surface et mieux contrôler la terminaison de surface de ces nanocristaux. La couche en nanodiamants ainsi réalisée présente une certaine porosité mais son intérêt réside surtout dans le fait qu'elle peut être fonctionnalisée de plusieurs façons compte tenu de la diversité offerte par la chimie du carbone. On parle alors de couche diamant fonctionnalisée. Ces fonctionnalisations vont permettre de moduler l'affinité chimique de la surface de la couche de diamant.

De nombreuses techniques de fonctionnalisation de surface de diamant sont reportées dans la littérature. Parmi elles, on peut notamment citer les techniques de traitement de surface par voies physiques [Chein et Tzeng, 1999, Simon *et al.*, 2009], par photochimie UV [Strother *et al.*, 2002] , par oxydation de surface [Boukherroub *et al.*, 2005] ainsi que les techniques de fonctionnalisation par voies chimiques [Bongrain *et al.*, 2011].

1.1.3 Approche multicapteurs

Les capteurs parfaitement sélectifs, c'est-à-dire ne répondant qu'à un seul type de molécules, sont rares. Concernant les capteurs SAW, le fait qu'ils répondent à tout type de molécule, du fait de la contribution massique, les rend très peu sélectifs. Pour remédier à ce problème, l'approche communément désignée sous le nom de nez électronique a été proposée [Gardner, 1994]. Cette dernière consiste à mettre en réseau plusieurs capteurs peu sélectifs mais possédant des affinités chimiques différentes les uns des autres. La réponse de chaque

capteur à un gaz donné sera donc différente et génèrera ainsi une signature chimique. On note également qu'il existe une approche alternative consistant à mettre en réseau des capteurs de technologies différentes [Mayoue *et al.*, 2013].

1.2 Protocoles expérimentaux

1.2.1 Présentation du dispositif utilisé

Les capteurs SAW utilisés dans le cadre de cette thèse ont été fournis par l' *Institute of Microstructure Technology* (IMT) du *Karlsruhe Institute of Technology* (KIT). Il s'agit de capteurs SAW de type résonateur à ondes de Rayleigh. Le matériau piézoélectrique utilisé est le quartz ST-cut dont la fréquence de résonance (f) est comprise entre 428 et 430 MHz. Le dispositif utilisé pour la mesure de la fréquence des capteurs est l'instrument SAGAS fourni par le KIT [Rapp *et al.*, 2000]. Cet instrument intègre non seulement la fluidique permettant de diriger les vapeurs sur les capteurs mais également l'électronique capable de traiter les signaux renvoyés par les 8 capteurs qu'il peut intégrer. Un capteur non exposé aux vapeurs permet alors, par soustraction, de numériser à une fréquence comprise entre 1 et 10 Hz le signal Δf. Ce système permet également de réguler la température des capteurs et de l'électronique de manière à réduire l'influence de ce paramètre.

Dans le cadre de cette thèse, un ensemble de 8 capteurs SAW du KIT ont été utilisés. Ils ont été recouverts d'une couche sensible en diamant fonctionnalisé. Les fonctionnalisations de ces capteurs sont : OH, $(OH)_+$, $\phi - NO_2$, CH_3OH, $\phi - Cl$, NHx, caproïque et butylamine. Ici la notation $(.)_+$ signifie que la molécule en question est à une concentration élevée. Les protocoles expérimentaux utilisés pour fonctionnaliser les capteurs sont décrits dans [Girard *et al.*, 2009, Girard *et al.*, 2010, Chevallier *et al.*, 2011, Tard, 2013].

Toutes les expériences menées ont été effectuées en régulant la température des capteurs à 22^o C. Les signaux ont été échantillonnés à une fréquence de 10 Hz.

1.2.2 Environnement de laboratoire : toxiques chimiques

Le Laboratoire Capteurs Diamant travaille depuis plusieurs années sur la détection de toxiques chimiques. Aussi, pour mener à bien mes travaux, un certain nombre d'expériences ont été conduites en utilisant ces derniers compte tenu de l'expérience du laboratoire dans ce domaine et de l'intérêt de ce sujet pour développer des technologies nouvelles pour leur détection et leur identification.

Les expériences menées sur la thématique de l'identification de toxiques chimiques ont consisté à réaliser plusieurs acquisitions de la réponse des capteurs

lorsque ceux-ci sont exposés à un flux de 200mL/min d'ammoniac (NH_3), de dioxyde de soufre (SO_2), de sulfure d'hydrogène (H_2S), de méthanol (CH_3OH) et de toluène (C_7H_8) à des concentrations de 2, 4, 6, 8 et 10 parties par million (ppm).

Un banc de dilution spécialement conçu a été utilisé pour générer des vapeurs par dilution, dans de l'azote, de gaz calibrés obtenus auprès de la société Messer. La figure 1.2 montre le banc de dilution utilisé.

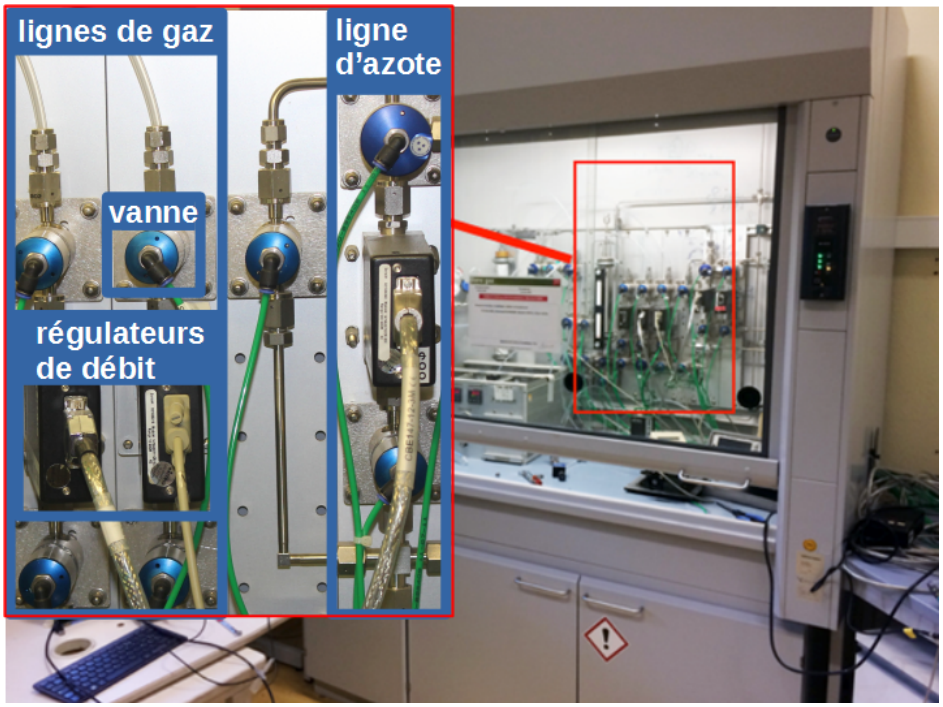


FIGURE 1.2 – Image du banc de gaz utilisé au Laboratoire Capteurs Diamant.

De manière à constituer une base de données suffisamment importante pour pouvoir appliquer les méthodes décrites dans le chapitre 2, plusieurs acquisitions ont été réalisées pour chaque composé et à chaque concentration. La chambre contenant les capteurs a été purgée à l'azote pendant près de 30 secondes entre chaque acquisition.

La figure 1.3 montre les réponses des capteurs exposés à 10 ppm d'ammoniac pendant 10 secondes.

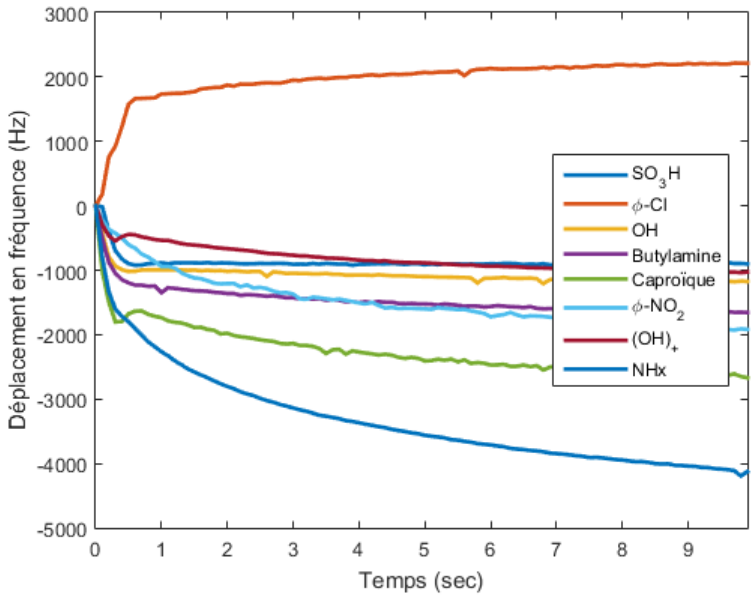


FIGURE 1.3 – Réponse des 8 capteurs lors d'une exposition à 10 ppm de sulfure d'hydrogène.

1.2.3 Environnement partiellement contrôlé : capsules de café commerciales

L'approche multicapteurs a également été utilisée en environnement partiellement contrôlé pour obtenir une deuxième base de données. Pour établir cette base, le choix s'est porté sur la contrefaçon de capsules de café.

Les expériences menées concernant le problème de l'identification de capsules de café contrefaites ont consisté à exposer les capteurs à 21 différents types de capsules de café disponibles dans le commerce et à 7 différents types de capsules contrefaites. Ces acquisitions ont été réalisées avec un système comportant uniquement 4 capteurs avec les fonctionnalisations suivantes : OH, CH_3OH, $\phi - Cl$, et butylamine.

Les différentes capsules ont été vidées dans un bécher fermé et les vapeurs dégagées ont été amenées à la chambre contenant les capteurs en utilisant une pompe. Plusieurs acquisitions ont été réalisées avec plusieurs capsules de chaque type. Les acquisitions concernant chaque type de café ont été faites les unes à la suite des autres de manière à ce que les variations de température et d'humidité de la salle d'expérimentation soient les plus faibles possibles. Après chaque acquisition, la cellule contenant les capteurs a été purgée à l'air pendant 30 secondes. La figure 1.4 illustre le montage expérimental utilisé pour acquérir les données.

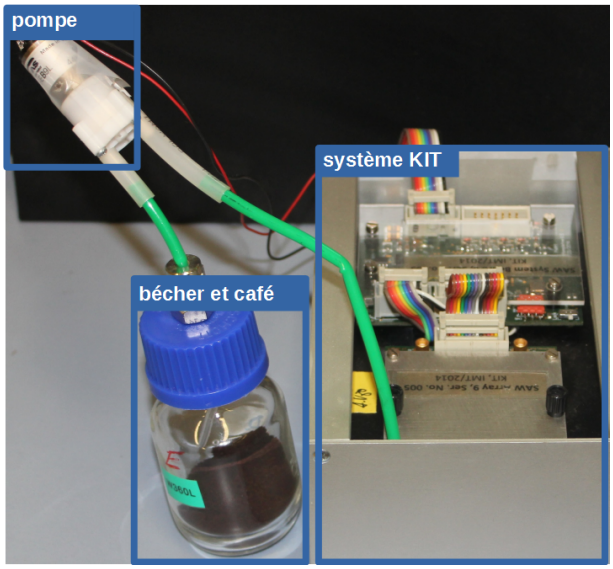


FIGURE 1.4 – Image du montage expérimental.

La figure 1.5 montre les réponses des capteurs exposés aux vapeurs générées par une capsule authentique.

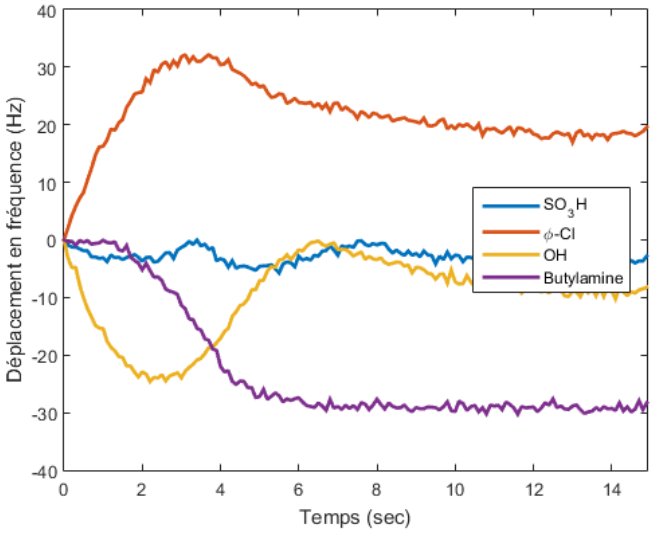


FIGURE 1.5 – Réponse des 4 capteurs en présence d'une capsule de café authentique.

1.2.4 Environnement non contrôlé : DMMP et 4-NT en sac à dos

Les dernières expériences menées s'inspirent de l'utilisation des nez électroniques dans le domaine de la sécurité. Il s'agit de détecter la présence de méthylphosphonate de diméthyle (dimethyl methylphosphonate, DMMP), un simulant du gaz Sarin, et de 4-nitrotoluène (4-NT), un explosif, en présence d'interférents : eau, éthanol, gel douche et engrais. Le DMMP, l'eau, l'éthanol et le gel douche étaient à l'état liquide tandis que le 4-NT et l'engrais étaient sous forme de poudre. Ces acquisitions ont été réalisées selon le protocole suivant : 10 mL de DMMP ou de 4-NT ainsi que 10 mL d'un interférent ont été répandus sur des feuilles de papier absorbant placées dans un sac à dos neuf de 10 L. Après une phase de diffusion de 10 minutes, les vapeurs émises ont été aspirées jusqu'à l'intérieur de la chambre contenant les capteurs en utilisant une pompe. Des acquisitions similaires ont été effectuées, en utilisant le même protocole, mais en considérant des couples d'interférents.

De manière à constituer une base de données suffisamment importante, plusieurs acquisitions ont été réalisées en utilisant différents sacs à dos pendant plusieurs jours. La chambre contenant les capteurs a été purgée à l'air pendant près de 30 secondes entre chaque acquisition. La figure 1.6 illustre le montage expérimental utilisé pour acquérir les données.

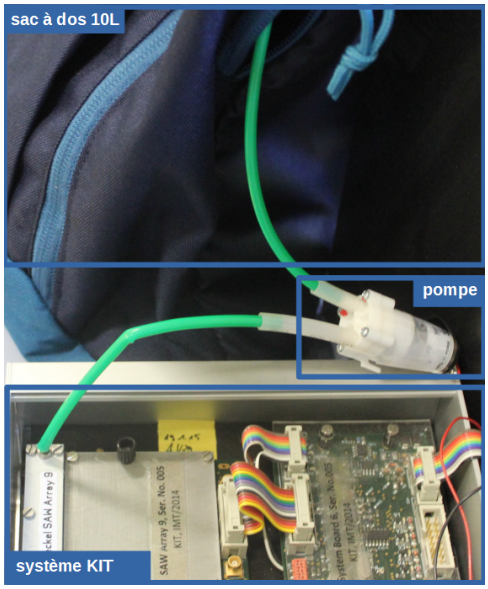


FIGURE 1.6 – Image du montage expérimental.

Pour illustrer ces nombreuses mesures, la figure 1.7 montre les réponses des capteurs à un mélange de DMMP et d'éthanol.

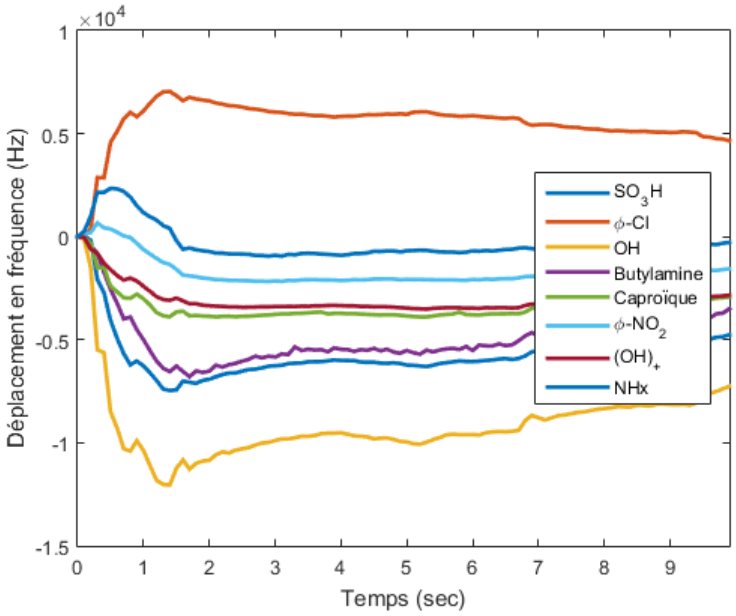
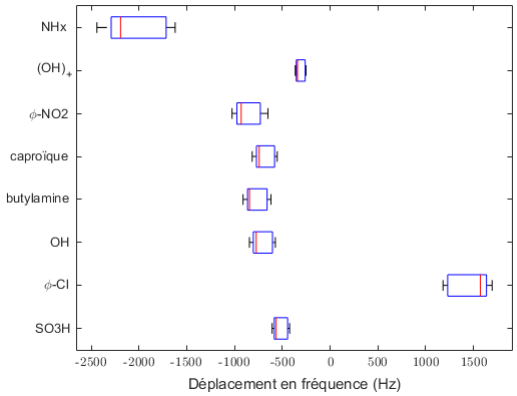


FIGURE 1.7 – Réponse des 8 capteurs en présence d'un mélange composé de DMMP et d'éthanol.

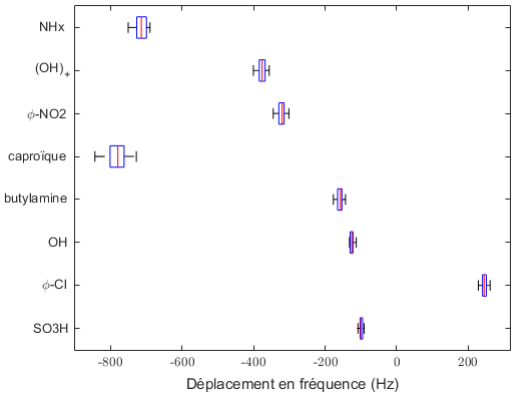
1.2.5 Répétabilité des protocoles expérimentaux

La figure 1.8 illustre la répétabilité des différentes acquisitions. Celle-ci représente la médiane, les quartiles et les extrema de l'amplitude en régime stationnaire des signaux pour toutes les acquisitions réalisées avec le même composé. Cette figure met en évidence que :

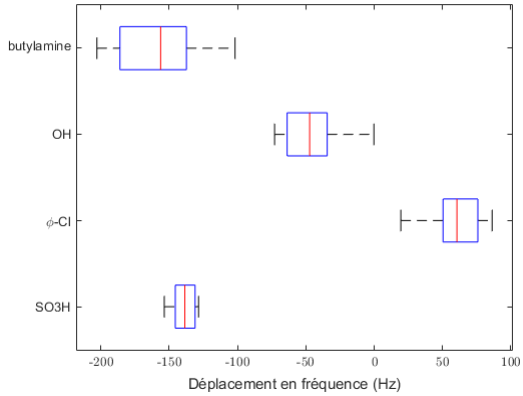
- Les mesures effectuées dans un environnement de laboratoire (figures (a) et (b)) sont répétables, la variance étant inférieure à 10%. Ceci s'explique par le fait que les concentrations sont maîtrisées, que les composés chimiques sont déshydratés et que la température des capteurs est régulée.
- Les acquisitions effectuées dans un environnement partiellement ou non contrôlé (figures (c), (d), (e) et (f)) présentent un peu plus de variabilité avec une variance de près de 20%. Ceci peut notamment s'expliquer par l'hégémonie des concentrations des vapeurs générées et par le fait que ni la température ni l'humidité de la salle dans laquelle ont été effectuées les expériences ne soient maîtrisées.



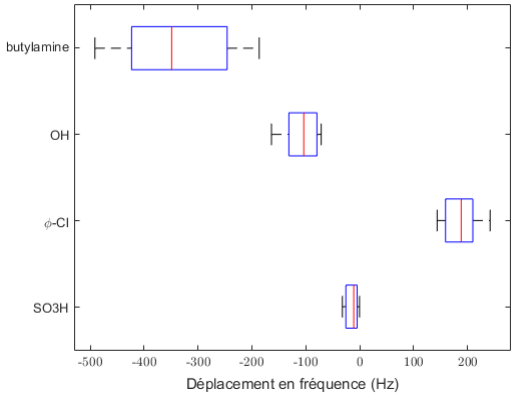
(a) Dioxyde de soufre à 6 ppm.



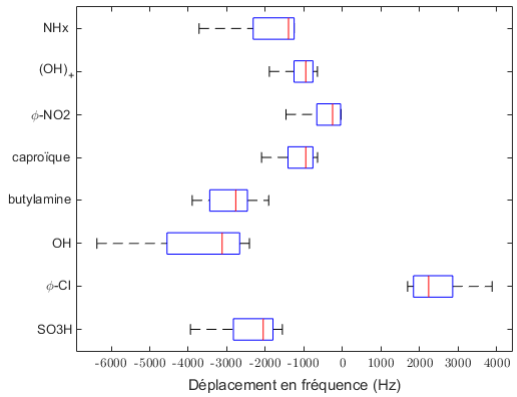
(b) Sulfure d'hydrogène à 2 ppm.



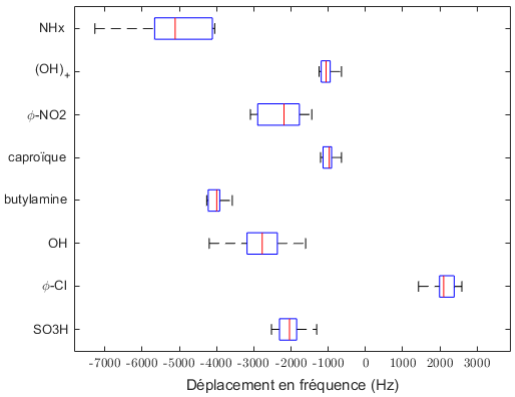
(c) Capsule authentique.



(d) Capsule contrefaite.



(e) Mélange 4-NT et engrais.



(f) Mélange eau et éthanol.

FIGURE 1.8 – Répétabilité des acquisitions pour la base des toxiques chimiques (a) et (b), pour la base des capsules de café (c) et (d) ; et pour la base constituée du DMMP et du 4-NT (e) et (f).

Bilan du chapitre

Après une description du principe de fonctionnement des capteurs SAW, nous avons mené une étude bibliographique sur ce type de capteurs et mis en évidence que le déplacement total en fréquence des capteurs SAW est principalement dû à la superposition de deux contributions, toutes deux étant modélisables par des équations linéaires différentielles du premier ordre. Ce chapitre a également présenté l'utilisation du diamant fonctionnalisé en tant que couche sensible et a brièvement introduit les approches de type nez électronique.

Pour finir, les différentes expériences effectuées et les différents jeux de données qui en découlent ont été présentés. Ces bases de données seront utilisées tout au long de ce document pour évaluer de manière quantitative les approches proposées.

Chapitre 2

Approches multiparamétriques pour l'identification de signatures chimiques

Introduction

Sans prétendre à une totale exhaustivité, ce second chapitre s'efforce de présenter un état de l'art des approches multiparamétriques appliquées à l'identification de signatures chimiques. Ainsi, ce chapitre ne se limite pas aux méthodes propres aux capteurs SAW mais décrit également des techniques utilisées pour d'autres technologies de capteurs chimiques. Nous décrivons dans un premier temps les différents descripteurs extraits de la réponse des capteurs. Ensuite, nous présentons les principales méthodes de changement de représentation et de réduction de dimension. Enfin, nous exposons les algorithmes d'apprentissage supervisé employés dans le cadre de l'identification de signatures chimiques. Notons que dans la littérature, l'analyse en composantes principales et les réseaux de neurones sont utilisés dans la très grande majorité des applications [Duval *et al.*, 2014], ce chapitre présente également des algorithmes moins répandus.

2.1 Extraction des descripteurs

2.1.1 Descripteurs directement issus de la réponse des capteurs

En général, la réponse temporelle d'un capteur SAW se compose de deux phases, comme l'illustre la figure 2.1 :

– le signal varie d'abord fortement : il s'agit du régime transitoire, puis

– le signal reste quasiment constant : il s'agit alors du régime stationnaire.

L'amplitude des signaux lors du régime stationnaire, noté A, est sans conteste le descripteur le plus utilisé pour construire des systèmes de reconnaissance d'odeurs. Cependant, de nombreux auteurs suggèrent que le régime transitoire est également porteur d'information. Par exemple, les auteurs de [Llobet *et al.*, 1997] proposent d'utiliser, en plus de l'amplitude en régime stationnaire, le temps de montée (RT, *rise time*). Celui-ci étant défini comme le temps nécessaire pour que le signal mesuré passe de 20% à 60% de l'amplitude en régime stationnaire.

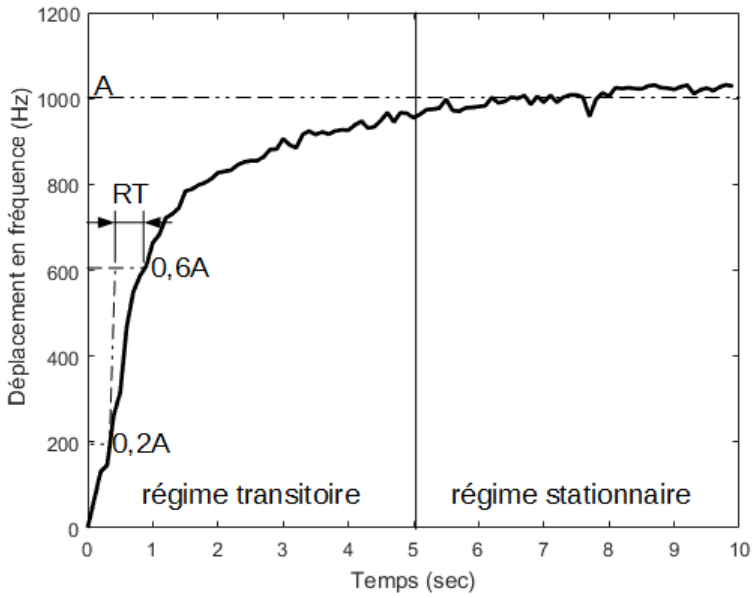


FIGURE 2.1 – Illustration des régimes transitoire et stationnaire de la réponse d'un capteur SAW exposé à 10 ppm d'ammoniac.

Une autre approche est proposée dans [Martinelli *et al.*, 2003] en utilisant comme descripteur l'aire des signaux et l'aire de l'espace de phase

$$A_{phase} = \int_0^T \frac{df}{dt} df,$$

où T représente la durée de la mesure.

2.1.2 Descripteurs issus de la modélisation de la réponse des capteurs

Une autre catégorie de descripteurs peut être obtenue en modélisant la réponse des capteurs par des fonctions paramétrées. Ces paramètres sont ensuite

estimés en minimisant un critère des moindres carrés avec l'algorithme de Levenberg Marquardt et ensuite utilisés comme descripteurs. Plusieurs modèles ont été proposés [Carmel *et al.*, 2003] dont un modèle exponentiel paramétré par β et τ :

$$R(t) = \beta\tau(1 - e^{-\frac{t}{\tau}}),$$

un modèle de Lorentz lui aussi paramétré par β et τ :

$$R(t) = \beta\tau tan^{-1}\Big(\frac{t}{\tau}\Big);$$

et un modèle double sigmoïde dont les paramètres sont α, β, γ, ϵ, δ, η , μ, ν et λ :

$$R(t) = \frac{\alpha}{\pi}\left(1 - e^{-\left(\frac{t-\beta}{\gamma}+\epsilon\right)^{\delta}}\right)^{\eta}\left(\frac{\pi}{2} - tan^{-1}\Big(\frac{t-\mu}{\nu}\Big)\right)^{\lambda}.$$

Ces descripteurs ont notamment l'avantage de décrire à la fois le régime stationnaire et le régime transitoire.

Descripteurs issus de techniques de traitement numérique du signal

La troisième famille de descripteurs est issue de travaux concernant l'analyse temps-fréquence. Les auteurs de [Naidoo et Broadhurst, 2000] proposent d'utiliser les coefficients d'une transformée discrète en cosinus (*Discrete Cosine Transform*, DCT) des signaux. Cette transformée permet, de manière similaire à la transformée de Fourrier, de représenter un signal comme une somme de cosinus oscillant à différentes fréquences. Singh et *al.* [Singh et Yadava, 2011] suggèrent d'utiliser les coefficients d'approximation d'une transformée discrète en ondelettes (*Discrete Wavelet Transform*, DWT) comme descripteurs.

2.2 Changement de représentation et réduction de dimension

2.2.1 Analyse en composantes principales

L'analyse en composantes principales (*Principal Component Analysis*, PCA) est une méthode dont l'objectif est de supprimer les potentielles colinéarités des variables. Il s'agit de déterminer une transformation linéaire qui effectue un changement de base de telle sorte que, dans la nouvelle base, la première coordonnée ait la plus grande variance, la seconde coordonnée ait la seconde plus grande variance et ainsi de suite. Plus formellement, si les données sont représentées par un ensemble de m observations d'un vecteur de dimension n

$$x = [x_1, \ldots, x_n],$$

la $k^{\text{ième}}$ composante principale est donnée par $g_k = a_{1,k}x_1 + ... + a_{n,k}x_n$ où les coefficients $a_{i,j}$ sont ceux de la matrice de la transformation linéaire recherchée. Pour les m observations nous avons le système :

$$\begin{bmatrix} g_{1,k} \\ \vdots \\ g_{i,k} \\ \vdots \\ g_{m,k} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \vdots & \vdots \\ \vdots & x_{i,j} & \vdots \\ \vdots & \vdots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} \begin{bmatrix} a_{1,k} \\ \vdots \\ a_{n,k} \end{bmatrix}$$

ou, sous forme matricielle $g_k = \mathbf{X}a_k$. L'objectif de la PCA est de déterminer le vecteur a_k minimisant la variance de la $k^{\text{ième}}$ composante principale g_k. Si les données sont centrées (de moyenne 0) et réduites (de variance 1), ce problème peut s'écrire comme un problème de valeurs propres. En effet, la variance de la composante g_k est $a_k^T\mathbf{V}a_k$ où $\mathbf{V}$ est la matrice de covariance des observations. Cette variance peut être écrite de la façon suivante :

$$J = a_k^T\mathbf{V}a_k - \lambda_k(a_k^T a_k - ||a_k||^2).$$

Comme le vecteur optimal a_k doit maximiser cette quantité, il annule la dérivée de J i.e. $2(\mathbf{V} - \lambda_k\mathbf{I})a_k = 0$. Ainsi, le vecteur a_k est le vecteur propre associé à la $k^{\text{ième}}$ valeur propre de la matrice de covariance $\mathbf{V}$. La PCA est une des techniques couramment employées pour prétraiter des données dans l'objectif d'identifier des signatures chimiques [Benedetti *et al.*, 2004, Tang *et al.*, 2010, Jha et Yadava, 2009, Cevoli *et al.*, 2011].

2.2.2 Autoencodeurs

Les autoencodeurs (AE) sont une architecture de réseaux de neurones (la section 2.3.2 décrit de manière plus approfondie les réseaux de neurones) dont l'objectif est de reconstruire leurs entrées [Bengio, 2009]. Ils ont une couche de sortie ayant le même nombre de neurones que leur couche d'entrée et un nombre arbitraire de neurones dans leur couche cachée. La figure 2.2 représente l'architecture d'un autoencodeur. Cette architecture a la capacité de changer la représentation des données en considérant la sortie des neurones de la couche cachée comme étant la nouvelle représentation des données d'entrée.

2.2.3 Réduction de dimension

Les deux algorithmes précédemment décrits sont également employés pour réduire la dimension de l'espace auquel appartiennent les données. Dans le cas de la PCA, cette opération est réalisée en ne gardant que les n premières

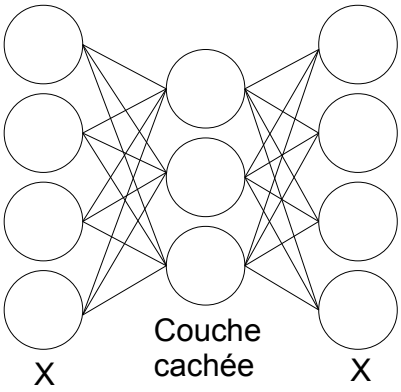


FIGURE 2.2 – Exemple d'un autoencodeur permettant de reconstruire des données de dimension 4 avec 3 neurones dans la couche cachée.

composantes expliquant $k\%$ de la variance, et dans les cas des autoencodeurs, cette opération est réalisée en ayant un nombre de neurones dans la couche cachée inférieur à la dimension des données.

2.3 Apprentissage supervisé pour l'identification de signatures chimiques

L'apprentissage supervisé est un champ d'études de l'intelligence artificielle dont l'objectif est de construire, de manière automatisée, une fonction f reproduisant une variable Y à partir d'une variable observée X :

$$Y = f(X) + \epsilon$$

où ϵ représente un bruit ou une erreur de mesure [Bishop, 2006]. Les problèmes d'identification de signatures chimiques sont des cas typiques d'apprentissage supervisé, cadre général dans lequel nous disposons d'un ensemble de n exemples étiquetés

$$\{(x_1, y_1)...(x_i, y_i)...(x_n, y_n)\}$$

avec $x_i \in X$ et $y_i \in Y \quad \forall i \in [\![1;\ n]\!]$. L'objectif est de construire, à partir de ces données, un modèle permettant de prédire la classe y d'un nouvel exemple x. La plupart du temps, le modèle est construit en choisissant une fonction paramétrée et en déterminant ces paramètres par la minimisation d'un critère sur l'ensemble d'apprentissage. Dans les sections suivantes, nous décrivons les modèles les plus couramment employés.

2.3.1 Classifieurs linéaires

Analyse discriminante par régression des moindres carrés partiels

L' analyse discriminante par régression des moindres carrés partiels (*Partial Least Squares Discriminant Analysis*, PLS-DA) est un classifieur linéaire binaire dont l'objectif est de déterminer un hyperplan séparateur. Il s'agit d'une variante de la technique de régression des moindres carrés partiels qui construit un modèle d'équation :

$$\begin{cases} X &= Tp + e \\ Y &= Tq + f \end{cases},$$

où e et f sont des résidus ; T, p et q sont les paramètres du modèle. Ces derniers sont estimés en utilisant l'algorithme PLS1 décrit entre-autre dans [Brereton et Lloyd, 2014] et qui a été adapté aux problèmes multiclasses dans [Huang *et al.*, 2013].

Analyse discriminante linéaire

L'analyse discriminante linéaire (*Linear Discrimination Analysis*, LDA) est une approche qui permet de déterminer un hyperplan séparateur entre deux classes d'exemples. Ce modèle peut s'exprimer sous la forme

$$f(x) = w^T x + w_0.$$

Une approche pour estimer les paramètres w a été proposée par Fisher dans [Fisher, 1936]. Dans cet article, il a été établi que ces derniers peuvent être estimés en résolvant le problème d'optimisation :

$$w = argmin \frac{w^T \boldsymbol{C_b} w}{w^T \boldsymbol{C_w} w},$$

où $\boldsymbol{C_b}$ est la matrice de covariance des deux classes, et $\boldsymbol{C_w}$ est la matrice de covariance de chaque classe. Bien que cet algorithme ait initialement été conçu pour ne traiter que les problèmes binaires, il a ensuite été étendu au problème multiclasses dans [Rao, 1948]. Dans le contexte plus général de la reconnaissance d'odeurs, cet algorithme a notamment trouvé des applications dans l'industrie viticole [Buratti *et al.*, 2004].

Séparateurs à vastes marges

Dans le contexte des problèmes de classification binaire, les machines à vecteurs de support (*Support Vector Machine*, SVM) sont une classe d'algorithmes qui reposent sur le principe d'hyperplan séparateur optimal. Dans la terminologie associée à cette approche, le terme vecteur de support désigne les

points, parmi ceux de l'ensemble d'apprentissage, qui sont, au sens de la distance Euclidienne, les plus proches d'un hyperplan séparateur. Si les données sont linéairement séparables, il existe souvent une infinité d'hyperplans séparateurs. On définit l'hyperplan optimal comme celui maximisant la distance entre lui-même et les vecteurs de support. La figure 2.3 illustre ce dernier. Plus formellement [Dreyfus *et al.*, 2008], il existe un hyperplan H d'équation $\omega^T x + \omega_0 = 0$ séparant les exemples x_i d'étiquette $y_i \in \{-1, 1\},\ i \in [\![1;\ n]\!]$ et vérifiant

$$\begin{cases} \omega^T x_i + \omega_0 & \geq & 1 & \text{si } y_i = 1 \\ \omega^T x_i + \omega_0 & \leq & -1 & \text{si } y_i = -1 \end{cases}$$

ou, en combinant ces deux équations, vérifiant $y_i(\omega^T x_i + \omega_0) \geq 1$. La distance entre cet hyperplan et l'hyperplan H_1 d'équation

$$H_1 \quad : \omega^T x + \omega_0 = 1 \text{ est } \frac{|1 - \omega_0|}{||\omega||}$$

et la distance avec l'hyperplan H_2 d'équation

$$H_2 \quad : \quad \omega^T x + \omega_0 = -1 \text{ est } \frac{|1 + \omega_0|}{||\omega||}.$$

La marge est la distance entre les hyperplans H_1 et H_2, et est égal à $\frac{2}{||\omega||}$. Maximiser cette quantité sous les contraintes $y_i(\omega^T x_i + \omega_0) \geq 1$ est équivalent à la résolution du problème d'optimisation

$$\begin{cases} \omega = \text{argmin} & \frac{1}{2}||\omega|| \\ \text{s. c. :} & \forall i \in [\![1;\ n]\!] \quad y_i(\omega^T x_i + \omega_0) \geq 1. \end{cases}$$

qui peut être résolu avec la technique des multiplicateurs de Lagrange.

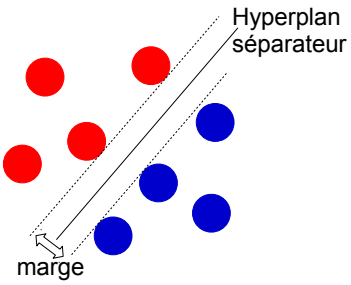


FIGURE 2.3 – Illustration de l'hyperplan séparateur optimal et des marges.

Cependant, dans la plupart des problèmes de classification, les données ne sont pas linéairement séparables et aucun hyperplan séparateur n'existe. Une

approche pour passer outre ce problème consiste à introduire des variables artificielles au problème d'optimisation :

$$\begin{cases} \omega = \text{argmin} & \frac{1}{2}||\omega|| + C\sum_{i=1}^{n}\epsilon_i \\ \text{s. c. :} & \forall i \in [\![1;\ n]\!]\ \ y_i(\omega^T x_i + \omega_0) \geq 1 - \epsilon_i \end{cases}$$

où $C \neq 0$ est une constante positive réelle.

