الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Hadj Lakhdar Batna
Faculté de la Technologie
Département d'électronique

جامعة الحاج لخضر باتنة
كلية التكنولوجيا
قسم الالكترونيك

Thèse présentée par

Leila DOUHA

En vue de l'obtention du titre de
Docteur en sciences
Spécialité
Electronique

Thème

Analyse des données en grande dimension en utilisant des modèles linéaires et non linéaires

Devant le jury

Dr. Ahmed LOUCHENE	M.C.A.	Président	Université de Batna
Dr. Nabil BENOUDJIT	Prof.	Rapporteur	Université de Batna
Dr. Farid MELGANI	Prof.	Co-rapporteur	Université de Trento (Italie)
Dr. Okba KAZAR	Prof.	Examinateur	Université de Biskra
Dr. Djamel SAIGAA	Prof.	Examinateur	Université de M'sila
Dr. Djemai ARAR	M.C.A.	Examinateur	Université de Batna

Ce travail de recherche rentre dans le cadre d'une collaboration entre
L'université Hadj Lakhdar Batna (Algérie) et l'université de Trento (Italie).

Année universitaire : 2012/2013

A mes parents à qui je dois beaucoup…

Remerciements

Toute éloge revient au tout puissant que nous remercions,

de par sa grâce, pour ce travail.

Il est toujours délicat de remercier l'ensemble des personnes qui ont contribué à l'aboutissement de ce travail de recherche. Que ceux qui ne sont pas mentionnés ne m'en tiennent pas rigueur.

Je tiens tout d'abord à exprimer ma profonde gratitude à mon encadreur Monsieur Nabil Benoudjit, Professeur à l'université de Batna, pour m'avoir honoré par son encadrement, pour toute la confiance qu'il m'a accordée, Pour avoir suivi mon travail avec beaucoup d'intérêt et pour ses conseils précieux. Je le remercie aussi pour sa présence et sa disponibilité continue.

Je témoigne toute ma reconnaissance à mon co-encadreur Monsieur Farid Melgani, Professeur à l'université de Trento, pour m'avoir accueilli chaleureusement dans son laboratoire, pour la qualité de ses conseils, ses lectures enrichissantes, sa grande disponibilité et son suivi continu.

Je souhaite remercier les membres du jury pour m'avoir fait l'honneur d'examiner ce travail:

Dr. Ahmed Louchene pour en avoir assuré la présidence.

Prof.Djamel Saigaa qui a bien voulu participer à ce jury.

Prof. Okba Kazar qui a bien voulu examiner ce travail.

Dr. Djemai Arar qui a bien voulu accepter de juger ce travail.

Je n'oublierai pas de remercier très cordialement le doctorant Fouzi Douak pour sa sympathie et son aide précieuse.

Enfin mes plus grands remerciements vont à ma famille sans qui je n'écrirais pas ces lignes aujourd'hui. Merci à ma mère pour l'amour qu'elle me donne et le soutien qu'elle m'apporte. Elle a été un élément clé dans la réussite de ma thèse.

Résumé

La spécificité des données modernes est sans aucun doute leurs grandes dimensions. Le traitement de ces données, nécessite soit d'adapter les méthodes existantes, soit de réduire la dimensionnalité des données. Le but visé à travers cette thèse est d'étudier l'effet de la dimension sur différentes algorithmes de régression.

Dans le contexte de régression où l'objectif est souvent de trouver une relation continue entre les observations et un ou plusieurs paramètres du phénomène physique étudié, le déséquilibrage entre le nombre d'observations et le nombre d'échantillons exigés pour exécuter la méthode de régression implique la prétendue malédiction de la dimensionnalité.

La technique des Machines à vecteurs supports (SVM) est une méthode d'apprentissage statistique destinée à résoudre les problèmes de classification et de régression qui est aujourd'hui considérée comme une des méthodes les plus performantes sur de nombreux problèmes réels, notamment pour les problèmes en grande dimension. Elle repose sur un fondement théorique solide basé sur le principe de maximisation de la marge, ce qui lui confie une grande capacité de généralisation.

Dans cette thèse, nous avons traité la question de l'utilisation de la méthode SVM pour la régression des données en grande dimension. Nous avons présenté une investigation sur les performances de cette méthode de différents points de vue. Les résultats obtenus sont comparés à ceux obtenus avec les réseaux de neurones artificiels RBF et MLP ainsi que la méthode de régression linéaire multiple (MLR) sur deux différentes bases de données spectrophotométriques.

L'efficacité d'une méthode de régression dépend fortement des caractéristiques du problème de régression considéré. En conséquence, ce qui rend difficile de choisir, a priori, l'algorithme le plus approprié pour une base de données. Cette question est à relever dans ce travail par une approche de régression basée sur la fusion d'un ensemble de différents régresseurs. Quatre différentes stratégies de fusion sont explorées. Dans ce contexte, nous avons proposé une nouvelle approche pour l'estimation de la concentration d'un composant chimique à partir des mesures spectrophotométriques, nommée SBS (Selection Based Strategy). Elle diffère des autres méthodes de fusion proposées par le fait qu'elle n'est pas basée sur une combinaison directe des estimations produites par tous les régresseurs inclus dans l'ensemble, mais sur un mécanisme de sélection qui identifie la meilleure estimation prévue disponible. L'évaluation expérimentale est réalisée sur trois bases de données réelles liées à l'industrie agroalimentaire. Les résultats expérimentaux obtenus montrent qu'en général la fusion d'un ensemble de différents algorithmes de régression conduit à un processus de régression qui en résulte plus robuste et parfois aussi plus précis que les méthodes de régression traditionnelles.

Les résultats obtenus dans ce travail sont encourageants et ouvrent de nouvelles perspectives de recherche.

***Mot –Clés**: Estimation de la concentration d'un composant chimique, réseaux de neurones artificiels, machines à vecteurs supports (SVM), modèle de régression, problème de dimensionnalité, fusion des données, spectroscopie infrarouge.*

Table des matières

Liste des figures

Liste des tableaux

Introduction Générale

1. Contexte scientifique

La spectroscopie proche infrarouge NIRS (Near Infrared Spectroscopy) a connu une croissance rapide depuis l'utilisation de sa première application dans les années 1960 dans l'industrie céréalière. Depuis lors, les applications réussies des technologies proches infrarouges ont été signalées dans plusieurs domaines. Dans les dernières années, NIRS s'est révélée être un outil analytique puissant. Elle a été largement appliquée dans le domaine de la chimie et en particulier pour l'analyse et le contrôle de qualité des aliments [1], pharmaceutique [2], les industries textiles [3] et bien d'autres. Dans l'agroalimentaire, la spectroscopie dans l'infrarouge proche et moyen est la technique la mieux adaptée à l'analyse et à la caractérisation des produits agroalimentaires et donne lieu à des applications analytiques très diverses. Elle repose sur l'acquisition rapide d'un grand nombre de données spectrales (plusieurs centaines, voir plusieurs milliers). L'information acquise nécessite le plus souvent, pour être exploitée de manière pertinente, le recours à des méthodes de traitements mathématiques et statistiques des données. Le développement de ces méthodes est un des thèmes de la chimiométrie. Le terme chimiométrie vient de l'anglais

(chemometrics). Plusieurs définitions de chimiométrie existent, mais la plus communément acceptée est la suivante [4]:

«La Chimiométrie une discipline de la chimie qui utilise des méthodes mathématiques, statistiques et informatiques pour extraire l'information utile présente dans des données de mesures chimiques. Un autre terme moins restrictif (par rapport à la chimie) utilisé dans la littérature est « l'analyse multivariable » (« multivariate analysis »).

Dans ce travail de recherche, nous nous sommes intéressés à l'application de la chimiométrie dans le domaine de la chimie analytique pour traiter les données spectrophotométriques qui sont souvent représentées en grande dimension.

Dans les applications analytiques, le cas le plus fréquent correspond à la prédiction d'une variable quantitative, telle que par exemple la concentration d'un composant présent dans un produit étudié à partir des données spectrales mesurées sur plusieurs longueurs d'ondes (plusieurs centaines, voir plusieurs milliers). D'un point de vue chimiométrique, les données spectrales ont des caractéristiques remarquables, qui nécessitent un traitement par des méthodes spécifiques [5]. La matrice des données peut comporter plus de variables (données spectrales) que d'observations (spectres). Ainsi, certaines colonnes (variables) de la matrice des données peuvent être représentées comme étant une combinaison linéaire d'autres colonnes de cette même matrice. Ce phénomène est connu sous le nom de colinéarité est la source de plusieurs problèmes dans l'application directe de plusieurs méthodes statistiques, telles que la régression linéaire multiple MLR (Multiple Linear Regression) [5, 6]. Pour éviter les problèmes liés à la redondance entre les variables, plusieurs méthodes linéaires ont été proposées, telles que La régression linéaire multiple pas à pas SMLR (Stepwise Multiple Linear Regression) [7, 8], la régression en composantes principales PCR (Principal Component Regression) [6, 9, 10] et la régression des moindres carrés partiels PLSR (Partial Least Square Regression) [6, 11, 12].

Dans certains cas cependant, la relation physique entre données spectrales et variable à prédire ne peut être approchée de façon linéaire. L'utilisation de modèles non

linéaires devient alors indispensable. La non linéarité en analyse infrarouge résulte de divers facteurs, à savoir les déviations par rapport à la loi de Beer-Lambert, la réponse non linéaire des capteurs, les dérives dans la source lumineuse etc... [13, 14].

Quand la non linéarité est importante, plusieurs techniques non linéaires de modélisation peuvent être utilisées, telles que les réseaux de neurones artificiels ANN (Artificial Neural Network) et les machines à vecteurs supports SVM (Support Vector Machine). Des exemples typiques des ANN sont les réseaux de neurones artificiels perceptrons multicouches MLP (Multi-Layered Perceptron) [15] et les réseaux de neurones artificiels à fonctions radiales de base RBF (Radial Basis Function) [16].

Dans ce travail de recherche nous nous sommes intéressés au traitement des données spectrophotomètriques qui possèdent un grand nombre de données spectrales (plusieurs centaines, voir plusieurs milliers), plus de variables (données spectrales) que d'observations (spectres), colinéarité entre les données spectrales et la relation non linéaire entre les données spectrales (variables indépendantes) et la concentration en analyte (variable dépendante).

2. Objectif de la thèse

Dans cette recherche notre objectif était d'étudier et d'extraire la relation existante entre les variables explicatives ou indépendantes et les variables expliquées ou dépendantes. L'accent est mis sur la prédiction de la concentration (variables dépendantes) présente dans un produit étudié à partir des variables indépendantes, qui sont des données spectrales mesurées sur plusieurs longueurs ou nombre d'ondes des produits agroalimentaires (bases de données réelles : wine, jus d'orange et pomme). Les données spectrophotométriques ont des caractéristiques spécifiques qui doivent être prises en compte pour obtenir des modèles prédictifs stables et performants. Pour ce faire, notre propre recherche a été motivée par l'objectif premier de tester l'effet de la dimensionnalité sur différents algorithmes de régression. Nous présentons une étude expérimentale détaillée de la méthode SVM pour l'estimation de la concentration des composants chimiques par des mesures spectrophotométriques. En

particulier, nous présentons une investigation sur les performances de cette méthode de différents points de vue, y compris: 1) l'influence du type de noyau dans la tâche de régression, 2) la sensibilité au nombre de variables d'entrée, 3) la sensibilité au nombre d'échantillons d'apprentissage disponibles, et 4) la stabilité globale [17].