Cette approche est souvent couplée avec l'astuce des noyaux qui consiste à projeter les données dans un espace de plus grande dimension grâce à une transformation ϕ. Cette astuce est motivée par le fait que, dans ce nouvel espace, les données pourront mieux être séparées et donc que l'erreur ϵ sera inférieure. L'hyperplan séparateur doit donc être déterminé dans ce nouvel espace, où il s'écrit en fonction de termes de la forme $\phi(x_i)^T\phi(x_j)$ qui correspondent à une fonction noyau $K(x_i, x_j)$. Parmi les noyaux, nous pouvons notamment citer :

- les noyaux linéaires
$$K(x, y) = x \cdot y;$$
- les noyaux polynomiaux de degré d
$$K(x, y) = (x \cdot y)^d \text{ ou } K(x, y) = (x \cdot y + 1)^d;$$
- les noyaux Gaussiens
$$K(x, y) = e^{-\gamma||x-y||^2} \text{ avec } \gamma > 0.$$

Bien que les SVM aient été initialement conçus pour les problèmes de classification binaires, deux approches ont été développées pour traiter les problèmes multiclasses [Hsu et Lin, 2002] :

- L'approche *un contre tous* consiste à construire un SVM par classe avec des exemples étiquetés 1 pour les exemples de la classe en question et -1 pour les autres. Un exemple de classe inconnue est associé à la classe du SVM dont il est le plus éloigné de l'hyperplan séparateur.
- L'approche *un contre un* consiste à construire $\frac{N(N-1)}{2}$ SVM binaires, où N est le nombre de classes, à partir de chaque couple de classes. Un exemple de classe inconnue est associé à la classe la plus représentée parmi les sorties des SVM. Le principal inconvénient de cette approche est qu'elle nécessite de construire un nombre de SVM s'accroissant de manière quadratique avec le nombre de classes.

Concernant les problèmes d'identification de signatures chimiques, les SVM ont été utilisés avec succès dans de nombreux domaines et notamment dans l'industrie agroalimentaire [Pardo et Sberveglieri, 2005, Brudzewski *et al.*, 2004] et dans l'industrie chimique [Gaudioso *et al.*, 2007].

2.3.2 Approches connexionnistes

Depuis leur introduction dans le milieu des années 50 suite aux travaux de Hebb, McCulloch, Pitts et Rosenblatt [Rosenblatt, 1958], les réseaux de neurones sont devenus un domaine de recherche à part entière dont plusieurs modèles ont émergé. Un neurone artificiel, ou symbolique, est une fonction qui cherche à modéliser le comportement des neurones biologiques [Bishop, 1995]. Pour un neurone ayant n entrées x_i, la sortie est donnée par

$$o_i = \phi\Big(\sum_{i=1}^{n} w_i x_i + b\Big)$$

w_i sont les poids de l'entrée, b est un biais et ϕ est une fonction de transfert. La figure 2.4 est une représentation graphique d'un neurone symbolique. Parmi

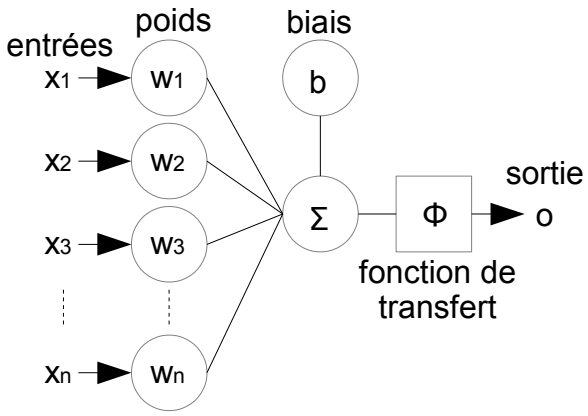


FIGURE 2.4 – Représentation graphique d'un neurone symbolique.

les fonctions de transfert ϕ couramment utilisées, nous pouvons citer :

- les fonctions linéaires
$$L(x) = x;$$
- les fonctions sigmoïdes
$$S(x) = \frac{1}{1+e^{-x}};$$
- et les fonctions Gaussiennes
$$G(x) = e^{-\frac{||\mu - x||^2}{2\sigma^2}}.$$

Pour traiter des problèmes d'apprentissage supervisé, les neurones sont organisés dans des structures plus importantes appelées réseaux de neurones. Une approche courante pour organiser des neurones consiste à les empiler sous forme de couches : les sorties de chaque neurone de chaque couche sont les entrées des neurones de la couche suivante, et ainsi de suite. De nombreux

modèles de tels réseaux ont été développés et ont été appliqués avec succès dans le contexte de la reconnaissance d'odeurs et dans de multiples domaines comme par exemple l'industrie agroalimentaire [Benedetti *et al.*, 2004, Olunloyo *et al.*, 2011, Lili *et al.*, 2012, Soh *et al.*, 2014] et l'industrie chimique [Roppel et Wilson, 2001, Chuanzhi *et al.*, 2007, Jha et Yadava, 2009]. Dans les sections suivantes, nous décrivons les modèles les plus utilisés.

Perceptron multicouche

Le perceptron multicouche (*MultiLayer Perceptron*, MLP) est l'une des architectures de réseaux de neurones les plus courantes. Cette architecture se compose d'au minimum trois couches : une couche d'entrée, au moins une couche cachée et une couche de sortie [Bishop, 1995], comme l'illustre la figure 2.5.

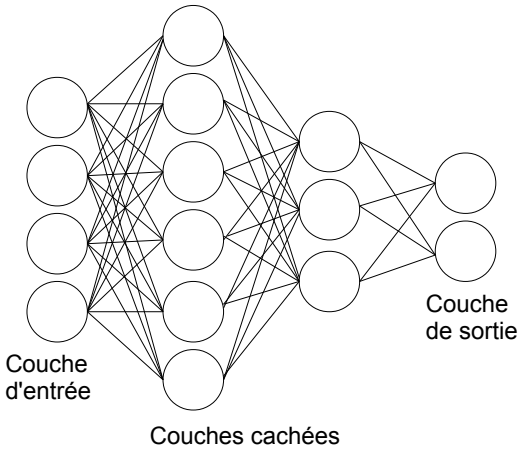


FIGURE 2.5 – Exemple d'un perceptron multicouches avec 4 neurones dans la couche d'entrée, deux couches cachées constituées de 6 et 3 neurones et 2 neurones dans la couche de sortie.

Topologie des MLP

L'architecture des perceptrons multicouches a un impact très important sur les performances en classification. Bien qu'il n'existe pas, à l'heure actuelle, de méthode pour déterminer la topologie la plus adaptée à un problème, plusieurs heuristiques ont été proposées dans la littérature :

- Le nombre de neurones de la couche d'entrée est égal à la dimension des descripteurs. Nous pouvons toutefois remarquer que certaines configurations ajoutent un neurone supplémentaire qui fait office de biais.
- Le nombre de couches cachées et le nombre de neurones par couche sont les paramètres qui ont le plus d'influence sur les performances du réseau.

Hormis le cas où les données sont linéairement séparables et où aucune couche cachée n'est nécessaire, il n'existe pas de méthode qui réponde à cette question. Néanmoins, de nombreuses heuristiques ont été proposées [Priddy et Keller, 2005]. Par exemple, Blum [Blum, 1992] suggère que le nombre de neurones des couches cachées doit être compris entre celui de la couche d'entrée et celui de la couche de sortie. Des méthodes itératives ont également été proposées. De telles méthodes consistent à ajouter de manière itérative des neurones à une petite architecture, ou, a contrario, de partir d'une grande structure et d'utiliser une stratégie d'élagage pour réduire sa taille. Ces approches sont décrites de manière détaillée dans [Reed, 1993]. Une troisième approche [Stanley, 2002] propose d'utiliser des algorithmes évolutionnistes pour déterminer la topologie la plus appropriée. Enfin, des techniques de recherche exhaustive sur une grille sont également fréquemment utilisées.

– Dans le cadre de l'apprentissage supervisé, la taille de la couche cachée est entièrement déterminée par le problème de classification, c'est-à-dire exactement un neurone par classe.

Apprentissage des paramètres des MLP

L'apprentissage des MLP est la phase lors de laquelle les différents poids et biais des neurones du réseau sont déterminés. L'algorithme dit de rétropropagation du gradient est la technique la plus couramment utilisée. Il s'agit d'une méthode itérative de descente de gradient. Plus formellement, si

$$e_j[n] = \hat{y}_j[n] - y_j[n]$$

représente l'erreur en la valeur du $j^{\text{ième}}$ neurone de sortie $\hat{y}_j[n]$ et la valeur réelle $y_j[n]$ alors, durant la phase d'apprentissage, les poids sont itérativement mis à jour de manière à minimiser l'erreur

$$E[n] = \sum_j e_j[n]^2.$$

Les poids $\omega_{i,j}$, entre le $i^{\text{ième}}$ neurone de la couche précédente et le $j^{\text{ième}}$ de la couche courante, sont mis à jour de la manière suivante :

$$\omega_{i,j}[n+1] = \omega_{i,j}[n] - \eta \frac{\partial E[n]}{\partial \omega_{i,j}[n]}$$

η étant une constante appelée taux d'apprentissage. On peut montrer que pour les neurones de sortie, on a

$$\frac{\partial E}{\partial \omega_{i,j}} = e_j \phi'(o_j)$$

et que, pour les neurones cachés, on a

$$\frac{\partial E}{\partial \omega_{i,j}} = \phi'(o_j) \sum_{k \in dest(j)} \frac{\partial E[n]}{\partial \omega_{j,k}[n]} \omega_{j,k}[n],$$

où o_j est la sortie du $j^{\text{ième}}$ neurone. De nombreuses variantes ont été proposées dans la littérature :

- Rétropropagation avec inertie : pour éviter que la descente de gradient ne converge vers un minimum local, Jacobs [Jacobs, 1988] propose d'ajouter un terme d'inertie :

$$\omega_{i,j}[n+1] = \omega_{i,j}[n] - \eta \frac{\partial E[n]}{\partial \omega_{i,j}[n]} + m(\omega_{i,j}[n] - \omega_{i,j}[n-1])$$

  où m est une constante.
- Rétropropagation avec taux d'apprentissage variable : cette variante propose d'utiliser un taux d'apprentissage μ changeant à chaque itération de l'algorithme. Une technique bien connue consiste à diminuer ce taux à chaque itération. Plagianakos et *al.* [Plagianakos *et al.*, 2002] ainsi que Barzilai et Borwein [Barzilai et Borwein, 1988] proposent d'exploiter des informations locales concernant la fonction de perte et de modifier le taux d'apprentissage en conséquence.

Bien qu'étant la méthode la plus utilisée, des alternatives à l' algorithme de rétropropagation du gradient existent [Livieris et Pintelas, 2008]. Parmi elles, on peut notamment citer : la méthode du gradient conjugué [Shewchuk, 1994], les approches de type quasi-Newton [Broyden, 1970], l'algorithme de Levenberg-Marquardt ainsi que des algorithmes évolutionnistes [Stanley, 2002]. Tous ces algorithmes ont en commun le fait qu'ils sont itératifs.

Un problème soulevé par de tels algorithmes est celui de leur arrêt. Plusieurs critères d'arrêt peuvent être considérés, par exemple lorsque :

- la durée de calcul dépasse un certain temps ;
- un nombre maximum d'itérations a été atteint ;
- la différence entre deux itérations successives est inférieure à un certain seuil $\epsilon > 0$, i.e.$||X_{k+1} - X_k|| \leq \epsilon$;
- l'erreur atteint un certain seuil $|Err(X_k)| \leq \epsilon$;
- l'erreur augmente sur un ensemble de validation.

Les algorithmes décrits ont été présentés de manière dite *online*, c'est-à-dire que les différents poids sont mis à jour après qu'un exemple ait été présenté au réseau. Une autre approche, dite *batch*, consiste à mettre les poids à jour après avoir accumulé l'erreur sur toute la base d'apprentissage. Ces deux approches peuvent être couplées en utilisant des mini-*batch* : l'erreur est alors accumulée sur une partie de la base d'apprentissage.

Minima locaux

De manière générale, tous les algorithmes présentés précédemment peuvent converger vers des minima locaux. Une technique classique pour éviter ce phénomène consiste à effectuer plusieurs phases d'apprentissage à partir de différents paramètres initiaux. Le nombre d'itérations de ces algorithmes soulève également plusieurs problèmes : si le nombre d'itérations est trop important, le phénomène de surapprentissage peut apparaître. Intuitivement, le surapprentissage apparaît lorsque le réseau a mémorisé les exemples d'apprentissage sans avoir acquis de capacité de généralisation, c'est-à-dire sans être capable, grâce aux exemples appris, de traiter des exemples encore non rencontrés. Ceci se traduit, lors de la phase d'apprentissage, par le fait que l'erreur sur la base d'apprentissage diminue tandis que celle sur la base de validation augmente. Ainsi, une méthode permettant d'éviter ce phénomène consiste à mesurer, lors de l'apprentissage, l'erreur sur un ensemble de validation, et à arrêter l'algorithme lorsque celle-ci augmente. Une autre approche consiste à ajouter à la fonction de perte un terme de régularisation qui pénalise les poids importants. Plus formellement, la fonction de perte devient $E(X) + \lambda N(X)$ où N est une fonction qui croît avec les composantes de X et λ une constante appelée coefficient de régularisation. Parmi les fonctions couramment utilisées pour N, on peut citer la norme ℓ_2 également connue sous le nom de norme Euclidienne. Une troisième approche, dite méthode du *dropout* [Srivastava *et al.*, 2014], consiste, lors de la phase d'apprentissage, à ne garder les neurones actifs qu'avec une certaine probabilité. Enfin, le problème de surapprentissage est également lié à la complexité du modèle utilisé ; dans ce cas, une topologie plus petite doit être considérée. Piotrowski et *al.* décrivent ces méthodes de manière plus approfondie dans [Piotrowski et Napiorkowski, 2013] et proposent une évaluation de celles-ci.

Réseaux à fonctions radiales

Les réseaux à fonctions radiales (*Radial Basis Function*, RBF) sont composés, comme le montre la figure 2.6 de trois couches : une couche d'entrée, une couche cachée qui effectue une transformation non linéaire de ces entrées en utilisant une fonction radiale et une couche de sortie [Schwenkar *et al.*, 2004].

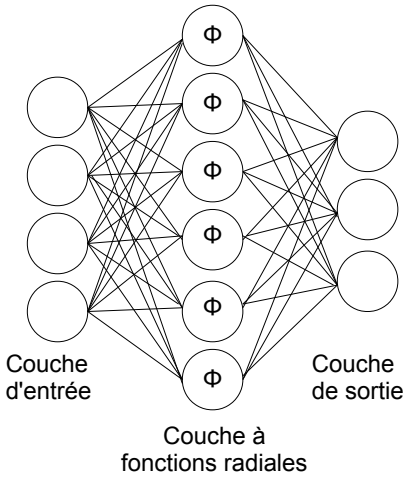


FIGURE 2.6 – Exemple d'un réseau RBF ayant 4 neurones dans la couche d'entrée, une couche radiale ayant 6 neurones, une couche de sortie de 3 neurones.

Une fonction de base radiale est une fonction ϕ symétrique autour d'un centre μ :

$$\phi(x) = \phi(||x - \mu||).$$

Un exemple typique de fonction radiale est la fonction Gaussienne. En général, ces fonctions sont également paramétrées par σ qui correspond à la largeur de la fonction. Un réseau RBF calcule une combinaison linéaire de fonctions radiales :

$$y(x) = \sum_i w_i \phi(||x - \mu_i||, \sigma_i).$$

L'apprentissage des modèles RBF utilise le fait que l'on peut diviser les paramètres en deux groupes : le premier est constitué des centres et des largeurs des fonctions radiales et le second des poids w_i. Cette caractéristique permet de réaliser l'apprentissage du modèle de manière séquentielle. Dans un premier temps, les paramètres μ_i et σ_i sont estimés, le plus souvent, à l'aide d'un algorithme de *clustering* et en considérant les centres et les largeurs des *clusters* comme les paramètres des fonctions radiales. Puis dans un second temps, les poids optimaux w_i sont calculés en résolvant, de manière analytique, le problème de régression linéaire suivant :

$$\omega = \text{argmin} \quad \sum_x \Big(y(x) - \sum_i w_i \phi(||x - \mu_i||, \sigma_i) \Big)^2.$$

Réseaux probabilistes

Enfin, une troisième architecture particulière de réseaux de neurones est celle des réseaux probabilistes (*Probabilistic Neural Network*, PNN) [Specht, 1990]. De tels réseaux possèdent, comme le montre la figure 2.7, 4 couches : une

couche d'entrée, une couche à fonctions radiales, une couche de sommation et une couche de décision.

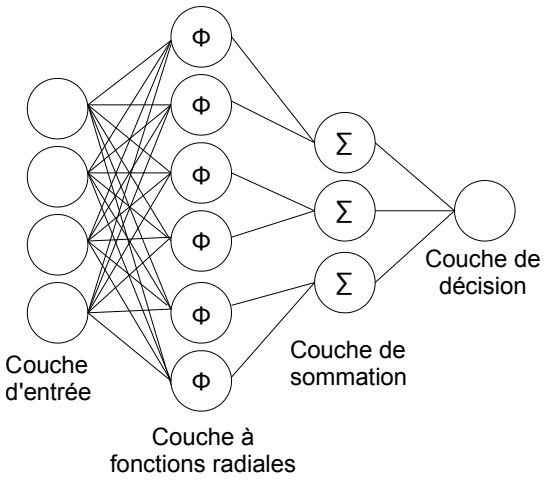


FIGURE 2.7 – Exemple d'un réseau probabiliste ayant 4 neurones dans la couche d'entrée, une couche radiale ayant 6 neurones, une couche de sommation de 3 neurones et un 1 neurone de décision.

Contrairement aux deux architectures précédemment décrites, l'architecture des PNN dépend uniquement de la base d'apprentissage : il y a autant de neurones dans la couche cachée qu'il y a d'exemples dans la base de données. Cette couche calcule la distance entre l'exemple d'entrée et chaque exemple de la base d'apprentissage. La couche de sommation est composée d'un neurone par classe et ses entrées sont les sorties des neurones de la couche radiale appartenant à cette même classe. Enfin, la couche de décision sélectionne le neurone de la couche de sommation ayant la plus grande sortie. La figure 2.7 représente l'architecture d'un réseau PNN.

2.3.3 Approches à base de voisinages

La méthode des k-plus proches voisins (k-*Nearest Neighbours*, KNN) est une technique de classification non supervisée. Pour classer un nouvel exemple, ses k plus proches voisins sont identifiés dans la base d'apprentissage, et cet exemple est associé à la classe la plus représentée parmi ces voisins. Plus formellement, pour une distance $d(\_,\_)$ et pour un exemple de classe inconnue x, les distances entre cet exemple et tous ceux de la base d'apprentissage sont calculées. Puis, si les statistiques d'ordre k sont notées

$$0 \leq d_{(1)}(x) \leq d_{(2)}(x) \leq \ \ldots \ \leq d_{(k)}(x) \leq \ \ldots \ \leq d_{(n)}(x)$$

alors l'ensemble des k plus proches voisins est

$$A_k(x) = \{x_i : d(x, x_i) < d_{(k)}(x)\}$$

et l'exemple x est assigné à la classe la plus représentée dans $A_k(x)$. La distance d est un paramètre critique, qui influence grandement les performances de KNN. Pour passer outre ce problème, les auteurs de [Weinberger *et al.*, 2006] proposent l'algorithme des k-plus proches voisins à vastes marges (Large Margin k-Nearest Neighbours, LMNN). Ils proposent d'effectuer une transformation linéaire des données et d'utiliser une distance Euclidienne :

$$D(x, y) = ||\mathbf{L}(x - y)||.$$

La matrice $\mathbf{L}$ peut être déterminée en résolvant, par une technique de descente de gradient, le problème d'optimisation convexe suivant :

$$\mathbf{L} = argmin_{\mathbf{L}} \ (1 - \mu) \sum_{i \to j} ||\mathbf{L}(x_i - x_j)||^2$$

$$+\mu \sum_{i,j \to i} \sum_{l} (1 - y_{i,l})(1 + ||\mathbf{L}(x_i - x_j)||^2 + ||\mathbf{L}(x_i - x_l)||^2)_+$$

où $y_{i,l} = 1$ ssi $y_i = y_j$, et $y_{i,l} = 0$ sinon, $(x)_+ = max(0, x)$ et $\mu \in [0;\ 1]$ qui permet de mitiger l'influence des deux termes de la fonction objectif :

- le premier terme pénalise les grandes distances entre les exemples appartenant à une même classe ;
- le second terme pénalise les petites distances entres les exemples de catégories différentes.

La notation $i \to j$ indique que x_i est un voisin cible de x_j, les voisins cibles de x_i sont les k plus proches exemples parmi ceux ayant la même classe que x_i. La principale différence entre KNN et LMNN est, comme le montre la figure 2.8, que les lignes isodistances de KNN (en utilisant une distance Euclidienne) sont des cercles, tandis que celles de LMNN sont des ellipses.

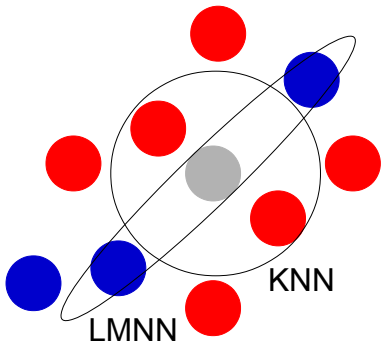


FIGURE 2.8 – Interprétation géométrique de la différence entre KNN et LMNN.

La valeur de k est un autre paramètre ayant un impact important sur les performances des approches de type plus proches voisins. Si une valeur élevée de k peut réduire les effets du bruit et des données aberrantes, une

valeur faible permet de construire des frontières de décision plus complexes. La valeur de k est communément fixée en effectuant une recherche sur une grille. Ces approches sont particulièrement sensibles à la représentativité de la base d'apprentissage. En effet, si l'une des classes est surreprésentée dans la base d'apprentissage alors celle-ci tend à dominer les résultats prédits. Malgré ce défaut, ces approches ont été appliquées avec succès dans plusieurs domaines : dans l'industrie agroalimentaire pour l'identification de fruits [Tang *et al.*, 2010], dans le secteur médical pour l'identification de bactéries [Schiffman *et al.*, 2000] et dans l'industrie chimique [Güney et Atasoy, 2012].

2.3.4 Approches à base de règles

Les arbres de décision (*Decision Tree*, DT) font partie des techniques non paramétriques d'apprentissage supervisé dont l'objectif est de prédire la classe d'une entrée par l'apprentissage de règles de décision à partir de la base d'apprentissage. Les nœuds d'un arbre de décision correspondent à une disjonction logique entre plusieurs alternatives disjointes représentées par ses branches, ces alternatives étant des propositions logiques. Les feuilles de l'arbre correspondent au résultat final de la combinaison de décision, c'est-à-dire, à la classe des exemples vérifiant la conjonction des différentes propositions.

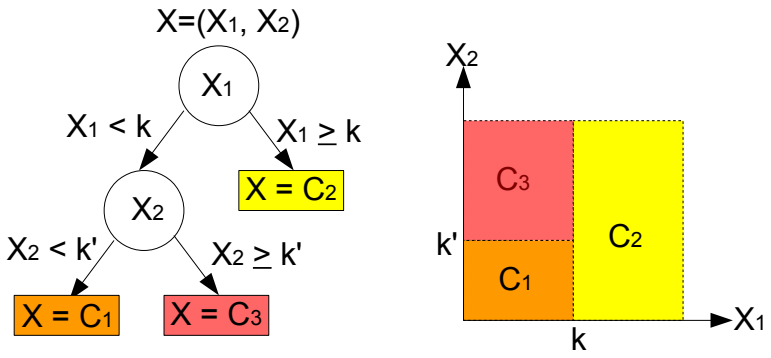


FIGURE 2.9 – Exemple d'un arbre de décision et de ses frontières de classification.

Le problème consistant à construire des arbres de décision ayant un nombre minimal de nœuds étant NP-complet, la construction des arbres de décision utilise des heuristiques n'ayant pas de garantie quant à l'optimalité de l'arbre déterminé. L'heuristique la plus courante est la méthode ID3 [Quinlan, 1986]. Celle-ci utilise une stratégie gloutonne descendante pour explorer l'espace de recherche : l'ensemble d'apprentissage est itérativement divisé en sous-ensembles tandis que l'arbre est développé. Pour diviser la base d'apprentissage, la méthode ID3 choisit le descripteur maximisant le gain d'information. Pour com-

mencer, l'entropie de chaque descripteur est calculée

$$H = -\sum p_i \ log(p_i) \ , \ p_i = \frac{\text{nombre d'éléments de l'ensemble } C_i}{\text{nombre total d'éléments}}$$

et l'ensemble de base est divisé en fonction des valeurs prises par le descripteur dont l'entropie est minimale. D'autres heuristiques, comme la méthode CART utilisant l'indice de Gini [Breiman *et al.*, 1984], ont été proposées mais leur utilisation est moins répandue. Le principal avantage des arbres de décision réside dans le fait qu'ils puissent être interprétés étant donné qu'ils consistent en un ensemble de règles. Cependant, comme leurs frontières de décision sont parallèles aux axes, comme le montre la figure 2.9, ils ont souvent des performances plus faibles que d'autres algorithmes.

Pour tenter d'améliorer leurs performances, la technique d'ensemble d'arbres de décision (*Bagged Tree*, BT) a été introduite dans [Breiman, 2001]. Cette méthode divise la base d'apprentissage en k sous-ensembles de n éléments en tirant ces derniers de manière aléatoire, construit un arbre de décision sur chacun de ces sous-ensembles et agrège leurs résultats par un vote majoritaire. Les hyperparamètres k et n sont communément déterminés en effectuant une recherche sur une grille de dimension 2. Ces algorithmes à base de règles ont trouvé leur intérêt dans le cadre des problèmes de reconnaissance de signatures chimiques. Ils ont notamment été employés dans le domaine de l'industrie agroalimentaire [Cho et Kurup, 2011] et pour l'identification de composés chimiques [Fujioka *et al.*, 2013].

2.3.5 Approches basées sur la modélisation des signaux

Les techniques de modélisation des signaux sont des outils fondamentaux pour l'analyse de séries temporelles. De nombreux modèles ont été développés [Carmona, 2004] :

- Les processus auto-régressifs (Auto-Regressive, AR) : un processus AR d'ordre p s'écrit

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

  où $\phi_0, ..., \phi_p$ sont des paramètres et ϵ_t du bruit blanc.
- Les processus à moyenne glissante (Moving Average, MA) : un processus MA d'ordre q s'écrit

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i},$$

  $\theta_0, ..., \theta_q$ sont ses paramètres, μ est la moyenne de X_t et ϵ_t du bruit blanc.

– Les processus auto-régressifs à moyenne glissante (Auto-Regressive-Moving-Average, ARMA) sont une combinaison des modèles AR et MA :

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}.$$

Les paramètres de ces modèles sont estimés par la méthode des moindres carrés ou par la technique de maximisation de la vraisemblance [Box *et al.*, 2015, Hamilton, 1994]. Bien que la recherche sur une grille puisse être envisagée, plusieurs approches ont été proposées pour déterminer la valeur des hyperparamètres p et q :

– Test statistique : cette approche consiste à surestimer les valeurs de p et q et à effectuer un test statistique sur ϕ_p et θ_q pour déterminer s'ils sont significatifs. Les tests statistiques les plus couramment employés sont les tests dits *F-statistic* et *t-statistic* [Maddala et Lahiri, 2009].
– Critère d'information : cette seconde approche consiste à effectuer une recherche sur une grille et à sélectionner les paramètres minimisant un critère d'information [Stoica et Selen, 2004]. Les critères d'information typiques sont :
  – Le critère d'information de Bayes (BIC) :

$$BIC_{p,q} = -2log(\mathcal{L}) + klog(n);$$

  – et le critère d'information d'Akaike (AIC) :

$$AIC_{p,q} = -2log(\mathcal{L}) + 2k.$$

  Ici, $\mathcal{L}$ est la vraisemblance du modèle, n le nombre d'échantillons de la série temporelle et k le nombre total de paramètres.

Les modèles précédemment décrits sont récursifs ; si l'on dispose d'une expression analytique du signal, une modélisation interprétable dans le domaine temporel est souvent préférable. Par exemple, les auteurs de [Mayoue *et al.*, 2013] proposent de modéliser la réponse des capteurs par un système du premier ordre avec une dérive linéaire :

$$f(t) = K(1 - e^{\frac{-t}{\tau}}) + \alpha t + \beta,$$

et d'estimer les paramètres en utilisant la technique des moindres carrés récursifs. Ces techniques peuvent être utilisées pour traiter des problèmes de classification : un modèle est construit pour chaque classe à partir de la base d'apprentissage et un nouvel exemple est assigné à la classe du modèle qui lui est le plus fidèle. Par exemple, [Mayoue *et al.*, 2013] utilise ces techniques pour détecter des explosifs (TNT et EDGN) en présence d'interférents.

Bilan du chapitre

Dans ce chapitre, les principales approches multiparamétriques proposées dans la littérature ont été présentées. Bien que certaines d'entre elles soient particulièrement utilisées dans le cadre de l'identification de signatures chimiques, chacune a su trouver son intérêt dans un domaine particulier.

Les réseaux de neurones font partie des algorithmes les plus répandus malgré le fait que leur entraînement et que la recherche de leur architecture soient particulièrement coûteux en temps de calcul. Toutefois, le récent développement d'architectures de calcul spécialisées tend à limiter cet inconvénient [Vainbrand et Ginosar, 2010]. Ces architectures permettent également d'embarquer des réseaux de neurones dans des systèmes mobiles.

Les SVM et les approches à base de voisinage sont également très employés bien que ces dernières requièrent de stocker en mémoire la base d'apprentissage intégralement. Bien qu'étant plus marginales, les approches à base de règles et celles basées sur des techniques de modélisation du signal ont tout de même été appliquées avec succès [Frenois *et al.*, 2014].

Chapitre 3

Positionnement par rapport à l'état de l'art

3.1 Évaluation et analyse des performances des techniques de l'état de l'art

3.1.1 Protocole d'évaluation

Les résultats présentés dans cette partie ont été obtenus en sélectionnant les hyperparamètres des différents algorithmes en effectuant une recherche sur une grille : des noyaux linéaire, quadratique, cubique et Gaussien ont été considérés pour les SVM ; les performances des réseaux de type MLP ont été obtenues en évaluant des architectures comprenant 1 ou 2 couches ayant entre 1 et 25 neurones ; et une recherche sur une grille allant de 1 à 25 a été utilisée pour évaluer les performances des approches à base de voisinage. L'algorithme de rétropropagation du gradient avec un terme de régularisation a été utilisé pour apprendre les différents paramètres des réseaux de neurones tout en évitant le phénomène de surapprentissage. Les descripteurs ont été normalisés entre 0 et 1. La dimension des descripteurs a été réduite de telle sorte qu'avec la PCA, 90% de la variance soit expliquée, et avec les autoencodeurs, de telle sorte que l'erreur de reconstruction soit 90% de celle obtenue avec un seul neurone dans la couche cachée.

Les résultats donnés dans les tables 3.1 et 3.3 représentent la moyenne des performances en classification des différentes méthodes tandis que les tables 3.2 et 3.4 représentent l'écart-type de ces dernières. Dans le cas de la base de données constituée des toxiques chimiques, les performances en classification sont définies comme étant le ratio des exemples correctement classés sur le nombre total d'exemples, tandis que dans le cas de celle constituée des capsules de café, les performances en classification sont définies comme étant le ratio de

vrais positifs. Cette même métrique sera également utilisée dans la suite de de document pour évaluer les performances des algorithmes proposés sur la base de données contenant le DMMP et le 4-NT. Ces résultats ont été obtenus lors d'un processus de validation croisée à 5 plis [Geman et Doursat, 1992].

Le processus de validation croisée consiste en la division de la base d'apprentissage en k sous-ensembles de tailles similaires. $k-1$ sous-ensembles sont utilisés pour apprendre le modèle tandis que le dernier est utilisé pour estimer ses performances. Ce processus est réitéré k fois de telle sorte que chacun des différents sous-ensembles ait été utilisé pour estimer les performances. Cette méthode de validation permet d'obtenir des résultats plus robustes puisqu'un classifieur peut avoir de bonnes performances sur un jeu de données particulier et en avoir de mauvaises sur un autre. Les résultats décrits dans la section suivante ont été obtenus avec $k = 5$. Ceux-ci sont compris entre 0 et 1. Un score de 1 signifie que tous les exemples ont été correctement classés tandis qu'un score de 0 signifie qu'ils ont tous été mal classés.

Le meilleur résultat obtenu pour chaque couple de technique de prétraitement et d'algorithme d'apprentissage supervisé est indiqué en gras. Dans le cas des tables 3.1 et 3.3 présentant les performances en classification, il s'agit du score le plus élevé. Tandis que dans le cas des tables 3.2 et 3.4, il s'agit du score le plus faible.

TABLE 3.1 – Taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données constituée des toxiques chimiques.

		Algorithmes d'apprentissage supervisé									
Descripteurs	Prétraitement	LDA	PLS-DA	SVM	MLP	RBF	PNN	KNN	LMNN	DT	BT
Amplitude	Auncun	0,903	0,902	**0,921**	0,905	0,902	0,882	0,898	0,913	0,905	0,906
	PCA	0,910	0,905	0,927	0,902	0,892	0,922	0,917	**0,930**	0,925	0,913
	AE	0,913	0,907	0,920	0,923	0,920	0,890	0,904	0,917	0,904	**0,935**
Amplitude et temps de monté	Auncun	0,891	0,893	0,894	0,895	0,893	0,883	0,899	**0,899**	0,887	0,894
	PCA	0,888	0,899	**0,908**	0,907	0,897	0,887	0,905	0,884	0,883	0,936
	AE	0,943	0,944	0,930	0,947	0,940	0,882	0,886	0,926	0,887	**0,959**
Aire	Auncun	0,838	0,855	0,875	**0,906**	0,899	0,876	0,901	0,892	0,866	0,891
	PCA	0,839	0,907	0,908	**0,936**	0,900	0,915	0,930	0,903	0,878	0,913
	AE	0,841	0,888	0,891	0,913	0,923	0,885	0,903	0,881	0,885	0,938
Aire de l'espace de phase	Auncun	0,805	0,805	0,844	0,868	0,883	0,855	0,848	**0,890**	0,822	0,865
	PCA	0,815	0,816	0,863	0,881	**0,893**	0,854	0,864	0,891	0,845	0,858
	AE	0,807	0,848	0,839	**0,911**	**0,911**	0,844	0,857	0,881	0,844	0,903
Modèle exponentiel	Auncun	0,883	0,879	**0,913**	0,893	0,897	0,877	0,895	0,906	0,896	0,889
	PCA	0,893	0,910	0,933	0,881	0,913	0,924	0,915	**0,945**	0,893	0,914
	AE	0,878	0,916	**0,939**	0,894	0,911	0,933	0,929	0,918	0,868	0,931
Modèle de Lorentz	Auncun	0,878	0,904	0,892	**0,915**	0,898	0,879	0,883	0,895	0,891	0,893
	PCA	0,907	0,925	0,940	0,941	0,919	0,931	0,897	0,916	0,927	**0,943**
	AE	0,920	0,933	0,911	**0,943**	0,907	0,896	0,907	0,925	0,899	0,899
Modèle double sigmoïdes	Auncun	0,814	0,860	0,850	**0,896**	0,794	0,838	0,830	0,853	0,818	0,888
	PCA	0,826	0,862	0,882	**0,904**	0,813	0,835	0,851	0,862	0,860	0,900
	AE	0,874	0,928	0,926	**0,934**	0,919	0,873	0,924	0,890	0,911	0,916
Transformée en cosinus	Auncun	0,818	0,806	0,832	0,819	**0,835**	0,822	0,809	0,806	0,811	0,816
	PCA	0,825	0,836	0,847	**0,876**	0,853	0,856	0,825	0,825	0,839	0,829
	AE	0,868	0,853	**0,883**	0,846	0,852	0,857	0,817	0,817	0,818	0,844
Transformée en ondelettes	Auncun	0,854	0,816	**0,895**	0,876	0,881	0,885	0,871	0,871	0,868	0,886
	PCA	0,879	0,842	0,909	**0,929**	0,920	0,898	0,909	0,877	0,866	0,905
	AE	0,905	0,917	0,914	**0,939**	0,906	0,879	0,915	0,914	0,905	0,898

TABLE 3.2 – Écarts-types des taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données constituée des toxiques chimiques.

		Algorithmes d'apprentissage supervisé									
Descripteurs	Prétraitement	LDA	PLS-DA	SVM	MLP	RBF	PNN	KNN	LMNN	DT	BT
Amplitude	Auncun	0,027	0,010	0,010	0,017	0,014	0,012	0,013	0,090	0,010	**0,007**
	PCA	0,018	0,014	0,032	**0,013**	**0,013**	0,019	0,013	0,028	0,015	0,019
	AE	0,018	0,019	0,017	**0,015**	**0,015**	**0,015**	0,022	**0,015**	0,017	0,021
Amplitude et temps de monté	Auncun	0,027	0,030	0,035	0,047	0,050	**0,023**	0,065	0,027	0,057	0,050
	PCA	0,031	0,025	**0,022**	0,025	0,046	0,059	0,060	0,056	0,032	0,059
	AE	0,065	0,035	0,045	0,061	0,056	0,037	0,057	**0,026**	0,057	0,038
Aire	Auncun	0,019	0,021	0,025	0,019	0,014	0,019	0,013	**0,012**	0,022	0,017
	PCA	0,029	0,024	0,019	**0,011**	0,021	0,027	0,021	0,020	0,019	0,015
	AE	0,022	0,019	0,024	0,015	0,012	0,022	**0,010**	0,022	0,029	0,022
Aire de l'espace de phase	Auncun	0,042	0,037	0,024	**0,021**	0,055	0,026	0,060	0,053	0,038	0,030
	PCA	0,020	0,035	0,059	**0,022**	0,048	0,047	0,067	0,057	0,038	0,039
	AE	0,065	0,053	0,065	0,053	**0,040**	0,044	0,054	0,049	0,054	0,048
Modèle exponentiel	Auncun	0,026	0,025	0,025	**0,011**	0,023	0,028	0,012	0,030	0,029	0,015
	PCA	0,024	0,014	0,029	0,021	**0,012**	0,013	0,025	0,021	0,017	0,024
	AE	0,010	0,015	**0,012**	0,025	0,018	0,018	0,023	0,017	0,027	0,022
Modèle de Lorentz	Auncun	0,027	0,023	0,014	0,016	**0,011**	0,029	0,027	0,022	0,025	0,022
	PCA	0,028	0,020	**0,012**	0,014	0,020	0,028	0,018	0,017	0,029	0,023
	AE	0,025	0,022	0,020	0,017	0,019	0,024	0,025	0,025	**0,011**	**0,011**
Modèle double sigmoïdes	Auncun	0,029	0,029	0,029	0,032	0,032	0,027	0,028	0,034	0,027	**0,015**
	PCA	0,032	0,027	0,030	0,032	0,027	0,029	0,031	0,034	0,029	**0,015**
	AE	0,031	0,030	0,034	0,027	0,029	0,030	0,028	0,033	0,026	**0,019**
Transformée en cosinus	Auncun	0,023	0,025	0,029	**0,021**	0,022	0,028	0,027	0,028	0,025	**0,021**
	PCA	0,029	0,022	0,029	0,027	0,020	0,026	0,025	0,028	0,024	**0,014**
	AE	0,027	0,025	0,028	0,025	0,027	0,022	0,025	0,027	0,028	**0,017**
Transformée en ondelettes	Auncun	0,025	0,026	**0,021**	0,028	0,025	0,022	**0,021**	**0,021**	**0,021**	0,028
	PCA	0,026	0,027	0,023	0,027	0,025	0,029	0,027	0,027	0,021	**0,012**
	AE	0,022	0,024	0,023	0,029	0,029	0,020	0,020	0,025	0,022	**0,011**

TABLE 3.3 – Taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données composée des capsules de café.

		Algorithmes d'apprentissage supervisé									
Descripteurs	Prétraitement	LDA	PLS-DA	SVM	MLP	RBF	PNN	KNN	LMNN	DT	BT
Amplitude	Auncun	0,536	0,542	0,600	0,612	0,593	0,593	0,582	**0,615**	0,521	0,536
	PCA	0,557	0,548	0,619	**0,638**	0,610	0,605	0,600	0,626	0,538	0,558
	AE	0,536	0,551	0,602	**0,629**	0,582	0,601	0,589	0,621	0,530	0,536
Amplitude et temps de monté	Auncun	0,553	0,559	0,615	**0,628**	0,615	0,599	0,592	0,623	0,522	0,563
	PCA	0,567	0,578	**0,627**	0,625	0,625	0,611	0,610	0,624	0,536	0,583
	AE	0,559	0,569	0,614	0,622	0,588	0,608	0,587	**0,626**	0,526	0,571
Aire	Auncun	0,532	0,525	0,563	0,559	**0,567**	0,564	0,538	0,552	0,536	0,542
	PCA	0,538	0,534	**0,578**	**0,578**	0,577	0,571	0,548	0,564	0,540	0,564
	AE	0,529	0,529	0,566	0,577	0,566	**0,579**	0,524	0,560	0,541	0,545
Aire de l'espace de phase	Auncun	0,492	0,506	0,521	0,522	**0,534**	0,526	0,529	0,518	0,509	0,513
	PCA	0,500	0,513	0,540	0,537	**0,552**	0,538	0,537	0,537	0,517	0,527
	AE	0,484	0,504	0,531	0,534	0,526	**0,537**	0,530	0,527	0,516	0,513
Modèle exponentiel	Auncun	0,554	0,571	0,595	0,581	0,598	**0,597**	0,576	0,610	0,534	0,562
	PCA	0,576	0,583	0,606	0,602	0,613	0,610	0,592	**0,618**	0,552	0,574
	AE	0,552	0,565	**0,610**	0,591	0,596	0,603	0,571	0,608	0,559	0,574
Modèle de Lorentz	Auncun	0,563	0,554	0,601	0,589	**0,603**	0,587	0,532	0,596	0,543	0,567
	PCA	0,579	0,576	**0,613**	**0,613**	0,630	0,595	0,546	0,608	0,549	0,588
	AE	0,561	0,557	**0,611**	0,594	0,600	0,606	0,535	0,607	0,551	0,571
Modèle double sigmoïdes	Auncun	0,532	0,521	0,556	0,541	0,556	0,536	0,521	0,532	0,546	**0,576**
	PCA	0,544	0,528	0,575	0,550	0,563	0,542	0,534	0,544	0,557	**0,589**
	AE	0,526	0,524	0,559	0,558	0,544	0,552	0,516	0,533	0,552	**0,591**
Transformée en cosinus	Auncun	0,526	0,529	0,568	**0,564**	0,561	0,549	0,536	0,561	0,548	0,551
	PCA	0,541	0,546	**0,581**	0,576	0,570	0,562	0,541	0,571	0,557	0,557
	AE	0,529	0,530	**0,584**	**0,584**	0,552	0,567	0,537	0,573	0,557	0,552
Transformée en ondelettes	Auncun	0,522	0,519	0,539	0,551	0,553	0,548	0,531	**0,543**	0,518	0,523
	PCA	0,531	0,532	0,540	0,571	**0,572**	0,558	0,548	0,547	0,527	0,533
	AE	0,520	0,513	0,542	**0,554**	0,546	0,557	0,526	0,547	0,529	0,523

TABLE 3.4 – Écarts-types des taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données composée des capsules de café.

		Algorithmes d'apprentissage supervisé									
Descripteurs	Prétraitement	LDA	PLS-DA	SVM	MLP	RBF	PNN	KNN	LMNN	DT	BT
Amplitude	Auncun	0,070	0,077	0,091	0,076	0,075	0,079	0,075	0,099	0,068	**0,062**
	PCA	0,069	0,090	0,091	0,063	0,070	0,086	**0,062**	0,087	0,073	0,087
	AE	0,069	0,077	0,087	**0,064**	0,073	0,089	0,087	0,082	0,085	0,078
Amplitude et temps de monté	Auncun	0,086	0,085	0,088	0,091	0,110	**0,075**	0,077	0,087	0,079	0,083
	PCA	0,079	0,087	0,092	0,089	0,100	**0,075**	0,082	0,095	0,078	0,110
	AE	0,067	0,082	0,081	0,063	0,099	0,077	**0,063**	0,110	0,068	0,065
Aire	Auncun	0,088	0,088	0,080	0,089	0,077	**0,069**	0,085	0,081	0,070	0,080
	PCA	0,076	0,085	**0,062**	0,090	0,078	0,080	0,072	0,087	0,067	0,089
	AE	0,082	0,085	0,072	0,086	**0,064**	0,080	0,075	0,082	0,070	0,082
Aire de l'espace de phase	Auncun	0,074	0,100	0,065	**0,063**	0,067	0,100	0,100	0,098	0,070	0,069
	PCA	0,087	0,074	0,067	0,100	**0,062**	0,088	0,066	0,100	0,110	0,090
	AE	0,095	0,070	0,068	0,110	0,110	0,100	0,110	**0,065**	0,094	0,075
Modèle exponentiel	Auncun	0,079	0,082	**0,058**	0,089	0,064	0,082	0,079	0,079	0,062	0,072
	PCA	0,084	0,070	0,077	0,076	**0,055**	0,072	0,085	0,084	0,056	0,062
	AE	0,061	0,070	0,073	**0,059**	0,085	0,077	0,073	0,067	0,069	0,088
Modèle de Lorentz	Auncun	0,079	0,085	0,077	0,066	0,054	0,086	**0,053**	0,069	0,064	0,089
	PCA	0,056	0,064	0,064	0,055	0,087	**0,052**	0,087	0,072	0,082	0,061
	AE	0,083	0,068	0,075	**0,061**	0,066	0,062	0,082	**0,061**	0,081	0,082
Modèle double sigmoïdes	Auncun	0,072	0,081	0,076	0,096	0,078	**0,065**	0,077	0,078	0,074	0,077
	PCA	0,075	0,100	**0,066**	0,086	0,098	0,086	0,067	0,097	0,071	0,069
	AE	0,090	0,100	0,110	0,110	0,091	0,087	0,089	0,080	0,100	**0,065**
Transformée en cosinus	Auncun	0,083	0,073	0,092	0,071	0,099	0,100	0,073	0,094	0,081	**0,059**
	PCA	0,083	0,091	0,084	0,065	0,110	0,084	0,062	0,095	0,100	**0,054**
	AE	0,093	0,089	0,092	0,065	0,110	0,080	0,098	0,082	0,080	**0,061**
Transformée en ondelettes	Auncun	0,099	0,087	0,087	0,063	0,070	0,094	0,072	0,061	0,098	**0,059**
	PCA	0,078	0,100	0,092	0,080	0,067	0,097	0,082	0,077	0,080	**0,058**
	AE	0,093	0,073	0,087	0,082	0,095	0,086	0,094	0,081	0,100	**0,061**

3.1.2 Analyse des résultats du *benchmark*

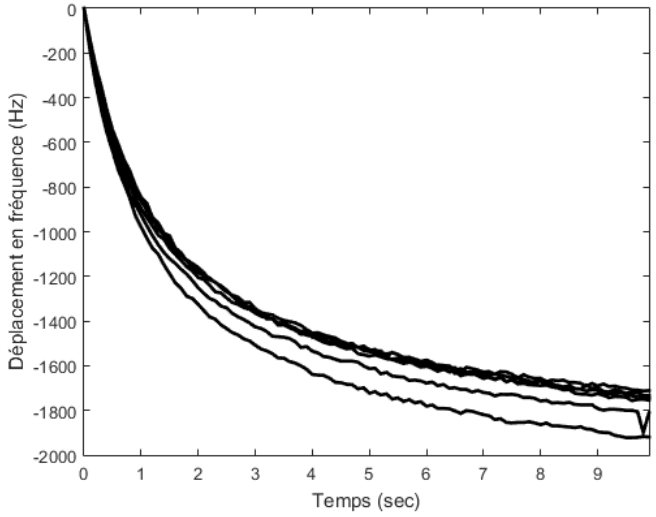
Concernant les techniques d'apprentissage supervisé, les résultats du *benchmark* mettent en évidence le caractère non linéairement séparable des problèmes d'identification de signatures chimiques puisque, à l'exception des SVM, les classifieurs linéaires sont les moins performants. Cette exception montre que l'astuce du noyau, qui permet de projeter les données dans un espace de dimension supérieure, permet de rendre le problème plus facilement séparable. Nous pouvons également remarquer que les algorithmes permettant d'apprendre des frontières de décision complexes permettent, de manière générale, d'obtenir les meilleures performances. On remarque également que les variantes BT des arbres de décision et LMNN de la méthode des k plus proches voisins permettent de remédier à certaines faiblesses des algorithmes originaux : en effet les performances de ces deux variantes sont supérieures à celles des algorithmes de base. Cependant, comme la plupart des algorithmes ne sont pas interprétables, toute analyse supplémentaire est ardue.

Concernant les techniques de prétraitement, les résultats montrent clairement l'intérêt des techniques de changement de représentation et de réduction de dimension. On peut notamment remarquer que les performances obtenues en utilisant ces techniques ont toujours été les meilleures. Les autoencodeurs semblent être préférables à l'analyse en composantes principales. Ce phénomène peut s'expliquer par leur caractère non linéaire contrairement à la PCA. Toutefois, les différences de performances étant très faibles, d'autres expériences devraient être menées pour étayer cette hypothèse.

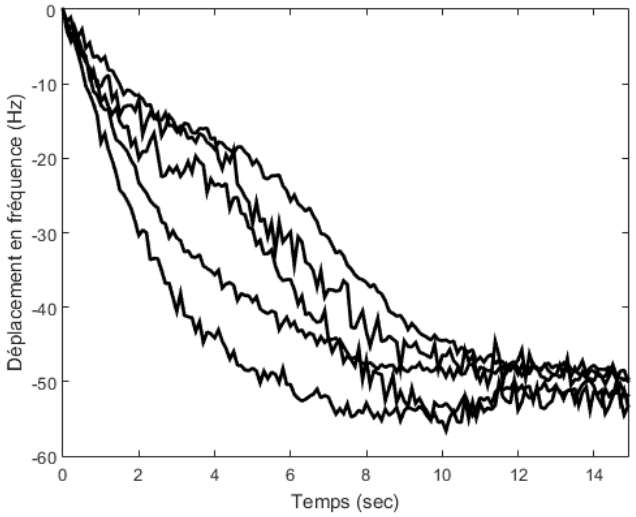
Les tables 3.2 et 3.4 montrent que l'écart-type est relativement faible : proche de 5% sur la base de données constituée des toxiques chimiques et proche de 8% sur celle constituée des capsules de café. Ceci indique que le phénomène de surapprentissage n'est pas apparu lors du processus d'apprentissage et donc que les taux de classification sont fiables.

Enfin, les résultats obtenus sur la base de données des toxiques chimiques montrent que le régime transitoire contient des informations révélatrices quant à la nature des composés chimiques tandis que ceux obtenus sur la base de données des capsules de café tendent à montrer le contraire. Ce phénomène peut s'expliquer par la non reproductibilité du processus d'échantillonnage des vapeurs des capsules de café. L'exemple de la figure 3.1 est révélateur. Celle-ci représente respectivement les réponses d'un capteur SO_3H exposé 5 fois à 10 ppm de sulfure d'hydrogène et à la même capsule de café. Cette figure permet de mettre en évidence que, contrairement au régime stationnaire, le régime transitoire n'est pas reproductible lorsque les vapeurs générées sont conduites sur les capteurs en utilisant une pompe. Ce phénomène n'apparaît pas sur la base de données des toxiques chimiques car les débits sont régulés et les concentrations maîtrisées. Ainsi, l'utilisation du régime transitoire des

réponses des capteurs ne doit être considérée que si le procédé et les conditions d'acquisition sont maîtrisés et reproductibles.



(a) Réponses d'un capteur fonctionnalisé SO_3H exposé 5 fois à 10 ppm de sulfure d'hydrogène.



(b) Réponses d'un capteur fonctionnalisé SO_3H exposé 5 fois à la même capsule de café.

FIGURE 3.1 – Non reproductibilité des régimes transitoires lorsque l'environnement n'est pas suffisamment contrôlé.

3.2 Proposition d'une nouvelle approche pour l'identification de signatures chimiques

L'analyse des résultats du *benchmark* permet de mettre en évidence qu'un descripteur idéal permettant d'identifier des signatures chimiques doit avoir les caractéristiques suivantes :

- Il doit porter des informations concernant non seulement le régime stationnaire mais également le régime transitoire puisque nous avons vu que, dans le cas de toxiques chimiques, l'utilisation de ces deux sources d'information permet d'obtenir les meilleurs résultats.
- Il doit être indépendant des variations temporelles relatives de la concentration de manière à ce que les problèmes concernant la reproductibilité du protocole expérimental n'aient pas d'impact sur lui.
- Il doit être indépendant de la valeur absolue de la concentration de telle sorte qu'un composé chimique de même nature mais à deux concentrations différentes ait le même descripteur.
- Il doit être insensible à la température.

L'étude bibliographique du chapitre 1 a mis en évidence que le déplacement en fréquence des SAW est la superposition d'une contribution massique, d'une contribution viscoélastique et d'une contribution électro-acoustique, toutes trois modélisables par des équations différentielles linéaires d'ordre 1. Puisqu'il a été expérimentalement montré que la contribution électro-acoustique est négligeable devant les deux autres, nous ne la considérerons pas dans cette étude. Nous supposerons donc que le déplacement en fréquence est la superposition des contributions massique et viscoélastique, toutes deux modélisables par les équations différentielles :

$$\tau_m \frac{\partial F_m}{\partial t} + F_m = K_m c,$$
$$\tau_v \frac{\partial F_v}{\partial t} + F_v = K_v c.$$

Nous proposons d'estimer, sans aucune hypothèse quant au profil de concentration c, les paramètres de ces équations différentielles, à savoir leurs gains et leurs constantes de temps et d'utiliser ces derniers pour non seulement identifier les signatures chimiques, mais également estimer la concentration de composés purs. Ces travaux ont plusieurs motivations :

- Ces paramètres portent des informations à la fois sur le régime transitoire et sur le régime stationnaire ce qui, comme expliqué dans la section 3.1.2, a un intérêt pour la reconnaissance de signatures chimiques.
- Ils sont également indépendants du profil de concentration c et donc de la manière dont les vapeurs sont conduites dans la chambre contenant les

capteurs. Ainsi, ils devraient être plus robustes aux problèmes concernant la reproductibilité du protocole expérimental comme expliqué dans la section 3.1.2.
– Ces paramètres permettent également de déterminer le profil temporel de concentration c apportant ainsi des informations quant à l'évolution de la concentration dans le temps.
– Ils appartiennent également à un espace de dimension plus grande que la majorité de ceux proposés dans la littérature et ainsi apportent davantage d'informations concernant les signatures chimiques. Par exemple, supposons que les amplitudes des contributions massique et viscoélastique de deux composés chimiques a et b soient respectivement

$$G_{m,a},\ G_{v,a} \text{ et } G_{m,b},\ G_{v,b}.$$

Alors, leurs amplitudes en régime stationnaire sont respectivement

$$S_a = G_{m,a} + G_{v,a} \text{ et } S_b = G_{m,b} + G_{v,b}.$$

Nous pouvons très bien avoir $S_a = S_b$ ce qui signifie que l'amplitude du régime stationnaire rend ces deux composés indifférentiables tandis que nous avons

$$G_{m,a} \neq G_{m,b} \text{ et } G_{v,a} \neq G_{v,b}$$

ce qui les rendrait différentiables.
– Enfin, ces paramètres ont un sens physique et permettront donc de proposer des algorithmes interprétables.

Le descripteur proposé vérifie ainsi les deux premiers points discutés précédemment. Néanmoins, celui-ci est dépendant de la valeur absolue de la concentration mais nous verrons dans la seconde partie de ce manuscrit que cette propriété peut être intéressante car elle permet d'estimer la concentration des composés chimiques. Il est également sensible à la température mais comme celle-ci est régulée au niveau de la chambre contenant les capteurs son impact est limité.

Enfin, nous veillerons le plus souvent à proposer des méthodes ayant un sens physique, permettant de les rendre interprétables de manière à pouvoir analyser au mieux les résultats obtenus.

Conclusion de la première partie

Après avoir décrit le principe de fonctionnement des capteurs SAW et les approches de type nez électronique pour l'identification de signatures chimiques, nous nous sommes intéressés aux couches sensibles en diamant qui permettent une grande sensibilité et de nombreuses terminaisons de surface qui sont autant de fonctionnalisations aux affinités chimiques différentes.

Cette première partie a également été l'occasion de décrire les bases de données utilisées tout au long de cette étude ainsi que les différentes techniques de l'état de l'art. Ces techniques ont d'ailleurs pu être comparées quantitativement sur les bases de données décrites dans cette même partie. Parmi les résultats remarquables, il convient de noter que les techniques de réduction de dimension et que les algorithmes d'apprentissage supervisé ayant des frontières de décision complexes permettent d'obtenir les meilleurs scores.

Ce *benchmark* a également permis de mettre en évidence l'impact du protocole expérimental sur les performances de ces méthodes. Un exemple révélateur est celui des informations portées par le régime transitoire des réponses des capteurs. Dans le cas où le protocole d'acquisition est reproductible, le régime transitoire est alors porteur d'informations révélatrices quant à la nature de la signature chimique. Dans le cas contraire, les résultats sont plus mitigés.

Enfin, une succincte description de la méthode proposée dans le cadre de cette étude et de ses motivations a terminé cette première partie.

Deuxième partie

Identification de signatures chimiques

Introduction

L'objectif de cette partie est d'introduire de nouveaux descripteurs se rapprochant du descripteur idéal décrit dans le chapitre 3 et de comparer leurs performances, à la fois sur des problèmes d'identification de composés chimiques et d'estimation de la concentration, avec les résultats des techniques de l'état de l'art.

Le chapitre 4, après avoir justifié le modèle utilisé dans la suite de ce manuscrit, s'intéresse à l'estimation des paramètres des équations différentielles modélisant les contributions massique et viscoélastique. Afin d'approcher ces derniers, une formulation de ce problème comme un problème d'optimisation sous contraintes est proposée, puis la résolution de ce dernier par l'utilisation de différentes métaheuristiques est discutée.

Le chapitre 5 met en évidence l'intérêt de ces paramètres en tant que descripteurs en entrée d'algorithmes d'apprentissage supervisé. Dans un premier temps, il décrit les résultats expérimentaux, obtenus dans le contexte de l'identification de composés chimiques, puis les compare avec les performances obtenues en utilisant les techniques de l'état de l'art. Il détaille également une méthode permettant de fusionner les descripteurs proposés et ceux de l'état de l'art. Dans un second temps, une méthode permettant d'estimer le profil temporel de concentration, basée sur des techniques de déconvolution et de régression non paramétrique, est exposée et ses performances expérimentales sont détaillées.

Pour finir, le chapitre 6 décrit un algorithme glouton permettant, pour un problème donné, de déterminer les fonctionnalisations des capteurs les plus appropriées. L'évaluation des résultats de cet algorithme conclut ce chapitre.

Chapitre 4

Estimation des paramètres des contributions massique et viscoélastique

Introduction

Dans la section 1.1.1, nous avons vu qu'il existe deux modélisations pour formaliser la réponse des capteurs SAW : les équations différentielles décrites dans [Ballantine *et al.*, 1997, Tard, 2013] et celles décrites dans [Manai, 2014] dans lesquelles les constantes de temps sont des fonctions linéaires de la concentration. Dans un premier temps, nous donnerons les arguments qui nous ont fait choisir la formalisation de Ballantine et Tard. Puis, nous nous intéresserons à l'estimation des paramètres de ces équations différentielles en les considérant comme des solutions d'un problème d'optimisation que nous poserons et résolverons en utilisant des métaheuristiques.

4.1 Justification du modèle utilisé

Dans le chapitre 1.1.1, nous avons vu, en négligeant la contribution électroacoustique, que la réponse d'un capteur SAW peut être modélisée par ce système d'équations :

$$\begin{cases} \tau_m \frac{\partial F_m}{\partial t} + F_m = & K_m c \\ \tau_v \frac{\partial F_v}{\partial t} + F_v = & K_v c \\ F_m + F_v = & F \end{cases} \tag{4.1}$$

où c est la concentration. Toutefois, un modèle alternatif a été proposé dans [Manai, 2014]. Dans ce modèle les constantes de temps τ_m et τ_v sont des fonctions linéaires de la concentration :

$$\tau_i = \alpha_i c + \beta_i \quad i \in \{m, v\}.$$

Pour juger de la représentativité de ce raffinement, le temps de réponse à 90% a été mesuré sur tous les exemples de la base de données constituée des toxiques chimiques. Le temps de réponse à 90%, noté $t_{90\%}$, est défini comme étant le temps nécessaire pour qu'un signal monotone atteigne 90% de sa valeur en régime stationnaire [Levine, 1996] ou, plus formellement, $t_{90\%}$ est défini tel que $F(t_{90\%}) = 0,9F_s$ où F_s représente la valeur du régime stationnaire de F. La figure 4.1 illustre les variations du temps de réponse à 90% d'un capteur exposé à de l'ammoniac. Entre 12 et 17 acquisitions ont été réalisées pour chaque concentration.

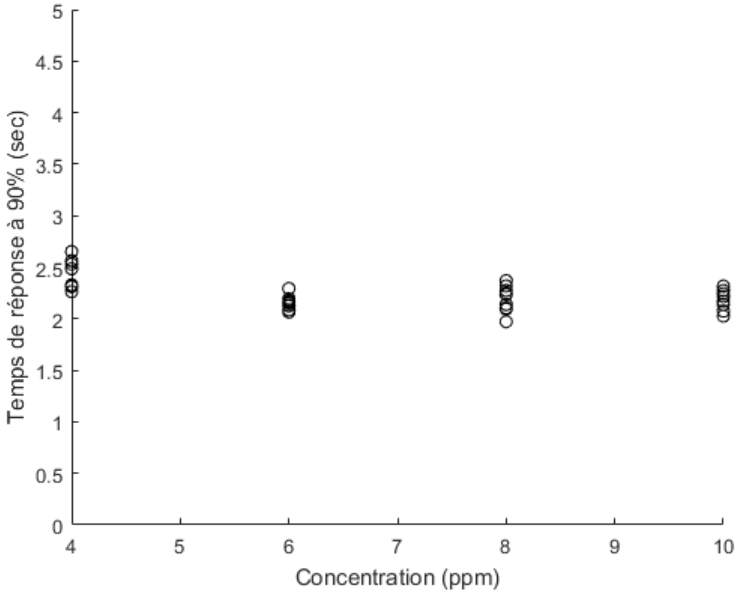


FIGURE 4.1 – Temps de réponse à 90% d'un capteur exposé à de l'ammoniac.

Cette figure montre non seulement que les variations du temps de réponse à 90% sont faibles mais également qu'elles sont négligeables devant l'écart-type (variant entre $0,2$ et $0,7$ seconde) des résultats obtenus pour une même concentration. Ces mêmes observations peuvent être réalisées quel que soit le composé chimique et quelle que soit la fonctionnalisation du capteur auquel nous nous intéressons. Nous pouvons donc en déduire que, dans la plage de concentrations qui nous intéresse, ce raffinement n'est que très peu représentatif. Par conséquent, nous ne considérerons pas ce raffinement dans cette étude et les équations seront ainsi simplifiées.

4.2 Formulation du problème d'optimisation

Dans cette section, nous formulons un problème d'optimisation dont les solutions sont les paramètres des contributions massique et viscoélastique.

La méthode d'Euler permet d'approcher une dérivée $\frac{\partial F}{\partial t}(t)$ par

$$\frac{\partial F}{\partial t}(t) \approx \frac{F[n] - F[n-1]}{T_s}$$

où T_s est le taux d'échantillonnage, $F[n]$ est le $n^{ième}$ échantillon numérisé de la séquence F et $c[n]$ le $n^{ième}$ échantillon numérisé du profil de concentration. En appliquant cette méthode aux équations (4.1), nous obtenons pour $i \in \{m, v\}$:

$$F_i[n] = \frac{\tau_i}{\tau_i + T_s} F_i[n-1] + K_i \frac{T_s}{\tau_i + T_s} c[n]$$

Ces équations ont deux propriétés intéressantes :

1. Elles sont linéaires puisque, pour $i \in \{m, v\}$ si

$$F_{a,i}[n] = \frac{\tau_i}{\tau_i + T_s} F_{a,i}[n-1] + K_i \frac{T_s}{\tau_i + T_s} c_a[n]$$

et

$$F_{b,i}[n] = \frac{\tau_i}{\tau_i + T_s} F_{b,i}[n-1] + K_i \frac{T_s}{\tau_i + T_s} c_b[n]$$

alors nous avons $\forall\ \alpha \in \mathbb{R}$

$$F_{a,i}[n] + \alpha F_{b,i}[n] = \frac{\tau_i}{\tau_i + T_s}(F_{a,i}[n-1] + \alpha F_{b,i}[n-1])$$

$$+K_i \frac{T_s}{\tau_i + T_s}(c_a[n] + \alpha c_b[n]),$$

2. Elles sont temps invariant puisque, pour $i \in \{m, v\}$, nous avons

$$F_i[n - n_0] = \frac{\tau_i}{\tau_i + T_s} F_i[n - n_0 - 1] + K_i \frac{T_s}{\tau_i + T_s} c[n - n_0].$$

L'annexe A (page 129) décrit les principales propriétés des processus linéaires et temps invariant et montre notamment que la sortie de tels processus n'est autre que le produit de convolution de leur réponse impulsionnelle avec leur entrée. Ainsi, nous avons :

$$\begin{cases} F_m &= h_m * c \\ F_v &= h_v * c \end{cases} \tag{4.2}$$

où h_m et h_v sont respectivement les réponses impulsionnelles des contributions massique et viscoélastique.

Détermination des réponses impulsionnelles

Comme expliqué dans l'annexe A (page 129), la réponse impulsionnelle d'un processus est sa sortie lorsque son entrée est la fonction delta de Dirac. Pour $i \in \{m, v\}$, nous avons donc

$$\begin{cases} h_i[0] = K_i \frac{T_s}{\tau_i + T_s} \\ h_i[n] = \frac{\tau_i}{\tau_i + T_s} h_i[n-1] \quad \forall n > 0. \end{cases}$$

Puisqu'il s'agit d'une série géométrique, nous avons

$$h_i[n] = K_i \frac{T_s}{\tau_i + T_s} \left(\frac{\tau_i}{\tau_i + T_s} \right)^n, \quad i \in \{m, v\}. \tag{4.3}$$

Nous posons :

$$T_i = \frac{\tau_i}{\tau_i + T_s} \text{ ou } \tau_i = \frac{T_i T_s}{1 - T_i} \text{ et} \tag{4.4}$$

$$A_i = K_i(1 - T_i) \text{ ou } K_i = \frac{A_i}{1 - T_i}. \tag{4.5}$$

La substitution des équations (4.3), (4.4) et (4.5) dans (4.2) donne

$$F[n] = \left(A_m {T_m}^n + A_v {T_v}^n \right) * c[n].$$

Cette relation met en évidence que les paramètres A_m et A_v ne peuvent être estimés qu'à une constante multiplicative $\alpha \neq 0$ près puisque nous avons :

$$F[n] = \left(\frac{A_m}{\alpha} {T_m}^n + \frac{A_v}{\alpha} {T_v}^n \right) * (\alpha c[n]).$$

Ainsi, et sans perte de généralité, nous pouvons faire l'hypothèse que le profil de concentration c^* est unitaire, c'est-à-dire que son amplitude lors du régime stationnaire est 1. Ceci implique que

$$F[n] = \left(A_m {T_m}^n + A_v {T_v}^n \right) * c^*[n] \text{ et} \tag{4.6}$$

$$\frac{A_m}{1 - T_m} + \frac{A_v}{1 - T_v} = F_S,$$

où F_S est l'amplitude du régime stationnaire de F.

La présence du produit de convolution dans l'équation (4.6) rendant tout développement supplémentaire complexe, nous proposons de le transformer en un produit scalaire en utilisant des fonctions génératrices. Une fonction génératrice est une série polynomiale dont les coefficients sont les termes successifs d'une séquence donnée [Meyer et Rubinfeld, 2005] :

$$G(s, x) = \sum_{n=0}^{+\infty} s[n] x^n.$$

De telles fonctions ne sont définies que pour les valeurs de x pour lesquelles la somme converge. Cet ensemble est communément appelé région de convergence (Region Of Convergence, ROC) de la fonction génératrice :

$$ROC_{G(s)} = \left\{ x : \sum_{n=0}^{+\infty} s[n]x^n \text{ converge} \right\}.$$

Les fonctions génératrices et leurs propriétés sont décrites dans l'annexe B (page 131). Dans les paragraphes suivants, nous calculons les fonctions génératrices associées respectivement au profil de concentration unitaire c^*, puis à la réponse impulsionnelle h et enfin à la réponse des capteurs.

Fonction génératrice associée à c^*

La fonction génératrice associée au profil de concentration c^* est

$$\forall x \in ROC_{G(c^*)} \quad G(c^*, x) = \sum_{n=0}^{+\infty} c^*[n]x^n.$$

Puisque le profil de concentration c^* est inconnu, il nous est non seulement impossible de simplifier cette équation, mais également de déterminer la région de convergence de cette même fonction génératrice. Cependant, le fait que c^* représente une concentration et soit donc bornée par c^*_{max}va nous permettre de trouver un intervalle sur lequel nous pourrons garantir la convergence de la fonction génératrice. Considérons les séries

$$s_{max}[n] = \sum_{i=0}^{n} c^*_{max}|x^i| = c^*_{max} \sum_{i=0}^{n} |x|^i \text{ et}$$

$$s[n] = \sum_{i=0}^{n} c^*[i]|x^i|.$$

La série s_{max} étant une série géométrique, elle converge vers une limite l_{max} sur l'intervalle $\{x : |x| < 1\}$ et nous avons donc :

$$\forall x \in]-1, 1[, \forall \epsilon > 0, \exists N \in \mathbb{N}, n \geq N \Rightarrow \left| \sum_{i=0}^{n} c^*_{max}|x^n| - l_{max} \right| \leq \epsilon.$$

De plus, nous pouvons remarquer que les séries s_{max} et s sont croissantes :

$$s_{max}[n] \leq s_{max}[n] + c^*_{max}|x^{n+1}| = s_{max}[n+1]$$

et

$$s[n] \leq s[n] + c^*[n]|x^{n+1}| = s[n+1],$$

et que

$$s_{max}[n] = \sum_{i=0}^{n} c^*_{max}|x^n| \geq s[n] = \sum_{i=0}^{n} c^*[i]|x^i|.$$

Ainsi, sur l'intervalle $]-1,1[$ où s_{max} converge vers l_{max} nous avons :

$$l_{max} \geq \sum_{i=0}^{+\infty} c^*[i]|x^i|.$$

Comme s est croissante et bornée sur l'intervalle $]-1,1[$, elle converge vers sa limite l. Ainsi, la série $\sum_{n\geq 0} c^*[n]x^n$ est absolument convergente sur ce même intervalle. L'absolue convergence impliquant la convergence simple, nous pouvons en conclure que la série $\sum_{n\geq 0} c^*[n]x^n$ converge sur l'intervalle $]-1,1[$ et ainsi que

$$]-1,1[\subset ROC_{G(c^*)}. \tag{4.7}$$

Fonction génératrice associée à h

La fonction génératrice associée à la réponse impulsionnelle est

$$\forall x \in ROC_{G(h)} \;\; G(h,x) = \sum_{n\geq 0}(A_m T_m^n + A_v T_v^n)x^n = \frac{A_m}{1-T_m x} + \frac{A_v}{1-T_v x}$$

et sa région de convergence est :

$$ROC_{G(h)} = \left\{x : |x| < min\left(\frac{1}{T_m}, \frac{1}{T_v}\right)\right\} \text{ car } T_m, T_v < 1. \tag{4.8}$$

Fonction génératrice associée à F

La fonction génératrice associée à la réponse F d'un capteur est

$$\forall x \in ROC_{G(F)} \;\; G(F,x) = G(c^* * h, x) = G(c^*, x)G(h,x),$$

et sa région de convergence est

$$ROC_{G(F)} = ROC_{G(c^*)} \cap ROC_{G(h)}.$$

Les équations (4.4), Eq. (4.7) et Eq. (4.8) nous permettent d'affirmer que

$$\{x : |x| < 1\} \subset ROC_{G(F)} \text{ car } T_m, T_v < 1.$$

Formulation du problème

Comme les nez systèmes multicapteurs sont constitués d'un ensemble de N_s capteurs, ils peuvent être modélisés par des systèmes ayant une entrée et plusieurs sorties (*Single Input Multiple Ouput system*, SIMO *system*). Leur entrée est le profil de concentration $c^*[n]$ et leurs sorties sont les réponses des différents capteurs F_i pour $i \in [\![1;\ N_s]\!]$. Le calcul des fonctions génératrices associées à chaque réponse F_i donne le système :

$$G(F_i, x) = \left(\frac{A_{m,i}}{1 - T_{m,i}x} + \frac{A_{v,i}}{1 - T_{v,i}x}\right) G(c^*, x).$$

Il est possible de déterminer $G(c^*, x)$ en utilisant la première équation (pour $i = 1$)

$$G(c^*, x) = \frac{G(F_1, x)}{\frac{A_{m,1}}{1-T_{m,1}x} + \frac{A_{v,1}}{1-T_{v,1}x}}$$

puis en substituant son expression dans les autres (pour $i \in [\![2;\ N_s]\!]$) nous obtenons

$$G(F_i, x) = \left(\frac{A_{m,i}}{1 - T_{m,i}x} + \frac{A_{v,i}}{1 - T_{v,i}x}\right) \frac{G(F_1, x)}{\frac{A_{m,1}}{1-T_{m,1}x} + \frac{A_{v,1}}{1-T_{v,1}x}}.$$

Ainsi, par construction, les différents paramètres $A_{i,v}$ et $T_{i,v}$ peuvent être estimés en résolvant le problème d'optimisation :

$$\begin{cases} \text{argmin} & \sum_{x \in X} \left\| GH(x) \frac{G(F_1,x)}{\frac{A_{m,1}}{1-T_{m,1}x} + \frac{A_{v,1}}{1-T_{v,1}x}} - F(x) \right\| \\ \text{s. c. :} & \\ & \forall i \in [\![1;\ N_s]\!] \quad T_{m,i} \in]0;\ 1[\text{ et } T_{v,i} \in]0;\ 1[\\ & \forall i \in [\![1;\ N_s]\!] \quad A_{m,i} < 0 \\ & \forall i \in [\![1;\ N_s]\!] \quad \frac{A_{m,i}}{1-T_{m,i}} + \frac{A_{v,i}}{1-T_{v,i}} = F_{S,i} \end{cases} \tag{4.9}$$

où

$$GH(x) = \begin{bmatrix} \frac{A_{m,2}}{1-T_{m,2}x} + \frac{A_{v,2}}{1-T_{v,2}x} \\ \vdots \\ \frac{A_{m,N_s}}{1-T_{m,N_s}x} + \frac{A_{v,N_s}}{1-T_{v,N_s}x} \end{bmatrix},$$

$$F(x) = \begin{bmatrix} G(F_2, x) \\ \vdots \\ G(F_{N_s}, x) \end{bmatrix},$$

l'ensemble X des éléments pour lesquels la somme est calculée est un sous-ensemble de dimension finie de

$$[-10^{\frac{-\epsilon-log(M)}{N}}, 10^{\frac{-\epsilon-log(M)}{N}}],$$

et M est la valeur maximale des réponses des capteurs. Cet ensemble permet, comme expliqué dans l'annexe B (page 131), de garantir que les fonctions génératrices $GF(F_i, x)$ sont estimées avec une précision $10^{-\epsilon}$.

Toute distance pourrait être utilisée dans le problème d'optimisation (4.9). Sans perte de généralité, nous avons utilisé la distance Euclidienne.

4.3 Comparaison de métaheuristiques pour la résolution du problème

Ce problème d'optimisation ne pouvant pas être résolu de manière analytique, n'étant pas convexe et ayant plusieurs minima locaux, l'utilisation d'algorithmes d'optimisation numériques, appelés métaheuristiques, est l'approche la plus appropriée pour le résoudre. Dans cette section, nous comparons les performances de techniques classiques de l'état de l'art sur ce problème : recuit simulé (*Simulated Annealing*, SA), stratégies d'évolution $(\lambda + \mu)$ ($(\lambda + \mu)$-*Evolution Strategy*, ES) et optimisation par essaim particulaire (*Particule Swarm Optimization*, PSO). Ces métaheuristiques sont décrites dans l'annexe C (page 135). Les performances de ces méthodes sont évaluées selon deux critères :

1. Leur précision, c'est-à-dire leur capacité à approcher la solution optimale. Les véritables valeurs des différents paramètres n'étant pas connues, nous considérons qu'une méthode est plus précise qu'une autre si, à la fin du processus d'optimisation, elle fournit une plus petite valeur de la fonction objectif (4.9). La métrique utilisée pour mesurer la précision d'une métaheuristique est donc la fraction d'exemples pour laquelle cette dernière a déterminé la solution ayant la plus petite valeur de fonction objectif.
2. Leur reproductibilité, c'est-à-dire leur capacité à trouver la même solution si plusieurs processus d'optimisation sont effectués. Ce critère est très important ici car d'une part les méthodes employées mettent en jeu des nombres aléatoires et d'autre part, pour maximiser les performances des algorithmes d'apprentissage, le bruit doit être le plus faible possible pour garantir les meilleures performances. La métrique utilisée pour quantifier ce critère est obtenue en moyennant, sur tous les exemples de la base de données, la variance obtenue après avoir effectué 10 fois le processus d'optimisation sur chaque exemple.

La méthode la plus appropriée est donc celle ayant la plus grande précision et la plus petite variance.