Dans ce qui précède, ont été présentées les méthodes de régression qui peuvent être appliqué aux données spectrales. Certes il existe dans la littérature, un grand nombre de méthodes de régression ayant déjà fait leurs preuves. L'utilisateur peut se trouver dans une situation difficile, c'est-à-dire laquelle des techniques à choisir parmi toutes les approches proposées dans la littérature ? Dans ce travail de recherche, nous proposons une solution possible pour résoudre ce problème : qui est la fusion des résultats fournis par un ensemble de différentes méthodes de régression [18]. A cet égard, notre deuxième objectif est double:

- Premièrement, évaluer l'intérêt de la fusion de différents types de regresseurs par rapport à l'usage d'un seul type de regresseur. Nous proposons une approche à l'estimation de la concentration d'un composant chimique à base d'un système robuste multiple RMS (Robust Multiple System) [18]. Trois méthodes différentes sont considérées pour la conception de RMS. La diversité dans l'ensemble des régresseurs et un choix approprié de la stratégie de fusion sont deux points essentiels que nous avons pris en compte, afin d'obtenir une meilleure robustesse et précision. La conception de RMS est obtenue en construisant un ensemble de regresseurs sur la base de trois méthodes de régression différentes, ce sont RBF, MLP et SVM. RMS intègre les estimations obtenues à partir des trois différents algorithmes de régression qui composent l'ensemble par des stratégies de combinaison (fusion) linéaire et non linéaire : une est basée sur l'utilisation de la moyenne simple ACS (Average Combination Strategy), tandis que la seconde effectue une moyenne pondérée après la détermination (de façon supervisée) des poids à attribuer à chaque membre de l'ensemble des régresseurs. Elle est nommée stratégie de la moyenne pondérée WCS (Weighted Combination Strategy). La troisième

méthode est la stratégie de fusion non linéaire NLFS (Non Linear Fusion Strategy) qui est accomplie via un réseau de neurone artificiel RBF.

- Deuxièmement, nous proposons d'analyser la précision de chaque régresseur inclut dans l'ensemble des régresseurs dans les différentes portions de l'espace des caractéristiques pour aboutir à un régresseur robuste. Cette tâche est réalisée via la quatrième stratégie de fusion SBS (Selection Based Strategy) qui fournit en sortie l'estimation obtenue par l'algorithme de régression (inclut dans l'ensemble), caractérisé par la plus haute précision attendue dans la région de l'espace des caractéristiques associée avec le modèle considéré [18].

3. Structure de la thèse et contributions

Ce manuscrit est organisé en quatre chapitres décrits comme suit:

Chapitre 1: Analyse des données en grande dimension

Le premier chapitre présente, dans un premier temps, une description sur les données en grande dimension. Ensuite, dans un deuxième temps, il propose un tour d'horizon des différents modèles linéaires et non linéaires utilisés dans l'analyse des données.

Chapitre 2 : Description des bases de données

Ce chapitre est consacré tout d'abord à la description des données spectrophotométriques, par la suite il donne une description de trois bases de données différentes liées à l'industrie agroalimentaire.

Chapitre 3: La régression en spectrophotométrie : étude expérimentale

Au cours de ce chapitre, une application de la méthode de régression par SVM au problème d'estimation de la concentration d'un composant chimique est présentée. Nous présentons une investigation sur les performances de cette méthode de différents points de vue et nous verrons en quoi elle est adaptée aux problèmes de grande dimension.

Chapitre 4: Système de régression robuste

Ce chapitre aborde tout d'abord l'intérêt de la fusion d'un ensemble de régresseurs. Trois stratégies de fusion sont présentées. Ensuite nous proposons l'analyse de la précision de chaque régresseur inclut dans l'ensemble des régresseurs. Nous présentons une nouvelle stratégie de fusion basée sur la sélection pour l'estimation de la concentration d'un composant chimique à partir des mesures spectrophotométriques dans les différentes portions de l'espace des caractéristiques pour aboutir à un régresseur robuste.

Conclusion générale

Le manuscrit se termine par une conclusion générale qui donne une synthèse de la recherche développée dans cette thèse ainsi que les perspectives de recherches qui s'ouvrent à nous.

Les contributions de cette thèse sont:

- Notre travail a d'abord porté sur l'étude de l'effet des données en grande dimension sur différents algorithmes de régression souvent utilisés dans la littérature. Cette étude a montré que SVM est une méthode efficace et adaptée à la régression dans les espaces de grande dimension [17].
- La deuxième contribution est la proposition de la fusion d'un ensemble de régresseurs. Cette dernière apporte une amélioration notable des performances du processus de régression en termes de précision. Une stratégie SBS basée sur un nouvel algorithme de fusion pour l'estimation de la concentration d'un composant chimique est proposée aussi [18].

Chapitre 1
Analyse des données en grande dimension

1.1. Introduction

L'objectif de ce chapitre introductif est de poser les bases nécessaires à la compréhension du travail de recherche réalisé dans le cadre de cette thèse. Dans un premier temps nous introduisons une description sur les données en grande dimension. Ensuite, nous présentons le problème de la malédiction de la dimensionnalité (Curse of dimensionality en anglais) et les intuitions géométriques en grande dimension. La partie suivante est consacrée à la description de la régression basée sur des modèles linéaires et la régression basée sur des modèles non linéaires. La méthode de régression linéaire multiple (MLR) est tout d'abord présentée. Le principe de la régression sur composantes principales (PCR) et la régression au sens des moindres carrées (PLSR) sont ensuite décrits. Concernant les modèles non linéaires, nous commençons par les réseaux de neurones artificiels. Nous présentons deux structures des réseaux de neurones artificiels à savoir les réseaux perceptrons multicouches MLP et les réseaux à fonctions radiales de base RBF. Ce chapitre

termine par la description de la méthode des machines à vecteurs supports (SVM). La méthode est tout d'abord exposée dans le cas de la classification ensuite une description est donnée dans le cas de la régression.

1.2. Données en grande dimension

1.2.1. Introduction

Les espaces de grande dimension possèdent des propriétés mathématiques particulières qui affectent le comportement des méthodes manipulant des données dans ces espaces. En effet, la plupart des données provenant de problèmes concrets réels notamment de problèmes d'apprentissage apparaissent souvent en grande dimension: c'est-à-dire qu'un grand nombre de variables ont été mesurées pour chaque exemple d'apprentissage et en conséquence on aura un nombre d'observations (échantillons) largement inférieur au nombre de variables. Le traitement de ces données, nécessite soit d'adapter les méthodes statistiques usuelles, soit de réduire la dimensionnalité des données. Une façon d'aborder ce problème de dimensionnalité est de réduire le nombre de variables, soit en sélectionnant un sous-ensemble représentatif, soit en extrayant de nouvelles variables qui résument au mieux l'information portée par l'ensemble des variables originales. La difficulté de l'analyse des données en grande dimension réside dans le fait que les intuitions géométriques valables en faible dimension peuvent se révéler fausses ou inutiles en grande dimension. C'est pourquoi certains algorithmes efficaces en faible dimension sont peu performants en grande dimension. Par ailleurs l'utilisation des outils d'analyse de données ayant un esprit intuitif dans les faibles dimensions est plus complexe lorsque les données sont en dimension élevée. Dans ces situations, l'intuition du comportement de ces outils est perdue et peut mener à des conclusions erronées quant à leurs résultats [19]. Une difficulté majeure rencontrée lors de l'utilisation des algorithmes d'apprentissage est la malédiction de la dimensionnalité ou fléau de la dimension (Curse of dimensionality en anglais) [20, 21]. Ce terme fait référence à

plusieurs propriétés des données en grande dimension qui rendent l'apprentissage difficile.

1.2.2. Malédiction de la dimensionnalité

La malédiction de la dimension est un terme abondamment utilisé pour caractériser les différentes manifestations de la grande dimension et fait référence aux difficultés de l'analyse et du traitement des données qui apparaissent dans les espaces de grande dimension notamment aux problèmes dont souffrent la majorité des méthodes d'apprentissage en grande dimension. Cette expression trouve ses origines dans les travaux de Bellman [21] et fait référence à la croissance exponentielle de l'espace explorable avec le nombre de dimension. Il est communément admis que les modèles construits par l'apprentissage ne sont valables que dans l'espace où les données d'apprentissage sont disponibles. La généralisation sur des données qui sont différentes de tous les points d'apprentissage est impossible. La généralisation pertinente est possible à partir de l'interpolation (mais pas à partir de l'extrapolation). L'un des ingrédients clés d'un développement réussi des algorithmes d'apprentissage est donc d'avoir suffisamment de données pour l'apprentissage de manière à remplir l'espace où le modèle doit être valide [20]. Le problème se pose lorsque la taille de l'ensemble d'apprentissage est trop faible comparée à la dimension de l'espace alors que le nombre de données d'apprentissage doit croître de façon exponentielle avec la dimension. Cette augmentation exponentielle est la première conséquence de ce qu'on appelle la malédiction de la dimensionnalité [20, 21].

Plus généralement, la malédiction de la dimensionnalité est l'expression de tous les phénomènes qui apparaissent avec les données en grande dimension, et qui ont souvent de mauvaises conséquences sur le comportement et les performances des algorithmes d'apprentissage [20].

1.2.3. Intuitions géométriques en grande dimension

Les problèmes qui se posent dans les espaces de grandes dimensions sont nombreux et il est généralement difficile d'étendre dans ces espaces les intuitions que l'on a dans

les espaces à deux ou à trois dimensions. De nombreux algorithmes d'apprentissage statistique classiques trouvent leur inspiration dans des intuitions géométriques en faible dimension, mais lorsqu'on tente de les appliquer tels quels sur des données en grande dimension, provenant de problèmes concrets, les performances peuvent être décevantes [22]. Une particularité est que l'information de la distance Euclidienne entre deux points en faible dimension est beaucoup plus informative qu'en grande dimension. Ce principe consiste à dire que si ces points appartiennent à un espace de grande dimension, la distance euclidienne perd son sens: on peut en effet montrer que tous les points sont à peu prés à la même distance l'un de l'autre, ce qui rend impossible leur comparaison par la distance. Un autre point qui vaut d'être mentionné est l'importance de l'extrapolation par rapport à l'interpolation. En faible dimension, on a tendance à penser en termes d'interpolation (entre un petit nombre de points d'une courbe par exemple). Mais en grande dimension, la probabilité qu'un point de test appartienne à la fermeture convexe des données d'apprentissage devient très faible. On se trouve donc la plupart du temps en situation d'extrapolation [20, 22].

1.3. Modèles linéaires et non linéaires

1.3.1. Modèles linéaires

1.3.1.1. Introduction

Dans beaucoup de domaines scientifiques, les problèmes les plus souvent rencontrés sont des systèmes à entrées et sorties. Les entrées sont des phénomènes mesurables et souvent contrôlables et les sorties dépendent de ces entrées. Les techniques de régression ont pour objectif d'évaluer les relations entre les entrées et les sorties en l'absence de modèles théoriques. La régression possède une terminologie très riche et très diverse pour désigner les grandeurs de sortie (grandeurs d'intérêt) et les grandeurs d'entrée comme le montre le tableau 1.1. Le but de ces méthodes est ainsi de prédire ou d'expliquer le comportement d'une ou plusieurs variables dites variables expliquées [5, 23], encore appelées variables indépendantes par des variables plus facile à mesurer ou à contrôler appelées variables explicatives ou variables prédictives.

Lorsque ces variables sont peu nombreuses, qu'elles ne sont pas trop redondantes et que leurs liaisons avec les réponses sont suffisamment bien connues, alors la régression linéaire multiple est la technique adéquate pour ajuster un modèle aux données. La technique est décrite dans le paragraphe suivant.

Grandeur d'intérêt y	**Variable x**
Variable expliquée Variable dépendante Variable endogène Réponse Observation Variable d'intérêt	Variable explicative Variable indépendante Variable exogène Facteur Régresseur Prédicteur Contrôle

Tableau 1.1: Principales terminologies des grandeurs d'intérêt et des variables [23].

1.3.1.2. Régression linéaire multiple

La régression linéaire multiple (MLR) est une généralisation à "p" variables explicatives de la régression linéaire simple. Dans la méthode MLR, il s'agit de prédire une variable quantitative "y" à partir de plusieurs variables explicatives $x_1, x_2, ... x_p$ qui sont des variables spectrales prises à différentes longueurs d'onde, ou "p" est le nombre de variables. Le modèle de régression linéaire multiple s'écrit donc ainsi:

$$y = b_0 + b_1 x_1 + ... + b_p x_p + e, \qquad (1.1)$$

où $b_0, b_1, ... b_p$ désignent les paramètres inconnus qui sont à estimer à partir des observations et "e" est le terme d'erreur. Le modèle de la méthode MLR peut s'écrire sous forme matricielle de la façon suivante [24, 25, 26]:

$$y = Xb + e, \qquad (1.2)$$

où X est la matrice des variables explicatives $x_1, x_2, ...x_p$ augmentée avec une colonne de 1. Dans l'équation (1.2), "b" est le vecteur des paramètres inconnus $b_0, b_1, ...b_p$ du modèle dont la méthode d'estimation la plus couramment utilisée est celle des moindres carrées résultant de la minimisation de la somme des carrées des résidus:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \qquad (1.3)$$

la résolution de l'équation (1.2) conduit à la solution optimale selon le critère des moindres carrées:

$$b = (X^T X)^{-1} X^T y, \qquad (1.4)$$

Cette estimation met en œuvre le calcul de la matrice $(X^T X)^{-1}$ qui est l'inverse de la matrice $(X^T X)$. Or l'inversion de cette matrice est impossible lorsque, le nombre d'observations "n" est supérieur au nombre de variables "p". Dans ce cas, la matrice $(X^T X)$ est singulière, c'est-à-dire non inversible [10 26]. Le modèle qui résulte de la régression linéaire multiple s'avère instable. Ces situations sont courantes en spectroscopie. Pour pallier aux problèmes liés à l'application de la régression linéaire multiple (MLR), on fait appel à d'autres méthodes alternatives à la régression linéaire multiple. Parmi celles-ci, la régression sur composantes principales (PCR) et la régression au sens des moindres carrées (PLSR). Les deux méthodes sont décrites dans ce qui suit.

1.3.1.3. Régression sur composantes principales

L'analyse en composantes principales (PCA) est la plus simple et la plus connue des techniques d'analyse de données multivariées. C'est une méthode statistique non supervisée bien adaptée à l'étude exploratoire des données spectrales. Lors de l'utilisation de la méthode PCA on cherche à remplacer les variables originales fortement redondantes, par des variables synthétiques; les composantes principales (scores en langue anglaise), de dimension très réduite par rapport à celle des variables

originales qui peuvent représenter au mieux les variations ou informations contenues dans ces variables d'origine, et qui ont l'avantage d'être non corrélées ou orthogonales entre elles. L'orthogonalité des variables rend les calculs simple et plus fiable. Les données condensées par PCA peuvent servir de variables de base à d'autres traitements statistiques tels que la régression ou l'analyse discriminante [5, 10, 26].

La régression en composantes principales (PCR) est une simple combinaison de l'analyse en composantes principales et de la régression linéaire multiple. Elle consiste à appliquer tout d'abord une PCA sur la matrice des variables indépendantes. Les composantes principales servent ensuite de variables prédictives au lieu des variables d'origine. L'objectif est de trouver les vecteurs propres de la matrice de covariance. Ces vecteurs propres correspondent aux directions des composantes principales des variables originales. Leur signification statistique est donnée par leurs valeurs propres correspondantes [5, 6, 11]. Usuellement une étape connue sous le nom de prétraitement (preprocessing) est effectuée pour que la matrice X soit centrée (moyenne nulle) et réduite (écart type =1). Une fois générée la matrice de covariance $C = \frac{1}{n} X^T X$ entre les variables originales, il suffit de rechercher les vecteurs propres associés aux valeurs propres de cette matrice. Les composantes principales sont classées selon leur importance décroissante: la première colonne de C véhicule, en principe, une information plus importante que la dernière. Les scores des plus importantes composantes principales sont alors employés comme entrées pour la régression linéaire multiple avec la variable dépendante y [5, 26].

Pour plus de détails sur la méthode PCR, le lecteur peut se référer à [5, 6, 11].

1.3.1.4. Régression au sens des moindres carrées

La régression au sens des moindres carrées partielles (PLSR) est la méthode d'estimation la plus communément utilisée en spectroscopie infrarouge, et à bien des égards la plus importante. Comme dans la méthode PCA, le modèle PLS (Partial Least Square) consiste à extraire des données originales des nouvelles variables non corrélées appelées variables latentes qui ne sont autres que des composantes

principales des variables d'origine. La méthode PLSR diffère de la méthode PCR sur la construction de ces nouvelles variables [5, 12, 26, 27, 28]. En effet, ces variables sont choisies pour représenter au mieux les variables originales tout en modélisant la relation entre X et y. En d'autres termes, la méthode PLSR réduit la dimension des variables (explicatives et réponses) en faisant une projection sur la direction suivant laquelle l'inter-covariance entre X et y est maximum.

Le lecteur peut se référer à [5, 6, 11, 12, 27, 28] pour plus de détails sur la méthode PLSR.

1.3.2. Modèles non linéaires

1.3.2.1. Réseaux de neurones artificiels

1.3.2.1.1. Introduction

Aujourd'hui de nombreux termes sont utilisés dans la littérature pour désigner le domaine des réseaux de neurones artificiels ou formels, comme connexionnisme ou neuromimétique. Les réseaux de neurones artificiels sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle avec une topologie spécifique d'interconnexions entre ces éléments et une loi d'apprentissage pour adapter les poids de connexions. Un réseau de neurones artificiel est parfaitement défini par le nombre de ses cellules élémentaires, la nature des fonctions d'activation ou de décision, le nombre de couches et les liaisons entre les différentes cellules. On distingue deux types d'architectures de réseaux de neurones artificiels: les réseaux de neurones statiques ou non bouclés et les réseaux de neurones dynamiques ou bouclés [29 30, 31]. Ce terme générique recouvre en fait une série de techniques de modélisation, aussi bien prédictive (réseaux supervisés) que descriptive (réseaux non supervisés). Les réseaux de neurones supervisés sont tout simplement les techniques de régression et de classification les plus puissantes à ce jour; à l'inverse des techniques classiques (régression linéaire multiple, analyse discriminante), ils se dispensent de l'hypothèse très contraignante sur le couplage linéaire entre variables explicatives et variable à expliquer, ou de frontières linéaires entre classes et pourtant,

malgré cette puissance de représentation, ils ont un remarquable pouvoir de généralisation [32].

1.3.2.1.2. Apprentissage des réseaux neuronaux

L'apprentissage ou l'entrainement est la propriété la plus intéressante des réseaux neuronaux. Elle peut être décrite comme étant une phase du développement d'un réseau de neurones artificiel durant laquelle le comportement du réseau est modifié jusqu'à l'obtention du comportement désiré. L'objectif essentiel d'un tel apprentissage est d'adapter les paramètres du réseau (les poids des connexions) pour qu'il soit capable de reconnaître des exemples n'appartenant pas à l'ensemble d'apprentissage [33]. L'apprentissage est dit supervisé lorsque les exemples sont constitués de couples du type : (entrée, sortie désirée). Ce type d'apprentissage se fait toujours par l'intermédiaire d'un critère à optimiser définissant la performnce du réseau à chaque étape. Les coefficients synaptiques (poids) sont alors ajustés dans le but de minimiser un critère de coût. Une fois l'apprentissage est achevé, le réseau peut être opérationnel [33, 34]. L'apprentissage est qualifié de non supervisé lorsque seules les valeurs d'entrée sont disponibles. Cet apprentissage se fait sur la base d'informations locales existant aux niveaux des neurones et découvre les propriétés collectives qui existent entre les données sur la base desquelles le réseau doit s'organiser de façon à optimiser une certaine fonction de coût. Cette propriété est dite auto-organisation (self organisation) [33, 35]. Les Réseaux de neurones supervisés les plus répandus sont les réseaux perceptrons multicouches MLP, et les réseaux à fonctions radiales de base RBF.

1.3.2.1.3. Réseaux perceptron multicouches

Les réseaux de neurones artificiels perceptrons multicouches MLP consistent à cascader un certain nombre de perceptrons en plusieurs couches avec des fonctions de décision différentiables reliées entre elles par des coefficients synaptiques (poids) et sont organisés de la manière suivante [33] : une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie (voir figure 1.1). Les neurones de deux

couches consécutives sont entièrement connectés. Le nombre total de couches d'un réseau est en général donné par celui des couches cachées plus la couche de sortie. L'apprentissage de ces réseaux consiste à trouver les valeurs optimales des différents vecteurs de poids. L'approche classique la plus utilisée est la rétro-propagation (back-propagation) [29, 30, 34]. Mathématiquement, cette méthode est basée sur l'algorithme de la descente du gradient et utilise les règles de dérivation des fonctions dérivables. Dans cette méthode, l'erreur commise en sortie du réseau est rétro-propagée vers les couches cachées, d'où le nom de rétro-propagation. En effet, la convergence de l'algorithme est basée sur l'ajustement de multiples variables, citons par exemple : la structure du réseau, le nombre maximum d'itérations d'apprentissage, le coefficient d'apprentissage, nature de la fonction d'activation, nombre de couches cachées, la taille de la couche cachée. Néanmoins, ces réseaux sont connus pour avoir un certain nombre d'inconvénients:1) la sensibilité aux minimums locaux lors de l'étape d'optimisation des poids qui empêche la convergence et cause l'oscillation de l'erreur, 2) l'absence de résultats théoriques satisfaisantes qui permettent de dimensionner correctement un réseau (combien de couches cachées faut-il utiliser? et combien doit-il y avoir de neurones dans chacune des couches cachées?), seule l'expérience permet de donner une idée sur ce choix [30].

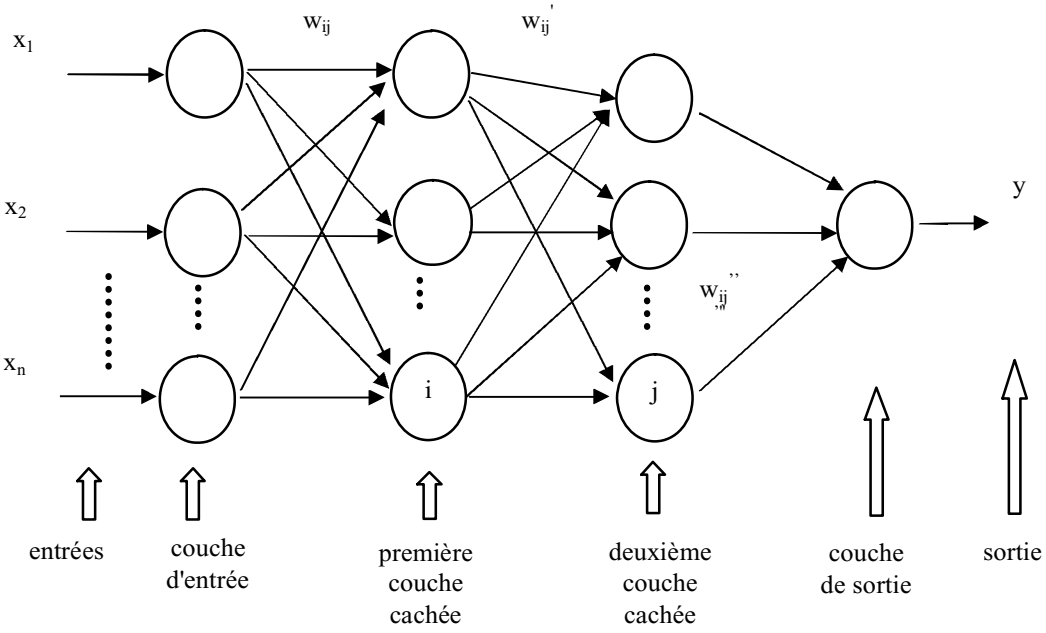


Figure 1.1 : Architecture d'un réseau de neurones artificiel perceptrons multicouches MLP avec deux couches cachées.