Les résultats expérimentaux décrits dans la table 4.1 ont été obtenus après 25000 évaluations de la fonction objectif. L'ensemble X est composé de 320 valeurs linéairement espacées dans l'intervalle

$$[-10^{\frac{-\epsilon-log(M)}{N}}, 10^{\frac{-\epsilon-log(M)}{N}}]$$

(320 correspond à 10 fois le nombre de paramètres : 4 paramètres pour chacun des 8 capteurs composant le système multicapteurs) avec $\epsilon = 16$ correspondant au nombre de chiffres significatifs des nombres codés en virgule flottante double précision [ANSI/IEEE, 2008]. Les différentes valeurs des paramètres des métaheuristiques ont été déterminées empiriquement. La température T de SA diminuait de 1% toutes les 100 évaluations de la fonction objectif, les paramètres λ et μ de SA valaient respectivement 100 et 40, le nombre de particules de PSO était de 100, les coefficients d'inertie, d'accélération cognitive et d'accélération sociale étaient respectivement de $0,729$, $1,1494$ et $1,1494$ et une topologie de voisinage en anneau a été utilisée. Si au cours du processus d'optimisation, une solution candidate ne satisfait pas les contraintes, celle-ci est projetée dans l'espace réalisable. Les constantes de temps sont projetées en leur affectant la valeur possible la plus proche tandis que les amplitudes sont projetées en les multipliant par une constante α_i

$$\alpha_i = \frac{F_{S,i}}{\frac{A_{m,i}}{1-T_{m,i}} + \frac{A_{v,i}}{1-T_{v,i}}}, \quad i \in [\![1;\ N_s]\!]$$

de telle sorte que

$$\frac{\alpha_i A_{m,i}}{1-T_{m,i}} + \frac{\alpha_i A_{v,i}}{1-T_{v,i}} = \alpha_i\Big(\frac{A_{m,i}}{1-T_{m,i}} + \frac{A_{v,i}}{1-T_{v,i}}\Big) = F_{S,i}.$$

À la fin du processus d'optimisation, les paramètres K_m, K_v, τ_m, τ_v sont obtenus en utilisant les équations (4.4) et (4.5).

La figure 4.2 montre l'évolution de la valeur de la fonction objectif de la meilleure solution trouvée pour 10 répétitions du processus d'optimisation sur un exemple de la base constituée des toxiques chimiques (voir section 1.2.2). Pour cette application, cette figure montre que PSO est non seulement la méthode la plus précise, mais également la plus reproductible. Des résultats plus quantitatifs, et surtout obtenus sur les bases de données complètes sont donnés dans la table 4.1. Ces résultats mettent en évidence que PSO est la méthode la plus précise sur les trois jeux de données utilisés et la plus reproductible sur les deux premiers, tandis que SA obtient de meilleurs résultats, qui restent

TABLE 4.1 – Performances des métaheuristiques.

	Précision	Variance	
SA	1,4%	K_i	64
		τ_i	0,14
$(\lambda + \mu)$-**ES**	16,3%	K_i	51
		τ_i	0,09
PSO	82,3%	K_i	49
		τ_i	0,08

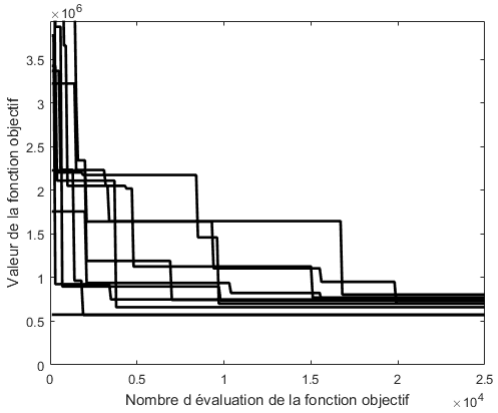
(a) Performances des métaheuristiques sur la base de données constituée des toxiques chimiques.

	Précision	Variance	
SA	4,4%	K_i	164
		τ_i	0,12
$(\lambda + \mu)$-**ES**	38,3%	K_i	150
		τ_i	0,07
PSO	57,3%	K_i	82
		τ_i	0,04

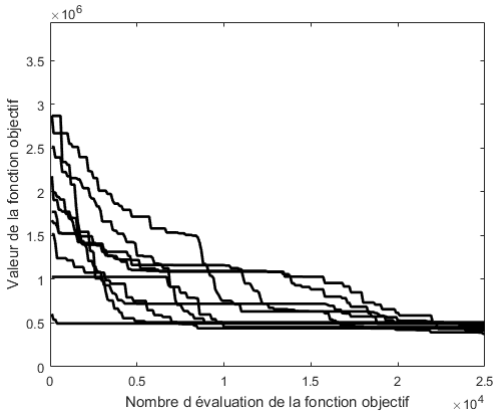
(b) Performances des métaheuristiques sur la base de données constituée des capsules de café.

	Précision	Variance	
SA	2,4%	K_i	123
		τ_i	0,17
$(\lambda + \mu)$-**ES**	36,6%	K_i	103
		τ_i	0,12
PSO	61,0%	K_i	105
		τ_i	0,14

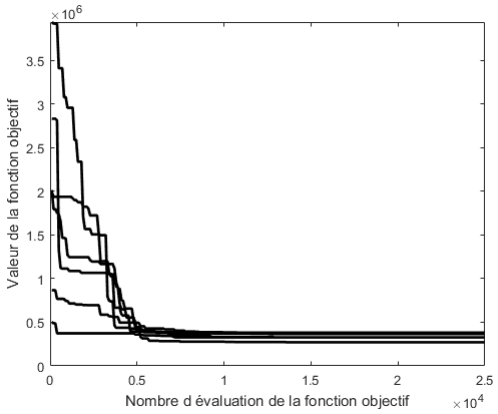
(c) Performances des métaheuristiques sur la base de données constituée du DMMP et du 4-NT.



(a) Recuit simulé



(b) Stratégie d'évolution $(\lambda + \mu)$



(c) Optimisation par essaim particulaire

FIGURE 4.2 – Évolution de la valeur de la fonction objectif de la meilleure solution trouvée.

néanmoins très proches de ceux de PSO en termes de reproductibilité sur la base de données constituée du DMMP et du 4-NT.

Sur la base de cette étude, nous choisissons PSO pour estimer les différents paramètres. Nous nous intéressons, dans le paragraphe suivant, au choix du nombre de particules et d'itérations.

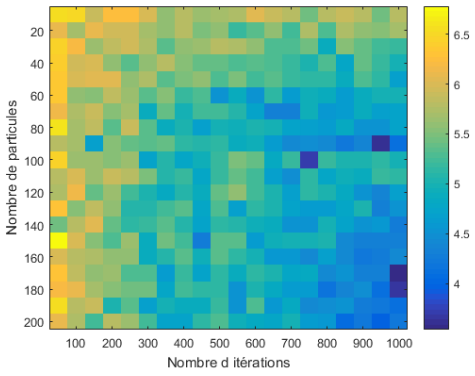
Choix du nombre de particules et d'itérations de PSO

Pour déterminer le nombre de particules et d'itérations de PSO, nous proposons de mesurer ses performances, à la fois en termes de précision et de reproductibilité, pour différents jeux de ces paramètres. La figure 4.3 présente les résultats expérimentaux obtenus et montre que le choix de 100 particules et de 500 itérations est un bon compromis en termes de précision, de reproductibilité et de temps de calcul. Le temps de calcul est d'environ 2 secondes pour une implémentation parallèle en C# sur un processeur Intel i7-4600U.

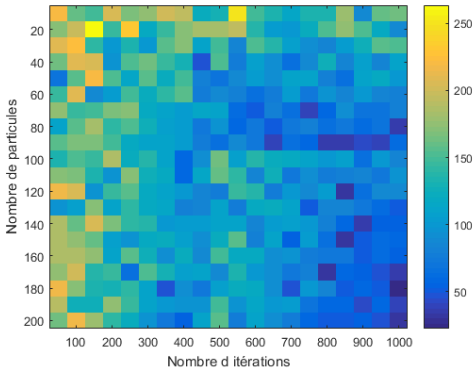
On remarque également, en s'intéressant aux constantes de temps, qu'une variance proche de $0,1$ correspondant à un écart-type proche de $0,3$ sur des variables d'amplitude proche de 1 correspond à un écart relatif de l'ordre de $0,3$. Ceci signifie qu'une erreur importante est potentiellement faite sur l'estimation de la valeur des constantes de temps. Concernant les gains, un écart-type de 10 sur des variables dont l'amplitude est de l'ordre de 100 correspond à un écart relatif de l'ordre de 10%. Ainsi, l'erreur potentiellement commise sur les amplitudes est bien moins importante que celle qui peut être faite sur les constantes de temps.

Bilan du chapitre

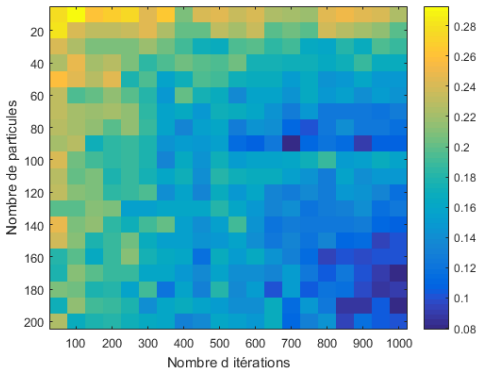
Nous avons décrit dans ce chapitre une méthode aveugle permettant d'estimer la valeur des paramètres des équations différentielles des contributions massique et viscoélastique. La méthode proposée repose sur la formulation d'un problème d'optimisation indépendant du profil de concentration c qui nous est inconnu. Ce problème d'optimisation ne pouvant être résolu de manière analytique, nous avons comparé à la fois en termes de précision et de reproductibilité, les performances du recuit simulé, de la stratégie d'évolution $(\lambda + \mu)$ et de l'optimisation par essaim particulaire. Les résultats expérimentaux ont montré que PSO est l'algorithme le plus approprié pour résoudre ce problème et ainsi estimer les paramètres des deux contributions.



(a) Précision des résultats (logarithmique de la valeur de la fonction objectif à la fin du processus d'optimisation).



(b) Reproductibilité des résultats pour A_m et A_v (variance).



(c) Reproductibilité des résultats pour T_m et T_v (variance).

FIGURE 4.3 – Performances de PSO pour différentes valeurs du nombre de particules et d'itérations. Les plus petites valeurs correspondent aux meilleurs résultats.

Chapitre 5

Application à l'identification de signatures chimiques et à l'estimation de leur concentration

Introduction

Dans ce chapitre, nous supposons que les paramètres K_m, K_v, τ_m et τ_v des contributions massique et viscoélastique sont estimés en utilisant la méthode proposée dans le chapitre 4. Nous décrivons dans un premier temps comment ces paramètres peuvent, en utilisant des techniques d'apprentissage supervisé, être utilisés pour identifier des signatures chimiques. Nous proposons ensuite une approche basée sur le principe de déconvolution et sur des techniques de régression non paramétrique afin de déterminer le profil de concentration.

5.1 Application à l'identification de composés chimiques

Dans cette section, nous mettons expérimentalement en évidence que les paramètres des contributions massique et viscoélastique peuvent être employés pour identifier des signatures chimiques. Nous donnons les résultats expérimentaux obtenus après un processus de validation croisée à 5 plis, en utilisant les paramètres

$$K_{i,m} \text{ et } K_{i,v} \text{ pour } i \in [\![1;\ N_s]\!]$$

de chaque capteur et les amplitudes en régime stationnaire

$$F_{i,s} \text{ pour } i \in [\![1;\ N_s]\!]$$

de chaque capteur comme descripteurs. N_s correspond au nombre de capteurs du réseau. Les constantes de temps n'ont pas été retenues car l'erreur potentiellement commise sur ces dernières est plus importante. Les tables 5.1, 5.2 et 5.3 représentent respectivement les résultats moyens obtenus après un processus de validation croisée à 5 plis sur les bases de données composées des toxiques chimiques (voir section 1.2.2), des capsules de café (voir section 1.2.3) et du DMMP et du 4-NT (voir section 1.2.4). Le meilleur résultat obtenu pour chaque couple technique de prétraitement - algorithme d'apprentissage est indiqué en gras.

Ces tables montrent que l'utilisation des paramètres des équations différentielles permet d'obtenir des résultats meilleurs de près de 2% sur les bases constituées des toxiques chimiques et du DMMP et du 4-NT et de près de 5% sur celle composée des capsules de café. Hormis sur la base de données constituée des toxiques chimiques, ces nouveaux descripteurs permettent également d'obtenir des écarts-types comparables.

TABLE 5.1 – Comparaison des performances moyennes obtenues en considérant les paramètres K_m et K_v comme descripteurs ainsi que les amplitudes en régime stationnaire sur la base composée des toxiques chimiques.

	K_m et K_v			Amplitude		
Prétraitement	SVM	LMNN	BT	SVM	LMNN	BT
Aucun	**0,939**	0,934	0,935	0,921	0,913	0,906
PCA	0,944	**0,958**	0,935	0,927	0,930	0,913
AE	0,937	**0,953**	0,947	0,920	0,917	0,935

(a) Taux de classification

	K_m et K_v			Amplitude		
Prétraitement	SVM	LMNN	BT	SVM	LMNN	BT
Aucun	0,009	0,011	0,021	0,010	0,09	**0,007**
PCA	0,051	0,048	0,028	0,032	0,028	**0,019**
AE	0,023	0,029	0,031	0,017	**0,015**	0,021

(b) Écarts-types

TABLE 5.2 – Comparaison des performances moyennes obtenues sur la base composée des capsules de café.

	K_m et K_v			Amplitude		
Prétraitement	SVM	LMNN	BT	SVM	LMNN	BT
Aucun	0,632	**0,651**	0,653	0,600	0,615	0,556
PCA	0,638	**0,698**	0,618	0,619	0,626	0,558
AE	0,613	**0,627**	0,568	0,620	0,621	0,556

(a) Taux de classification

	K_m et K_v			Amplitude		
Prétraitement	SVM	LMNN	BT	SVM	LMNN	BT
Aucun	0,073	0,071	**0,069**	0,091	0,099	0,076
PCA	0,084	0,079	0,091	0,091	0,087	**0,066**
AE	**0,071**	0,074	**0,071**	0,087	0,082	0,073

(b) Écarts-types

TABLE 5.3 – Comparaison des performances moyennes obtenues sur la base composée du DMMP et du 4-NT.

	K_m et K_v			Amplitude		
Prétraitement	SVM	LMNN	BT	SVM	LMNN	BT
Aucun	0,948	**0,966**	0,914	0,931	0,942	0,931
PCA	**0,966**	0,933	0,942	0,966	0,962	0,931
AE	**0,983**	0,942	0,933	0,966	0,931	0,966

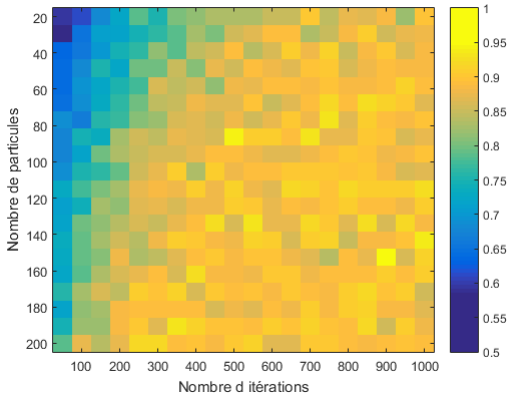
(a) Taux de classification

	K_m et K_v			Amplitude		
Prétraitement	SVM	LMNN	BT	SVM	LMNN	BT
Aucun	0,049	**0,041**	0,053	0,060	0,066	0,049
PCA	0,029	0,048	0,037	0,032	0,029	**0,021**
AE	**0,011**	0,038	0,033	0,026	0,028	0,026

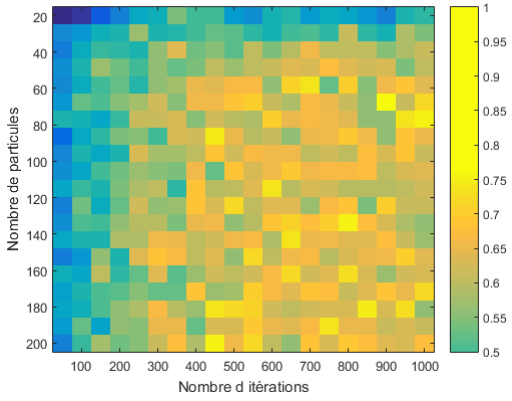
(b) Écarts-types

5.1.1 Impact des performances du processus d'optimisation sur le taux de classification

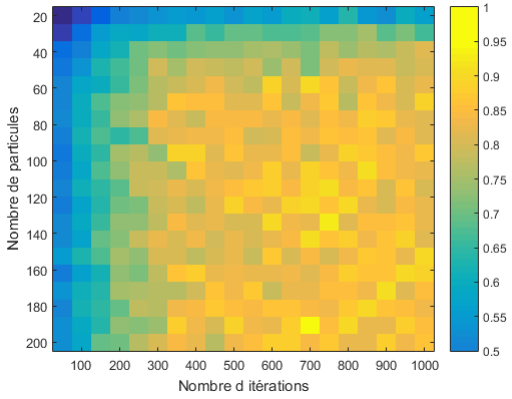
Dans cette section, nous proposons de mesurer l'impact du processus d'optimisation décrit dans la section 4.3 sur les performances en classification. La figure 5.1 montre l'évolution du taux moyen de classification, obtenu après un processus de validation croisée à 5 plis, en fonction de ces deux paramètres.



(a) Performances sur la base de données constituée des toxiques chimiques.



(b) Performances sur la base de données constituée des capsules de café.



(c) Performances sur la base de données constituée du DMMP et du 4-NT.

FIGURE 5.1 – Performances de PSO obtenues sans aucun prétraitement et en utilisant LMNN pour différentes valeurs du nombre de particules et d'itérations.

Ces courbes ont une forme proche de celles de la figure 4.3. Ceci montre que les performances en classification sont corrélées à la précision du processus d'optimisation et dépendent donc de celui-ci.

Il serait intéressant d'aller plus loin dans l'étude de l'impact du processus d'optimisation sur les performances en classification et notamment d'étudier l'effet de la précision, en termes de distance avec la solution optimale, de l'estimation des paramètres des deux contributions. Néanmoins, le fait que les véritables valeurs de ces paramètres ne soient pas connues est un obstacle à la mise en place d'études supplémentaires.

5.1.2 Fusion des descripteurs

Une analyse des exemples mal classés montre, comme l'illustre la table 5.4, que quelle que soit la base considérée, ceux-ci ne sont pas les mêmes suivant que les coefficients K_m et K_v ou que les amplitudes en régime stationnaire aient été utilisés pour entraîner les algorithmes d'apprentissage supervisé.

TABLE 5.4 – Matrices de confusion obtenues, après un processus de validation croisée à 5 plis, sans prétraitement avec LMNN en considérant respectivement les amplitudes et les gains comme descripteur.

		Classe prédite				
		H_2S	CH_3OH	NH_3	SO_2	C_7H_8
Classe réelle	H_2S	**0,900**	0,000	**0,100**	0,000	0,000
	CH_3OH	0,000	**0,920**	0,000	0,000	**0,080**
	NH_3	0,000	0,000	**0,994**	0,000	**0,006**
	SO_2	0,000	**0,093**	**0,047**	**0,767**	**0,093**
	C_7H_8	0,000	0,000	0,000	0,000	**1,000**

(a) Matrice de confusion obtenue en utilisant les amplitudes

		Classe prédite				
		H_2S	CH_3OH	NH_3	SO_2	C_7H_8
Classe réelle	H_2S	**1,000**	0,000	0,000	0,000	0,000
	CH_3OH	0,000	**0,993**	**0,007**	0,000	0,000
	NH_3	0,000	0,000	**1,000**	0,000	0,000
	SO_2	0,000	0,070	0,000	**0,967**	**0,027**
	C_7H_8	0,000	0,000	0,000	**0,036**	**0,0964**

(b) Matrice de confusion obtenue en utilisant les gains

Tenter d'expliquer ce phénomène requerrait une analyse plus approfondie des phénomènes physiques qui ne rentre pas dans le cadre de cette étude.

Néanmoins, cette observation motive le fait de considérer comme descripteur l'union de ces deux ensembles :

$$K_{i,m} \text{ et } K_{i,v} \text{ et } F_{i,s} \text{ pour } i \in [\![1;\ N_s]\!],$$

et de sélectionner les variables parmi celles-ci menant aux meilleurs résultats. Bien qu'il soit envisageable de tester les 2^{3N_s} possibilités pour de petites valeurs de N_s, une telle approche devient rapidement impraticable. Par exemple pour $N_s = 8$, il existe plus de 16 millions de sous-ensembles possibles. C'est pourquoi de nombreuses heuristiques ont été développées. Parmi elles, on peut notamment citer : les algorithmes génétiques [Leardi et Gonzales, 1998] et les techniques basées sur un diagramme de Hasse qui sont généralement préférées aux algorithmes génétiques du fait qu'elles sont déterministes
[Derksen et Keselman, 1992].

Un diagramme de Hasse est un graphe acyclique dirigé représentant la structure d'un ensemble. Chaque sous-ensemble de l'ensemble de référence est représenté par un nœud. Un arc relie deux nœuds si l'un est inclus dans l'autre et que la différence de cardinal des ensembles qu'ils représentent vaut 1. La figure 5.2 illustre le diagramme de Hasse associé à un ensemble $\{V_1, \ldots, V_4\}$.

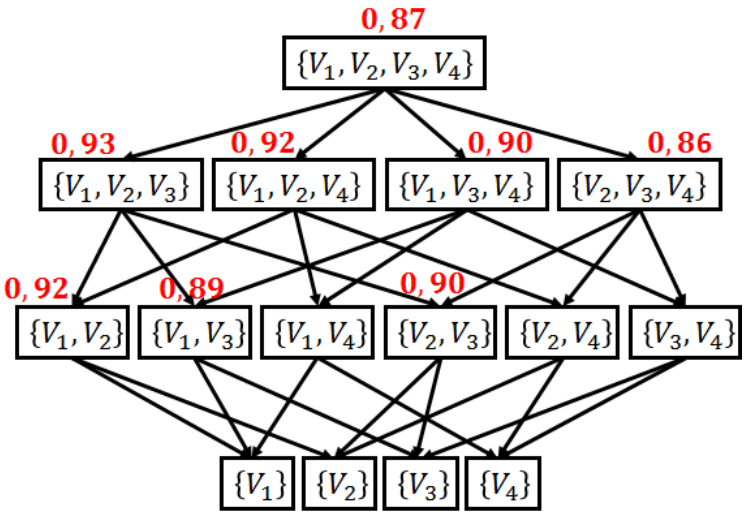


FIGURE 5.2 – Diagramme de Hasse pour la sélection de variables.

L'heuristique décrite dans [Derksen et Keselman, 1992] consiste en l'évaluation du taux d'exemples correctement classés sur la racine et de manière récursive sur les enfants du nœud ayant produit le meilleur taux. Ce processus récursif est arrêté lorsqu'aucun enfant n'a permis d'obtenir de meilleures performances que son parent. Sur l'exemple de la figure 5.2, les performances de la racine et de ses enfants sont d'abord évaluées (les performances sont en rouge) ; comme le sous-ensemble $\{V_1, V_2, V_3\}$ a permis d'obtenir de meilleures performances que son parent, alors ses enfants sont à leur tour évalués. Puisqu'aucun d'entre eux n'améliore le résultat, l'algorithme est arrêté et l'ensemble $\{V_1, V_2, V_3\}$ est renvoyé. Bien que cette méthode n'ait aucune garantie de trouver le sous-ensemble optimal, cette heuristique est déterministe (si la méthode d'apprentissage supervisé l'est également) et requiert l'évaluation, dans le pire des cas, de

$$1 + \sum_{i=2}^{N} \mathrm{C}_i^{i-1} = 1 + \sum_{i=2}^{N} i = \frac{N(N+1)}{2} \text{ sous-ensembles.}$$

Dans notre cas, avec $N_s = 8$, cette heuristique requerra au plus l'évaluation de 36 de sous-ensembles. Les tables 5.5, 5.6 et 5.7 représentent respectivement les résultats moyens obtenus, après un processus de validation croisée à 5 plis, sur les bases de données composées des toxiques chimiques, des capsules de café et du DMMP et du 4-NT, en considérant la fusion des deux descripteurs et en ayant utilisé l'heuristique de Hasse pour sélectionner les variables les plus appropriées. Le meilleur résultat est indiqué en gras.

TABLE 5.5 – Performances moyennes obtenues après sélection des descripteurs par l'heuristique de Hasse sur la base de données constituée des toxiques chimiques lors d'un processus de validation croisée à 5 plis.

	K_m, K_v et F_s		
Prétraitement	SVM	LMNN	BT
Aucun	0,963	0,962	0,959
PCA	0,969	**0,975**	0,955
AE	0,933	0,943	0,941

(a) Taux de classification

	K_m,K_v et F_s		
Prétraitement	SVM	LMNN	BT
Aucun	0,015	0,013	0,011
PCA	0,015	0,017	**0,009**
AE	0,017	0,011	0,012

(b) Écarts-Types

TABLE 5.6 – Performances moyennes obtenues après sélection des descripteurs par l'heuristique de Hasse sur la base de données constituée des capsules de café lors d'un processus de validation croisée à 5 plis.

	K_m, K_v et F_s		
Prétraitement	SVM	LMNN	BT
Aucun	0,695	0,721	0,704
PCA	0,742	0,748	0,732
AE	0,736	**0,753**	0,717

(a) Taux de classification

	K_m,K_v et F_s		
Prétraitement	SVM	LMNN	BT
Aucun	0,064	0,059	0.045
PCA	0,059	0,051	**0,048**
AE	0,071	0,061	0,051

(b) Écarts-Types

TABLE 5.7 – Performances moyennes obtenues après sélection des descripteurs par l'heuristique de Hasse sur la base de données constituée du DMMP et du 4-NT lors d'un processus de validation croisée à 5 plis.

	K_m, K_v et F_s		
Prétraitement	SVM	LMNN	BT
Aucun	0,981	0,971	0,971
PCA	0,978	0,957	0,934
AE	0,981	**0,984**	0,933

(a) Taux de classification

	K_m,K_v et F_s		
Prétraitement	SVM	LMNN	BT
Aucun	0,011	0,009	**0,006**
PCA	0,013	0,013	0,011
AE	0,011	0,009	0,019

(b) Écarts-Types

Ces résultats sont révélateurs de l'intérêt de fusionner les deux types de descripteurs et de sélectionner les variables les plus appropriées. Plus quantitativement, cette approche permet, tout en réduisant l'écart-type, d'améliorer les résultats de près de 5% sur les bases constituées des toxiques chimiques et

du DMMP et du 4-NT par rapport à ceux obtenus en utilisant les techniques de l'état de l'art et de près de 7% sur la base constituée des capsules de café.

5.2 Application à l'estimation du profil de concentration

5.2.1 Déconvolution

L'estimation des paramètres K_m, K_v, τ_m et τ_v permet de calculer la série temporelle correspondant à la réponse impulsionnelle des capteurs en utilisant l'équation (4.3). Étant donné que la réponse F_i d'un capteur et que sa réponse impulsionnelle h_i sont connues et que l'on a pour $i \in [\![1;\ N_s]\!]$

$$F_i = h_i * c^*,$$

il est possible de déterminer la série temporelle c^* correspondant au profil de concentration unitaire en effectuant une opération de déconvolution. Remarquons tout d'abord que

$$F_i[n] = h_i[0]c^*[n] + h_i[1]c^*[n-1] + ... + h_i[N]c^*[n-N].$$

Cette équation peut être écrite pour chaque capteur sous forme matricielle :

$$F_i = \mathbf{H_i} c^*$$

où

$$\mathbf{H_i} = \begin{bmatrix} h_i[0] & 0 & \cdots & \cdots & \cdots & 0 \\ h_i[1] & h_i[0] & 0 & & & \vdots \\ h_i[2] & h_i[1] & h_i[0] & 0 & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & 0 \\ h_i[N] & \dots & \dots & \dots & h_i[1] & h_i[0] \end{bmatrix}.$$

L'empilement de ces équations pour chaque capteur donne :

$$\begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_{N_s} \end{bmatrix} = \begin{bmatrix} \mathbf{H_1} \\ \mathbf{H_2} \\ \vdots \\ \mathbf{H_M} \end{bmatrix} c^* \text{ ou } F = \mathbf{H} c^*$$

et

$$c^* = (\mathbf{H^T H})^{-1} \mathbf{H^T} F.$$

Cependant, comme les réponses des capteurs F sont bruitées, l'utilisation de cette relation pour déterminer c^* produit également des résultats bruités. Pour augmenter le rapport signal sur bruit de c^*, nous proposons d'ajouter un terme de régularisation consistant à minimiser l'amplitude des dérivées d'ordre 2. Dans certaines circonstances, il peut également être intéressant d'ajouter des connaissances *a priori* sur le profil de concentration, par exemple $c^*[0] = 0$. Ainsi, une approche préférable pour estimer c^* consiste à résoudre le problème d'optimisation :

$$\begin{cases} \text{argmin} & ||F - \mathbf{H}c^*||^2 + k||\mathbf{D}c^*||^2 \\ \text{s. c. :} & \\ & sc^* = d \end{cases} \tag{5.1}$$

où $k > 0$ est le coefficient de régularisation, s est un vecteur permettant d'exprimer les connaissances *a priori*. Par exemple, si M désigne le nombre d'échantillons disponibles des séries F_i, pour $c^*[0] = 0$ on a

$$s = [1\ 0\ \ldots\ 0]_{1\times M} \text{ et } d = 0.$$

$\mathbf{D}$ est quant à elle la matrice de différenciation d'ordre 2. Cette matrice est la matrice triangulaire supérieure de Toeplitz générée par le vecteur

$$[1\ -2\ 1\ 0\ \ldots\ 0]_{1\times M}.$$

Ce problème d'optimisation peut être résolu de manière analytique en utilisant la méthode des multiplicateurs de Lagrange [Rockafellar, 2006]. Le Lagrangien est :

$$L(c^*, \lambda) = ||F - \mathbf{H}c^*||^2 + k||\mathbf{D}c^*||^2 + \lambda(sc^* - d)$$

$$= F^{\mathbf{T}}F - 2c^{*\mathbf{T}}\mathbf{H}^{\mathbf{T}}F + c^{*\mathbf{T}}\Big(\mathbf{H}^{\mathbf{T}}\mathbf{H} + k\mathbf{D}^{\mathbf{T}}\mathbf{D}\Big)c^* + \lambda(sc^* - d)$$

où λ est un multiplicateur de Lagrange. Le calcul de son gradient donne [Peterson et Pederson, 2012] :

$$\begin{cases} \frac{\partial L(c^*,\lambda)}{\partial c^*} = -2\mathbf{H}^{\mathbf{T}}F + 2\Big(\mathbf{H}^{\mathbf{T}}\mathbf{H} + k\mathbf{D}^{\mathbf{T}}\mathbf{D}\Big)c^* + \lambda s^{\mathbf{T}} \\ \frac{\partial L(c^*,\lambda)}{\partial \lambda} = sc^* - d \end{cases}$$

et ce gradient s'annule pour :

$$\begin{bmatrix} c^* \\ \lambda \end{bmatrix} = \begin{bmatrix} 2\mathbf{H}^{\mathbf{T}}\mathbf{H} + 2k\mathbf{D}^{\mathbf{T}}\mathbf{D} & s^T \\ s & 0 \end{bmatrix}^{-1} \begin{bmatrix} 2\mathbf{H}^{\mathbf{T}}F \\ d \end{bmatrix}. \tag{5.2}$$

Aspects algorithmiques

L'inconvénient de l'équation (5.2) réside dans le fait qu'elle nécessite de calculer la matrice de Gram $2\mathbf{H^T H} + 2k\mathbf{D^T D}$. Or, de manière générale de telles matrices ont souvent un important coefficient de conditionnement (*condition number*). Ce coefficient est défini comme étant le rapport de la plus grande valeur propre sur la plus petite valeur propre d'une matrice. Une règle empirique affirme que la précision perdue lors de l'inversion d'une matrice dont le coefficient de conditionnement est 10^k est de k décimales [Higham, 1996]. Sur les exemples de la base de données des toxiques chimiques, ce coefficient peut prendre des valeurs jusqu'à plus de 10^{14}. Ainsi, si les calculs sont effectués en utilisant des nombres à virgule flottante double précision, la formule (5.2) ne permet pas d'obtenir plus de 1 décimale, et ainsi les concentrations les plus faibles ne peuvent pas être estimées avec moins de 10% d'erreur.

Si M désigne le nombre d'échantillons disponibles des séries F_i et N_c représente le nombre de contraintes, alors nous pouvons remédier à ce problème en posant :

$$\mathbf{A} = \begin{bmatrix} \sqrt{2}\mathbf{H} \\ \sqrt{2k}\mathbf{D} \end{bmatrix} \text{ et } \mathbf{b} = \begin{bmatrix} F \\ 0_{M\times 1} \end{bmatrix},$$

de telle sorte que :

$$\mathbf{A^T A} = 2\mathbf{H^T H} + 2k\mathbf{D^T D}.$$

Une décomposition QR [Golub, 1983] donne :

$$s^T = \mathbf{Q} \begin{bmatrix} \mathbf{R_{N_c \times N_c}} \\ \mathbf{0_{(M-N_c)\times N_c}} \end{bmatrix}. \tag{5.3}$$

Posons :

$$\mathbf{AQ} = \begin{bmatrix} \mathbf{A_1}_{(M+N_c)\times N_c} & \mathbf{A_2}_{(M+N_c)\times(M-N_c)} \end{bmatrix} \text{ et } \mathbf{Q^T} c^* = \begin{bmatrix} y_{N_c\times 1} \\ z_{(M-N_c)\times 1} \end{bmatrix}. \tag{5.4}$$

L'équation (5.3) donne

$$s = \begin{bmatrix} \mathbf{R^T} \\ \mathbf{0} \end{bmatrix} \mathbf{Q^T} \text{ d'où } sc^* = \begin{bmatrix} \mathbf{R^T} \\ \mathbf{0} \end{bmatrix} \mathbf{Q^T} c^*,$$

la substitution de l'équation (5.4) mène à

$$sc^* = \begin{bmatrix} \mathbf{R^T} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \mathbf{R^T} y. \tag{5.5}$$

Par ailleurs, nous avons

$$\mathbf{AQQ^T} c^* = \mathbf{A_1} y + \mathbf{A_2} z. \tag{5.6}$$

La substitution des équations (5.5) et (5.6) dans le problème d'optimisation (5.1) donne :

$$\begin{cases} \text{argmin} & ||F - \mathbf{A_1}y + \mathbf{A_2}z||^2 \\ \text{sous la contrainte :} & \\ & \mathbf{R^T}y = d \end{cases}$$

Ainsi, y est déterminé à partir de $\mathbf{R^T}y = d$: $y = \mathbf{R^{T^{-1}}}d$ et z est obtenu en résolvant le problème d'optimisation :

$$z = argmin||\mathbf{A_2}z - (\mathbf{A_1}y - F)||^2$$

soit

$$z = \mathbf{A_2}^{\dagger}(\mathbf{A_1}y - F)$$

où $\mathbf{A_2}^{\dagger}$ représente la pseudo-inverse de la matrice $\mathbf{A_2}$. Pour finir, y et z étant connus, on a

$$c^* = \mathbf{Q}\begin{bmatrix} y \\ z \end{bmatrix}.$$

Cette méthode, bien que plus complexe que le calcul direct, ne nécessite pas le calcul de la matrice de Gram et est moins sujette aux problèmes de stabilité numérique.

Influence du coefficient de régularisation k

La figure 5.3 montre l'estimation du profil de concentration c^* lorsque les capteurs sont exposés à 10 ppm d'ammoniac pour plusieurs valeurs du coefficient de régularisation k.

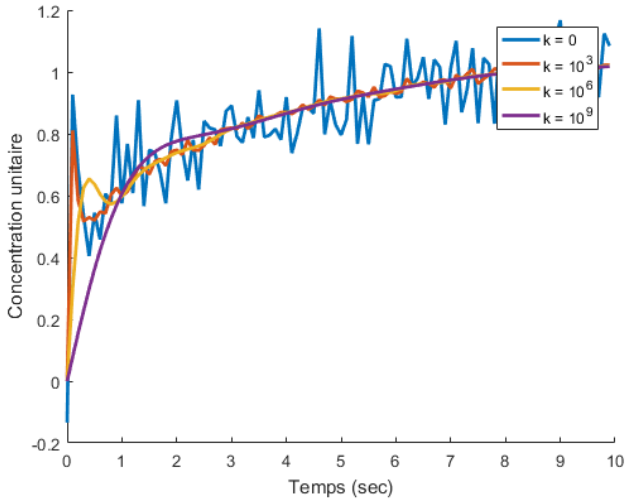


FIGURE 5.3 – Résultats de l'opération de déconvolution pour différentes valeurs de k lorsque les capteurs sont exposés à 10 ppm d'ammoniac.

Cette figure est révélatrice de l'influence du coefficient de régularisation k : une faible valeur de k donne un résultat bruité tandis qu'une grande valeur produit un résultat très lisse non représentatif de la réalité. Il est donc légitime de se poser la question du choix optimal de k. Dans le cadre plus général des problèmes des moindres carrés régularisés, l'auteur de [Hansen, 2000] tente d'y apporter une réponse. L'heuristique proposée, appelée méthode de la courbe L (*L curve*) consiste à tracer la courbe paramétrée par k donnée par

$$\begin{cases} x(k) &= ||F - \mathbf{H}c^*(k)|| \\ y(k) &= ||Dc^*(k)|| \end{cases}$$

et à déterminer la valeur optimale de k correspondant au coin de cette même courbe. Un exemple de la courbe L est donné par la figure 5.4.

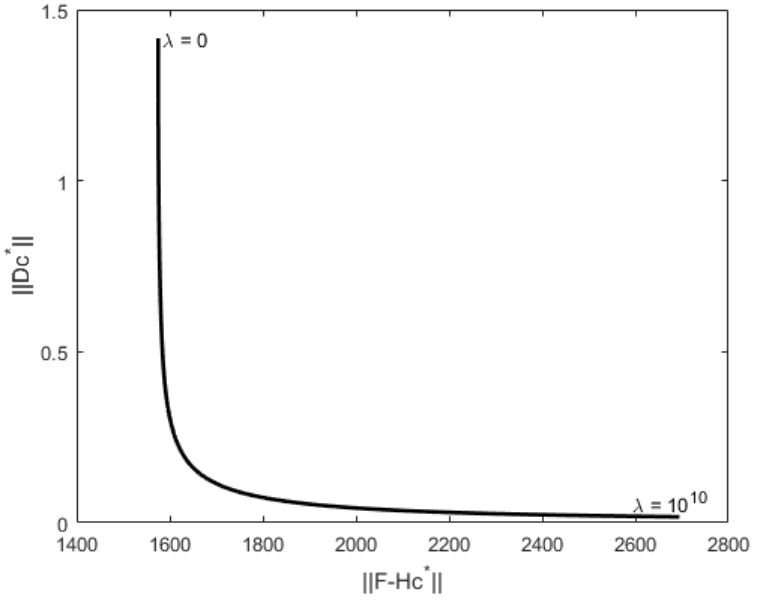


FIGURE 5.4 – *L curve* associée à l'exemple précédent.

Ce point est caractérisé comme étant celui où la courbure de cette courbe est maximale. La figure 5.5 représente la courbure de la courbe 5.4 obtenue en utilisant les équations (5.7).

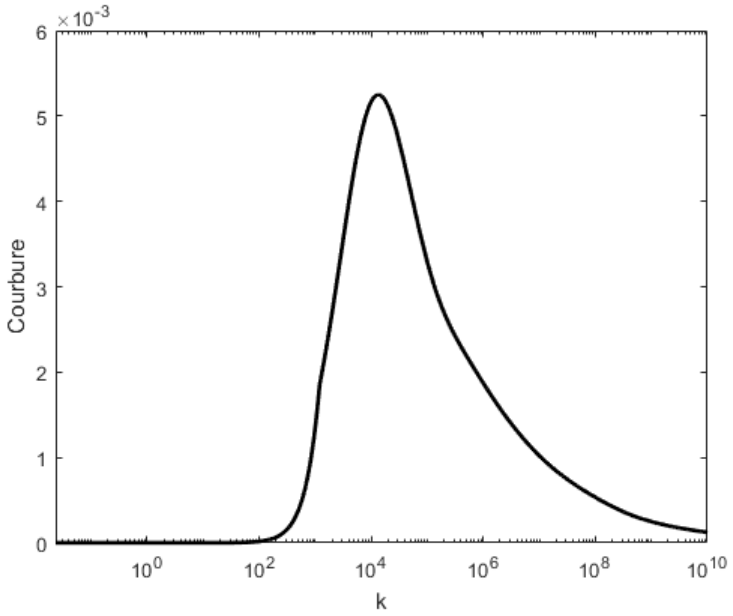


FIGURE 5.5 – Courbure de la figure 5.4.

La courbure, donnée par

$$\gamma(k) = \frac{x'(k)y''(k) - x''(k)y'(k)}{(x'(k)^2 + y'(k)^2)^{\frac{3}{2}}},$$

peut être numériquement estimée, comme le montre la figure 5.5, en prenant :

$$\begin{aligned} x'(k) &\approx \frac{x(k+\epsilon) - x(k)}{\epsilon}, \\ x''(k) &\approx \frac{x'(k+\epsilon) - x'(k)}{\epsilon}, \\ y'(k) &\approx \frac{y(k+\epsilon) - y(k)}{\epsilon}, \\ y''(k) &\approx \frac{y'(k+\epsilon) - y'(k)}{\epsilon}, \end{aligned} \tag{5.7}$$

pour de petites valeurs de ϵ. La valeur optimale de k peut donc être déterminée en maximisant $\gamma(k)$ à l'aide d'une montée de gradient :

$$\nabla\gamma \approx \frac{\gamma(k+\epsilon) - \gamma(k)}{\epsilon}.$$

La figure 5.6 montre le résultat de la méthode de déconvolution, sur le même exemple que la figure 5.3, en utilisant la valeur de k déterminée par la méthode de la courbe L.

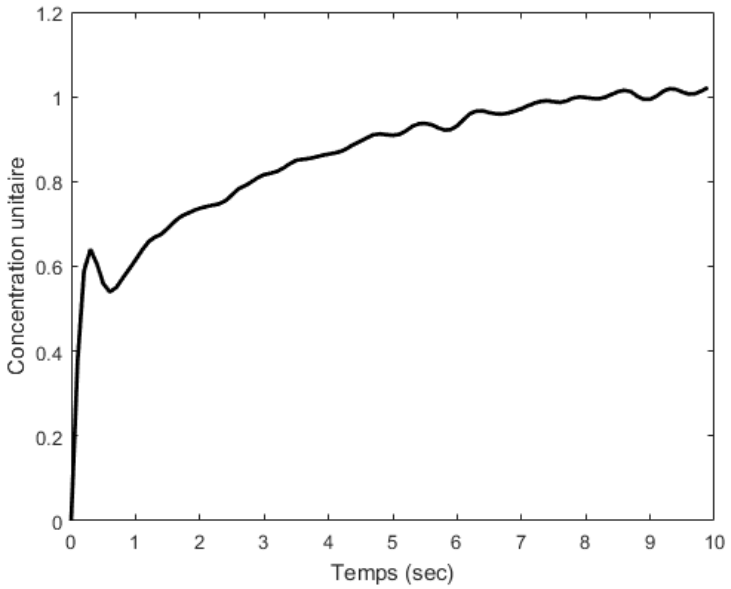


FIGURE 5.6 – Résultats de l'opération de déconvolution lorsque les capteurs sont exposés à 10 ppm d'ammoniac avec le coefficient de régularisation optimal $k = 11840$.

5.2.2 Techniques de régression non paramétrique

Le profil de concentration unitaire c^* étant connu, pour déterminer le véritable profil de concentration

$$c = \alpha c^*,$$

il est nécessaire de déterminer le coefficient α introduit dans l'équation (4.2) qui n'est autre que la concentration en régime stationnaire (puisque dans ce cas $c^* = 1$). Il s'agit d'un problème de régression consistant, à partir des paramètres K_m, K_v et F_s de chaque capteur, à construire une fonction f telle que

$$c = f(K_{i,m}, K_{i,v}, F_{i,s}) + \epsilon \text{ pour } i \in [\![1;\ N_s]\!]$$

où ϵ est un terme représentant une erreur. La fonction f nous étant inconnue, les techniques de régression non paramétrique sont les seules envisageables. Parmi elles, on peut notamment citer :

- Les réseaux de neurones, tels que décrits dans la section 2.3.2, n'ayant qu'un neurone de sortie correspondant à la valeur estimée de c.
- Les méthodes de régression à noyaux et notamment le modèle de Nadaraya-Waston [Bierens, 1994]. Ce modèle consiste, à partir de n exemples d'apprentissage (X_i, Y_i) pour $i \in [\![1;\ n]\!]$, à estimer f comme étant une moyenne

pondérée des exemples d'apprentissage :

$$f(x) = \frac{\sum_{i=1}^{n} K(X_i, x) Y_i}{\sum_{i=1}^{n} K(X_i, x)}$$

où K est une fonction noyau. Les noyaux couramment utilisés dans le cadre des problèmes de régression sont :

– les noyaux Gaussiens

$$K_h(y, x) = e^{-h||y-x||};$$

– les noyaux logistiques

$$K_h(y, x) = \frac{1}{2 + e^{-h||y-x||} + e^{h||y-x||}};$$

– les noyaux sigmoïdes

$$K_h(y, x) = \frac{1}{e^{-h||y-x||} + e^{h||y-x||}},$$

h étant un paramètre influençant la largeur des noyaux.

La table 5.8 compare les performances moyennes de ces différentes méthodes sur la base constituée des toxiques chimiques (1.2.2). Chaque résultat a été obtenu, après un processus de validation croisée à 5 plis, en effectuant une recherche sur une grille afin de déterminer la meilleure architecture du réseau ou le paramètre h optimal. La métrique utilisée pour juger de l'optimalité de ces paramètres n'est autre que l'erreur relative moyenne accumulée sur les N_t exemples de la base de test :

$$err = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(\frac{c_i - f(x_i)}{c_i} \right)^2.$$

Les descripteurs utilisés sont les paramètres K_m, K_v et F_s de chaque capteur. Une étape de sélection de variables, utilisant la méthode décrite dans la section 5.1.2, a également été effectuée de manière à trouver le sous-ensemble le plus approprié.

TABLE 5.8 – Performances moyennes des méthodes de régression non paramétrique sur la base de données constituée des toxiques chimiques.

		Erreur relative	Écart-type
Réseau de neurones		11,7%	6,2%
Régression à noyaux	Noyau Gaussien	9,3%	3,1%
	Noyau logistique	6,2%	3,4%
	Noyau sigmoïde	6,5%	3,4%

Ces résultats expérimentaux mettent en évidence l'intérêt des méthodes de régression à noyaux pour l'estimation de la concentration. Bien que les résultats dépendent de la fonction noyau choisie, ceux-ci sont supérieurs à ceux obtenus en utilisant un réseau de neurones tandis que leur écart-type est inférieur. La figure 5.7 illustre les résultats des étapes successives du processus d'estimation du profil de concentration :

a) Acquisition des signaux : les capteurs sont exposés à 10 ppm d'ammoniac les 10 premières secondes et à 6 ppm d'ammoniac les 10 dernières secondes.

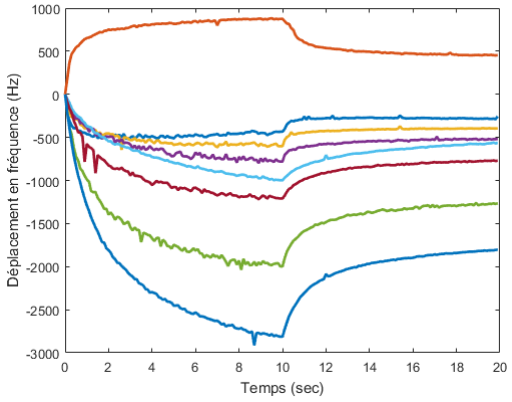
b) Estimation des réponses impulsionnelles en utilisant la méthode décrite dans le chapitre 4 et en n'utilisant que les 10 premières secondes des signaux.

c) Détermination du profil de concentration c en utilisant la méthode décrite dans la section 5.2.

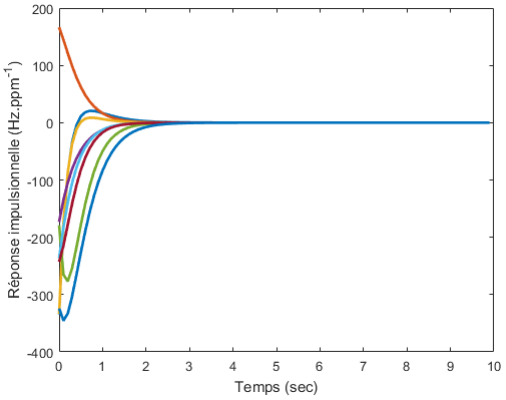
Bilan du chapitre

Dans ce chapitre, nous avons montré que les paramètres K_m, K_v, τ_m et τ_v des contributions massique et viscoélastique, estimés en utilisant la méthode proposée dans le chapitre 4, peuvent être appliqués à l'identification de signatures chimiques et à l'estimation de la concentration. Utilisés comme descripteurs, ceux-ci ont permis, sur les trois bases de données, d'obtenir des résultats meilleurs que ceux obtenus en utilisant l'amplitude en régime stationnaire des signaux seule. La fusion de ces deux types de descripteurs et la sélection des variables les plus appropriées grâce à une méthode basée sur un diagramme de Hasse a permis une augmentation plus significative de ces résultats.

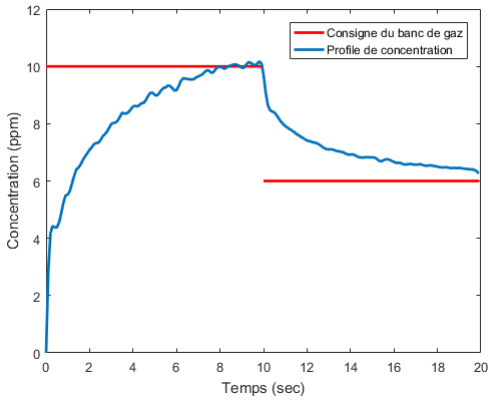
Plus quantitativement, l'approche proposée dans ce chapitre permet d'améliorer les résultats de près de 5% sur les bases constituées des toxiques chimiques et du DMMP et du 4-NT par rapport à ceux obtenus en utilisant les techniques de l'état de l'art et de près de 7% sur la base constituée des capsules de café. Ces paramètres permettent également, en utilisant des techniques de déconvolution et de régression non paramétrique, de déterminer, avec une erreur relative inférieure à 10%, le profil temporel de concentration.



(a) Acquisition des signaux.



(b) Estimation des réponses impulsionnelles.



(c) Détermination du profil de concentration.

FIGURE 5.7 – Processus d'estimation du profil de concentration.

Chapitre 6

Méthode de sélection des fonctionnalisations des capteurs

Introduction

Dans la section 1.1.3 (page 12), nous avons vu que l'approche multicapteurs a pour objectif d'augmenter la sélectivité du dispositif utilisé, et donc ses possibilités d'identification. Dans un contexte industriel, le coût ou le temps de fabrication de tels dispositifs pouvant être élevé, nous proposons dans cette section une méthode permettant de sélectionner les fonctionnalisations les plus appropriées pour identifier des composés chimiques sous contrainte de coût. Le nombre de capteurs utilisés dans les approches de type nez électrique pouvant potentiellement être élevé, rendant les recherches exhaustives impossibles, nous proposons dans ce chapitre un algorithme glouton interprétable, de complexité linéaire, permettant de répondre à ce problème.

6.1 Définition d'un critère de séparabilité

Les équations (4.1) montrent que, comme l'illustre la figure 6.1, lorsque le régime stationnaire est atteint, les amplitudes des contributions massique et viscoélastique sont des fonctions linéaires de la concentration. On a en effet :

$$\begin{cases} F_m = & K_m c \\ F_v = & K_v c \end{cases} .$$

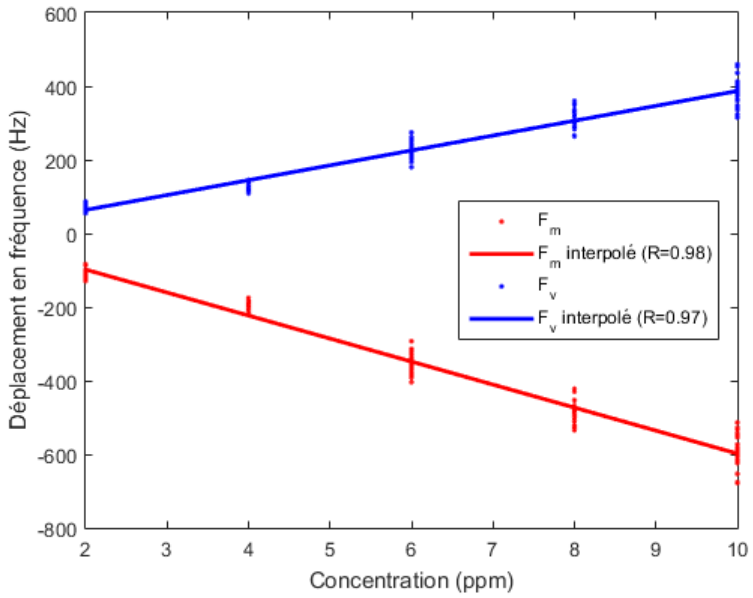


FIGURE 6.1 – Évolution des amplitudes des contributions massique et visco-élastique d'un capteur fonctionnalisé $(OH)_+$ et exposé à du toluène.

L'inconvénient de ce modèle théorique est qu'il ne prend pas en compte la variabilité des réponses des capteurs. Cette variabilité est également mise en évidence par la figure 6.1. Cette dernière montre que, pour une concentration donnée, les amplitudes des deux contributions ont une légère dispersion. Par exemple, pour un capteur fonctionnalisé $(OH)_+$ exposé à du toluène à 6 ppm la dispersion est de près de 100 Hz. Pour remédier à ce problème, nous proposons de modéliser l'amplitude des contributions en régime stationnaire en utilisant des intervalles :

$$\begin{cases} \beta_m c + \gamma_{m,-} \leq F_m \leq \beta_m c + \gamma_{m,+} \\ \beta_v c + \gamma_{v,-} \leq F_v \leq \beta_v c + \gamma_{v,+} \end{cases} . \tag{6.1}$$

Les paramètres β_m et β_v sont obtenus en effectuant une régression linéaire de degré 1 sur les couples (c, F_m) et (c, F_v), tandis que les paramètres

$$\gamma_{m,-}, \quad \gamma_{m,+} \text{ et } \gamma_{v,-}, \quad \gamma_{v,+}$$

sont choisis de telle sorte que tous les exemples vérifient les équations (6.1), hormis les potentielles valeurs aberrantes détectées en employant des méthodes d'identification d'*outliers* [Kriegel *et al.*, 2010]. La figure 6.1 illustre ce modèle.

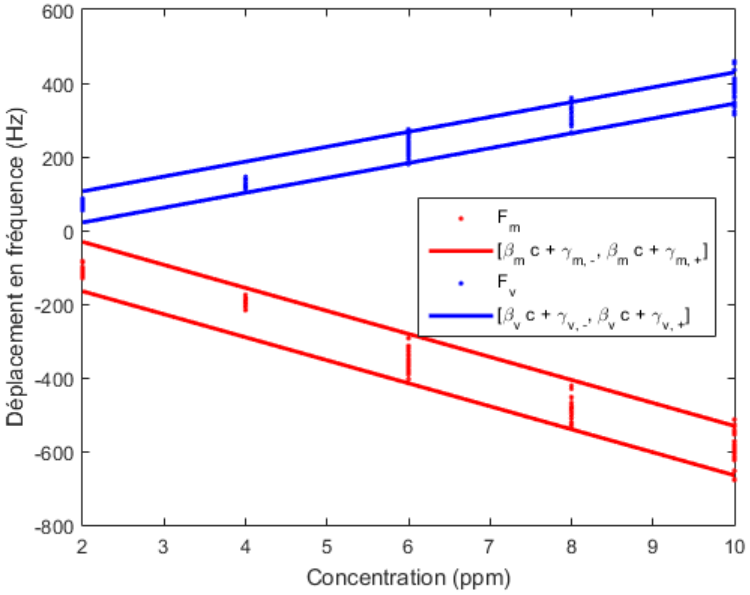


FIGURE 6.2 – Intervalles dans lesquels évoluent les amplitudes des contributions massique et viscoélastique d'un capteur fonctionnalisé $(OH)_+$ et exposé à du toluène.

Pour deux composés chimiques a et b, nous définissons la région de confusion massique comme étant l'intervalle

$$Conf_m = [min(c_1, c_2);\ max(c_1, c_2)]$$

où c_1 et c_2 sont respectivement les solutions positives, si elles existent, des équations

$$\beta_{m,a}c_1 + \gamma_{m,-,a} = \beta_{m,b}c_1 + \gamma_{m,+,b} \tag{6.2}$$

et

$$\beta_{m,b}c_2 + \gamma_{m,-,b} = \beta_{m,a}c_2 + \gamma_{m,+,a}. \tag{6.3}$$

De manière similaire, nous définissons également la région de confusion viscoélastique comme étant l'intervalle

$$Conf_v = [min(c_3, c_4);\ max(c_3, c_4)]$$

où c_3 et c_4 sont respectivement les solutions positives, si elles existent, des équations

$$\beta_{v,a}c_3 + \gamma_{v,-,a} = \beta_{v,b}c_3 + \gamma_{v,+,b} \tag{6.4}$$

et

$$\beta_{v,b}c_4 + \gamma_{v,-,b} = \beta_{v,a}c_4 + \gamma_{v,+,a}. \tag{6.5}$$

S'il n'existe pas de solutions positives à ces équations, nous avons alors $Conf_m = \emptyset$ et $Conf_v = \emptyset$. Ces deux régions peuvent être interprétées, comme l'illustre la figure 6.3, comme étant l'intervalle de concentration sur lequel les intervalles (6.1) s'intersectent. Puisque les composés chimiques a et b sont identifiables si l'une de leurs contributions n'est pas identique, on en déduit le critère de séparabilité :

$$Conf_m \cap Conf_v = \emptyset.$$

L'intervalle $Conf_m \cap Conf_v$ représente l'intervalle de concentration sur lequel les deux contributions sont identiques : il représente ainsi l'intervalle sur lequel les deux composés chimiques ne peuvent pas être identifiés. La figure 6.3 illustre les régions de confusion massique et viscoélastique associées aux réponses d'un capteur fonctionnalisé $(OH)_+$ exposé à du toluène (indice 1) et à du dioxyde de soufre (indice 2). Dans cet exemple, nous avons

$$Conf_m = [2,30;\ 3,53] \text{ et } Conf_v = [2,41;\ 3,68]$$

d'où $Conf_m \cap Conf_v = [2,41;\ 3,53]$. Ainsi un capteur fonctionnalisé $(OH)_+$ ne permet pas d'identifier du toluène et du dioxyde de soufre pour des concentrations dans l'intervalle $[2,41;\ 3,53]$. Dans la section suivante, nous montrerons comment utiliser cet intervalle pour sélectionner les fonctionnalisations les plus appropriées à un problème donné.

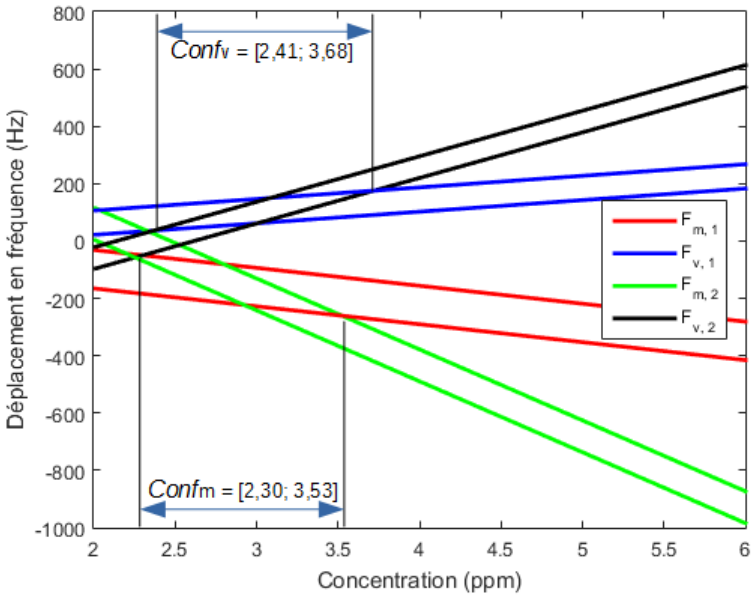


FIGURE 6.3 – Régions de confusion massique et viscoélastique associées aux réponses d'un capteur fonctionnalisé $(OH)_+$ exposé à du toluène (indice 1) et à du dioxyde de soufre (indice 2).

6.2 Proposition d'un algorithme glouton pour la sélection de capteurs

6.2.1 Formulation du problème sous la forme d'un problème d'optimisation

Dans cette section, nous formulons la question de la sélection des fonctionnalisations sous contrainte de coût sous la forme d'un problème d'optimisation en variable binaire. L'objectif consiste à déterminer les fonctionnalisations les plus appropriées pour identifier un ensemble G de composés chimiques parmi un ensemble F de N_f fonctionnalisations distinctes notées f_i pour $i \in [\![1;\ N_f]\!]$ sans connaître les taux de classification obtenables avec chacune d'entre elles. Soient les nombres c_i représentant les coûts des fonctionnalisations f_i et soit C_{max} le coût maximal autorisé. Nous supposerons également que le système employé peut contenir jusqu'à N_s capteurs et que, bien qu'il existe des systèmes où les fonctionnalisations sont doublées pour augmenter leur robustesse, les fonctionnalisations sont différentes deux à deux. Dans le cas contraire, il suffit de diviser le nombre de capteurs par 2. Pour indiquer si une fonctionnalisation est choisie ou non, nous utilisons un codage binaire : l'état de la $i^{ième}$ fonctionnalisation x_i vaut 1 si celle-ci est choisie ou 0 dans le cas contraire. Ainsi, les fonctionnalisations choisies sont complètement décrites par le vecteur $X = [x_1\ \ldots x_{N_f}]$ avec $\forall\ i \in [\![1;\ N_f]\!],\ x_i \in \{0, 1\}$.

Avec ces notations et ces définitions, la question de la sélection des fonctionnalisations sous contrainte de coût peut se formuler sous la forme du problème d'optimisation :

$$\begin{cases} \text{argmax } z(X) \\ \text{s. c. :} \\ \quad \sum_{i=1}^{N_f} x_i c_i \leq C_{max} \\ \quad \sum_{i=1}^{N_f} x_i \leq N_s \end{cases}$$

La fonction z est une fonction objectif permettant de mesurer les performances du vecteur X. Par exemple, z peut être le taux de classification obtenu en employant les méthodes décrites dans la section sur les fonctionnalisations correspondant à $x_i\ \forall\ i \in [\![1;\ N_f]\!]$.

6.2.2 Algorithme glouton

Dans cette section, nous proposons un algorithme glouton de complexité linéaire permettant de résoudre ce problème. La complexité linéaire est ici pri-

mordiale car l'évaluation d'un vecteur X nécessite d'estimer les paramètres des deux contributions, de sélectionner les variables les plus appropriées et d'entraîner, pour chaque ensemble de variables considéré, au moins 5 fois l'algorithme d'apprentissage, et peut ainsi prendre plusieurs heures. Une recherche exhaustive, nécessitant $\sum_{i=1}^{N_f} \mathrm{C}_i^{N_s}$ évaluations de la fonction objectif est donc inenvisageable.

Dans la suite de ce manuscrit, nous notons $Conf_{i,a,b}$ l'intersection des régions de confusion d'une fonctionnalisation i exposée aux composés chimiques a et b telle que décrite dans la section 6.1 ; et nous définissons

$$Conf_i = \bigcup_{\substack{(a,b)\in G \\ a\neq b}} Conf_{i,a,b}.$$

Cet intervalle représente les concentrations pour lesquelles un composé chimique de l'ensemble G peut être confondu avec un autre de ce même ensemble. Si $Conf_i = \emptyset$ alors il est possible d'identifier, sans erreur, les composés de G. Nous notons

$$L(A) = sup(A) - inf(A)$$

la largeur d'un intervalle A et, pour un ensemble fini d'intervalles A_i indexé par i et disjoints deux à deux, nous avons

$$L(\bigcup_i A_i) = \sum_i L(A_i).$$

Si les intervalles ne sont pas disjoints deux à deux, alors il est nécessaire de simplifier l'union avant de calculer sa largeur. Par exemple :

$$L([2;\ 5] \cup [3;\ 4]) = L([2;\ 5]) = 3 \neq L([2;\ 5]) + (L[3;\ 4]) = 4.$$

L'algorithme glouton proposé est motivé par le fait, comme l'illustre la figure 6.4, que plus la largeur de l'intervalle $Conf_i$ est faible, plus les performances que permet d'obtenir cette fonctionnalisation sont importantes.

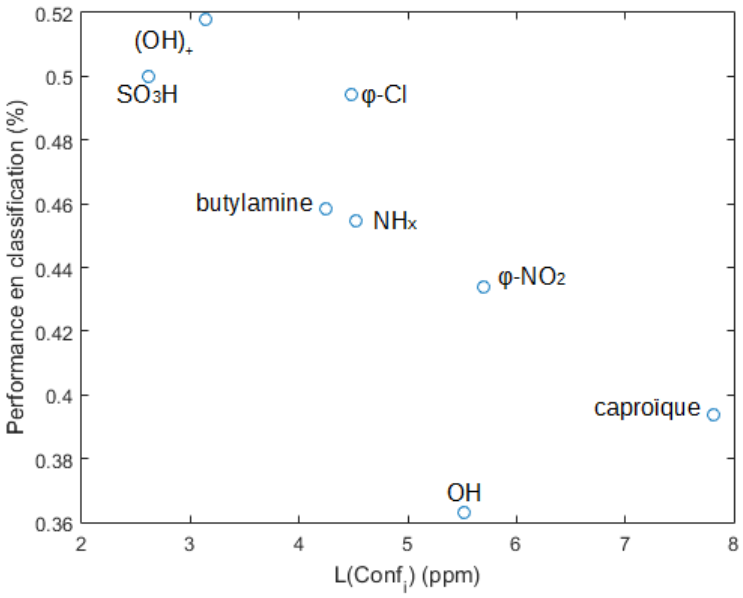


FIGURE 6.4 – Taux de classification en fonction de la largeur de l'intervalle de confusion.

L'algorithme glouton proposé consiste à

1. déterminer l'intervalle de confusion pour chaque fonctionnalisation,
2. ajouter à l'ensemble des fonctionnalisations déjà sélectionnées celle ayant le plus petit intervalle de confusion et vérifiant les contraintes ; et à
3. réitérer ce processus tant qu'une amélioration de la fonction objectif est observée.

L'algorithme 1 donne le pseudo-code de la méthode proposée.

6.2.3 Analyse de l'algorithme

Le nombre de calculs nécessaires pour déterminer les $L(Conf_i)$ croît de manière linéaire avec le nombre de capteurs et de manière quadratique avec le nombre de composés chimiques puisqu'il faut considérer tous les couples de ces derniers. Cependant, le temps de calcul nécessaire pour résoudre les équations linéaires à une inconnue (6.2), (6.3), (6.4) et (6.5) étant négligeable devant le temps requis pour évaluer les performances d'un ensemble de capteurs, l'intérêt de cet algorithme réside dans le fait qu'il ne nécessite dans le pire des cas que N_s évaluations de la fonction objectif, contre $2^{N_s} - 1$ pour la recherche exhaustive.

Procédure SélectionFonctionnalisations(N_f, c, C_{max}, N_s)
- ▷ N_f *: nombre de fonctionnalisations*
- ▷ c *: tableau contenant le coût de chaque fonctionnalisation*
- ▷ C_{max} *: coût maximal autorisé*
- ▷ N_s *: nombre de capteurs que le dispositif peut accueillir*
- $X = [0 \ \dots \ 0]_{1 \times N_f}$
- $C_{tot} = 0$
- $N_{tot} = 0$
- $Score = 0$
- **Pour** i **de** 1 **à** N_f **faire**
  - Calcul de $Conf[i]$
- **Fin Pour**
- Trier $Conf$ par ordre croissant
- **Pour** i **de** 1 **à** N_f **faire**
  - **Si** $N_{tot} + 1 \leq N_s$ **ET** $C_{tot} + c[i] \leq C_{max}$ **Alors**
    - $X_{eval} = X$
    - $X_{eval,i} = 1$
    - **Si** $z(X_{eval}) > Score$ **Alors**
      - $Score = z(X_{eval})$
      - $X_i = 1$
      - $C_{tot} = C_{tot} + c[i]$
      - $N_{tot} = N_{tot} + 1$
    - **Fin Si**
  - **Fin Si**
- **Fin Pour**
- **Retourner** X

Fin

Algorithme 1: Algorithme de sélection des fonctionnalisations.

6.3 Résultats expérimentaux

Les résultats décrits dans cette section ont été obtenus en utilisant la base de données composée des toxiques chimiques. Celle-ci est constituée des réponses de capteurs ayant pour fonctionnalisation : OH, $(OH)_+$, $\phi - Cl$, $\phi - NO_2$, SH_3OH, NHx, caproïque et butylamine, en présence de 5 composés chimiques (voir section 1.2.2 page 13). Les coûts de ces fonctionnalisations, correspondant aux temps de fabrication en heures, sont donnés dans la table 6.1. Le problème consiste ici à choisir au maximum 8 fonctionnalisations permet-

tant d'identifier ces 5 composés.

TABLE 6.1 – Coût des différentes fonctionnalisations

Fonctionnalisation	Coût en heures
OH	5
$(OH)_+$	5
$\phi - Cl$	5,25
$\phi - NO_2$	15
caproïque	5,5
SO_3H	5,5
NHx	4
butylamine	5,75

La table 6.2 donne les fonctionnalisations choisies et les performances moyennes en classification obtenues, après un processus de validation croisée à 5 plis, avec LMNN pour différentes valeurs du coût C exprimé en heures.

TABLE 6.2 – Fonctionnalisations sélectionnées pour différentes contraintes de coût sur la base de données constituée des toxiques chimiques.

C_{max}	Fonctionnalisations	Score moyen
15	*caproïque*, $(OH)_+$ et NHx	0,738
25	SO_3H, $(OH)_+$, OH et $\phi - Cl$	0,913
35	*caproïque*, $\phi - NO2$, $(OH)_+$, NHx et $\phi - Cl$	0,966

Ces résultats expérimentaux montrent notamment qu'il est possible d'obtenir des résultats proches de ceux décrits dans la section 5.1.2 en utilisant moins de capteurs et en réduisant de ce fait le coût du système.

Il peut être mis en évidence que, comme la plupart des algorithmes gloutons, celui-ci n'est pas optimal. Dans le cas $C_{max} = 15$, une recherche exhaustive montre que le choix des fonctionnalisations OH, $(OH)_+$ et NHx permet d'obtenir un score supérieur à celui obtenu avec caproïque, $(OH)_+$ et NHx de près de 10%. Néanmoins, l'intérêt de cet algorithme réside dans sa complexité linéaire : une implémentation en Matlab sur un processeur Intel i7-4600U montre que celui-ci est en pratique jusqu'à 50 fois plus rapide qu'une recherche exhaustive. Cette propriété à d'autant plus d'intérêt que la riche chimie du carbone permet la création de plusieurs dizaines de fonctionnalisations différentes [Tard, 2013] rendant toutes recherches exhaustives inenvisageables.

L'algorithme proposé dans cette section nécessite de connaître la concentration des exemples de la base d'apprentissage. Lorsque cette concentration n'est pas connue, comme c'est le cas pour les bases de données constituées des capsules de café et du DMMP et du 4-NT, la solution envisageable consiste à effectuer une recherche exhaustive : dans le pire des cas il n'y a respectivement que 15 et 255 possibilités. Les résultats expérimentaux moyens, obtenus après un processus de validation croisée à 5 plis, sont donnés dans les tables 6.2 et 6.4.

TABLE 6.3 – Fonctionnalisations sélectionnées pour différentes contraintes de coût sur la base de données constituée des capsules de café.

C_{max}	Fonctionnalisations	Score moyen
5	OH	0,302
10	$\phi - Cl$	0,493
15	SO_3H, $\phi - Cl$	0,563

TABLE 6.4 – Fonctionnalisations sélectionnées pour différentes contraintes de coût sur la base de données constituée du DMMP et du 4-NT.

C_{max}	Fonctionnalisations	Score moyen
15	*caproïque*, $\phi - Cl$	0,742
25	*caproïque*, $\phi - Cl$, et *butylamine*	0,791
35	*caproïque*, $\phi - Cl$, $(OH)_+$, et *butylamine*	0,811

Bilan du chapitre

Dans ce chapitre, nous avons décrit un algorithme glouton, interprétable et de complexité linéaire permettant, si la concentration des exemples de la base d'apprentissage est connue, de déterminer les fonctionnalisations les plus appropriées pour résoudre des problèmes d'identification de composés chimiques sous contrainte de coût. Cet algorithme permet notamment de déterminer l'ensemble optimal avec une erreur de moins de 10% tout en s'exécutant près de 50 fois plus rapidement qu'une recherche exhaustive. Les résultats expérimentaux ont également montré qu'il est possible d'obtenir des résultats proches de ceux décrits dans la section 5.1.2 en utilisant moins de capteurs et en réduisant de ce fait le coût du système.

Conclusion de la deuxième partie

Dans cette partie, nous avons exploité les équations différentielles (4.1) pour proposer un nouveau type de descripteur, se rapprochant des caractéristiques du descripteur idéal, permettant à la fois d'identifier des composés chimiques et d'estimer leur concentration.

Après avoir formulé l'estimation des gains et des constantes de temps des contributions massique et viscoélastique comme étant la solution d'un problème d'optimisation sous contraintes, l'utilisation de métaheuristiques de l'état de l'art pour résoudre ce problème a été examinée.

Il a été expérimentalement établi que, sur les bases de données considérées dans cette étude, l'optimisation par essaim particulaire est la méthode la plus appropriée, à la fois en termes de précision et de reproductibilité, pour résoudre ce problème et donc pour estimer les paramètres des deux contributions. L'utilisation de ces paramètres comme descripteurs en entrée d'algorithmes d'apprentissage supervisé, dans le contexte de l'identification de signatures chimiques, a permis une amélioration de plus de 2% des performances par rapport à celles obtenues en utilisant les techniques de l'état de l'art. Il a également été mis en évidence expérimentalement que la fusion de ces descripteurs avec les traditionnelles amplitudes des signaux en régime stationnaire a permis, en utilisant une heuristique de sélection de variables basée sur un diagramme de Hasse, une amélioration de près de 5% des performances.

Une technique permettant d'estimer le profil de concentration a également été proposée. Celle-ci, basée sur des techniques de déconvolution et de régression non paramétrique, permet d'obtenir non seulement des informations quantitatives sur la concentration, mais également des informations concernant ses variations temporelles. Il a été montré expérimentalement que cette méthode permet d'estimer la concentration d'un composé chimique pur avec moins de 10% d'erreur relative.

Enfin, un algorithme glouton interprétable et de complexité linéaire a été décrit pour déterminer les fonctionnalisations les plus appropriées pour résoudre un problème donné. Celui-ci a notamment montré que les performances obtenues sur la base de données constituée du DMMP et 4-NT peuvent être maintenues en utilisant 7 fonctionnalisations différentes tandis que ceux obtenus sur la base de données composée des toxiques chimiques peuvent être maintenues en n'utilisant que 6 fonctionnalisations, réduisant de ce fait le coût du dispositif.

Troisième partie

Vers l'identification des composés d'un mélange de composés chimiques

Introduction

Dans cette partie, nous introduisons les principaux problèmes liés à l'identification de mélanges de composés chimiques et nous évaluons expérimentalement les approches proposées pour y répondre.

Le chapitre 7 introduit les principales typologies des problèmes d'identification de mélanges ainsi que les approches existantes pour les résoudre et les performances de ces dernières sont ensuite expérimentalement évaluées. Puis, des approches phénoménologiques, propres aux capteurs SAW, sont proposées pour modéliser leur réponse lorsque ceux-ci sont exposés à un mélange de gaz. Enfin, la vérification expérimentale de la validité de ces modèles conclut ce chapitre.

Le chapitre 8 s'appuie fortement sur les résultats établis expérimentalement dans le chapitre 7. Il décrit une méthode permettant de déterminer le nombre de constituants d'un mélange. Cette méthode consiste à linéariser le modèle de la réponse des capteurs à un mélange, à estimer ces paramètres en les considérant comme les solutions d'un problème d'optimisation et à utiliser les résidus comme entrées d'un algorithme d'apprentissage supervisé. Les performances de cette méthode sont ensuite expérimentalement évaluées dans le cas de mélanges simples constitués d'au plus trois composés. Pour finir, nous montrons comment l'utilisation de la connaissance du nombre de composés d'un mélange permet d'améliorer, dans le contexte des problèmes de reconnaissance, le taux de classification.

Chapitre 7

Problématiques liées à l'identification des mélanges de composés chimiques

Introduction

Dans ce chapitre, nous décrivons qualitativement les problématiques liées à l'identification des mélanges de gaz. Nous discuterons successivement de la complétude de la base d'apprentissage et de la potentielle connaissance *a priori* du nombre de composés présents dans le mélange. Les différentes approches décrites seront ensuite comparées expérimentalement. Pour finir, une modélisation phénoménologique de la réponse des SAW à un mélange sera proposée.

7.1 Typologie des problèmes et des approches pour l'identification de mélanges

7.1.1 Exhaustivité de la base d'apprentissage

Le premier élément identifié comme étant capital pour identifier des mélanges est l'exhaustivité de la base d'apprentissage, c'est-à-dire la présence d'exemples correspondant à tous les mélanges envisageables. Ainsi, deux approches peuvent être formulées selon que des acquisitions des réponses des capteurs aux mélanges d'intérêt aient été réalisées ou non.

De manière générale, la première approche nécessite un nombre d'expériences conséquent d'autant plus que chaque expérience requiert plusieurs acquisitions de signaux . Par exemple, une borne supérieure peut être donnée si l'on désire identifier tous les mélanges possibles composés d'au plus N_g compo-

sés chimiques purs, la première approche nécessite d'effectuer N_g expériences tandis que la seconde en nécessite $2^{N_g} - 1$. Si de plus des acquisitions à N_c différentes concentrations doivent être réalisées de manière à avoir une base de données représentative, alors la première approche nécessite $N_g N_c$ expériences tandis que la seconde en nécessite $(N_c + 1)^{N_g} - 1$.