La sortie de l'unité j par exemple de la couche q est donnée par :

$$y_j^q = f\left(\sum_{i=1}^{N_{q-1}} w_{ij}^q y_i^{q-1}\right), \qquad (1.5)$$

où w_{ij}^q représente le poids entre le neurone "i" de la couche "q-1" et le neurone "j" de la couche "q", y_i^{q-1} est la sortie de l'unité i dans la couche q-1, et "f" représente la fonction d'activation ou de décision pour toutes les unités des couches cachées. C'est une fonction différentiable non linéaire. Le plus souvent, la fonction d'activation utilisée est la fonction sigmoïde définie pour tout réel x par [30] :

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (1.6)$$

Le lecteur peut se référer à [30, 31, 35, 36, 37] pour plus de détails sur les réseaux de neurones artificiels perceptrons multicouches MLP.

1.3.2.1.4. Réseaux à fonctions radiales de base

Les réseaux de neurones artificiels perceptrons multicouches MLP présentés ci-dessus offrent l'avantage de pouvoir tenir compte de relations non linéaires entre les données spectrales et les variables à prédire. Parmi les méthodes alternatives permettant de tenir compte de la non linéarité, mais conceptuellement plus simples sont les réseaux de neurones artificiels à fonctions radiales de base RBF. La construction d'un réseau de neurones artificiel RBF est rapide et facile, et c'est là le principal avantage de ce type de réseau de neurones. Un réseau de neurones artificiel RBF comporte deux couches de neurones (voir figure 1.2); une couche cachée dont les neurones sont connectés à ceux de la couche d'entrée par des connexions non pondérées, et une couche de sortie dont les neurones sont connectés à ceux de la couche cachée par des connexions pondérées. Dans la couche de sortie, la fonction d'activation est linéaire. La sortie réalisée par le réseau est exprimée sous la forme d'une combinaison linéaire de fonctions radiales [30]:

$$y(x) = \sum_{j=1}^{k} \lambda_j \Phi\left(\left\| x - c_j \right\|\right), \qquad (1.7)$$

où k représente le nombre de neurones dans la couche cachée, λ_j est le poids associé à la connexion entre le neurone j de la couche cachée et la couche de sortie et c_j est la valeur des centres.

Dans la couche cachée, la fonction radiale la plus couramment utilisée est le noyau gaussien [30]:

$$\Phi\left(\left\|x - c_j\right\|\right) = \exp-\left(\frac{\left\|x - c_j\right\|^2}{2\delta_j^2}\right), \qquad (1.8)$$

où δ_j est la largeur associée au noyau j, $x=(x_1,\ldots.x_n)$ et $c_i=(c_{i1}\ldots..c_{in})$ représentent le vecteur d'entrée et le vecteur "centre" propre à chaque neurone respectivement. Dans un réseau de neurones artificiel RBF il y a quatre paramètres principaux à régler.

- Le nombre de neurones k dans la couche cachée.
- La position des centres des gaussiennes c de chacun des neurones.
- La largeur de ces gaussiennes δ_j (généralement l'écarts-types des noyaux gaussiens)
- Les poids des connexions λ_i entre les neurones de la couche cachée et les neurones de la couche de sortie.

La procédure d'apprentissage dans ce cas est divisée en deux étapes, un apprentissage non supervisé dans la couche cachée suivi par un apprentissage supervisé dans la couche de sortie [38]. L'algorithme des K-means et la méthode des moindres carrés sont souvent utilisées pour la sélection des centres "c" et l'adaptation des poids de connexions λ respectivement. Comparés aux réseaux multicouches, l'apprentissage des réseaux RBF est plus rapide. En effet cette rapidité vient du fait que ce type de réseaux possède deux couches seulement ($c+\delta$ et λ) [26].

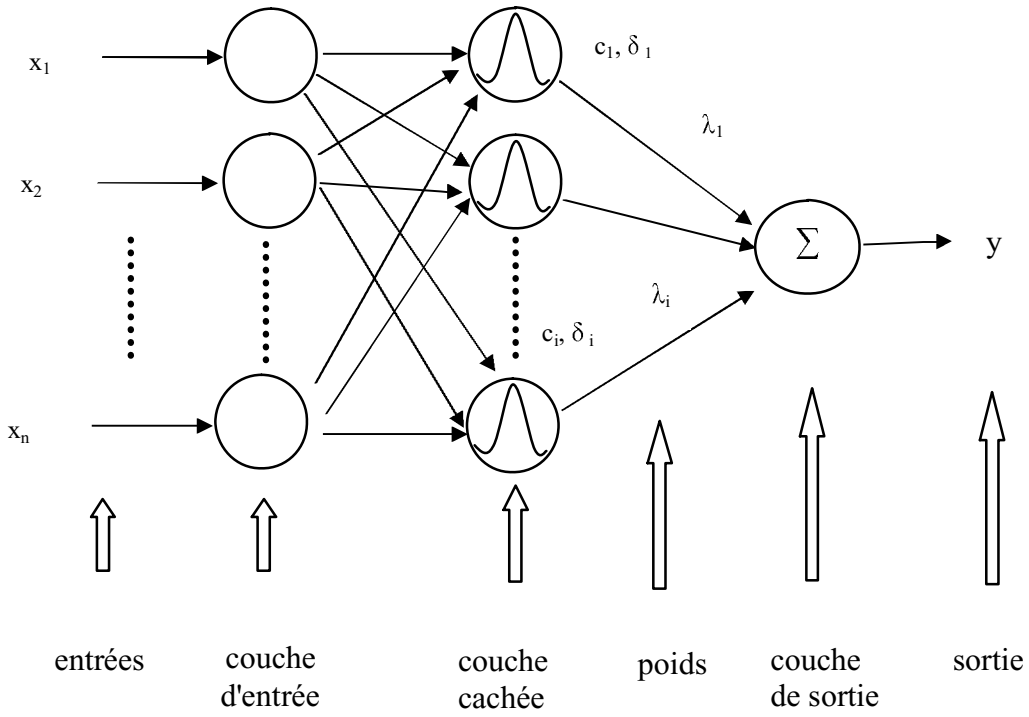


Figure 1.2: Architecture d'un réseau de neurones artificiel à fonctions radiales de base RBF.

Le lecteur pourra se référer à [16, 39, 40, 41, 42] pour plus de détails sur les réseaux de neurones artificiels à fonctions radiales de base RBF.

1.3.2.2. Machine à Vecteurs Supports

1.3.2.2.1. Introduction

Les machines à vecteurs de support SVM (Support Vector Machine) développées au cours des années 90 par Vapnik, sont une famille d'algorithmes d'apprentissage destinées à résoudre les problèmes de classification et de régression qui sont aujourd'hui considérées comme une des méthodes les plus performantes sur de nombreux problèmes réels, notamment pour les problèmes en grande dimension. L'efficacité de SVM est due à sa base théorique solide qui constitue l'un de ses atouts. Les SVM sont appelées aussi « séparateurs à vaste marge » dans le cas de la classification dont le principe est décrit dans ce qui suit.

1.3.2.2.2. Classification par Machines à Vecteurs de Support

A. Principe

L'algorithme des machines à vecteurs de support a initialement été développé comme un algorithme de classification binaire par apprentissage supervisé. L'algorithme sous sa forme initiale revient à chercher une frontière de décision linéaire appelé hyperplan qui permet de regrouper les données similaires dans une même classe et de séparer les données hétérogènes, en garantissant que la marge entre l'hyperplan et les points les plus proches de chaque classe soit maximale. Le problème de recherche de l'hyperplan séparateur optimal possède une formulation duale. Ceci est particulièrement intéressant car, sous cette formulation duale, le problème peut être résolu au moyen de méthodes d'optimisation quadratique standard. L'intérêt de cette méthode est la sélection de vecteurs supports qui représentent les vecteurs discriminants grâce auxquels est déterminé l'hyperplan. Les exemples utilisés lors de la recherche de l'hyperplan ne sont alors plus utiles et seuls ces vecteurs supports sont utilisés pour classer un nouveau cas, ce qui peut être considéré comme un avantage pour cette méthode [43].

B. SVM linéaire: cas linéairement séparable

Etant donné un ensemble d'apprentissage $\{(x_i, y_i)\}_{i=1}^{N}$ où $x_i \in R^d$ et $y_i \in \{-1,1\}$ y_i définissant la classe d'un exemple donné. L'objectif de SVM est de trouver un hyperplan permettant de séparer l'ensemble d'apprentissage de sorte que tous les points d'une même classe soient d'un même côté de l'hyperplan. Cela équivaut à trouver un hyperplan $f(x)$ définit comme [44]:

$$f(x) = w.x + b, \qquad (1.9)$$

où w et b sont les paramètres à estimer. Pour cet hyperplan la marge vaut $1/\|w\|$, et donc la recherche de l'hyperplan optimal parmi les hyperplans valides revient à minimiser $\|w\|$, soit à résoudre le problème suivant:

$$\begin{cases} \text{minimiser} & \frac{1}{2}\|w\|^2 \\ \text{tel que} & y_i(w.x_i + b) \geq 1 \qquad i = 1...N \end{cases}, \tag{1.10}$$

cette formulation primale possède une forme duale dans le cas ou la fonction objective et les contraintes sont strictement convexes. Dans ce cas, la résolution de l'expression duale du problème est équivalente à la solution du problème original. Le problème devient celui de la recherche de paramètres a vérifiant le système d'équations suivant:

$$\begin{cases} \text{maximiser} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i.x_j) \\ \text{tel que} & \sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{et} \quad \alpha_i \geq 0, \qquad i = 1...N \end{cases}, \tag{1.11}$$

où les "α" représentent les multiplicateurs de Lagrange. L'hyperplan solution correspondant peut alors être écrit:

$$f(x) = \sum_{i \in S} \alpha_i y_i (x_i \cdot x) + b, \tag{1.12}$$

où S est le sous ensemble d'échantillons d'apprentissage correspondant aux multiplicateurs de Lagrange non nuls.

C. SVM linéaire: cas non linéairement séparable

La formulation SVM décrite dans le paragraphe précédent n'est valable que si les données sont linéairement séparables. Une telle condition est difficile à satisfaire dans la classification des données réelles. Afin de gérer les données non séparables, on part du problème primal linéaire et on introduit des variables ξ appelées variables d'écart (slack variables en anglais). La nouvelle fonction de coût est définie comme [44]:

$$\begin{cases} \text{minimiser} \quad \psi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{tel que} \quad y_i(w.x_i + b) \geq 1 - \xi_i \quad \text{et} \quad \xi_i \geq 0, \quad i = 1...N \end{cases}, \tag{1.13}$$

où C est un paramètre de régularisation doit être choisi par l'utilisateur. On obtient alors, comme dans le cas séparable, une formulation duale donnée par :

$$\begin{cases} \text{max imiser} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i.x_j) \\ \text{tel que} \quad \sum_{i=1}^{N}\alpha_i y_i = 0 \quad \text{et} \quad \forall i, \quad 0 \leq \alpha_i \leq C, \end{cases}, \tag{1.14}$$

Il est intéressant de noter que, dans le cas non séparable, deux types de vecteurs de support coexistent: 1) vecteurs de support de marge qui se trouvent sur la marge hyperplan et 2) vecteurs de support qui tombent sur le mauvais côté de cette marge (voir figure 1.3).

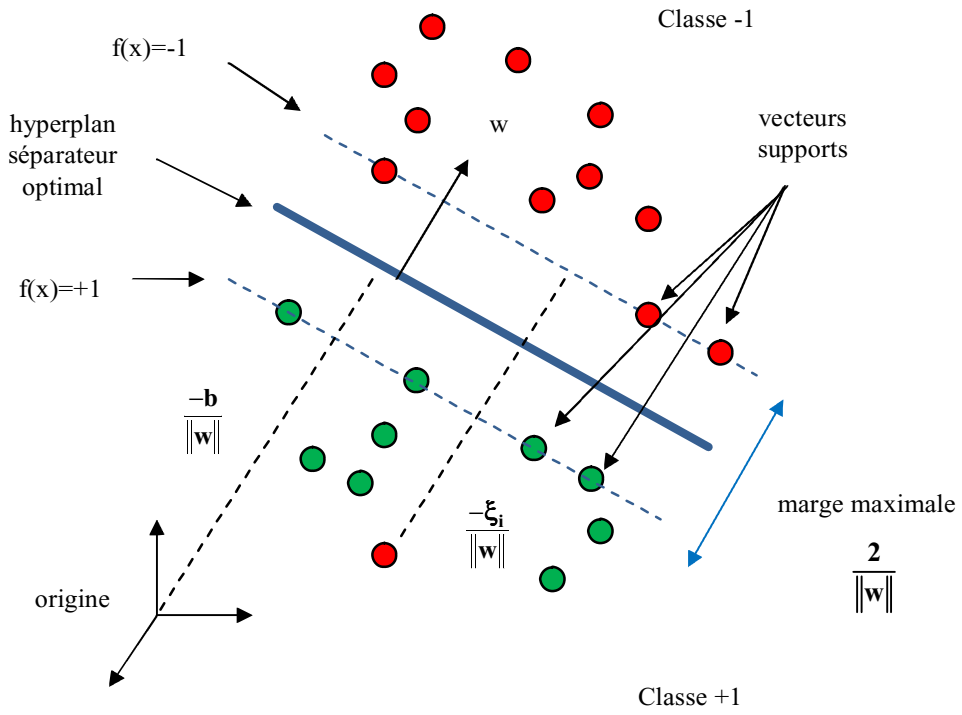


Figure 1.3: Hyperplan séparateur optimal dans SVM pour un cas non linéairement séparable. Les cercles verts et rouges renvoient aux classes "+1" et "-1," respectivement [44].

D. Méthode de classification non linéaire

1. Principe

Séparer n'importe quel jeu de données par un simple hyperplan est un problème qui n'est pas évident. La section précédente décrit le principe des SVM dans le cas ou les données sont linéairement séparables. C'est une limitation sévère qui condamne à ne pouvoir résoudre que des problèmes particuliers. Cependant, dans la plupart des problèmes réels, ce n'est pas toujours le cas et il est donc nécessaire de contourner ce problème. Dans ce but, l'idée clé de la méthode SVM est de changer l'espace des données. Le principe consiste à projeter les données d'apprentissage x_i dans un espace de dimension (d'), plus élevée que celle de l'espace d'origine (d) grâce à une

fonction non linéaire $\Phi(x)$ qu'on appelle fonction noyau, choisie a priori et de réaliser une séparation linéaire dans le nouvel espace. L'espace ainsi obtenu est appelé espace des caractéristiques ou aussi espace transformé ou espace d'attributs ou encore espace de redescription [45]. La figure 1.4 donne une illustration de ce principe.

En d'autre terme, L'idée est de transformer un problème de séparation non linéaire dans l'espace original en un problème de séparation linéaire dans l'espace de redescription de plus grande dimension. En effet, intuitivement, plus la dimension de l'espace de redescription est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée.

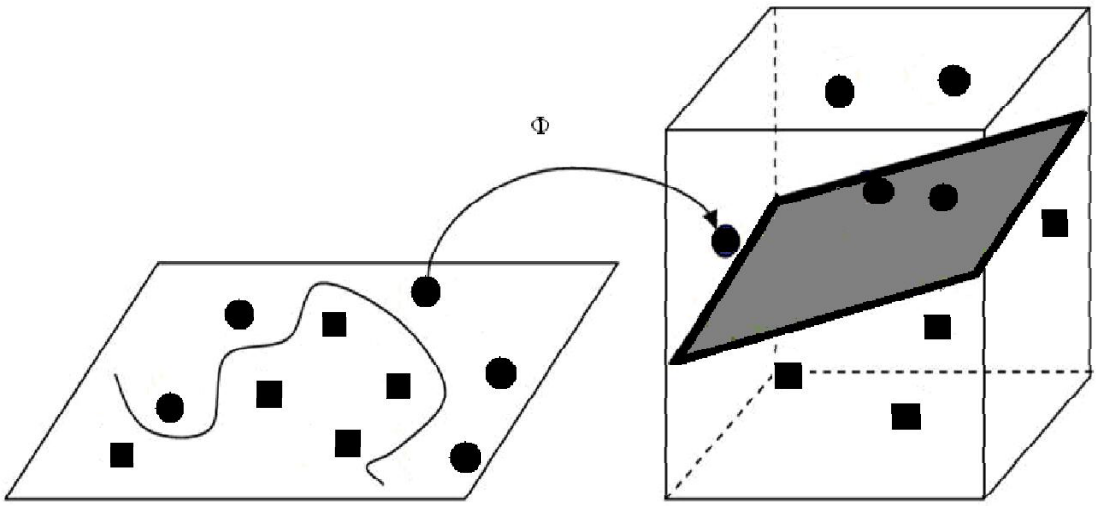


Figure 1.4: Passage des données de l'espace d'entrée vers un espace de redescription où les données sont linéairement séparables [46].

2. Formulation mathématique

La résolution du problème dual tel que celui défini dans les équations (1.11) pour le cas linéairement séparable se fait par substitution des produits scalaires dans l'espace d'origine $(x_i \cdot x_j)$ avec des produits scalaires dans l'espace transformé $(\Phi(x_i) \cdot \Phi(x_j))$. À ce stade, le principal problème est le calcul explicite de $\Phi(x)$, ce qui peut s'avérer coûteux et parfois irréalisables. La méthode du noyau ou encore appelée astuce du

noyau (kernel trick) fournit un moyen efficace pour traiter ce problème. Le noyau K (·,·) doit être choisi de telle sorte qu'il satisfait la condition imposée par le théorème de Mercer [47], de sorte qu'il peut correspondre à un certain type de produit scalaire dans le nouvel espace des caractéristiques de dimension supérieur. Le théorème de Mercer explicite les conditions que le noyau K(·,·) doit satisfaire pour être une fonction noyau : elle doit être symétrique, semi-définie positive, c'est à dire,

$$K(x_i, x) = \Phi(x_i) \cdot \Phi(x), \tag{1.15}$$

ce type de fonction noyau permet de simplifier la solution du problème dual considérablement, car elle évite le calcul des produits scalaires dans l'espace transformé comme dans:

Le problème d'optimisation dans ce cas est donné par:

$$\begin{cases} \text{maximiser} & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{tel que} & 0 \le \alpha_i \le C \quad \text{et} \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}, \tag{1.16}$$

où C est un paramètre de régularisation prédéfini et les "α" sont les multiplicateurs de Lagrange ou encore appelées variables duales. L'équation de l'hyperplan séparateur dans le nouvel espace devient:

$$f(x) = \sum_{i \in S} \alpha_i y_i K(x_i, x) + b. \tag{1.17}$$

L'intérêt de la fonction noyau est que le calcul se fait dans l'espace d'origine, ceci est beaucoup moins coûteux qu'un produit scalaire en grande dimension. En plus la transformation Φ n'a pas besoin d'être connue explicitement, seule la fonction noyau intervient dans les calculs. Les fonctions noyaux les plus courantes sont [38]:

Fonction linéaire:

$$K(x_i, x) = x_i \cdot x. \tag{1.18}$$

Fonction polynomiale:

$$K(x_i, x) = (x_i.x + 1)^n , \tag{1.19}$$

où "n" représente le degré du polynôme choisi a priori par l'utilisateur.

Fonction radiale de base:

$$K(x_i, x) = \exp\left(-\gamma \|x_i - x\|^2\right), \tag{1.20}$$

où γ est un paramètre inversement proportionnelle à la largeur du noyau Gaussien, spécifié a priori par l'utilisateur.

1.3.2.2.3. Régression par Machines à Vecteurs de Support

A. Principe

A l'origine conçue pour les tâches de classification ou reconnaissance de formes, les SVM permettent également de traiter les problèmes de régression non linéaire. L'idée principale d'un SVR (Support Vector Regression) est de trouver une fonction qui ajuste les données avec un écart inférieur à une quantité donnée ε pour chaque couple (x_i, y_i) de l'ensemble d'apprentissage. Le problème de régression peut être définie comme suit: étant donné un ensemble de données d'apprentissage $\{(x_i, y_i)\}_{i=1}^{N}$, des vecteurs d'entrée x_i ($x_i \in R^d$) et des sorties correspondantes y_i. L'objectif consiste à ajuster une fonction plate (flat function) $f(x)$ aux points de l'ensemble des données. Tout algorithme de régression pratique possède une fonction de perte (loss function) L (y,f(x)) = $|\xi|_\varepsilon$ qui décrit la déviation de la fonction estimée par rapport à la fonction réelle. De nombreuses formes de la fonction de perte peuvent être trouvées dans la littérature: par exemple, linéaire, quadratique, exponentielle, etc…. Dans ce travail, la

fonction de perte de Vapnik est utilisée, qui est connu comme la fonction de perte ε-insensible et définie comme :

$$|\xi|_\varepsilon = \begin{cases} 0, & \text{si } |y - f(x)| \le \varepsilon \\ |y - f(x)| - \varepsilon, & \text{autrement} \end{cases}, \qquad (1.21)$$

où ε représente une constante positive prédéfinie.

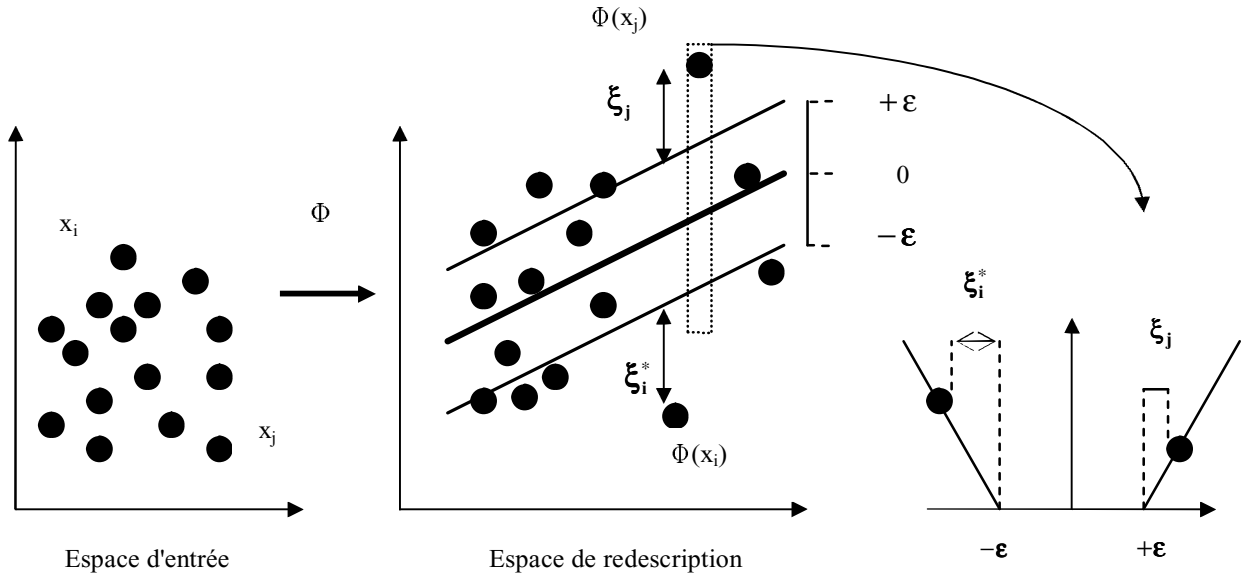


Figure 1.5 : Exemple de ε-insensible tube et la fonction d'erreur utilisée dans la technique [48].