Il faut toutefois remarquer que bien qu'une base de données ne contienne pas d'exemple de toutes les combinaisons de composés chimiques possibles, elle peut tout de même être exhaustive et contenir des exemples de tous les mélanges envisageables pour une application donnée. Un exemple révélateur est celui de la base de données constituée des capsules de café (voir section 1.2.3 page 15). Les auteurs de [Flament, 2002, Clarke, 2013] font remarquer que le café est constitué de plus de 1000 molécules différentes, et bien que notre base d'apprentissage ne contienne qu'un certain nombre de mélanges d'entre elles, elle est tout de même représentative des capsules d'intérêt.

7.1.2 Connaissance *a priori* du nombre de composés du mélange

Le second élément identifié est la connaissance *a priori* du nombre de composés constituant un mélange : si celui-ci est connu, il est alors possible de réduire le nombre de possibilités. Par exemple, si l'on désire identifier un mélange de n composés appartenant à un ensemble en contenant N_g, alors la connaissance de N_g permet de réduire le nombre de possibilités de $2^n - 1$ à $\mathrm{C}_n^{N_g}$.

7.1.3 Problèmes et approches pour l'identification des mélanges

Les observations précédemment décrites permettent de formuler quatre différents problèmes d'identification de mélanges :

1. Le premier problème consiste à identifier les composés d'un mélange en ayant une base de données incomplète et aucune connaissance quant au nombre de composés constituant le mélange. Ce problème est, à notre connaissance, actuellement toujours ouvert.
2. Le second problème a pour objectif d'identifier les composés d'un mélange en ayant une base de données incomplète et une connaissance du nombre de composés du mélange. Ce problème est également, à notre connaissance, actuellement toujours ouvert.
3. Le troisième problème consiste, à partir d'une base de données exhaustive, à identifier des mélanges ayant un nombre inconnu de composés chimiques, comme c'est souvent le cas des problématiques rencontrées dans

l'industrie agroalimentaire. L'approche la plus répandue permettant de répondre à cette question consiste à créer une classe par mélange et à entraîner, grâce au caractère exhaustif de la base de données, des algorithmes d'apprentissage supervisé [Tang *et al.*, 2010, Cevoli *et al.*, 2011, Singh *et al.*, 2014, Kong *et al.*, 2016].

4. Le quatrième problème consiste à identifier un mélange en ayant non seulement une base de données exhaustive, mais également une connaissance du nombre de composés présents dans le mélange. Deux approches peuvent être envisagées pour tenter de répondre à cette problématique. La première consiste à créer une classe par mélange et à entraîner un classifieur pour chaque valeur possible du nombre de composés. Ainsi, si l'on souhaite identifier des mélanges ayant k différentes valeurs du nombre de composés, cette approche nécessite d'entraîner k algorithmes d'apprentissage supervisé multiclasses. La seconde approche consiste quant à elle à ajouter aux descripteurs utilisés le nombre k et à entraîner un unique classifieur.

La table 7.1 résume les principaux problèmes liés à l'identification des mélanges et les approches associées.

TABLE 7.1 – Typologie des problèmes et des approches pour l'identification de mélanges.

		Exhaustivité de la base d'apprentissage	
		Base complète	Base incomplète
Connaissance de N_g	Oui	Entraînement de N_g classifieurs Ajout de N_g aux descripteurs	Problème ouvert
	Non	Création d'une classe par mélange	Problème ouvert

7.1.4 Résultats expérimentaux

Pour évaluer les performances des méthodes décrites dans le cas où la base de données est exhaustive, la base de données constituée des toxiques chimiques (voir section 1.2.2 page 13) a été complétée en y ajoutant les mélanges binaires $H_2S - CH_3OH$, $CH_3OH - NH_3$, $H_2S - NH_3$, $NH_3 - C_7H_8$ et $SO_2 - C_7H_8$ à des concentrations de 4 ppm - 8 ppm et 8 ppm - 4 ppm ; et des mélanges ternaires $H_2S - CH_3 - NH_3$, $H_2S - NH_3 - SO_2$ et $H_2S - SO_2 - C_7H_8$ à des concentrations de 5 ppm et en suivant le protocole expérimental décrit dans la section 1.2.2. Dans cette étude, nous nous sommes limités à des mélanges ternaires car le banc de gaz décrit dans la section 1.2.2 (page 13) ne permet pas de réaliser des mélanges contenant plus de 3 composés.

La figure 7.1 montre la réponse des capteurs à un mélange composé de sulfure d'hydrogène à 8 ppm et d'ammoniac à 4 ppm.

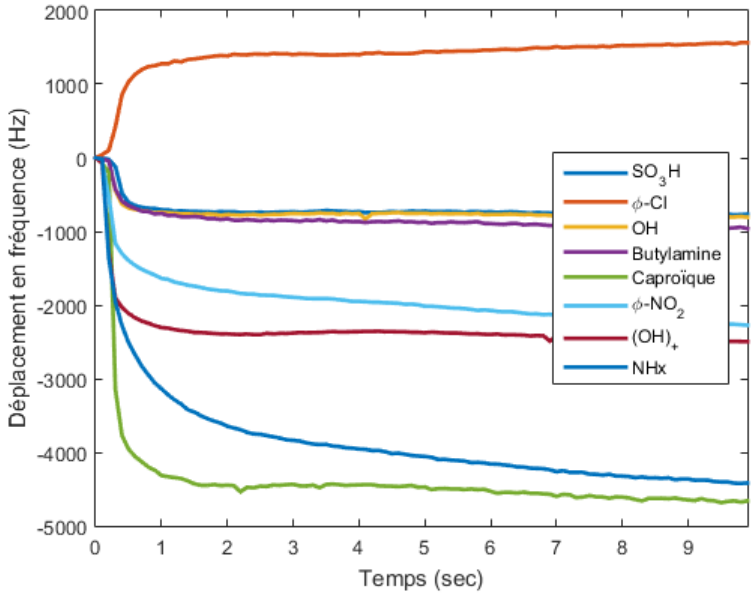


FIGURE 7.1 – Réponse des capteurs à un mélange composé de sulfure d'hydrogène à 8 ppm et d'ammoniac à 4 ppm.

La table 7.2 donne les résultats moyens, obtenus après un processus de validation croisée à 5 plis, des différentes approches proposées dans le cas d'une base de données exhaustive. Ces résultats montrent que la connaissance de N_g permet non seulement une amélioration de plus de 5% du taux de classification, mais également une réduction de près de 2% de la variance. De plus, ces résultats mettent en évidence que dans le cas où N_g est connu, l'ajout de ce dernier aux descripteurs utilisés pour entraîner l'algorithme d'apprentissage supervisé est préférable à l'entraînement de N_g classifieurs.

TABLE 7.2 – Performances des algorithmes proposés dans le cas d'une base de données exhaustive.

		Score moyen	Écart-type
N_g connu	Entraînement de N_g classifieurs	0,922	0,053
	Ajout de N_g aux descripteurs	0,946	0,054
N_g inconnu	Création d'une classe par mélange	0,871	0,073

7.2 Proposition d'approches phénoménologiques pour la modélisation de la réponse des SAW à un mélange

Une autre problématique soulevée est celle de la modélisation de la réponse des capteurs. L'étude des différentes interactions physico-chimiques des molécules d'un mélange avec une couche sensible en diamant fonctionnalisé est un problème complexe sur lequel il n'y a pas de publication à ce jour. Aussi, nous proposerons dans cette section plusieurs modèles empiriques pour modéliser la réponse des capteurs à un mélange et nous vérifierons leur validité expérimentalement.

7.2.1 Proposition de modèles empiriques

Dans cette section, nous notons $F_i(c_i)$ l'amplitude en régime stationnaire d'un capteur exposé à un composé chimique i à la concentration c_i et F l'amplitude en régime stationnaire d'un capteur exposé à un mélange de N_g composés chimiques.

Nous proposons de tester la validité d'un modèle de type somme pondérée dont les coefficients sont représentés par les variables β :

$$F = \sum_{i=1}^{N_g} \beta_i F_i(c_i)$$

et d'un modèle ayant en plus un terme d'interaction :

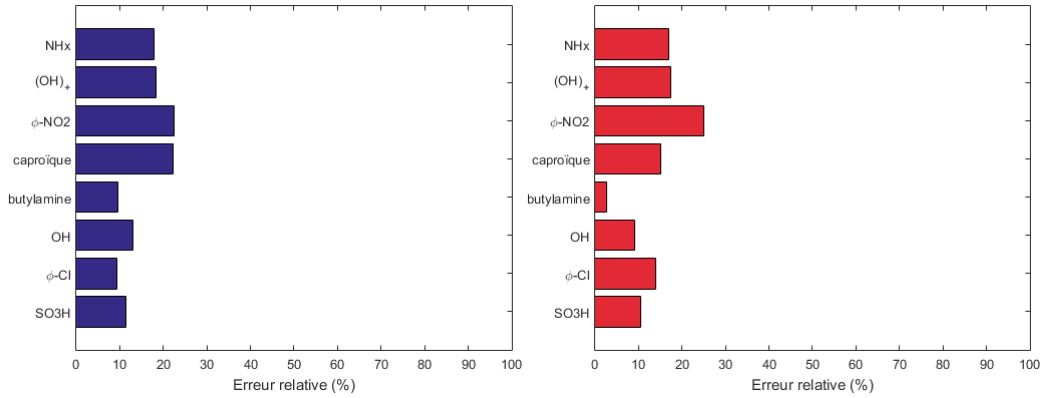
$$F = \sum_{i=1}^{N_g} \beta_i F_i(c_i) + \sum_{\substack{i,j \\ i<j}}^{N_g} \beta_{i,j} F_i(c_i) F_j(c_j).$$

7.2.2 Validation expérimentale des modèles

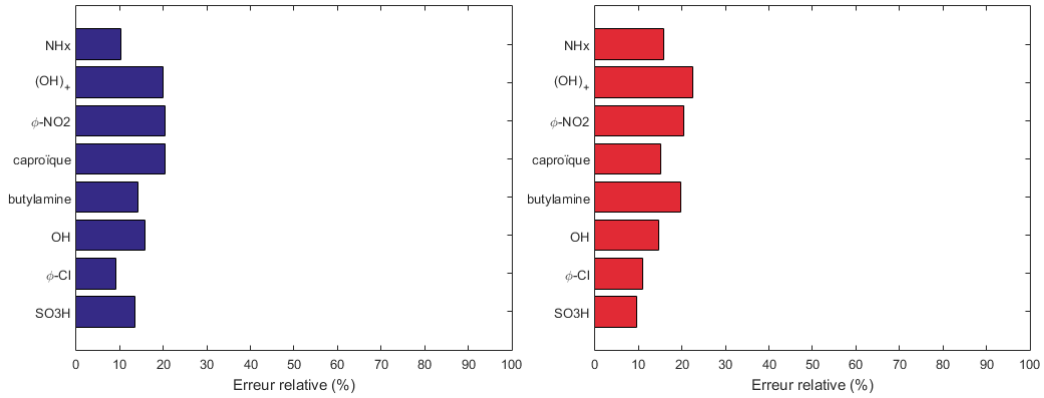
Pour juger de la validité de ces modèles, nous proposons d'identifier, par régression linéaire, leurs paramètres β et ensuite de mesurer l'erreur relative moyenne de ces modèles et ce pour tous les mélanges présents dans la base de données et pour l'ensemble des 8 fonctionnalisations utilisées.

La figure 7.2 présente une partie des résultats obtenus concernant les mélanges binaires tandis que la figure 7.3 présente une partie de ceux obtenus concernant les mélanges ternaires.

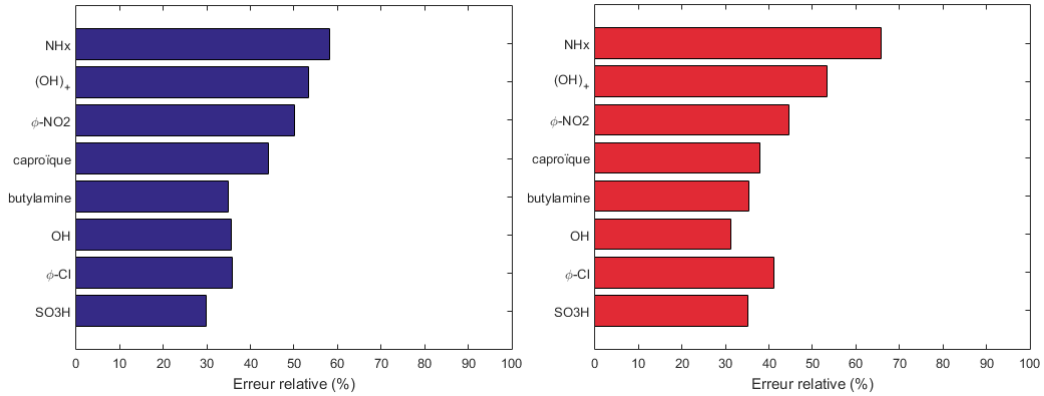
Plusieurs conclusions peuvent être tirées de ces résultats :



(a) Mélange d'ammoniac (4 ppm) et de sulfure d'hydrogène (6 ppm).

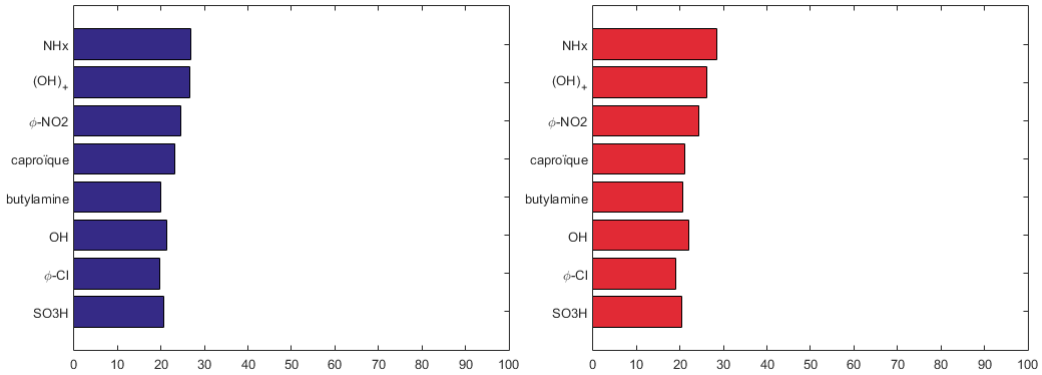


(b) Mélange d'ammoniac (6 ppm) et de méthanol (4 ppm).

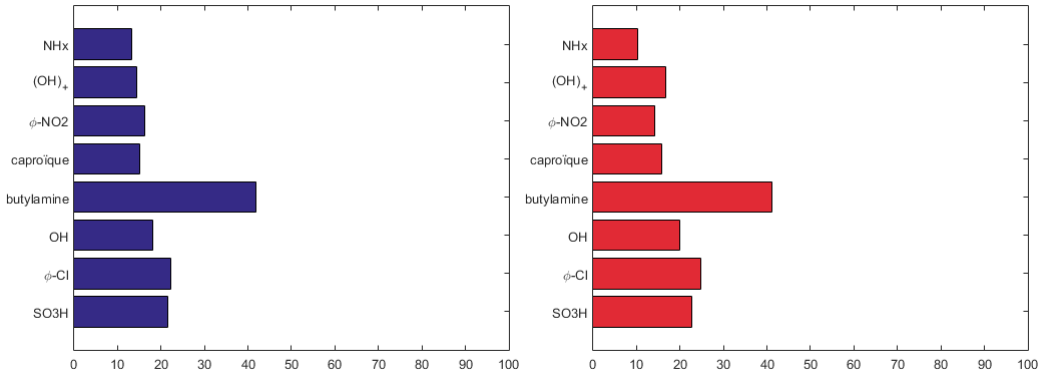


(c) Mélange de dioxyde de soufre (4 ppm) et de toluène (6 ppm).

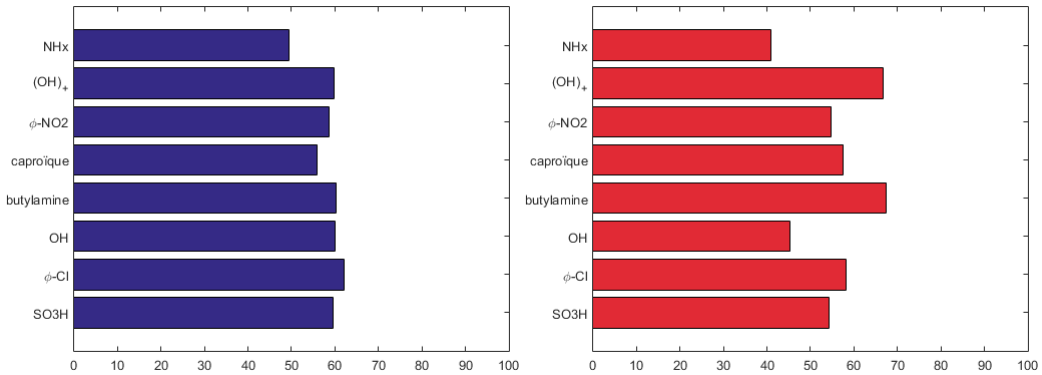
FIGURE 7.2 – Erreur relative pour les mélanges binaires en utilisant le modèle de type somme pondérée (en bleu) et celui ayant un terme d'interaction (en rouge).



(a) Mélange d'ammoniac, de sulfure d'hydrogène et de méthanol à 5 ppm.



(b) Mélange d'ammoniac, de sulfure d'hydrogène et dioxyde de soufre à 5 ppm.



(c) Mélange de sulfure d'hydrogène, dioxyde de soufre et de toluène à 5 ppm.

FIGURE 7.3 – Erreur relative pour les mélanges ternaires en utilisant le modèle de type somme pondérée (en bleu) et celui ayant un terme d'interaction (en rouge).

– Les coefficients β de chaque modèle dépendent de la nature des composés présents dans le mélange.
– Les résultats obtenus en utilisant le modèle avec le terme d'interaction sont très proches de ceux obtenus en utilisant le modèle de type somme pondérée. De plus, les différents coefficients $\beta_{a,b}$ sont faibles (de l'ordre de 10^{-5}) ce qui indique que le modèle ayant le terme d'interaction n'est pas plus représentatif de la réalité que celui qui n'en a pas.
– À l'exception de la fonctionnalisation butylamine dans le cas des mélanges ternaires, les résultats sont similaires pour chaque fonctionnalisation. Néanmoins, les coefficients β dépendent également de cette dernière.
– À l'exception des mélanges contenant du toluène, l'erreur relative moyenne est de 19,7% et l'écart-type de 4,5%. Puisque jusqu'à 10% de cette erreur peut être due à la variabilité des réponses des capteurs discutée dans la section 1.2.5 (page 18), ces résultats semblent donc indiquer que le modèle de type somme pondérée est le plus approprié.

Le cas du toluène est ici problématique car les mélanges le contenant ne vérifient pas les modèles proposés. Une hypothèse pouvant expliquer cette observation concerne le caractère apolaire du toluène.

En chimie, la polarité est une caractéristique décrivant la répartition des charges négatives et positives d'une molécule. Si le barycentre des charges positives coïncide avec celui des charges négatives alors la molécule est dite apolaire. Dans le cas contraire, la molécule est dite polaire et présente certaines spécificités. La polarité des molécules influe entre autres sur leur capacité à se mélanger : les molécules apolaires sont miscibles entre elles, les molécules polaires sont également miscibles entre elles, mais les molécules polaires et les molécules apolaires ne le sont pas. La table 7.3 donne le moment dipolaire en debyes (noté D) des molécules considérées dans ces expériences [Wallard *et al.*, 2005].

Table 7.3 – Moment dipolaire des molécules.

Molécule	Moment dipolaire
Sulfure d'hydrogène	1,47 D
Méthanol	1,70 D
Ammoniac	1,47 D
Dioxyde de soufre	1,63 D
Toluène	0,37 D

Le moment dipolaire du toluène étant très inférieur à celui des autres molécules, celui-ci peut être considéré comme apolaire. Ainsi, le toluène n'est pas miscible avec les autres molécules et il est probable que les interactions

de surface au niveau de la couche sensible en diamant soient pilotées par des phénomènes de compétition entre les molécules de toluène et les autres. Ces phénomènes, propres aux mélanges contenant du toluène, pourraient rendre compte du fait que le modèle proposé soit inadapté dans ces cas. D'autres expérimentations, impliquant d'autres molécules apolaires, ne rentrant pas dans le cadre de cette étude, seraient à mener pour infirmer ou valider cette hypothèse.

Bilan du chapitre

Dans cette section, nous avons décrit les principales problématiques liées à l'identification de mélanges à partir desquelles nous avons pu identifier différentes typologies. Le caractère exhaustif de la base d'apprentissage est d'une importance primordiale. Le cas où la base n'est pas complète reste, à l'heure actuelle et à notre connaissance, un problème encore largement ouvert tandis que dans le cas contraire des approches ont été proposées. Ces dernières peuvent, si le nombre d'entités chimiques présentes dans le mélange est connu, tirer parti de cette information. Les expériences réalisées ont montré que l'ajout du nombre de composés constituant le mélange aux descripteurs permet d'augmenter les performances de près de 5%.

Dans un second temps, nous nous sommes intéressés à la modélisation de la réponse des capteurs. Deux modèles ont été proposés : le premier étant de type somme pondérée et le second ajoutant au premier un terme d'interaction. Les expériences menées ont permis de mettre en évidence le fait que le terme d'interaction n'apportait pas de gain et que dans le cas de mélanges sans toluène, le modèle de type somme pondérée permet d'obtenir une erreur relative moyenne inférieure à 20% et semble ainsi le plus approprié pour décrire les réponses des capteurs à un mélange.

Chapitre 8

Estimation du nombre de composés dans un mélange et application à l'identification des mélanges

Introduction

Dans le chapitre 7 nous avons vu que, dans le cas d'une base d'apprentissage exhaustive, les meilleurs résultats sont obtenus lorsque le nombre de constituants des mélanges est connu. Afin de nous rapprocher de cette situation optimale, nous proposons dans ce chapitre une méthode permettant d'estimer le nombre de constituants d'un mélange. Le modèle de type somme pondérée décrit dans le chapitre 7 sera d'abord étendu pour qu'il représente aussi le régime transitoire, puis celui-ci sera linéarisé, ses paramètres seront estimés par régression linéaire et les erreurs quadratiques moyennes des résidus seront utilisées pour estimer le nombre de composés en utilisant des algorithmes d'apprentissage supervisé. Pour finir, une présentation des résultats expérimentaux obtenus conclura ce chapitre.

8.1 Estimation du nombre de constituants d'un mélange

8.1.1 Extension du modèle de type somme pondérée

Nous avons vu dans la section 4.2 que la discrétisation des équations différentielles (4.1) modélisant les contributions massique et viscoélastique donne

pour $i \in \{m, v\}$:

$$F_i[n] = \frac{\tau_i}{\tau_i + T_s} F_i[n-1] + K_i \frac{T_s}{\tau_i + T_s} c[n]$$

La solution de telles équations aux différences est donnée par la somme :

1. d'une solution de l'équation sans second membre

$$F_i[n] = \frac{\tau_i}{\tau_i + T_s} F_i[n-1]$$

de la forme

$$F_i[n] = A_i \left(\frac{\tau_i}{\tau_i + T_s} \right)^n,$$

où A_i est une constante, et

2. d'une solution particulière de cette dernière. Puisque le profil de concentration c n'est pas connu, aucune solution formelle ne peut être donnée. Cette solution sera notée P_i pour $i \in \{m, v\}$ dans le reste de cette étude.

Ainsi, la réponse F d'un capteur exposé à un composé chimique peut s'écrire

$$F[n] = A_m \left(\frac{\tau_m}{\tau_m + T_s} \right)^n + A_v \left(\frac{\tau_v}{\tau_v + T_s} \right)^n + P_m[n] + P_v[n].$$

En posant

$$T_m = \frac{\tau_m}{\tau_m + T_s} \text{ et } T_v = \frac{\tau_v}{\tau_v + T_s},$$

cette expression peut s'écrire

$$F[n] = A_m T_m^n + A_v T_v^n + P_m[n] + P_v[n]. \tag{8.1}$$

Nous avons par ailleurs vu dans la section 7.2 qu'un modèle de type somme pondérée de la forme

$$F[n] = \sum_{i=1}^{N_g} \beta_i F_i[n] \tag{8.2}$$

est approprié pour modéliser la réponse d'un capteur SAW à un mélange. On peut d'ailleurs remarquer que la non-validité du modèle pour les mélanges contenant du toluène est ici informative. En effet, si le modèle n'est pas vérifié, c'est-à-dire si l'erreur commise est importante, alors nous pouvons en conclure que nous sommes en présence d'un mélange. La substitution de l'équation (8.1), associée à chaque constituant du mélange dans l'équation (8.2) donne

$$F[n] = \sum_{i=1}^{N_g} \beta_i \Big(A_{m,i} T_{m,i}^n + A_{v,i} T_{v,i}^n + P_{m,i}[n] + P_{v,i}[n] \Big). \tag{8.3}$$

En enlevant tout sens physique aux différents coefficients de cette équation, cette dernière peut s'écrire sous la forme

$$F[n] = \sum_{i=1}^{2N_g} A_i T_i^n + P[n], \tag{8.4}$$

où $P[n] = \sum_{i=1}^{N_g} \beta_i \Big(P_{m,i}[n] + P_{v,i}[n] \Big)$ tandis que les variables A_i et T_i correspondent respectivement aux coefficients $\beta_i A_{m,i}$ ou $\beta_i A_{v,i}$ et aux coefficients $T_{m,i}$ ou $T_{v,i}$ de l'équation (8.3).

L'annexe D (page 139) montre que l'équation (8.4) peut se réécrire sous la forme :

$$F[n] = \sum_{i=1}^{2N_g} \alpha_i S_F^i[n] + P^{2N_g-1}[n] + P_s[n]. \tag{8.5}$$

Les coefficient α_i sont des constantes réelles, P^{2N_g-1} est un polynôme de degré $2N_g - 1$ et P_s est une série inconnue. S_F^i représente la somme cumulée d'ordre i de la série F :

$$S_F^i = \begin{cases} F & \text{si } i = 0 \\ S_{S_F^{i-1}} & \text{sinon} \end{cases}.$$

8.1.2 Formulation et résolution d'un problème de régression linéaire

Dans cette section, nous supposons que N échantillons $F[0], \ldots, F[N-1]$ ont été numérisés à une fréquence d'échantillonnage constante. Nous définissons les vecteurs F et C de taille respective N et $4N_g + N$:

$$F = \begin{bmatrix} F[0] \\ \vdots \\ F[N-1] \end{bmatrix} \qquad C = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{2N_g} \\ p_0 \\ \vdots \\ p_{2N_g-1} \\ P_s[0] \\ \vdots \\ P_s[N-1] \end{bmatrix}.$$

Nous définissons également les matrices $\mathbf{S}$ et $\mathbf{P}$ de taille $N \times 2N_g$

$$\mathbf{S} = \begin{bmatrix} S_F^1[0] & \ldots & S_F^{2N_g}[0] \\ \vdots & & \vdots \\ S_F^1[N-1] & \ldots & S_F^{N_g}[N-1] \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} 1 & 0^1 & \dots & 0^{2N_g-1} \\ \vdots & \vdots & & \vdots \\ 1 & (N-1)^1 & \dots & (N-1)^{2N_g-1} \end{bmatrix}$$

et nous formons la matrice $\mathbf{X}$ de taille $N \times (4N_g + N)$:

$$\mathbf{X} = \begin{bmatrix} \mathbf{S} & \mathbf{P} & \mathbf{I_N} \end{bmatrix}.$$

Avec ces définitions, l'équation (8.5) peut être écrite sous la forme matricielle

$$F = \mathbf{X}C.$$

Puisque le nombre de variables est plus important que le nombre d'équations ($4N_g + N$ variables contre N équations), il est impératif d'ajouter un terme de régularisation pour éviter le phénomène de surapprentissage. Nous définissons la matrice $\mathbf{D}$ de différenciation d'ordre 2 et de taille $(N-2) \times N$:

$$\mathbf{D} = \begin{bmatrix} -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2 & -1 \end{bmatrix}$$

et nous formons la matrice de régularisation $\mathbf{\Gamma}$ de taille $(4N_g + N - 2) \times (4N_g + N)$:

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{I_{4N_g}} & \mathbf{0_{4N_g \times N}} \\ \mathbf{0_{(N-2) \times 4N_g}} & \mathbf{D} \end{bmatrix}.$$

La matrice identité permet d'éviter que les coefficients α et p ne deviennent trop grands tandis que la matrice $\mathbf{D}$ a pour effet de lisser la séquence P_S. Les différents paramètres peuvent être estimés en résolvant le problème d'optimisation suivant :

$$\widehat{C} = argmin||F - \mathbf{X}C||^2 + \lambda||\mathbf{\Gamma}C||^2 \tag{8.6}$$

où λ désigne le coefficient de régularisation. Il s'agit d'un problème dit de Tikhonov dont la solution est [Boyd et Vandenberghe, 2014] :

$$\widehat{C} = (\mathbf{X^T X} + \lambda \mathbf{\Gamma^T \Gamma})^{-1} \mathbf{X^T} F.$$

En pratique, comme la matrice de Gram $\mathbf{X^T X} + \lambda \mathbf{\Gamma^T \Gamma}$ a un très haut coefficient de conditionnement (de l'ordre de 10^{16}), son inverse ne peut pas être calculée avec précision. Pour éviter le calcul de cette matrice, nous pouvons remarquer, comme l'indique Golub [Golub, 1983], que :

$$||F - \mathbf{X}C||^2 + \lambda||\mathbf{\Gamma}C||^2 = \left|\left| \begin{bmatrix} F \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{\Gamma} \end{bmatrix} C \right|\right|^2$$

ainsi, le problème d'optimisation (8.6) est équivalent à

$$\widehat{C} = argmin \left\| \begin{bmatrix} F \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{\Gamma} \end{bmatrix} C \right\|^2.$$

De plus, comme les éléments de $\mathbf{X}$ peuvent avoir des valeurs potentiellement très élevées, le coefficient de conditionnement de la matrice $\begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{\Gamma} \end{bmatrix}$ peut être diminué en normalisant les lignes de $\mathbf{X}$. Plus formellement, si $x_{i,j}$ désigne les éléments de cette matrice, nous définissons le vecteur

$$w_{N\times 1} = \begin{bmatrix} \frac{1}{max(|x_{1,\_}|)} \\ \vdots \\ \frac{1}{max(|x_{N,\_}|)} \end{bmatrix}$$

dont les éléments sont l'inverse du maximum en valeur absolue de chaque ligne de la matrice $\mathbf{X}$. Nous définissons également la matrice de normalisation

$$\mathbf{W} = \mathbf{1}_{\mathbf{1}\times(\mathbf{4N_g+N})} \otimes w.$$

Puis, $\widehat{C}$ peut être estimé en résolvant le problème d'optimisation suivant :

$$\widehat{C} = argmin \left\| \begin{bmatrix} F \odot \mathbf{W} \\ 0 \end{bmatrix} - \begin{bmatrix} \mathbf{X} \odot \mathbf{W} \\ \sqrt{\lambda}\mathbf{\Gamma} \end{bmatrix} C \right\|^2.$$

dont la solution est donnée par :

$$\widehat{C} = \begin{bmatrix} \mathbf{X} \odot \mathbf{W} \\ \sqrt{\lambda}\mathbf{\Gamma} \end{bmatrix}^{\dagger} \begin{bmatrix} F \odot \mathbf{W} \\ 0 \end{bmatrix}. \tag{8.7}$$

Dans ces équations, $\otimes$ désigne le produit de Kronecker et $\odot$ désigne le produit d'Hadamard.

Dans le reste de cette étude, nous notons R les résidus et nous les définissons par

$$R = F - \mathbf{X}\widehat{C}. \tag{8.8}$$

L'erreur quadratique moyenne (*Mean squared Error*, MSE) est ensuite donnée par

$$MSE = \frac{1}{N} \sum_{i=0}^{N-1} R[i]^2.$$

8.2 Résultats expérimentaux

Dans cette section, nous donnons les résultats obtenus concernant l'estimation du nombre de composés d'un mélange en utilisant l'amplitude en régime stationnaire des capteurs et l'ensemble des erreurs quadratiques moyennes calculées à partir des équations (8.7) et (8.8) pour $N_g = 1, 2, 3$, et pour chaque capteur comme entrées d'algorithmes d'apprentissage supervisé. Dans le premier cas, les descripteurs sont

$$x = (F_{S,1}, \ldots, F_{S,N_s})$$

où N_s désigne le nombre de capteurs et F_S l'amplitude de la réponse des capteurs en régime stationnaire. Dans le second cas, les descripteurs sont

$$x = (MSE_{1,N_g=1}, \ldots, MSE_{1,N_g=3}, \ldots, MSE_{N_s,N_g=1}, \ldots, MSE_{N_s,N_g=3})$$

où $MSE_{i,N_g=k}$ est l'erreur quadratique moyenne du $i^{ième}$ capteur calculée à partir des équations (8.7) et (8.8) pour $N_g = k$. Les tables 8.1 et 8.2 donnent les performances moyennes obtenues après un processus de validation croisé à 5 plis en utilisant respectivement les amplitudes en régime stationnaire et les erreurs quadratiques moyennes. Dans le premier cas, les performances moyennes sont de 90,6% et l'écart-type de 3,5% tandis que dans le second, la moyenne est de 82,6% et l'écart-type de 6,6%.

TABLE 8.1 – Performances moyennes obtenues lors d'un processus de validation croisée à 5 plis en utilisant les amplitudes en régime stationnaire et un SVM Gaussien.

		Nombre estimé de composés		
		1	2	3
Nombre réel de composés	1	0,880	0,050	0,070
	2	0,120	0,860	0,020
	3	0,000	0,020	0,980

TABLE 8.2 – Performances moyennes obtenues lors d'un processus de validation croisée à 5 plis en utilisant les erreurs quadratiques moyennes et un SVM Gaussien.

		Nombre estimé de composés		
		1	2	3
Nombre réel de composés	1	0,830	0,080	0,090
	2	0,100	0,810	0,090
	2	0,010	0,060	0,840

Ces résultats montrent que les descripteurs proposés dans ce chapitre sont moins performants que les amplitudes en régime stationnaire des signaux. On peut s'interroger sur les origines de ces résultats. Un élément de réponse est apporté par le rapport signal sur bruit (*Signal to Noise Ratio*, SNR). En effet, de manière à rendre les erreurs quadratiques moyennes comparables, les signaux considérés doivent avoir un SNR similaire. La table 8.3 présente les résultats d'une analyse statistique du SNR en décibel (dB) des signaux.

TABLE 8.3 – Analyse statistique du rapport signal sur bruit des signaux en dB.

	Minimum	Moyenne	Maximum	Écart-type
$N_g = 1$	9,5	14,1	17,3	2,1
$N_g = 2$	8,6	16,7	17,0	2,2
$N_g = 3$	12,2	15,1	16,3	1,7

Ces résultats montrent que, bien que les signaux soient en moyenne de bonne qualité (SNR $>$ 10 dB), la dispersion du SNR ainsi que son écart-type est relativement important puisqu'une augmentation de 3 dB signifie que l'amplitude du bruit est doublée.

Toutefois, une analyse des exemples mal classés montre que ceux -ci ne sont pas les mêmes dans les deux cas. La non interprétabilité des descripteurs et des algorithmes d'apprentissage supervisé rend l'analyse de cette observation complexe et sort du cadre de cette étude. Elle motive cependant le fait de fusionner ces descripteurs. La table 8.4 donne les résultats ainsi obtenus en utilisant un autoencodeur et un ensemble d'arbres de décision : la moyenne est de 94,7% et l'écart-type de 2,9%.

TABLE 8.4 – Performances moyennes obtenues lors d'un processus de validation croisée à 5 plis en utilisant les amplitudes en régime stationnaire et les erreurs quadratiques moyennes avec un autoencodeur et un ensemble d'arbres de décision.

		Nombre estimé de composés		
		1	2	3
Nombre réel de composés	1	0,940	0,040	0,020
	2	0,040	0,950	0,010
	2	0,048	0,000	0,952

Ces résultats montrent que les performances peuvent tout de même être améliorées de près de 4%.

Influence du coefficient de régularisation

Les résultats décrits jusqu'à présent ont été obtenus avec $\lambda = 10$. Dans cette section, nous évaluons l'influence du coefficient de régularisation sur le taux de classification. La méthode de la courbe L telle que décrite dans la section 5.2.1 page 82 n'est pas exploitable ici, car pour pouvoir comparer les erreurs quadratiques moyennes, celles-ci doivent toutes avoir été calculées avec la même valeur de λ. La figure 8.1 illustre l'évolution du taux moyen de classification en fonction de λ et montre qu'excepté pour $\lambda < 5$, ce coefficient n'a que très peu d'influence sur le taux de classification. Celui-ci est donc fixé à 10.

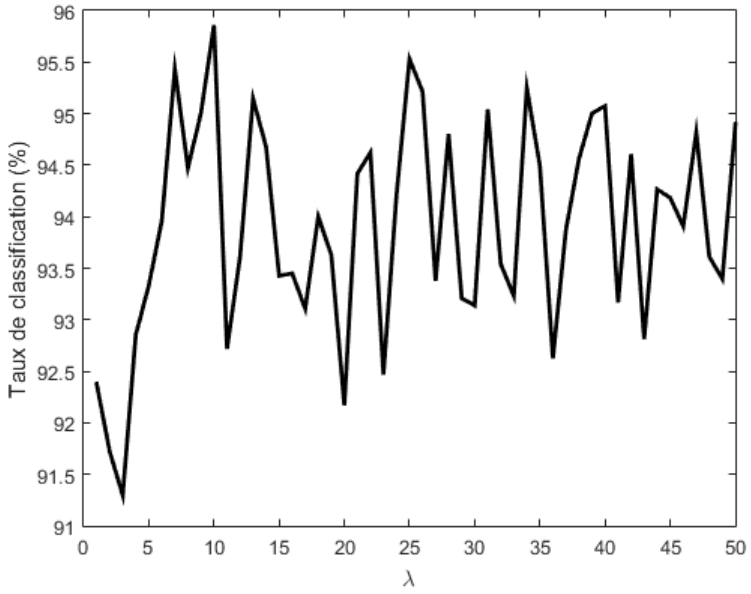


FIGURE 8.1 – Influence de λ sur le taux de classification.

8.3 Application à l'identification de mélanges

Nous avons vu dans la section 7.1.4 que la connaissance du nombre de constituants d'un mélange permet, en ajoutant cette connaissance aux descripteurs, d'améliorer le taux d'identification. Dans cette section, nous vérifions si cette observation est encore valable lorsque le nombre de constituants est estimé. La base de données utilisée dans cette section est celle constituée des toxiques chimiques et des mélanges binaires et ternaires de ces derniers telle que décrite dans la section 7.1.4. Cette approche n'a été validée que pour des mélanges binaires ou ternaires car le banc de gaz décrit dans la section 1.2.2 page 13 ne permet pas de réaliser des mélanges ayant plus de 3 composés. Le

nombre de composés N_g a d'abord été estimé en utilisant la méthode décrite dans la section 8.1 et un classifieur a ensuite été entraîné, en créant une classe par mélange, pour identifier ces derniers à partir des descripteurs

$$x = (N_g,\ F_{S,1}, \ldots, F_{S,N_s}).$$

La table 8.5 montre les résultats ainsi obtenus après un processus de validation croisée à 5 plis, et les compare à ceux obtenus lorsque N_g est inconnu. Elle met en évidence que l'estimation de N_g permet d'améliorer les performances de près de 2% tout en réduisant la variance de près de 1% et ainsi de se rapprocher de ceux obtenus lorsqu'il est connu.

TABLE 8.5 – Comparaison des performances moyennes lorsque N_g est estimé.

	Score moyen	Écart-type
N_g inconnu	0,871	0,073
N_g estimé	0,895	0,066
N_g connu	0,946	0,054

Bilan du chapitre

Étant donné que l'intérêt de la connaissance du nombre N_g de constituants d'un mélange a été démontré dans le chapitre 7, nous avons proposé une méthode permettant d'estimer ce dernier. Expérimentalement, et dans le cas des mélanges simples ayant au plus 3 constituants, il a été montré que la fusion de erreurs quadratiques moyennes et des amplitudes en régime stationnaire permet d'atteindre des performances de près de 94% contre 90% pour l'utilisation des amplitudes en régime stationnaire seules. Il a également été expérimentalement mis en évidence que l'utilisation de l'estimation de N_g permet d'améliorer les performances en identification de près de 2%.

Conclusion de la troisième partie

Dans cette section, nous avons d'abord décrit les typologies des différents problèmes d'identification de mélanges. Nous avons vu que ces problèmes peuvent être caractérisés par deux propriétés : le caractère exhaustif de la base d'apprentissage et la connaissance du nombre de constituants du mélange. Si les problèmes ne disposant pas de base d'apprentissage exhaustive restent encore ouverts, plusieurs approches ont été proposées pour résoudre ceux en ayant une. Parmi ces derniers, nous avons pu mettre en évidence l'intérêt qu'a la connaissance du nombre de composés du mélange : celui-ci permet en effet d'améliorer les performances de près de 5%.

Nous avons également proposé des modèles phénoménologiques pour représenter la réponse d'un capteur SAW à un mélange. Il a été mis en évidence expérimentalement que le modèle de type somme pondérée est plus représentatif que celui ayant en plus un terme d'interaction, car bien qu'ayant moins de paramètres, il est tout aussi précis que le second. Il permet notamment de représenter la réponse des capteurs avec une erreur moyenne de près de 20%. Toutefois, leur validité est moins affirmée pour les mélanges contenant du toluène.

Enfin, puisque l'intérêt de la connaissance du nombre de constituants d'un mélange a été démontré, nous avons proposé une méthode permettant d'estimer le nombre de composés d'un mélange. Cette méthode consiste à linéariser le modèle de la réponse des capteurs à un mélange, à estimer ces paramètres en les considérant comme les solutions d'un problème d'optimisation et à utiliser les résidus comme entrée d'un algorithme d'apprentissage supervisé. Expérimentalement, et dans le cas des mélanges simples ayant au plus 3 constituants, nous avons établi que ces nouveaux descripteurs, fusionnés avec les amplitudes en régime stationnaire, permettent d'estimer le nombre de constituants avec une précision de près de 95%. Enfin, nous avons montré que l'utilisation de cette connaissance permet d'améliorer, dans le contexte des problèmes d'identification, de près de 2% les résultats et ainsi de se rapprocher de ceux obtenus lorsque le nombre de constituants des mélanges est connu.

Conclusion générale

Les capteurs à ondes acoustiques de surface offrent à la fois des seuils de sensibilité extrêmement bas et la possibilité de détecter un grand nombre de composés chimiques du fait de leur transduction gravimétrique. Leur sélectivité est néanmoins très dépendante de la couche sensible utilisée. L'utilisation de films minces de diamant comme couche sensible offre la possibilité d'obtenir des fonctionnalisations de surface très variées compte tenu du fait que la chimie du carbone est très riche. La variété des fonctionnalisations de surface permet ainsi de créer des réseaux de capteurs possédant des affinités chimiques différentes. Cette approche multicapteurs et multiparamétrique est au centre des techniques de reconnaissance de signatures chimiques développée dans ces travaux.

Les limitations des techniques multiparamétriques de l'état de l'art identifiées dans le chapitre 3 ont motivé les travaux consistant à estimer les paramètres des contributions massique et viscoélastique. Bien qu'étant dépendant de la concentration, ces paramètres, en plus de l'avantage d'avoir un sens physique et d'être interprétables, portent des informations à la fois sur le régime transitoire et sur le régime stationnaire. Ils sont également indépendants du profil de concentration, appartiennent également à un espace de dimension plus grand que la majorité des descripteurs proposés dans la littérature et permettent de déterminer le profil temporel de concentration.

L'approche proposée et validée expérimentalement consiste à estimer ces paramètres en les considérant comme des solutions d'un problème d'optimisation. La formulation de ce dernier fait l'objet du chapitre 4 de ce manuscrit. Ce problème ne pouvant être résolu de manière analytique, les performances en termes de précision et de reproductibilité du recuit simulé, de la stratégie d'évolution $(\lambda + \mu)$ et de l'optimisation par essaim particulaire ont été comparées. Sur la base de cette étude, il s'est révélé que l'optimisation par essaim particulaire est la métaheuristique la plus appropriée parmi les trois pour résoudre ce problème et ainsi pour estimer les paramètres des deux contributions.

Pour traiter la problématique de l'identification de signatures chimiques,

nous avons montré dans le chapitre 5 l'intérêt de l'utilisation de ces paramètres comme descripteurs en entrée d'algorithmes d'apprentissage supervisé sur des jeux de données obtenus dans un environnement de laboratoire, dans un environnement partiellement contrôlé et dans un environnement non contrôlé. Nous avons ainsi mis en évidence que l'utilisation des paramètres K_m et K_v permet d'obtenir de meilleurs résultats que ceux de l'état de l'art, et ce pour les trois bases de données considérées dans cette étude. De plus, la fusion de ces paramètres avec les amplitudes en régime stationnaire des réponses des capteurs et la sélection des variables les plus appropriées grâce à une méthode basée sur un diagramme de Hasse a permis une augmentation plus significative de ces résultats. Plus quantitativement, l'approche proposée dans ce chapitre permet d'améliorer les résultats de près de 5% sur les bases constituées des toxiques chimiques et du DMMP et du 4-NT par rapport à ceux obtenus en utilisant les techniques de l'état de l'art et de près de 7% sur la base constituée des capsules de café. Par ailleurs, et contrairement aux descripteurs proposés dans la littérature, ceux introduits dans ces travaux apportent une information temporelle concernant l'évolution de la concentration des composés chimiques. Ils permettent notamment, en utilisant des techniques de déconvolution et de régression à noyaux, de déterminer avec une erreur relative inférieure à 10% le profil temporel de concentration.

Le choix des couches sensibles a un impact majeur sur les performances des systèmes multicapteurs. Ceci a donc motivé les travaux portant sur la sélection de fonctionnalisations décrits dans le chapitre 6. Ces travaux ont consisté en la formulation d'un algorithme glouton, interprétable et de complexité linéaire permettant, si la concentration des exemples de la base d'apprentissage est connue, de déterminer les fonctionnalisations les plus appropriées pour résoudre des problèmes d'identification de composés chimiques sous contrainte de coût. Nous avons appliqué cet algorithme sur les bases de données utilisées dans cette étude et montré que celui-ci est jusqu'à 50 fois plus rapide qu'une recherche exhaustive. Notons également qu'il est possible d'obtenir des résultats satisfaisant en n'utilisant qu'un faible nombre de capteurs appropriés, réduisant de ce fait le coût du système.

Cette étude se termine en abordant les problématiques liées à l'identification des mélanges. Dans le chapitre 7, nous proposons un modèle de type somme pondérée pour représenter la réponse d'un capteur à un mélange et, bien que la polarité des molécules semble avoir un impact sur celui-ci, le validons expérimentalement. Ce chapitre détaille également les typologies des différents problèmes de reconnaissance de mélanges et met expérimentalement en évidence, dans le cas où la base d'apprentissage est exhaustive, l'intérêt de la connaissance a priori du nombre de constituants du mélange. Celui-ci per-

met en effet d'améliorer les performances en classification de près de 5%. Cette observation a motivé les travaux décrits dans le chapitre 8. Dans ce chapitre, nous proposons une méthode permettant d'estimer le nombre de constituants d'un mélange. Cette dernière consiste à linéariser le modèle introduit dans le chapitre 7, puis à formuler un problème de régression linéaire et à utiliser les erreurs quadratiques moyennes comme entrées d'algorithmes d'apprentissage supervisé. Dans le cas des mélanges simples ayant au plus trois constituants, il a été expérimentalement montré que la fusion de ces erreurs quadratiques et des amplitudes en régime stationnaire permet d'atteindre des performances de près de 94% contre 90% pour l'utilisation des amplitudes en régime stationnaire seules. De plus, nous avons montré que l'utilisation de cette connaissance permet d'améliorer de près de 2% les résultats dans le contexte des problèmes d'identification, et ainsi de se rapprocher de ceux obtenus lorsque le nombre de constituants des mélanges est connu.

Les résultats des travaux de recherche objet de cette thèse ouvrent de nombreuses perspectives pour l'utilisation de capteurs à ondes acoustiques de surface ayant une couche sensible en diamant fonctionnalisé pour l'identification de signatures chimiques. En effet, plusieurs axes de travaux futurs peuvent être envisagés. Les méthodes de l'état de l'art décrites dans le chapitre 3 concernent uniquement celles ayant trouvé, à l'heure actuelle, des applications à des problématiques d'identification de composés chimiques. Une comparaison plus approfondie, incluant notamment des algorithmes d'apprentissage supervisé encore non appliqué à ce problème tels que les forêts aléatoires [Breiman, 2001], XGBoost [Chen et Guestrin, 2016] ou des réseaux de neurones profonds [Goodfellow *et al.*, 2016] devrait être conduite. Pour rendre possible l'application des réseaux de capteurs à des problématiques réelles, hors de conditions de laboratoires, il serait nécessaire de modéliser l'impact des variables d'environnement que sont la température et l'humidité sur la réponse des capteurs SAW. Ces travaux peuvent être soit d'ordre physique et consister en la formulation d'un modèle physique permettant de modéliser l'impact de la température et de l'humidité sur la réponse des capteurs, soit d'ordre mathématique et informatique et consister en la formulation d'un modèle de type boîte noire. L'objectif est ici de pouvoir introduire des méthodes plus robustes aux variations des variables de l'environnement. De plus, une étude des variations des propriétés physiques de la couche sensible dans le temps est à conduire. Ces variations, dues à des réactions chimiques entre la couche sensible et les molécules cibles, provoquent un changement de la réponse des capteurs dans le temps. Pour passer outre ce problème, les systèmes existants [Rousier *et al.*, 2012] nécessitent de changer les capteurs à intervalle de temps régulier. Une analyse physique ou de type boîte noire pourrait permettre de modéliser ces phénomènes et de concevoir des algorithmes robustes

à ceux-ci permettant ainsi de conserver les mêmes capteurs, ce qui, dans un contexte industriel, permettrait de réduire les coûts de maintenance. Le chapitre 7 ayant mis en évidence que l'identification de mélanges reste un problème ouvert lorsque la base d'apprentissage n'est pas exhaustive, des études plus approfondies des différentes interactions physico-chimiques des molécules d'un mélange avec la couche sensible et la formulation d'un modèle associé permettront d'étendre le problème d'estimation des paramètres des différentes contributions au cas des mélanges. Un autre axe de recherche envisageable consiste à développer des couches sensibles plus sélectives de manière à augmenter les capacités de discrimination du système multicapteurs.

Annexe A

Processus linéaires temps invariant et réponse impulsionnelle

Dans cette annexe, nous décrivons les principales propriétés des processus [1] linéaires temps invariant et de leur réponse impulsionnelle. Il s'agit d'un résumé de [Oppenheim et Schafer, 2009].

A.1 Processus linéaires temps invariant

Soit S un processus transformant un signal d'entrée discret $e_i[n]$ en un signal de sortie discret $s_i[n]$ tel que

$$S(e_i[n]) = s_i[n].$$

Ce processus est dit linéaire si, pour tout $i,\ j \in \mathbb{N}$, il vérifie :

$$S(e_i[n] + ke_j[n]) = s_i[n] + ks_j[n]$$

pour toute constante $k \in \mathbb{R}$. Ce processus est dit temps invariant si, pour tout i, il vérifie :

$$S(e_i[n - n_0]) = s_i[n - n_0]$$

pour tout entier naturel n_0. Un processus est dit linéaire temps invariant s'il vérifie ces deux propriétés.

1. Dans la littérature, on trouve également le terme système.

A.2 Réponse impulsionnelle

La réponse impulsionnelle $h[n]$ d'un processus S est sa sortie lorsque son entrée est la fonction δ de Dirac : $h[n] = S(\delta[n])$ avec

$$\delta[n] = \begin{cases} 1 & \text{si } n = 0 \\ 0 & \text{sinon} \end{cases} .$$

A.3 Réponse impulsionnelle et processus linéaires temps invariant

La réponse impulsionnelle joue un rôle fondamental dans la théorie des processus linéaires temps invariant. En effet, en remarquant que tout signal d'entrée e_i peut s'écrire sous la forme :

$$e_i[n] = \sum_{k=0}^{+\infty} e_i[k]\delta[n-k],$$

nous avons

$$s_i[n] = S\left(\sum_{k=0}^{+\infty} e_i[k]\delta[n-k]\right),$$

Le fait que le processus soit linéaire conduit à :

$$s_i[n] = \sum_{k=0}^{+\infty} e_i[k]S(\delta[n-k]),$$

et la propriété d'invariance à :

$$s_i[n] = \sum_{k=0}^{+\infty} e_i[k]h[n-k] = e_i[n] * h[n].$$

Ainsi, tout système linéaire temps invariant peut être caractérisé par sa réponse impulsionnelle et sa sortie n'est autre que le produit de convolution de son entrée avec sa réponse impulsionnelle.

Annexe B

Définition et propriétés des fonctions génératices

B.1 Fonctions génératrices

Une fonction génératrice, telle qu'introduite par Abraham de Moivre, est une série polynomiale dont les coefficients sont les termes successifs d'une séquence donnée [Meyer et Rubinfeld, 2005]. Plus formellement, l'expression analytique de la fonction génératrice G associée à une séquence s est :

$$G(s,x) = \sum_{n=0}^{+\infty} s[n]x^n.$$

De telles fonctions ne sont définies que pour les valeurs de x pour lesquelles la somme converge, c'est-à-dire, pour les valeurs de $x \in \mathbb{R}$ vérifiant

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, n \geq N \Rightarrow \left| \sum_{i=0}^{n} s[i]x^i - l \right| \leq \epsilon.$$

L'ensemble de ces valeurs est communément appelé région de convergence (Region Of Convergence, ROC) de la fonction génératrice :

$$ROC_{G(s)} = \left\{ x : \sum_{n=0}^{+\infty} s[n]x^n \text{ converge} \right\}.$$

B.2 Produit de convolution

Une propriété intéressante des fonctions génératrices réside dans leur capacité à transformer un produit de convolution en un produit scalaire

[Meyer et Rubinfeld, 2005]. En effet, pour deux séquences a et b, nous avons :

$$\begin{aligned} G(a*b,x) &= \sum_{n=0}^{+\infty}\sum_{m=0}^{+\infty} a[m]b[n-m]x^n = \sum_{m=0}^{+\infty}\sum_{n=0}^{+\infty} a[m]b[n-m]x^n \\ &= \sum_{m=0}^{+\infty} a[m]\sum_{n=0}^{+\infty} b[n-m]x^n = \sum_{m=0}^{+\infty} a[m]\sum_{r=0}^{+\infty} b[r]x^{r+m} \\ &= \sum_{m=0}^{+\infty} a[m]x^m \sum_{r=0}^{+\infty} a[r]x^r = G(a,x)G(b,x), \end{aligned}$$

et

$$ROC_{G(a*b)} = ROC_{G(a)} \cap ROC_{G(b)}.$$

B.3 Approximation numérique

Une des problématiques fréquemment rencontrées lorsque l'on considère des signaux numérisés est le fait que l'on ne dispose que de N échantillons. Dans ce cas, la valeur d'une fonction génératice ne peut pas être calculée puisque seul un nombre fini d'échantillons n'est disponible. Néanmoins, la fonction génératrice associée à toute séquence s dont la valeur absolue est bornée par M peut être estimée avec une précision ϵ sur l'ensemble

$$[-x_{max}, x_{max}] \text{ avec } x_{max} = 10^{\frac{-\epsilon - log(M)}{N}}.$$

En réalité, cette heuristique garantit que $s[N]x_{max}^N = 10^{-\epsilon}$. La figure B.1 représente la quantité

$$\left| \sum_{n=0}^{N+1} s[n]x^n - \sum_{n=0}^{N} s[n]x^n \right|$$

sur une échelle logarithmique pour différentes valeurs de N et x pour la séquence $s[n] = 5(1 - e^{-n/25})$.

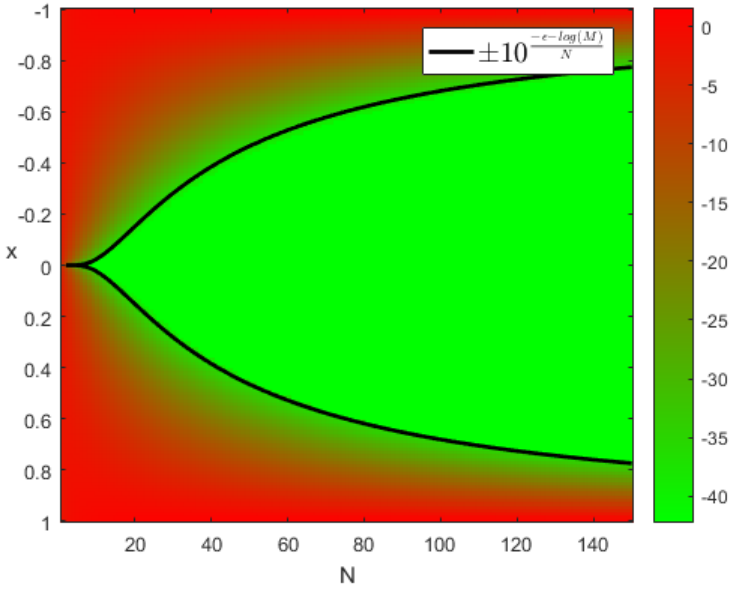


FIGURE B.1 – Précision, en échelle logarithmique, de l'estimation de la fonction génératrice associée à la séquence $s[n] = 5(1 - e^{-n/25})$.

Cette figure met en évidence que l'heuristique proposée garantit bien une bonne approximation de la valeur de la fonction génératrice. Ici, $\epsilon = 16$ ce qui correspond à la précision atteignable avec des nombres à virgule flottante double précision.

Annexe C

Métaheuristiques d'optimisation

Les métaheuristiques, telles que présentées dans ce document, sont des algorithmes génériques, conçus pour résoudre de manière numérique et approchée des problèmes d'optimisation de la forme :

$$x_{opt} = \begin{cases} \text{argmin } f(x) \text{ ou argmax } f(x) \\ \text{s. c. :} \\ \quad g_i(x) = 0 \\ \quad h_i(x) \leq 0 \end{cases}$$

f étant la fonction à minimiser, g_i et h_i étant respectivement les contraintes d'égalités et d'inégalités. Une solution candidate x est dite faisable, ou réalisable, si celle-ci vérifie toutes les contraintes, l'ensemble des solutions faisables est noté $\mathcal{F}$. Bien qu'étant critiquées pour leur absence de garantie quant à la solution trouvée [Glover, 1997], les métaheuristiques ont montré leur intérêt dans de nombreux domaines [Gogna et Tayal, 2013]. On distingue généralement deux familles de métaheuristiques : les premières sont basées sur l'évolution itérative d'une solution unique comme le recuit simulé (Simulated Annealing, SA), tandis que les secondes manipulent un ensemble de solutions candidates comme les stratégies d'évolution (Evolution Statregy, ES) ou encore l'optimisation par essaim particulaire (Particle Swarm Optimization, PSO). Bien qu'il existe actuellement plusieurs dizaines de métaheuristiques, nous avons limité notre étude aux trois précédemment citées.

C.1 Recuit simulé

La méthode du recuit simulé, telle qu'initialement formulée dans [Kirkpatrick *et al.*, 1983], trouve ses origines dans le processus de recuit utilisé en métallurgie. A partir d'une solution initiale x_0, l'algorithme génère, à chaque nouvelle itération, une solution perturbée x_p de la solution courante x_t en lui

ajoutant une variable aléatoire Gaussienne centrée :

$$x_p = x_t + \mathcal{N}(0,\, \sigma^2),$$

le plus souvent avec $\mu = 0$. Si la solution perturbée x_p est meilleure que la solution courante x_t, celle-ci est retenue. Dans le cas contraire, celle-ci peut quand même être retenue avec une certaine probabilité. Plus formellement, on a :

$$x_{t+1} = \begin{cases} x_p & \text{si } f(x_p) < f(x_t) \\ x_p & \text{si } \mathcal{U}(0,1) \leq e^{\frac{f(x_t)-f(x_p)}{T}} \\ x_t & \text{sinon} \end{cases}$$

T est un paramètre, décroissant au cours des itérations, de telle sorte que plus sa valeur est élevée, plus la probabilité d'acceptation de la solution perturbée est importante.

C.2 Stratégie d'évolution $\lambda + \mu$

La stratégie d'évolution $\lambda+\mu$, telle que définie dans [Schwefel, 1995], considère un ensemble de μ solutions candidates. A chaque itération, l'algorithme génère λ nouvelles solutions en ajoutant une variable aléatoire Gaussienne centrée à des solutions choisies parmi les λ premières :

$$x_\mu = x_\lambda + \mathcal{N}(0,\, \sigma^2),$$

et à ne garder, à l'itération suivante, que les λ meilleures solutions candidates parmi les $\lambda + \mu$ solutions courantes. Plusieurs méthodes permettant de choisir les solutions à perturber ont été développées. Parmi elles, nous pouvons notamment citer [Siarry, 2014] :

- les techniques de sélection proportionnelle ;
- les techniques de sélection par rang ;
- les techniques de sélection par tournois.

C.3 Optimisation par essaim particulaire

La méthode de l'optimisation par essaim particulaire, telle que formulée dans [Kennedy et Eberhart, 1995], s'inspire des déplacements collectifs observés chez certains animaux sociaux tels que les poissons et les oiseaux migrateurs. Cette méthode considère un ensemble de solutions candidates se déplaçant dans l'espace de recherche en quête de l'optimum global. Leur déplacement est influencé par trois composantes :

- Un terme d'inertie : chaque solution tend à suivre sa direction courante.

– Une composante cognitive : chaque solution tend à se diriger vers le meilleur emplacement par lequel elle est déjà passée.
– Une composante sociale : chaque solution tend à se diriger vers le meilleur emplacement de ses voisins.

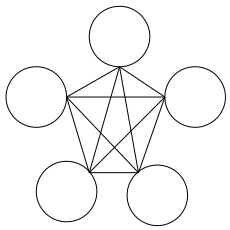
Plus formellement, si l'on note X_t une solution courante à l'itération t et V_t la vitesse de cette solution alors, les équations régissant le déplacement des solutions sont les suivantes :

$$V_{t+1} = \alpha_1 V_t + \alpha_2 \mathcal{U}(0,1)(P_{c,t} - X_t) + \alpha_3 \mathcal{U}(0,1)(P_{s,t} - X_t)$$

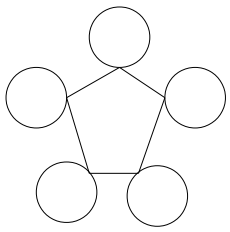
et

$$X_{t+1} = X_t + V_{t+1}.$$

α_1, α_2 et α_3 étant respectivement les coefficients d'inertie, d'accélération cognitive et d'accélération sociale, $P_{c,t}$ est la meilleure position connue de la particule et $P_{s,t}$ est la meilleure position de ses voisines. Le voisinage d'une particule est défini par l'introduction d'un graphe, chaque particule correspond à un nœud du graphe et ses voisines sont les particules qui lui sont adjacentes sur ce graphe. La figure C.1 illustre des topologies de voisinage communément utilisées.



(a) Topologie totalement connectée



(b) Topologie en anneau

FIGURE C.1 – Exemples de topologie de voisinage.

C.4 Optimisation sous contraintes

Plusieurs stratégies ont été développées pour étendre ces algorithmes aux problèmes avec contraintes [Siarry, 2014] :

– La majorité des méthodes de prise en compte des contraintes sont fondées sur le concept de fonctions de pénalité. Le problème initial avec contraintes est transformé en un problème sans contrainte

$$x_{opt} = \text{argmin } f(x) + p(x)$$

$$p(x) \begin{cases} = 0 & \text{si} \quad x \in \mathcal{F} \\ > 0 & \text{sinon} \end{cases} .$$

Le principal défaut de cette approche réside en la définition de la fonction de pénalité $p(x)$. [Siarry, 2014] propose une revue des fonctions de pénalité proposées dans la littérature.
– Une solution alternative, uniquement envisageable si l'algorithme d'optimisation met en jeu des nombres aléatoires, consiste à rééchantillonner ces derniers tant que la mise à jour de la solution candidate n'appartient pas à l'ensemble faisable.
– Une troisième méthode consiste à projeter les solutions candidates sur l'espace des solutions faisables.

C.5 Critères d'arrêt

Les algorithmes décrits dans cette section sont itératifs. Un problème soulevé par de tels algorithmes est celui de leur arrêt. De nombreux critères d'arrêt ont été proposés dans la littérature, parmi ceux-ci, on peut notamment citer :
– la durée de calcul dépasse un certain temps ;
– un nombre maximum d'itérations a été atteint ;
– la différence entre deux itérations successives est inférieure à un certain seuil $\epsilon > 0$, i.e.$||x_{k+1} - x_k|| \leq \epsilon$;
– la valeur de la fonction objectif atteint un certain seuil $f(x_k) \leq \epsilon$ ou $f(x_k) \geq \epsilon$.

[Zielinski et Laur, 2007] propose une revue complète des critères d'arrêt proposés dans la littérature.

C.6 Optimisation de l'implémentation

Une caractéristique des métaheuristiques travaillant sur un ensemble de solutions candidates est le fait que ces dernières peuvent être évaluées en parallèle. Ainsi, afin de maximiser les performances de ces algorithmes, il convient de considérer une implémentation parallèle de ces derniers, ceci est d'autant plus vrai qu'à l'heure actuelle, les processeurs multicœurs se sont démocratisés.

Annexe D

Linéarisation de la réponse des SAW à un mélange

Dans cette section, nous notons :
- $P^d[.]$ tout polynôme de degré d.
- $\mathcal{P}_k(E)$ l'ensemble des parties de E de cardinal égal à k.
- S_a désigne la somme cumulée de la suite a, c'est-à-dire la suite des sommes partielles de a. Par exemple, la somme cumulée de la suite $\{a[0], a[1], a[2], \ldots\}$ est la suite $\{a[0], a[0]+a[1], a[0]+a[1]+a[2], \ldots\}$. Plus formellement, le $i^{ième}$ terme de S_a est $\sum_{n=0}^{i} a[n]$. Nous notons aussi S_a^k la somme cumulée d'ordre k :

$$S_a^k = \begin{cases} a & \text{si } k = 0 \\ S_{S_a^{k-1}} & \text{sinon} \end{cases}.$$

Dans cette section, nous proposons une méthode pour linéariser l'équation modélisant la réponse d'un capteur SAW à un mélanges en faisant l'hypothèse d'un modèle de type somme pondérée :

$$F[n] = \sum_{i=1}^{2N_g} A_i T_i^n + P[n]. \tag{D.1}$$

Avant de linéariser cette équation, nous établissons les deux propriétés suivantes.
Propriété 1 : le terme général de la somme cumulée d'une suite dont le terme général est un polynôme de degré d est un polynôme de degré $d+1$.
Preuve : cette propriété est une conséquence directe de la formule de Faulhaber qui établit

$$\sum_{j=1}^{n} j^i = \frac{1}{i+1} \sum_{j=0}^{i} (-1)^j \mathrm{C}_{i+1}^j \mathcal{B}_j n^{i+1-j}$$

où $\mathcal{B}_j$ est le $j^{ième}$ nombre de Bernoulli [Ireland et Rosen, 1990]. Ainsi, pour une suite

$$a[n] = P^d[n] = \sum_{i=0}^{d} p_i n^i$$

donnée, nous avons

$$\begin{aligned} S_a[n] &= \sum_{j=0}^{n}\sum_{i=0}^{d} p_i j^i = \sum_{i=0}^{d}\sum_{j=0}^{n} p_i j^i = \sum_{i=0}^{d}\left(p_i 0^i + \sum_{j=1}^{n} p_i j^i\right) \\ &= p_0 + \sum_{i=0}^{d} p_i \frac{1}{i+1}\sum_{j=0}^{i}(-1)^j \mathrm{C}_{i+1}^{j}\mathcal{B}_j n^{i+1-j} \\ &= \sum_{i=0}^{d+1} p_i n^i = P^{d+1}[n]. \quad \square \end{aligned}$$

Propriété 2 : pour $k \geq 1$, la somme cumulée d'ordre k d'une suite dont le terme général est de la forme $\sum_{i=1}^{N} a_i r_i^n$ peut s'écrire

$$S_g^k[n] = \sum_{i=1}^{N} a_i \left(\frac{r_i}{r_i - 1}\right)^k r_i^n + P^{k-1}[n].$$

Preuve : Cette propriété peut être établie par récurrence.
Initialisation de la récurrence : cette propriété est vraie pour $k = 1$ puisque

$$\begin{aligned} S_g^1[n] &= \sum_{j=0}^{n}\sum_{i=1}^{N} a_i r_i^j = \sum_{i=1}^{N}\sum_{j=0}^{n} a_i r_i^j = \sum_{i=1}^{N} a_i \frac{r_i^{n+1} - 1}{r_i - 1} \\ &= \sum_{i=1}^{N} a_i \frac{r_i}{r_i - 1} r_i^n - \sum_{i=1}^{N} \frac{a_i}{r_i - 1} \\ &= \sum_{i=1}^{N} a_i \left(\frac{r_i}{r_i - 1}\right)^1 r_i^n + P^0[n]. \end{aligned}$$

Induction : supposons que la propriété est vraie pour $k = l \geq 2$. Alors

$$\begin{aligned} S_g^{l+1}[n] &= \sum_{j=0}^{n} S_g^l[i] \\ &= \sum_{j=0}^{n}\sum_{i=1}^{N} a_i \left(\frac{r_i}{r_i - 1}\right)^l r_i^j + P^{l-1}[j] \\ &= \sum_{i=1}^{N}\sum_{j=0}^{n} a_i \left(\frac{r_i}{r_i - 1}\right)^l r_i^j + \sum_{j=0}^{n} a_i P^{l-1}[j]. \end{aligned}$$

La propriété 1 amène à

$$\begin{aligned} S_g^{l+1}[n] &= \sum_{i=1}^{N} a_i \left(\frac{r_i}{r_i-1}\right)^l \frac{r_i^{n+1}-1}{r_i-1} + P^l[n] \\ &= \sum_{i=1}^{N} a_i \left(\frac{r_i}{r_i-1}\right)^{l+1} r_i^n - a_i \frac{r_i^l}{(r_i-1)^{k+1}} + P^l[n] \\ &= \sum_{i=1}^{N} a_i \left(\frac{r_i}{r_i-1}\right)^{l+1} r_i^n + P'^l[n]. \end{aligned}$$

Ainsi, cette propriété est aussi vraie pour $k = l + 1$.
Conclusion : ainsi, un raisonnement par récurrence permet d'établir que la propriété 2 est vraie pour $k \geq 1$. $\square$

Ces deux propriétés vont nous permettre de proposer une forme linéaire de l'équation (8.4) en calculant

$$F[n] + \sum_{i=1}^{2N_g} (-1)^i S_F^i[n] \sum_{I \in \mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j \in I} \frac{T_j - 1}{T_j}.$$

La propriété 1, la propriété 2 et le fait que l'opérateur somme cumulée soit linéaire amènent à :

$$\begin{aligned} &F[n] + \sum_{i=1}^{2N_g} (-1)^i S_F^i[n] \sum_{I \in \mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j \in I} \frac{T_j - 1}{T_j} \\ &= F[n] + \sum_{i=1}^{2N_g} (-1)^i \sum_{k=1}^{2N_g} A_k \left(\frac{T_k}{T_k - 1}\right)^i T_k^n \sum_{I \in \mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j \in I} \frac{T_j - 1}{T_j} \\ &+ \sum_{i=1}^{2N_g} (-1)^i P^{i-1}[n] \sum_{I \in \mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j \in I} \frac{T_j - 1}{T_j} \\ &+ \sum_{i=1}^{2N_g} (-1)^i S_{Sp}^i[n] \sum_{I \in \mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j \in I} \frac{T_j - 1}{T_j}. \end{aligned}$$

Regardons plus en détail les différents termes de l'équation D. Premièrement, nous pouvons remarquer que le terme

$$\sum_{i=1}^{2N_g}(-1)^i P^{i-1}[n] \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j\in I} \frac{T_j-1}{T_j}$$

est une somme pondérée de polynômes de degré inférieur ou égal à $2N_g - 1$. Il s'agit donc d'un polynôme $P^{2N_g-1}[n]$ de degré $2N_g - 1$. Deuxièmement, le terme

$$\sum_{i=1}^{2N_g}(-1)^i S^i_{Sp}[n] \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j\in I} \frac{T_j-1}{T_j}$$

est une suite indéterminée que nous considérerons par la suite comme étant une inconnue. Pour simplifier les notations, ce terme sera noté $P_s[n]$ dans la suite de ce chapitre. Pour finir, nous pouvons remarquer que le terme

$$\sum_{i=1}^{2N_g}(-1)^i \sum_{k=1}^{2N_g} A_k \left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j\in I} \frac{T_j-1}{T_j}$$

est égal à 0. En effet, celui-ci peut être simplifié en remarquant que la somme indexée par I peut être décomposée en deux pour $i \in [\![1;\ 2N_g]\!]$:

$$\begin{aligned}
&A_k \left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j\in I} \frac{T_j-1}{T_j} \\
&= A_k \left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_{i-1}([\![1;\ 2N_g]\!]\setminus\{k\})\setminus\{\emptyset\}} \frac{T_k-1}{T_k} \prod_{j\in I} \frac{T_j-1}{T_j} \\
&+ A_k \left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!]\setminus\{k\})\setminus\{\emptyset\}} \prod_{j\in I} \frac{T_j-1}{T_j} \\
&= A_k \left(\frac{T_k}{T_k-1}\right)^{i-1} T_k^n \sum_{I\in\mathcal{P}_{i-1}([\![1;\ 2N_g]\!]\setminus\{k\})\setminus\{\emptyset\}} \prod_{j\in I} \frac{T_j-1}{T_j} \\
&+ A_k \left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!]\setminus\{k\})\setminus\{\emptyset\}} \prod_{j\in I} \frac{T_j-1}{T_j}
\end{aligned}$$

Ici, nous avons dû priver l'ensemble des parties de l'ensemble vide pour, dans les cas $i = 1$ et $i = 2N_g$ où l'ensemble des parties privé de k est l'ensemble vide, ne pas introduire de constante du fait que, par convention, un produit vide est égal à 1.

Deux cas particuliers doivent être traités :

1. $i = 1 \Rightarrow \mathcal{P}_{i-1}([\![1;\ 2N_g]\!]\backslash\{k\})\backslash\{\emptyset\} = \emptyset$ ainsi

$$\begin{aligned} &A_k\left(\frac{T_k}{T_k-1}\right)^1 T_k^n \sum_{I\in\mathcal{P}_1([\![1;\ 2N_g]\!])\backslash\{\emptyset\}} \prod_{j\in I}\frac{T_j-1}{T_j} \\ &= A_k\left(\frac{T_k}{T_k-1}\right)^1 T_k^n \sum_{I\in\mathcal{P}_1([\![1;\ 2N_g]\!]\backslash\{k\})} \prod_{j\in I}\frac{T_j-1}{T_j} \end{aligned}$$

2. $i = 2N_g \Rightarrow \mathcal{P}_i([\![1;\ 2N_g]\!]\backslash\{k\})\backslash\{\emptyset\} = \emptyset$ et donc

$$\begin{aligned} &A_k\left(\frac{T_k}{T_k-1}\right)^{2N_g} T_k^n \sum_{I\in\mathcal{P}_{2N_g}([\![1;\ 2N_g]\!])\backslash\{\emptyset\}} \prod_{j\in I}\frac{T_j-1}{T_j} \\ &= A_k\left(\frac{T_k}{T_k-1}\right)^{2N_g-1} T_k^n \sum_{I\in\mathcal{P}_{2N_g-1}([\![1;\ 2N_g]\!]\backslash\{k\})} \prod_{j\in I}\frac{T_j-1}{T_j} \end{aligned}$$

Par conséquent, nous avons

$$\begin{aligned} &\sum_{i=1}^{2N_g}(-1)^i \sum_{k=1}^{2N_g} A_k\left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!])} \prod_{j\in I}\frac{T_j-1}{T_j} \\ &= -\sum_{k=1}^{2N_g} A_k\left(\frac{T_k}{T_k-1}\right)^1 T_k^n \sum_{I\in\mathcal{P}_1([\![1;\ 2N_g]\!]\backslash\{k\})} \prod_{j\in I}\frac{T_j-1}{T_j} \\ &+ \sum_{i=2}^{2N_g-1}(-1)^i \sum_{k=1}^{2N_g} A_k\left(\frac{T_k}{T_k-1}\right)^{i-1} T_k^n \sum_{I\in\mathcal{P}_{i-1}([\![1;\ 2N_g]\!]\backslash\{k\})} \prod_{j\in I}\frac{T_j-1}{T_j} \\ &+ A_k\left(\frac{T_k}{T_k-1}\right)^i T_k^n \sum_{I\in\mathcal{P}_i([\![1;\ 2N_g]\!]\backslash\{k\})} \prod_{j\in I}\frac{T_j-1}{T_j} \\ &+ \sum_{k=1}^{2N_g} A_k\left(\frac{T_k}{T_k-1}\right)^{2N_g-1} T_k^n \sum_{I\in\mathcal{P}_{2N_g-1}([\![1;\ 2N_g]\!]\backslash\{k\})} \prod_{j\in I}\frac{T_j-1}{T_j} \end{aligned} \tag{D.2}$$

La somme $\sum_{i=2}^{2N_g-1}(-1)^i$ peut être décomposée en deux : la partie correspondant aux i pairs et celle correspondant aux i impairs, les sommes obtenues

étant télescopiques, nous avons :

$$
\begin{aligned}
&\sum_{i=2}^{2N_g-1}(-1)^i\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{i-1}T_k^n\sum_{I\in\mathcal{P}_{i-1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&+A_k\left(\frac{T_k}{T_k-1}\right)^{i}T_k^n\sum_{I\in\mathcal{P}_{i}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&=\sum_{i=1}^{N_g-1}\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{2i-1}T_k^n\sum_{I\in\mathcal{P}_{2i-1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&+\sum_{i=1}^{N_g-1}A_k\left(\frac{T_k}{T_k-1}\right)^{2i}T_k^n\sum_{k=1}^{2N_g}\sum_{I\in\mathcal{P}_{2i}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&-\sum_{i=1}^{N_g-1}\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{2i}T_k^n\sum_{I\in\mathcal{P}_{2i}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&-\sum_{i=1}^{N_g-1}\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{2i+1}T_k^n\sum_{I\in\mathcal{P}_{2i+1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&=\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{1}T_k^n\sum_{I\in\mathcal{P}_{1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&-\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{2N_g-1}T_k^n\sum_{I\in\mathcal{P}_{2N_g-1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}
\end{aligned}
$$

La substitution de cette expression dans l'équation (D.2) mène à :

$$
\begin{aligned}
&\sum_{i=1}^{2N_g}(-1)^i\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{i}T_k^n\sum_{I\in\mathcal{P}_{i}(\llbracket 1;\ 2N_g\rrbracket)}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&=-\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{1}T_k^n\sum_{I\in\mathcal{P}_{1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&+\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{1}T_k^n\sum_{I\in\mathcal{P}_{1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&-\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{2N_g-1}T_k^n\sum_{I\in\mathcal{P}_{2N_g-1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&+\sum_{k=1}^{2N_g}A_k\left(\frac{T_k}{T_k-1}\right)^{2N_g-1}T_k^n\sum_{I\in\mathcal{P}_{2N_g-1}(\llbracket 1;\ 2N_g\rrbracket\setminus\{k\})}\prod_{j\in I}\frac{T_j-1}{T_j}\\
&=0
\end{aligned}
$$

Ainsi, l'équation (D.1) peut s'écrire sous la forme

$$\begin{aligned} &F[n] + \sum_{i=1}^{2N_g} (-1)^i S_F^i[n] \sum_{I \in \mathcal{P}_i(\llbracket 1;\ 2N_g \rrbracket)} \prod_{j \in I} \frac{T_j - 1}{T_j} \\ &= P^{2N_g - 1}[n] + P_s[n] \end{aligned}$$

Posons

$$\alpha_i = -(-1)^i \sum_{I \in \mathcal{P}_i(\llbracket 1;\ 2N_g \rrbracket)} \prod_{j \in I} \frac{T_j - 1}{T_j}$$

de manière à obtenir une forme linéaire de F :

$$F[n] = \sum_{i=1}^{2N_g} \alpha_i S_F^i[n] + P^{2N_g - 1}[n] + P_s[n].$$

Bibliographie

[Afzal, 2011] Afzal, A. (2011). *Chemical sensors : Comprehensive Sensor Techologies*. Momentum Press. Cité page 8.