B. Description de l'algorithme SVR

La méthode de régression par SVM ε-insensible [37, 38], est basée sur l'idée de trouver une fonction $f(x)$: 1) qui présente au plus une déviation maximale de ε par rapport aux sorties désirées y_i (i=1,…, N) et 2) et qui soit aussi lisse (smooth) que possible. En d'autres termes, l'algorithme de régression ne se soucie pas des erreurs aussi longtemps qu'ils sont moins de ε mais n'acceptera pas tout écart plus grand que cela. Ce qui est généralement effectué par le mappage des données de l'espace

d'origine à un espace de dimension supérieure c'est à dire $\Phi(x) \in \Re^{d'}$ (d' > d), pour augmenter la platitude (flatness) de la fonction et, par conséquent, de l'approximer d'une manière linéaire, comme suit:

$$f(x) = w^*.\Phi(x) + b^*, \tag{1.22}$$

où w^* et b^* sont les paramètres de la fonction à estimer. La fonction linéaire optimale dans l'espace des caractéristiques de dimension supérieure est celle qui minimise une fonction de coût et qui exprime une combinaison de deux critères: minimisation de la norme euclidienne (ce qui équivaut à la maximisation de la platitude) et la minimisation d'erreur. La fonction de coût est définie comme [50]:

$$\Psi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*), \tag{1.23}$$

cette minimisation de la fonction de coût est soumise sous les contraintes suivantes:

$$\begin{cases} y_i - (w.\Phi(x_i) + b) \le \varepsilon + \xi_i \\ (w.\Phi(x_i) + b) - y_i \le \varepsilon + \xi_i^* \qquad i = 1...N, \\ \text{et} \quad \xi_i, \xi_i^* \ge 0 \end{cases} \tag{1.24}$$

où, ξ_i et ξ_i^* sont les variables d'écart ou encore appelées variables ressort (slack variables) introduites pour tenir compte des échantillons non séparables (qui ne résident pas dans ε-déviation tube). La constante C représente un paramètre de régularisation qui permet d'ajuster le compromis entre la platitude (flatness) de la fonction $f(x)$ et la tolérance des déviations supérieur à ε. Cela correspond à traiter la fonction de perte dite ε-insensible décrite précédemment. Cela signifie que les différences entre les valeurs estimées et les cibles sont tolérées à l'intérieur du tube

(les erreurs les plus petites de ε), tandis qu'une pénalité linéaire est affecté à des estimations situées en dehors de ε- insensibles tube (voir l'exemple sur la figure 1.5).

Le problème d'optimisation ci-dessus peut être reformulé à travers une fonctionnelle de Lagrange. Les multiplicateurs de Lagrange peuvent être trouvés par une optimisation duale conduisant à une solution QP (quadratic programming) [50, 51].

Nous pouvons reformuler le problème dual de la régression non linéaire par SVM comme suit: étant donné l'ensemble d'apprentissage $\{(x_i, y_i)\}_{i=1}^{N}$, trouver les multiplicateurs de Lagrange $\{\alpha_i\}_{i=1}^{N}$ et $\{\alpha_i^*\}_{i=1}^{N}$ qui maximisent la fonction objective:

$$\begin{aligned} Q(\alpha_i, \alpha_i^*) &= \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{N} y_i(\alpha_i + \alpha_i^*) \\ &- \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j), \end{aligned} \tag{1.25}$$

sous les contraintes suivantes:

$$\begin{aligned} &\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0 \\ &0 \le \alpha_i \le C, \qquad i=1, 2\ldots N, \\ &0 \le \alpha_i^* \le C, \end{aligned} \tag{1.26}$$

Le résultat final est la fonction exprimée par l'équation suivante :

$$f(x) = \sum_{i \in S} (\alpha_i - \alpha_i^*) K(x_i, x) + b^*, \tag{1.27}$$

où, K (·,·) est une fonction noyau. S est le sous-ensemble des indices (i = 1, 2, ..., N) correspondant aux multiplicateurs de Lagrange non nuls αi ou αi*. Les échantillons d'apprentissage associés aux poids non nuls sont appelés *vecteurs de support*. Le noyau K (·,·) doit être choisi de telle sorte qu'il satisfait la condition imposée par le théorème de Mercer [47], de sorte qu'il peut correspondre à un certain type de produit scalaire dans le nouvel espace des caractéristiques de dimension supérieur.

Nous renvoyons le lecteur aux références [49, 50, 51, 52, 53] pour plus de détails sur la théorie de l'estimation par SVM.

1.4. Conclusion

Dans ce présent chapitre nous avons vu une présentation des données en grande dimension, dans laquelle une idée s'est donnée sur les principales notions de base qui peuvent nous aider à comprendre les problèmes des données en grande dimension. Nous avons passé en revue certaines techniques utilisées dans le cadre de l'analyse de ces données à savoir les modèles linéaires et les modèles non linéaires. Nous avons d'abord présenté la régression linéaire multiple (MLR) puis, nous avons abordé la régression sur composantes principales (PCR) et la régression au sens des moindres carrées (PLSR). A la suite de cette partie, nous avons entamé les modèles non linéaires. Nous avons présenté deux structures de réseaux de neurones artificiels, les réseaux de neurones artificiels perceptrons multicouches (MLP) et les réseaux de neurones artificiels à fonctions radiales de base (RBF). A la fin de ce chapitre l'approche des machines à vecteurs supports (SVM) a été présentée pour les deux cas de classification et de régression.

Chapitre 2
Description des bases de données

2.1. Introduction

Dans de nombreux domaines, en particulier dans l'agroalimentaire, la spectroscopie dans l'infrarouge proche et moyen est la technique la mieux adaptée à l'analyse et à la caractérisation des produits agroalimentaires et donne lieu à des applications analytiques très diverses. Une raison de son efficacité provient de la nature de l'information spectrale qu'elle permet d'acquérir. Elle repose en effet sur l'acquisition rapide d'un grand nombre de données spectrales.

L'objectif de ce chapitre est tout d'abord donner une introduction sur les données spectrophotométriques. Par la suite, les régions spectrales d'intérêt analytique sont succinctement exposées. La description des données spectrales et le principe de la loi de Beer-Lambert sont ensuite présentés. La dernière partie de ce chapitre est consacrée à la description des bases de données utilisées dans ce travail.

2.2. Données spectrophotométriques

2.2.1. Introduction

La spectroscopie est la science de l'étude de l'interaction des ondes électromagnétiques (lumière) avec la matière. La spectrométrie est l'application de techniques spectroscopiques pour caractériser des échantillons. Historiquement, ce terme s'appliquait à la décomposition, par exemple par un prisme, de la lumière visible émise (spectrométrie d'émission) ou absorbée (spectrométrie d'absorption) par l'objet à étudier. Aujourd'hui, ce principe est décliné en une multitude de techniques expérimentales spécialisées qui trouvent des applications dans quasiment tous les domaines de la physique au sens large. La spectrophotométrie est une méthode analytique quantitative qui consiste à mesurer l'absorbance ou la densité optique d'une substance chimique donnée, généralement en solution. Plus l'échantillon est concentré, plus il absorbe la lumière dans les limites de proportionnalité énoncées par la loi de Beer-Lambert [13]. Dans le domaine de l'analyse chimique, les informations recueillies sont utilisées pour caractériser la matière, sa nature, sa composition, sa structure. Le plus souvent (sauf en spectrométrie de masse), on décompose le rayonnement électromagnétique en ses différentes longueurs d'onde pour effectuer ensuite des mesures d'intensité sur le spectre obtenu [13].

2.2.2. Régions spectrales d'intérêt analytique

La lumière est caractérisée par sa longueur d'onde. De nombreuses plages de longueurs d'onde présentent un intérêt analytique. Parmi celles-ci, les plages spectrales de l'infrarouge et du proche infrarouge sont particulièrement adaptées à l'analyse des aliments. Si l'on parcourt le spectre électromagnétique dans l'ordre des énergies décroissantes on rencontre successivement :

Plage spectrale	Longueurs d'onde	Caractéristiques
Région des rayons X	0.05 nm-10 nm	Met en jeu les changements

		énergétiques des électrons des couches internes des atomes et des molécules.
Ultraviolet lointain	10 nm- 200 nm	Plage de l'émission électronique à partir des orbitales de valence et peut être utilisé en spectroscopie photo-électronique,
Ultraviolet proche	200 nm-400 nm	On observe les transitions électroniques des niveaux énergétiques des orbitales de valence. Cette plage spectrale est d'un intérêt particulier pour l'analyse des molécules insaturées. cette plage est le siège principal des phénomènes de luminescence, qui peuvent présenter un grand intérêt pour l'analyse des aliments.
Visible	400nm-800 nm	Plage spectrale où se produisent des transitions électroniques. La spectroscopie dans le visible est évidemment d'une grande importance pratique pour la caractérisation de la couleur des denrées alimentaires. Elle est également mise en œuvre dans de nombreuses méthodes colorimétriques.

Proche infrarouge	800 nm -2500 nm	Première plage spectrale en relation avec les vibrations des molécules. Il est le siège des harmoniques et des bandes de combinaison. il présente un grand intérêt pratique pour les applications analytiques dans l'agroalimentaire.
Moyen infrarouge	2500nm-25000nm	Région principale de la spectroscopie vibrationnelle. Sur le plan théorique, cette plage spectrale est particulièrement bien adaptée à l'identification des composés organiques et à l'étude de la conformation des molécules.
Région des micro-ondes	100 µm et 1cm	L'énergie absorbée est en liaison avec les rotations des molécules. Cette plage spectrale n'a reçu qu'un nombre limité d'applications agroalimentaires. Elle peut cependant être exploitée pour étudier les milieux aqueux hétérogènes et pour le dosage de l'eau dans les aliments.
Fréquences radio	1cm à 10 m	Est le siège de la résonance magnétique nucléaire et de la résonance électronique de spin. Les changements énergétiques sont ici liés au renversement des spins des

		noyaux ou des électrons.

Tableau 2.1: Régions spectrales présentant un intérêt pour le développement d'applications analytiques [5].

Il convient de noter ici que la plage spectrale du proche infrarouge présente de nombreux avantages pratiques et un grand intérêt pour les applications analytiques dans le domaine de l'agroalimentaire. Les bases de données utilisées dans ce travail de recherche sont représentées par les plages spectrales en proche et moyen infrarouge.

2.2.3. Description des données spectrales

Les données spectrales possèdent des caractéristiques particulières qui doivent être prises en compte pour obtenir des modèles prédictifs stables et performants [26, 54].

L'étape de calibration ou d'étalonnage du modèle est en général effectuée expérimentalement. On dispose alors d'une matrice X, avec "n" observations (spectres) et "p" variables (points de mesure) formant les données prédictives, et un vecteur "y" de "n" éléments correspondants aux valeurs de référence (voir figure 2.1).

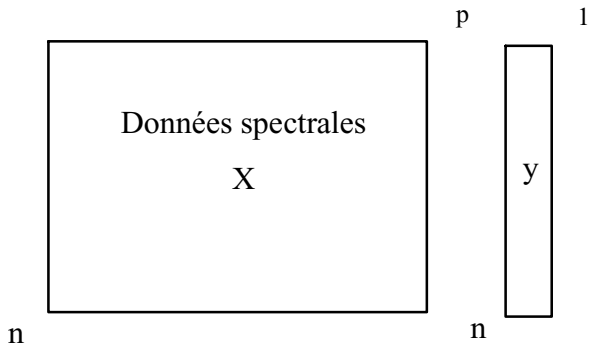


Figure 2.1 : Structure d'une collection de données spectrales utilisée pour un étalonnage.

Le spectre est le signal formé par l'acquisition de mesures à différentes longueurs d'onde. Les spectres sont par principe, continus, mais ils ne peuvent être enregistrés et

stockés que sous la forme d'une succession de mesures effectuées en un nombre limité de points. Avant d'appliquer une méthode chimiométrique, il est nécessaire de créer une collection de spectres sur laquelle portera l'étude et qui servira à établir un modèle. La figure 2.2 montre une collection de 218 spectres de jus d'orange dans la plage de longueurs d'onde allant de 1100 à 2500 nm, avec un pas de 2 nm. Ces données sont regroupées dans la matrice X comprenant dans cet exemple 218 lignes et 700 colonnes.