[ANSI/IEEE, 2008] ANSI/IEEE (2008). 754-2008 - ieee standard for floating-point arithmetic. Rapport technique, ANSI/IEEE. Cité page 65.

[Arsat *et al.*, 2009] Arsat, R., Breedon, M., Shafiei, M., Spizziri, P., Gilje, S., Kaner, R., Kalantarzadeh, K. et Woldarski, W. (2009). Optimization of ain thin layers on diamond substrates for hugh frequency saw resonators. *Chemical Physics Letters*, vol. 467(1): p. 344 – 347. Cité page 12.

[Ballantine *et al.*, 1997] Ballantine, B., White, R., Martin, S., Ricco, J., Zellers, E., Frye, G. et Wohltjen, H. (1997). *Acoustic wave sensors Theory, Design and Physico-Chemical Applications*. Academic Press. Cité pages 9 et 57.

[Ballantine et Wohltjen, 1989] Ballantine, D. et Wohltjen, H. (1989). Surface acoustic wave devices for chemical analysis. *Analytical Chemistry*, vol. 61(1): p. 704–715. Cité page 8.

[Barzilai et Borwein, 1988] Barzilai, J. et Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, vol. 8(1): p. 141–148. Cité page 32.

[Benedetti *et al.*, 2004] Benedetti, S., Mannina, S., Sabatini, A. G. et Marcazzan, G. L. (2004). Electronic nose and neural network use for the classification of honey. *Apidologie*, vol. 35(4): p. 1–6. Cité pages 24 et 30.

[Bengio, 2009] Bengio, Y. (2009). Foundations and trends in machine learning. Rapport technique, Université de Montréal. Cité page 24.

[Bierens, 1994] Bierens, H. J. (1994). *The Nadaraya–Watson kernel regression function estimator*. Topics in Advanced Econometrics : Cambridge University Press. Cité page 85.

[Bishop, 1995] Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press. Cité pages 29 et 30.

[Bishop, 2006] Bishop, C. M. (2006). *Patter recognition and Machine Learning*. Springer. Cité page 25.

[Blum, 1992] Blum, A. (1992). *Neural networks in C++ : an object-oriented framework for building connectionist systems*. John Wiley & Sons, Inc. Cité page 31.

[Bongrain *et al.*, 2011] Bongrain, A., Agnès, C., Rousseau, L., Scorsone, E., Arnault, J., Ruffinatto, S., Omnès, F., Mailley, P., Lissorgues, G. et Bergonzo, P. (2011). High sensitivity of diamond resonant microcantilevers for direct detection in liquids as probed by molecular electrostatic surface interactions. *Langmuire*, vol. 27(1): p. 12226 – 12234. Cité page 12.

[Boukherroub *et al.*, 2005] Boukherroub, R., Wallart, X., Szunerits, S., Marcus, B., Bouvier, P. et Mermoux, M. (2005). Photochemical oxidation of hydrigenated boron-doped diamond surfaces. *Electrochemistry Communications*, vol. 7(1): p. 937 – 940. Cité page 12.

[Box *et al.*, 2015] Box, G., Jenkins, G., Reinsel, G. et Ljung, G. (2015). *Time series analysis : forecasting and control*. John Wiley & Sons. Cité page 39.

[Boyd et Vandenberghe, 2014] Boyd, S. et Vandenberghe, L. (2014). *Convex Optimization*. Cambridge University Press. Cité page 118.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine learning*, vol. 45(1): p. 5–32. Cité pages 38 et 127.

[Breiman *et al.*, 1984] Breiman, L., Friedman, J., Stone, C. J. et Olshen, R. A. (1984). *Classification and regression trees*. CRC Press. Cité page 38.

[Brereton et Lloyd, 2014] Brereton, R. G. et Lloyd, G. R. (2014). Partial least squares discriminant analysis : taking the magic away. *Journal of Chemometrics*, vol. 28(4): p. 213–225. Cité page 26.

[Broyden, 1970] Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, vol. 6(1): p. 76–90. Cité page 32.

[Brudzewski *et al.*, 2004] Brudzewski, K., Osowski, S. et Markiewicz, T. (2004). Classification of milk by means of an electronic nose and svm neural network. *Sensors and Actuators B : Chemical*, vol. 98(2): p. 291–298. Cité page 28.

[Buratti *et al.*, 2004] Buratti, S., Benedetti, S., Scampicchio, M. et Pangerod, E. C. (2004). Characterization and classification of italian barbera wines by using an electronic nose and an amperometric electronic tongue. *Analytica Chimica Acta*, vol. 525(1): p. 133–139. Cité page 26.

[Carmel *et al.*, 2003] Carmel, L., Levy, S., Lancet, D. et Harel, D. (2003). A feature extraction method for chemical sensors in electronic noses. *Sensors and Actuators B : Chemical*, vol. 93(1): p. 67 – 76. Cité page 23.

[Carmona, 2004] Carmona, R. A. (2004). Time series models : Ar, ma, arma, and all that. *In Statistical Analysis of Financial Data in S-Plus*, pages 239–309. Springer New York. Cité page 38.

[Cevoli *et al.*, 2011] Cevoli, C., Cerretani, L., Gori, A., Caboni, M. F., Toschi, T. G. et Fabbri, A. (2011). Classification of pecorino cheeses using electronic nose combined with artificial neural network and comparison with gc–ms analysis of volatile compounds. *Food Chemistry*, vol. 129(1): p. 1315–1319. Cité pages 24 et 107.

[Chein et Tzeng, 1999] Chein, T. H. et Tzeng, Y. (1999). Cvd diamond grown by microwave plasma in mixtures of aceton/oxygen and acetone/carbon dioxide. *Diamond and Related Materials*, vol. 8(1): p. 1393 – 1401. Cité page 12.

[Chen et Guestrin, 2016] Chen, T. et Guestrin, C. (2016). Xgboost : A scalable tree boosting system. *In Proceedings of the 2016 International Conference on Knowledge Discovery and Data Mining (KDD)*. Cité page 127.

[Chevallier *et al.*, 2011] Chevallier, E., Scorsone, E. et Bergonzo, P. (2011). New sensitive coating based on modified diamond nanoparticles for chemical saw sensors. *Sensors and Actuators B : Chemical*, vol. 154(2): p. 238–244. Cité page 13.

[Cho et Kurup, 2011] Cho, J. H. et Kurup, P. U. (2011). Decision tree approach for classification and dimensionality reduction of electronic nose data. *Sensors and Actuators B : Chemical*, vol. 160(1): p. 542–548. Cité page 38.

[Chuanzhi *et al.*, 2007] Chuanzhi, C., Jinyi, M., Boli, Z. et Hongmin, J. (2007). A novel toxic gases detection system based on saw resonator array and probabilistic neural network. *In Electronic Measurement and Instruments*, pages 499–503. Cité page 30.

[Clarke, 2013] Clarke, R. J. (2013). *Coffe Chemistry*. Springer. Cité page 106.

[Derksen et Keselman, 1992] Derksen, S. et Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms : Frequency of obtaining authentic and noise variables. *Bristish Journal of Mathematical and Statistical Psychology*, vol. 45(1): p. 265–282. Cité pages 76 et 77.

[Dreyfus *et al.*, 2008] Dreyfus, G., Martinez, J., Samuelides, M., Gordon, M., Badran, F. et Thiria, S. (2008). *Apprentissage statistique*. Eyrolles. Cité page 27.

[Duval *et al.*, 2014] Duval, L., Duarte, L. T. et Jutten, C. (2014). An overview of signal processing issues in chemical sensing. *In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Cité page 21.

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, vol. 7(2): p. 179–188. Cité page 26.

[Flament, 2002] Flament, I. (2002). *Coffe Flavor Chemistry*. Wiley. Cité page 106.

[Frenois *et al.*, 2014] Frenois, C., Barthet, C., Peirera, F., Minot, B., Veignal, F., Besnard, S., Roussier, R. et Mayoue, A. (2014). Detection of vapour explosives by a multi-sensor prototype-performance evaluation under laboratory and real conditions. *In Proceedings of the 2014 IEEE Sensors Symposium*. Cité page 40.

[Fujioka *et al.*, 2013] Fujioka, K., Shimizu, N., Manome, Y., Ikeda, K., Yamamoto, K. et Tomizawa, Y. (2013). Discrimination method of the volatiles from fresh mushrooms by an electronic nose using a trapping system and statistical standardization to reduce sensor value variation. *Sensors*, vol. 13(11): p. 15532–15548. Cité page 38.

[García et Aparicio, 2002] García, D. L. et Aparicio, R. (2002). Sensors : From biosensors to the electronic nose. *Grasas y Aceites*, vol. 53(1): p. 96–114. Cité page 8.

[Gardner, 1994] Gardner, J. W. (1994). A brief history of electronic noses. *Sensors and Actuators B : Chemical*, vol. 18(1): p. 210 – 211. Cité page 12.

[Gaudioso *et al.*, 2007] Gaudioso, M., Khalaf, W. et Pace, C. (2007). On the use of the svm approach in analyzing an electronic nose. *In International Conference on Hybrid Intelligent Systems (HIS)*, pages p. 42–46. Cité page 28.

[Geman et Doursat, 1992] Geman, S. et Doursat, E. B. R. (1992). Neural networks and the biais/variance dilemma. *Neural computation*, vol. 4(1): p. 1–58. Cité page 42.

[Girard *et al.*, 2010] Girard, H., Arnault, J. C., Perruchas, S., Saada, S., Gacoin, T., Boilot, J. P. et Bergonzo, P. (2010). Hydrogenation of nanodiamonds using mpcvd : A new route toward organic functionalization. *Diamond and Related Materials*, vol. 19(7): p. 1117–1123. Cité page 13.

[Girard *et al.*, 2009] Girard, H., Perruchas, S., Gesset, C., Chaigneau, M., Vieille, L., Arnault, J. C., Bergonzo, P., Boilot, J. P. et Gacoin, T. (2009). Electrostatic grafting of diamond nanoparticles : a versatile route to nanocrystalline diamond thin films. *ACS applied materials & interfaces*, vol. 1(12): p. 2738–2746. Cité page 13.

[Glover, 1997] Glover, F. (1997). Heuristics for integer programming using surrogate constraonts. *Decision Sciences*, vol. 8(1). Cité page 135.

[Güney et Atasoy, 2012] Güney, S. et Atasoy, A. (2012). Multiclass classification of n-butanol concentrations with k-nearest neighbor algorithm and support vector machine in an electronic nose. *Sensors and Actuators B : Chemical*, vol. 166(1): p. 721 – 725. Cité page 37.

[Gogna et Tayal, 2013] Gogna, A. et Tayal, A. (2013). Metaheuristics : review and application. *Journal of Experimental and Theorical Artificial Intelligence*, vol. 4(1). Cité page 135.

[Golub, 1983] Golub, G. H. (1983). *Matrix Computation*. John Hopkins University Press. Cité pages 81 et 118.

[Goodfellow *et al.*, 2016] Goodfellow, I., Bengio, Y. et Courville, A. (2016). *Deep Learning*. The MIT Press. Cité page 127.

[Grate, 2000] Grate, J. (2000). Acoustic wave microsensor arrays for vapor sensing. *Chemical Reviews*, vol. 100(1): p. 2627 – 2648. Cité page 12.

[Grate et McGill, 1995] Grate, J. et McGill, R. (1995). Dewetting effects on polymer-coated surface acoustic wave vapor sensors. *Analytical Chemistry*, vol. 67(1): p. 4015 – 4019. Cité page 12.

[Hamilton, 1994] Hamilton, J. D. (1994). *Time series analysis*. Princeton university press. Cité page 39.

[Hansen, 2000] Hansen, P. C. (2000). The l-curve and its use in the numerical treatment of inverse problems. Rapport technique, Technical University of Denmark. Cité page 83.

[Harsanyi, 1995] Harsanyi, G. (1995). *Polymer Films in Sensor Applications : Technology, Materials, Devices and Their Characterisitcs*. Technomic Publishing. Cité page 12.

[Hietala *et al.*, 2001] Hietala, S., Hietala, V. et Brinker, C. (2001). Dual saw sensor technique for determining mass and modulus changes. *IEEE Transactions Ultrasonics Ferroelectrics Frequency Control*, vol. 49(1): p. 262 – 266. Cité page 10.

[Higham, 1996] Higham, N. J. (1996). *Accuracy and stability of numerical algorithms*. SIAM. Cité page 81.

[Hsu et Lin, 2002] Hsu, C. W. et Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, vol. 13: p. 415–425. Cité page 28.

[Huang *et al.*, 2013] Huang, C. C., Tuand, S. H., Huangand, C. S., Lien, H. H., Lai, L. C. et Chuang, E. Y. (2013). Multiclass prediction with partial least square regression for gene expression data : applications in breast cancer intrinsic taxonomy. *BioMed research international*, vol. 2013. Cité page 26.

[Ireland et Rosen, 1990] Ireland, K. et Rosen, M. (1990). *A Classical Introduction to Modern Number Theory*. Springer. Cité page 140.

[Jacobs, 1988] Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, vol. 1(4): p. 295 – 307. Cité page 32.

[Janata, 2009] Janata, J. (2009). *Principles of Chemical Sensors*. Springer. Cité pages 7 et 8.

[Jha et Yadava, 2009] Jha, S. K. et Yadava, R. D. S. (2009). Preprocessing of saw sensor array data and pattern recognition. *IEEE Sensors Journal*, vol. 9(10): p. 1202–1208. Cité pages 24 et 30.

[Kennedy et Eberhart, 1995] Kennedy, J. et Eberhart, R. (1995). Particle swarm optimization. *Neural Networks*, vol. 4: p. 1942 – 1948. Cité page 136.

[Kirkpatrick *et al.*, 1983] Kirkpatrick, S., Gelatt, C. D. et Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, vol. 220(4598). Cité page 135.

[Kong *et al.*, 2016] Kong, C., Chen, W. et Pan, M. (2016). A qualitative analysis algorithm and its application in mixed gas identification. *International Journal of Electronics and Electrical Engineering*, vol. 4(1): p. 91–95. Cité page 107.

[Kriegel *et al.*, 2010] Kriegel, H., Kroger, P. et Zimek, A. (2010). Outliers detection techniques. *In Proceedings of the 2010 SIAM International Conference on Data Mining (ICDM)*. Cité page 90.

[Lange *et al.*, 2008] Lange, K., Rapp, B. et Rapp, M. (2008). Surface acoustic wave biosensors : a review. *Analytical and Bioanalytical Chemistry*, vol. 391(1): p. 1509 – 1519. Cité page 8.

[Leardi et Gonzales, 1998] Leardi, R. et Gonzales, A. M. (1998). Genetic algorithms in variable selection. Rapport technique, University of Genoa. Cité page 76.

[Levine, 1996] Levine, W. S. (1996). *The Control Handbook*. CRC Press. Cité page 58.

[Lili *et al.*, 2012] Lili, W., Chao, Y., Aiying, L. et Baozhou, Z. (2012). Identification of early moldy rice samples by pca and pnn. *Communications and Information Processing*, vol. 288(1): p. 506–514. Cité page 30.

[Liu *et al.*, 2012] Liu, X., Cheng, S., Liu, H., Hu, S., Zhang, D. et Ning, H. (2012). A survey on gas sensing technology. *Sensors*, vol. 12(7): p. 9635–9665. Cité page 7.

[Livieris et Pintelas, 2008] Livieris, I. E. et Pintelas, P. (2008). A survey on algorithms for training artificial neural networks. Rapport technique, University of Patras. Cité page 32.

[Llobet *et al.*, 1997] Llobet, E., Brezmes, J., Vilanova, X., Sueiras, J. E. et Correig, X. (1997). Qualitative and quantitative analysis of volatile organic compounds using transient and steady-state responses of a thick-film tin oxide gas sensor array. *Sensors and Actuators B : Chemical*, vol. 41(1): p. 13 – 21. Cité page 22.

[Maddala et Lahiri, 2009] Maddala, G. et Lahiri, K. (2009). *Introduction to econometrics*. Wiley. Cité page 39.

[Manai, 2014] Manai, R. (2014). *Réseaux de biocapteurs de type MEMS en diamant pour la reconnaissance d'odeurs*. Thèse de doctorat, Université Pierre et Marie Curie. Cité pages 11 et 57.

[Martinelli *et al.*, 2003] Martinelli, E., Falconi, C., d'Amico, A. et Natale, C. D. (2003). Feature extraction of chemical sensors in phase space. *Sensors and Actuators B : Chemical*, vol. 95(1): p.132 – 139. Cité page 22.

[Mayoue *et al.*, 2013] Mayoue, A., Martin, A., Lebrun, G. et Larue, A. (2013). Recursive least squares algorithm dedicated to early recognition of explosive compounds thanks to multi-technology sensors. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages p. 8761–8765. Cité pages 13 et 39.

[Meyer et Rubinfeld, 2005] Meyer, A. R. et Rubinfeld, R. (2005). Generating functions. Rapport technique, Massachusetts Institue of Technology. Cité pages 60, 131, et 132.

[Naidoo et Broadhurst, 2000] Naidoo, B. et Broadhurst, A. D. (2000). Sensor array data processing using a 2-d discrete cosine transform. *In International Symposium on Olfaction and Electronic Noses (ISOEN)*, pages 153–158. Cité page 23.

[Olunloyo *et al.*, 2011] Olunloyo, V. O. S., Ibidapo, T. A. et Dinrifo, R. R. (2011). Neural network based electronic nose for cocoa beans quality assessment. *Agricultural Engineering Journal*, vol. 13(4): p. 1521–1533. Cité page 30.

[Oppenheim et Schafer, 2009] Oppenheim, A. V. et Schafer, R. W. (2009). *Digital Signal Processing*. Prentice Hall, 3 édition. Cité page 129.

[Pardo et Sberveglieri, 2005] Pardo, M. et Sberveglieri, G. (2005). Classification of electronic nose data with support vector machines. *Sensors and Actuators B : Chemical*, 107(2):730 – 737. Cité page 28.

[Penza *et al.*, 2007] Penza, M., Aversa, P., Cassano, G., Woldarski, W. et Kalantarzadeh, K. (2007). Layered saw gas sensor with single-walled carbon nanotube-based nanocomposite coating. *Sensors and Actuators B : Chemical*, vol. 127(1): p. 168 – 178. Cité page 12.

[Peterson et Pederson, 2012] Peterson, K. B. et Pederson, M. S. (2012). The matrix cookbook. Rapport technique, Technical University of Denmark. Cité page 80.

[Philip et Hess, 2003] Philip, J. et Hess, P. (2003). Elastic, mechanical, and thermal properties of nanocrystalline diamond films. *Journal of Applied Physics*, vol. 93(1): p. 2164 – 2171. Cité page 10.

[Piotrowski et Napiorkowski, 2013] Piotrowski, A. P. et Napiorkowski, J. J. (2013). A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. *Journal of Hydrology*, vol. 476(1): p. 97 – 111. Cité page 33.

[Plagianakos *et al.*, 2002] Plagianakos, V. P., Magoulas, G. D. et Vrahatis, M. N. (2002). Deterministic nonmonotone strategies for effective training of multilayer perceptrons. *Neural Networks, IEEE Transactions on*, vol. 13(6): p. 1268–1284. Cité page 32.

[Priddy et Keller, 2005] Priddy, K. L. et Keller, P. E. (2005). *Artificial Neural Networks : An Introduction*. SPIE- International Society for Optical Engineering. Cité page 31.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, vol. 1(1): p. 81–106. Cité page 37.

[Raj *et al.*, 2012] Raj, V. B., Nimal, A. T., Parmar, Y., Sharma, M. et Gupta, V. (2012). Investigations on the origin of mass and elastic loading in the time varying distinct response of zno saw ammonia sensor. *Sensors and Actuators B : Chemical*, vol. 166(1): p. 573–585. Cité page 11.

[Rao, 1948] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 10(2): p. 159–203. Cité page 26.

[Rapp *et al.*, 2000] Rapp, M., Reibel, J., Voigt, A., Balzer, M. et Bullow, O. (2000). New miniaturized saw-sensor array for organic gas detection driven by multiplexed oscillators. *Sensors and Actuators B : Chemical*, vol. 65(1): p. 169 – 172. Cité page 13.

[Reed, 1993] Reed, R. (1993). Pruning algorithms-a survey. *Neural Networks, IEEE Transactions on*, vol. 4(5): p. 740–747. Cité page 31.

[Rockafellar, 2006] Rockafellar, R. T. (2006). Lagrange multipliers and optimality. *SIAM Journal on Control and Optimization*, vol. 35(1): p. 183–238. Cité page 80.

[Roppel et Wilson, 2001] Roppel, T. A. et Wilson, D. (2001). Improved chemical identification from sensor arrays using intelligent algorithms. *Advanced Environmental and Chemical Sensing Technology*, vol. 4205(1): p. 260–266. Cité page 30.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, vol. 65(8): p. 386–408. Cité page 29.

[Rousier *et al.*, 2012] Rousier, R., Bouat, S., Grateau, H., Darboux, M., Hue, J., Gaillard, G., Besnard, S., Veignal, F., Montméat, P., Lebrun, G. et Larue, A. (2012). T-rex : a portable device to detect and identify explosives vapors. *In Proceedings of the 2012 Eurosensors Conference*. Cité page 127.

[Schiffman *et al.*, 2000] Schiffman, S. S., Wyrick, D., Gutierrez-Osuna, R. et Nagle, H. T. (2000). Effectiveness of an electronic nose for monitoring bacterial and fungal growth. *In International Symposium on Olfaction and Electronic Noses (ISOEN)*, pages p. 173–180. Cité page 37.

[Schwefel, 1995] Schwefel, H. P. (1995). *Evolution and Optimum Seeking*. Wiley & Sons, 1 édition. Cité page 136.

[Schwenkar *et al.*, 2004] Schwenkar, F., Kestler, H. A. et Palm, G. (2004). Radial-basis-function networks : learning and applications. *Knowledge-Based Intelligent Engineering Systems and Allied Technologies*. Cité page 33.

[Shewchuk, 1994] Shewchuk, J. R. (1994). An introduction to the conjugate gradient method without the agonizing pain. Rapport technique, Canergie Mellon University. Cité page 32.

[Siarry, 2014] Siarry, P. (2014). *Métaheuristiques*. Eyrolles, 1 édition. Cité pages 136, 137, et 138.

[Simon *et al.*, 2009] Simon, N., Decorse-Pascanut, C., Gonçalves, A. M., Ballutaud, D., Charrier, G. et Etcheberry, A. (2009). Nitrogenation of boron doped siamond : Comparison of an electro-chemical treatment in liquid ammonia and a nh3/n2 plamsa. *Diamond and Related Materials*, vol. 18(1): p. 890 – 894. Cité page 12.

[Singh *et al.*, 2014] Singh, H., Raj, V. B., Kumar, J., Mittal, U., Mishra, M., Nimal, A., Sharma, M. et Gupta, V. (2014). Metal oxide saw e-nose employing pca and ann for theidentification of binary mixture of dmmp and methanol. *Sensors and Actuators B : Chemical*, vol. 200(1): p. 147–156. Cité page 107.

[Singh et Yadava, 2011] Singh, P. et Yadava, R. D. S. (2011). Wavelet based fuzzy inference system for simultaneous identification and quatitation of volatile organic compounds using saw sensor transients. *In International Conference on Swarm, Evolutionary, and Memetic Computing (SEMCCO)*, pages 319–327. Cité page 23.

[Soh *et al.*, 2014] Soh, A. C., Chow, K. K., Yusuf, U. K. M., Ishak, A. J., Hassan, M. et Khamis, S. (2014). Development of neural network ba-

sed electronic nose for herbs recognition. *International Journal on Smart Sensing and Intelligent Systems*, vol. 7(2): p. 584–609. Cité page 30.

[Specht, 1990] Specht, D. F. (1990). Probabilistic neural network. *Neural Networks*, vol. 3(1): p. 109–118. Cité page 34.

[Srivastava *et al.*, 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. et Salakhutdinov, R. (2014). Dropout : A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, vol. 15(1): p. 1929–1958. Cité page 33.

[Stanley, 2002] Stanley, K. O. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, vol. 10(2): p. 99–127. Cité pages 31 et 32.

[Stoica et Selen, 2004] Stoica, P. et Selen, Y. (2004). Model-order selection : a review of information criterion rules. *IEEE Signal Processing Magazine*, vol. 21(4): p. 36–47. Cité page 39.

[Strother *et al.*, 2002] Strother, T., Knickerbocker, T., Russel, J., Butler, J., Smith, L., Hamers, R. et Washington, D. (2002). Photochemical functionalization of diamond films. *Langmuir*, vol. 18(24): p. 968 – 971. Cité page 12.

[Tang *et al.*, 2010] Tang, K. T., Chiu, S. W., Pan, C. H., Hsieh, H. Y., Liang, Y. S. et Liu, S. C. (2010). Development of a portable electronic nose system for the detection and classification of fruity odors. *Sensors*, vol. 10(1): p. 9179–9193. Cité pages 24, 37, et 107.

[Tard, 2013] Tard, B. (2013). *Etudes des Interactions Gaz-Surfaces Diamant par Gravimétrie sur Résonateur à Onde Acoustique*. Thèse de doctorat, Université Pierre et Marie Curie. Cité pages 8, 9, 11, 13, 57, et 97.

[Vainbrand et Ginosar, 2010] Vainbrand, D. et Ginosar, R. (2010). Network-on-chip architectures for neural networks. *In Proceedings of the Fourth ACM/IEEE International Symposium on Networks-on-Chip (NOCS)*. Cité page 40.

[Voiculescu et Nordin, 2012] Voiculescu, I. et Nordin, A. (2012). Acoustic wave based mems devices for biosensing application. *Biosensors and Bioelectronics*, vol. 33(1): p. 1 – 9. Cité page 8.

[Wallard *et al.*, 2005] Wallard, A., Sené, M., Craston, D., Williams, J. et Milton, M. (2005). Tables of physical and chemical constants. Rapport technique, National Physical Laboratory of Netherland. Cité page 112.

[Weinberger *et al.*, 2006] Weinberger, K. Q., Blitzer, J. et Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in Neural Information Processing Systems*, vol. 18: p. 1473–1480. Cité page 36.

[Wen *et al.*, 2007] Wen, W., Shitang, H., Shunzhou, L., L.Minghua et Yong, P. (2007). Enhanced sensitivity of saw gas sensor coated molecularly imprinted polymer incorporating high frequency stability oscillator. *Sensors and Actuactors B : Chemical*, vol. 125: p. 422 – 427. Cité page 8.

[Zielinski et Laur, 2007] Zielinski, K. et Laur, R. (2007). Stopping criteria for constrained single-objective optimization algortihm. *Informatica*, vol. 31(1): p. 51–59. Cité page 138.

Table des figures

1.1 Principe de fonctionnement d'un capteur SAW de type résonateur à onde de Rayleigh. 9
1.2 Image du banc de gaz utilisé au Laboratoire Capteurs Diamant. 14
1.3 Réponse des 8 capteurs lors d'une exposition à 10 ppm de sulfure d'hydrogène. 15
1.4 Image du montage expérimental. 16
1.5 Réponse des 4 capteurs en présence d'une capsule de café authentique. 16
1.6 Image du montage expérimental. 17
1.7 Réponse des 8 capteurs en présence d'un mélange composé de DMMP et d'éthanol. 18
1.8 Répétabilité des acquisitions pour la base des toxiques chimiques (a) et (b), pour la base des capsules de café (c) et (d) ; et pour la base constituée du DMMP et du 4-NT (e) et (f). 19

2.1 Illustration des régimes transitoire et stationnaire de la réponse d'un capteur SAW exposé à 10 ppm d'ammoniac. 22
2.2 Exemple d'un autoencodeur permettant de reconstruire des données de dimension 4 avec 3 neurones dans la couche cachée. . . . 25
2.3 Illustration de l'hyperplan séparateur optimal et des marges. . . 27
2.4 Représentation graphique d'un neurone symbolique. 29
2.5 Exemple d'un perceptron multicouches avec 4 neurones dans la couche d'entrée, deux couches cachées constituées de 6 et 3 neurones et 2 neurones dans la couche de sortie. 30
2.6 Exemple d'un réseau RBF ayant 4 neurones dans la couche d'entrée, une couche radiale ayant 6 neurones, une couche de sortie de 3 neurones. 34
2.7 Exemple d'un réseau probabiliste ayant 4 neurones dans la couche d'entrée, une couche radiale ayant 6 neurones, une couche de sommation de 3 neurones et un 1 neurone de décision. 35
2.8 Interprétation géométrique de la différence entre KNN et LMNN. 36
2.9 Exemple d'un arbre de décision et de ses frontières de classification. 37

3.1 Non reproductibilité des régimes transitoires lorsque l'environnement n'est pas suffisamment contrôlé. 48

4.1 Temps de réponse à 90% d'un capteur exposé à de l'ammoniac. 58
4.2 Évolution de la valeur de la fonction objectif de la meilleure solution trouvée. 67
4.3 Performances de PSO pour différentes valeurs du nombre de particules et d'itérations. Les plus petites valeurs correspondent aux meilleurs résultats. 69

5.1 Performances de PSO obtenues sans aucun prétraitement et en utilisant LMNN pour différentes valeurs du nombre de particules et d'itérations. 74
5.2 Diagramme de Hasse pour la sélection de variables. 76
5.3 Résultats de l'opération de déconvolution pour différentes valeurs de k lorsque les capteurs sont exposés à 10 ppm d'ammoniac. 82
5.4 *L curve* associée à l'exemple précédent. 83
5.5 Courbure de la figure 5.4. 84
5.6 Résultats de l'opération de déconvolution lorsque les capteurs sont exposés à 10 ppm d'ammoniac avec le coefficient de régularisation optimal $k = 11840$. 85
5.7 Processus d'estimation du profil de concentration. 88

6.1 Évolution des amplitudes des contributions massique et viscoélastique d'un capteur fonctionnalisé $(OH)_+$ et exposé à du toluène. 90
6.2 Intervalles dans lesquels évoluent les amplitudes des contributions massique et viscoélastique d'un capteur fonctionnalisé $(OH)_+$ et exposé à du toluène. 91
6.3 Régions de confusion massique et viscoélastique associées aux réponses d'un capteur fonctionnalisé $(OH)_+$ exposé à du toluène (indice 1) et à du dioxyde de soufre (indice 2). 92
6.4 Taux de classification en fonction de la largeur de l'intervalle de confusion. 95

7.1 Réponse des capteurs à un mélange composé de sulfure d'hydrogène à 8 ppm et d'ammoniac à 4 ppm. 108
7.2 Erreur relative pour les mélanges binaires en utilisant le modèle de type somme pondérée (en bleu) et celui ayant un terme d'interaction (en rouge). 110
7.3 Erreur relative pour les mélanges ternaires en utilisant le modèle de type somme pondérée (en bleu) et celui ayant un terme d'interaction (en rouge). 111

8.1 Influence de λ sur le taux de classification. 122

B.1 Précision, en échelle logarithmique, de l'estimation de la fonction génératrice associée à la séquence $s[n] = 5(1 - e^{-n/25})$. . . . 133

C.1 Exemples de topologie de voisinage. 137

Liste des tableaux

3.1 Taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données constituée des toxiques chimiques. 43

3.2 Écarts-types des taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données constituée des toxiques chimiques. 44

3.3 Taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données composée des capsules de café. 45

3.4 Écarts-types des taux de classification obtenus après un processus de validation croisée à 5 plis sur la base de données composée des capsules de café. 46

4.1 Performances des métaheuristiques. 66

5.1 Comparaison des performances moyennes obtenues en considérant les paramètres K_m et K_v comme descripteurs ainsi que les amplitudes en régime stationnaire sur la base composée des toxiques chimiques. 72

5.2 Comparaison des performances moyennes obtenues sur la base composée des capsules de café. 73

5.3 Comparaison des performances moyennes obtenues sur la base composée du DMMP et du 4-NT. 73

5.4 Matrices de confusion obtenues, après un processus de validation croisée à 5 plis, sans prétraitement avec LMNN en considérant respectivement les amplitudes et les gains comme descripteur. . 75

5.5 Performances moyennes obtenues après sélection des descripteurs par l'heuristique de Hasse sur la base de données constituée des toxiques chimiques lors d'un processus de validation croisée à 5 plis. 77

5.6 Performances moyennes obtenues après sélection des descripteurs par l'heuristique de Hasse sur la base de données constituée des capsules de café lors d'un processus de validation croisée à 5 plis. 78
5.7 Performances moyennes obtenues après sélection des descripteurs par l'heuristique de Hasse sur la base de données constituée du DMMP et du 4-NT lors d'un processus de validation croisée à 5 plis. 78
5.8 Performances moyennes des méthodes de régression non paramétrique sur la base de données constituée des toxiques chimiques. 86

6.1 Coût des différentes fonctionnalisations 97
6.2 Fonctionnalisations sélectionnées pour différentes contraintes de coût sur la base de données constituée des toxiques chimiques. . 97
6.3 Fonctionnalisations sélectionnées pour différentes contraintes de coût sur la base de données constituée des capsules de café. . . . 98
6.4 Fonctionnalisations sélectionnées pour différentes contraintes de coût sur la base de données constituée du DMMP et du 4-NT. . 98

7.1 Typologie des problèmes et des approches pour l'identification de mélanges. 107
7.2 Performances des algorithmes proposés dans le cas d'une base de données exhaustive. 108
7.3 Moment dipolaire des molécules. 112

8.1 Performances moyennes obtenues lors d'un processus de validation croisée à 5 plis en utilisant les amplitudes en régime stationnaire et un SVM Gaussien. 120
8.2 Performances moyennes obtenues lors d'un processus de validation croisée à 5 plis en utilisant les erreurs quadratiques moyennes et un SVM Gaussien. 120
8.3 Analyse statistique du rapport signal sur bruit des signaux en dB. 121
8.4 Performances moyennes obtenues lors d'un processus de validation croisée à 5 plis en utilisant les amplitudes en régime stationnaire et les erreurs quadratiques moyennes avec un autoencodeur et un ensemble d'arbres de décision. 121
8.5 Comparaison des performances moyennes lorsque N_g est estimé. 123

Communications personnelles

Publications

1. **O. Hotel**, J.P. Poli, C. Mer-Calfati, E. Scorsone and S. Saada, SAW sensor's frequency shift characterization for odour recognition and concentration estimation, IEEE Sensors, vol. 17(21) : p. 7011-7018.
2. **O. Hotel**, J.P. Poli, C. Mer-Calfati, E. Scorsone and S. Saada, A review of algorithms for SAW sensors e-nose based volatile compound identification, Sensors and Actuators B : Chemical, vol. 255(3) : p. 2472-2482.

Communications orales

Communications internationales

1. **O. Hotel**, J.P. Poli, C. Mer-Calfati, E. Scorsone and S. Saada, Estimation of the number of volatile compounds in simple mixture, Eurosensors, September 2017, Paris, France.
2. **O. Hotel**, J.P. Poli, C. Mer-Calfati, E. Scorsone and S. Saada, Estimation of the parameters of SAW sensor's Frequency shift : application to odour recognition and concentration evaluation, International Symposium on Olfaction and Electronic Nose, May 2017, Montréal, Canada.

Communication nationale

1. **O. Hotel**, J.P. Poli et S. Saada, Métaheuristiques pour l'identification de composés chimiques, Conférence ROADEF de la Société Française de Recherche Opérationnelle et Aide à la Décision, Février 2017, Metz, France.

Olivier Hotel **Décembre 2017**

Sujet : Algorithmes, Méthodes et Modèles pour l'Application des Capteurs à Ondes Acoustiques de Surface à la Reconnaissance de Signatures de Composés Chimiques

Résumé : Récemment, les systèmes multicapteurs ont trouvé de nombreuses applications dans des domaines tels que l'industrie agroalimentaire, l'environnement, la médecine et la défense. Parmi les technologies existantes, les capteurs à ondes acoustiques de surface sont l'une des plus prometteuses et fait l'objet de nombreuses recherches. Les travaux décrits dans ce manuscrit concernent le développement d'algorithmes permettant la reconnaissance de composés chimiques et l'estimation de leur concentration. Cette étude décrit une méthode permettant d'estimer les paramètres des phénomènes de transduction. L'intérêt de ces derniers est mis en évidence expérimentalement dans des applications consistant à identifier des toxiques chimiques, des capsules de café contrefaites et à détecter la présence de DMMP et de 4-NT en présence d'interférents.

Mots clés : Capteur SAW, Nez Électronique, Reconnaissance d'Odeurs, Estimation de Concentration

Subject : Algorithms, Methods and Models for Surface Acoustic Wave Sensors Application to Chemical Compound Recognition

Abstract: Recently, gas sensor arrays have found numerous applications in areas such as the food, the environment, the medicine and the defense industries. Among the existing technologies, the surface acoustic wave technology is one of the most promising and has been the subject of abundant research. The work described in this manuscript concerns the development of algorithms allowing the recognition of chemical compounds and the estimation of their concentration. This study describes a method for estimating the parameters of transduction phenomena. Their interest is demonstrated experimentally in applications consisting in identifying toxic chemical compounds, counterfeit coffee capsules and in detecting the presence of DMMP and 4-NT in the presence of interfering compounds.

Keywords : SAW Sensor, Electronic Nose, Odour Recognition, Concentration Estimation