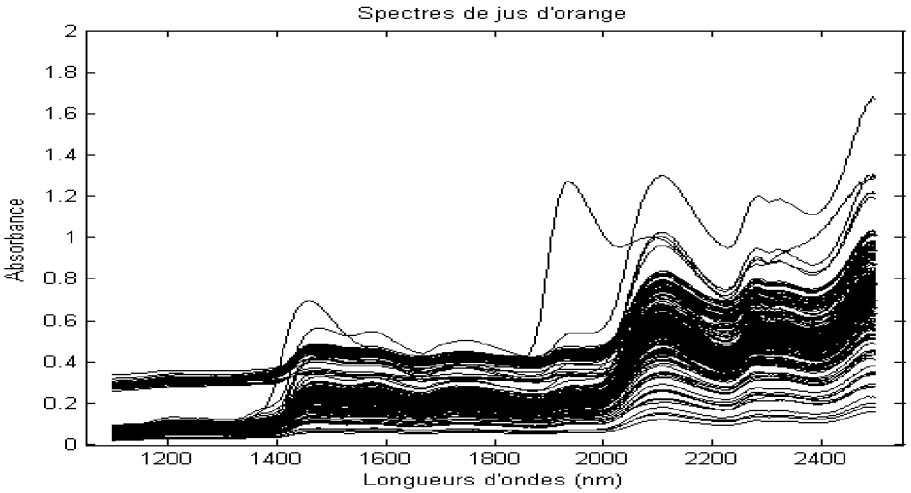
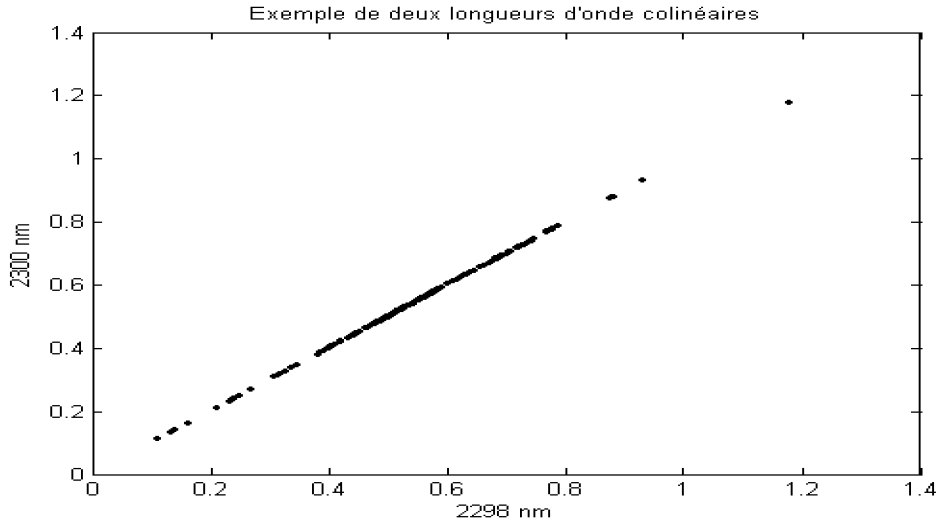


Figure 2.2 : Exemple de collection spectrale : 218 spectres de jus d'orange dans le proche infrarouge [55].

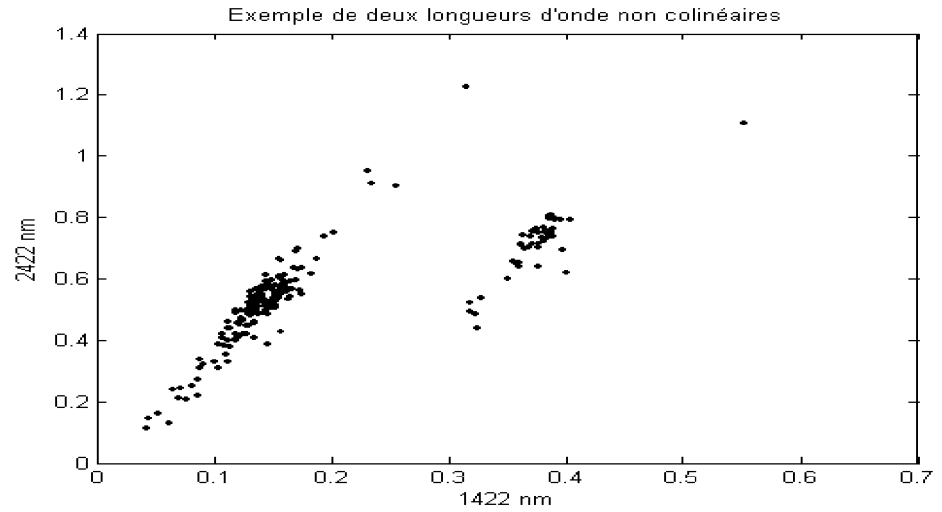
Le nombre de variables "p" d'un spectre est en général très élevé et peut atteindre plusieurs milliers. Il est très souvent supérieur au nombre de spectres "n". Par conséquent, certaines colonnes de la matrice X peuvent être représentées comme étant une combinaison linéaire d'autres colonnes de cette même matrice. Ce phénomène est assez peu couramment rencontré dans d'autres applications statistiques, connu sous le nom de colinéarité est la source de plusieurs problèmes dans l'application directe de plusieurs méthodes statistiques.

La figure 2.3 (a) montre le graphe des absorbances (log 1/R) (où R est le facteur de réflexion de la lumière sur la surface de l'échantillon) à deux longueurs d'onde pour les 218 spectres de la collection de la figure 2.2. Un exemple du phénomène de colinéarité est montré dans cette figure pour deux longueurs d'onde consécutives dans

le spectre (2298 et 2300 nm), les points représentant les échantillons se placent presque exactement sur une droite. Un exemple de deux longueurs d'ondes qui ne sont pas colinéaires est illustré dans la figure 2.3 (b). Pour un écart de 1000 nm entre les longueurs d'onde (1422 et 2422 nm) les points sont moins alignés.



(a) : Exemple des deux longueurs d'onde consécutives 2298 et 2300 nm qui sont très colinéaires



(b): Exemple des deux longueurs d'onde 1422 et 2422 nm qui ne sont pas colinéaires.

Figure 2.3: Illustration de l'effet de colinéarité des longueurs d'onde.

2.3. Loi de Beer-Lambert

L'analyse chimique quantitative par spectrophotométrie d'absorption dans le domaine de l'infrarouge repose sur la loi de Beer-Lambert [13, 26]. Elle est très utilisée en

chimie pour déterminer la concentration de produits en solution. Cette loi exprime la relation de proportionnalité existant entre l'absorbance A et les trois paramètres qui sont l'absorptivité (ε) de l'analyte absorbant la lumière, l'épaisseur de la cellule de mesure (x) et la concentration (C) de l'analyte. L'équation (2.1) est la forme mathématique sous laquelle est habituellement présentée cette loi.

$$A = \varepsilon.C.x \; . \qquad (2.1)$$

Dans un sens plus pratique, l'absorbance A est définie comme étant le logarithme décimal de l'inverse de la transmittance T, elle est exprimée mathématiquement par la relation :

$$A = \log 1/T = (\log I_0/I), \qquad (2.2)$$

où I_0 est l'intensité du faisceau lumineux sur l'échantillon, I est l'intensité de la lumière après avoir traversé l'échantillon. La proportionnalité entre l'absorbance et la concentration permet d'utiliser la spectroscopie d'absorption infrarouge comme méthode de dosage et d'analyse quantitative, du moins dans la limite de linéarité de la loi de Beer-Lambert (pour des absorbances comprises entre 0,3 et 2).

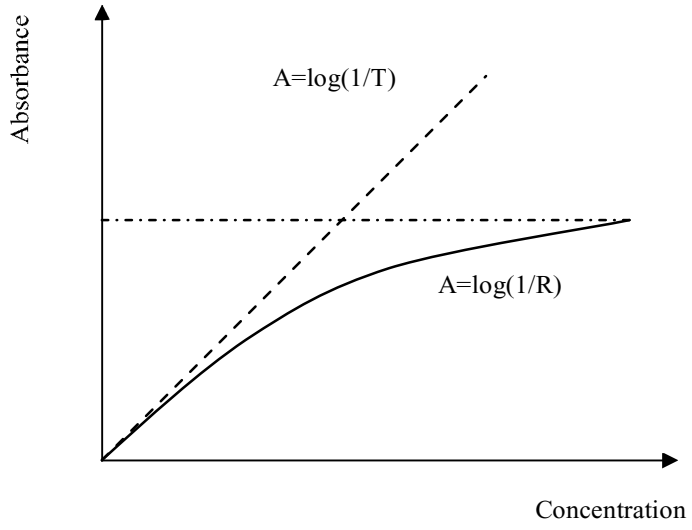


Figure 2.4 : L'absorbance en fonction de la concentration

On voit que dans la loi de Beer-Lambert, l'absorbance est directement proportionnelle à la concentration. Toutefois, on constate que la relation n'est plus linéaire au-delà d'une certaine concentration, comme indiqué dans la figure 2.4. On voit que la relation est linéaire pour le moyen infrarouge et pour le proche infrarouge elle est linéaire et non linéaire.

2.4. Bases de données réelles en spectrophotométrie

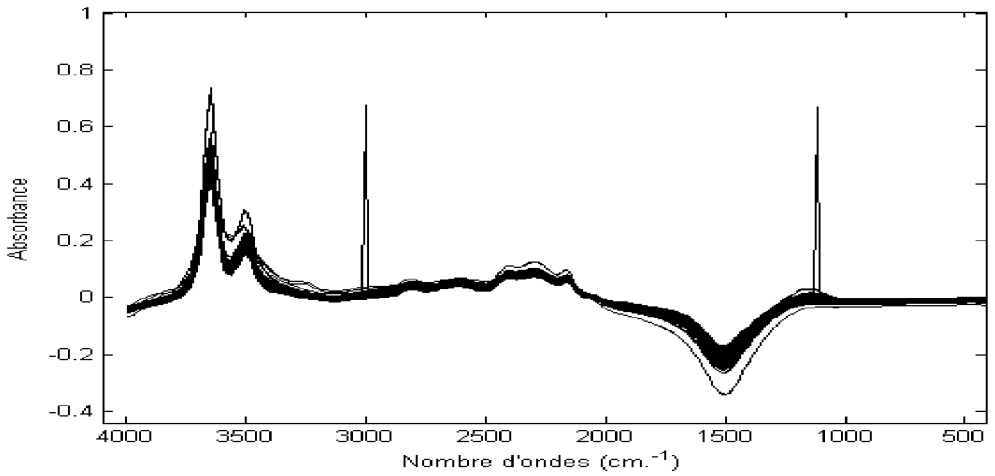
La validation de toutes les méthodes utilisées dans ce travail est réalisée sur trois bases de données réelles différentes liées à l'industrie agroalimentaire. La première et la deuxième base de données: base de données wine et base de données de jus d'orange respectivement ont été fournies par le laboratoire de la spectrophotométrie de l'unité de recherche AGRO / BNUT de l'UCL [55], tandis que la troisième base de données: base de données de pomme a été fourni par l'ENITIAA / INRA (unité de sensométrie et Chimiométrie) à Nantes, France.

Avant d'établir un modèle, l'ensemble des données disponibles est subdivisé en trois sous ensembles séparés : un ensemble d'apprentissage, un ensemble de validation et un ensemble de test. L'ensemble d'apprentissage est utilisé pour estimer les paramètres de chaque modèle. L'ensemble de validation est utilisé pour comparer les modèles et sélectionner le meilleur entre eux. L'ensemble de test est construit uniquement pour estimer l'erreur commise par le meilleur modèle sélectionné dans l'étape de validation. Les distinctions entre ces sous-ensembles sont cruciales, car les termes d'ensembles de validation et de test sont souvent confondus dans la littérature [26]. La section suivante est dédiée à la description des trois bases de données utilisées dans ce travail.

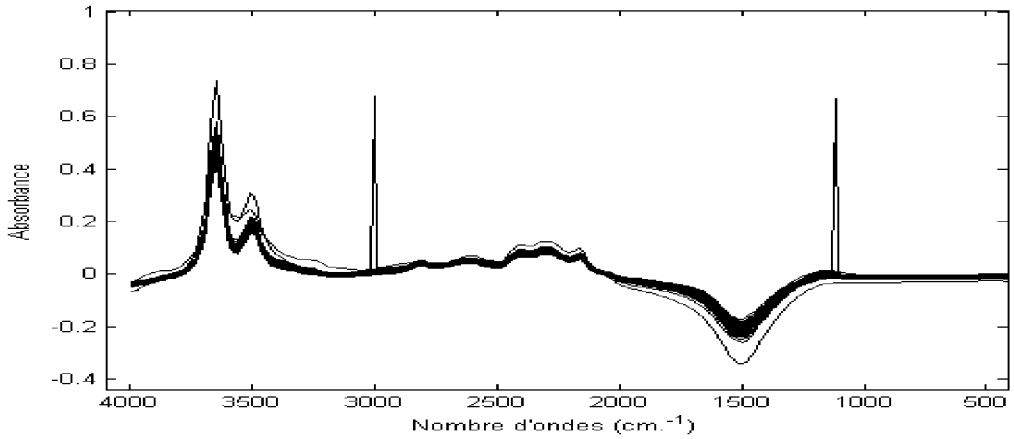
2.4.1. Base de wine

Pour la base de données wine, le problème consiste à prédire la concentration en alcool dans les échantillons de wine, à partir de 124 spectres mesurés en moyen

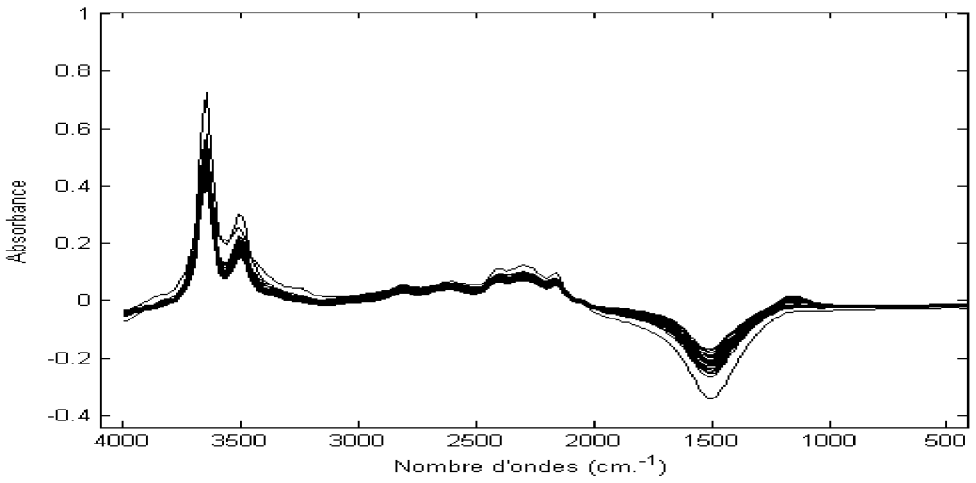
infrarouge. Une collection des 124 spectres est montrée dans la figure 2.5 (a) avec 256 variables spectrales qui sont l'absorbance (log (1/T)) à 256 nombre d'ondes entre 4000 et 400 cm-1 (où T est la transmission de la lumière à travers l'épaisseur de l'échantillon). 60 spectres sont choisis pour l'apprentissage, 34 spectres sont choisis pour la validation et 30 spectres étant réservés pour le test (voir les figures 2.5 (b, c et d)). Les données d'absorbance de la base de données wine varient de -0,3410 à 0,7350 comme l'illustre bien la figure 2.5 (a), et la teneur en alcool (variable de sortie) varie de 7,48% à 18,5% en volume.



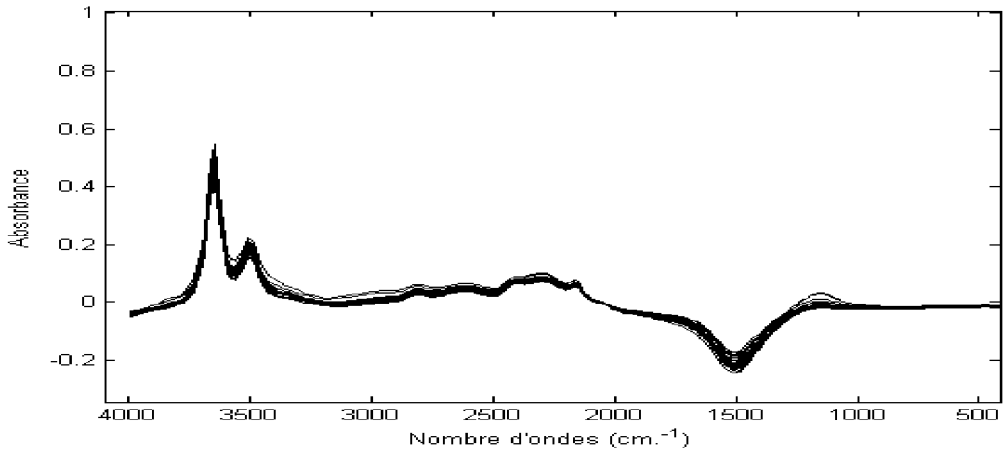
(a) collection des spectres infrarouge moyen de wine



(b) ensemble d'apprentissage



(c) ensemble de validation



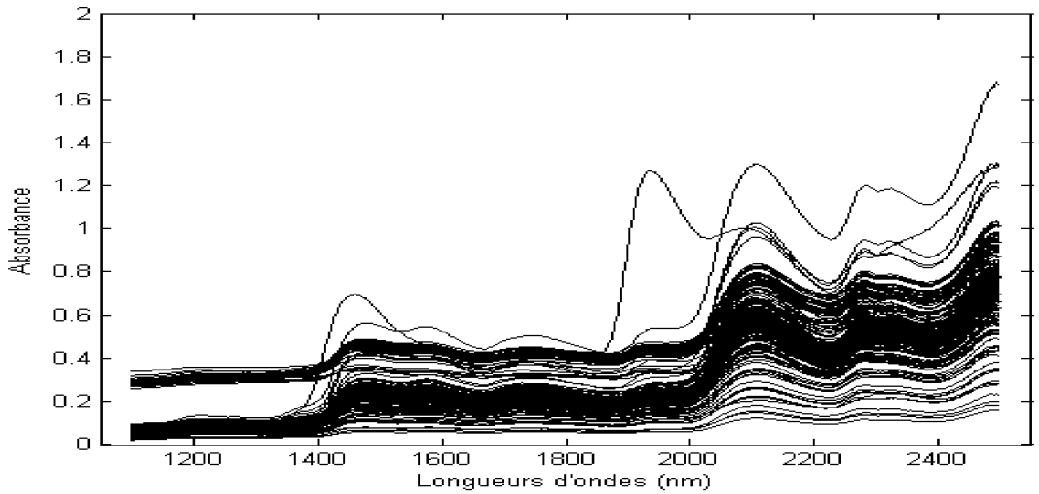
(d) ensemble de test

Figure 2.5 : Spectres de transmittance dans l'infrarouge moyen de wine

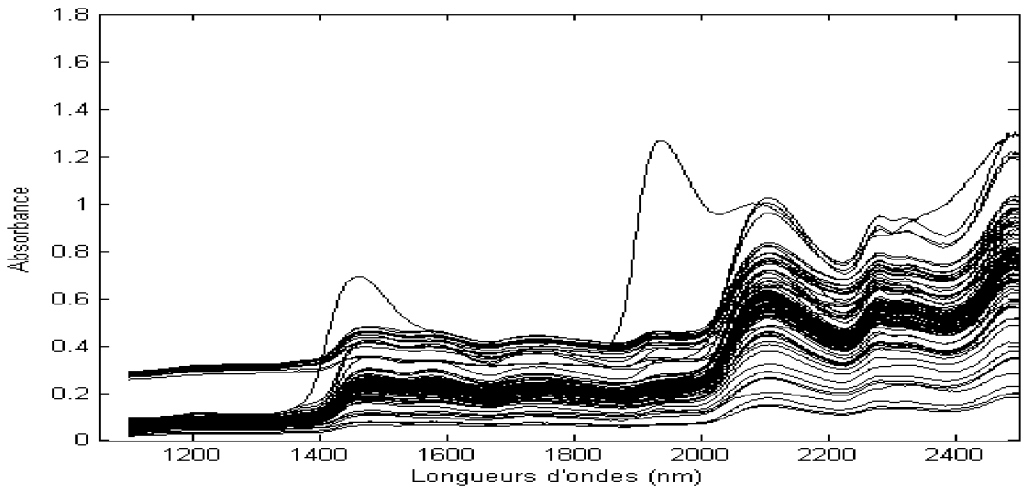
2.4.2. Base de jus d'orange

La base de données du jus d'orange consiste en 218 spectres proches infrarouges d'échantillons de jus d'orange pour lesquels la concentration en sucre (saccharose) doit être prédite. La figure 2.6 (a) montre une collection des 218 spectres avec 700 variables spectrales qui sont l'absorbance (log (1/R)) à 700 longueurs d'ondes entre 1100 et 2500 nm (où R est le facteur de réflexion de la lumière sur la surface de l'échantillon). Les 218 spectres ont été répartis en 100 spectres dans l'ensemble

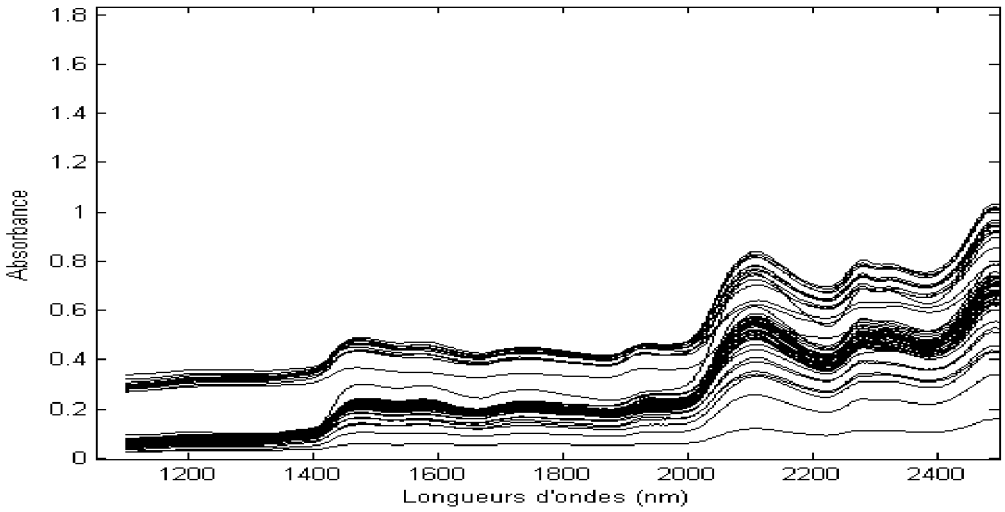
d'apprentissage (voir figure 2.6 (b)), 50 spectres dans l'ensemble de validation (voir figure 2.6 (c)) et 68 spectres dans l'ensemble de test (voir figure 2.6 (d)). Il convient de noter ici, que la variation de l'absorbance de cette base de données est de 0,0199 à 1,6814, et la concentration du saccharose (grandeur de sortie) est comprise entre 0% à 95,2% en poids.



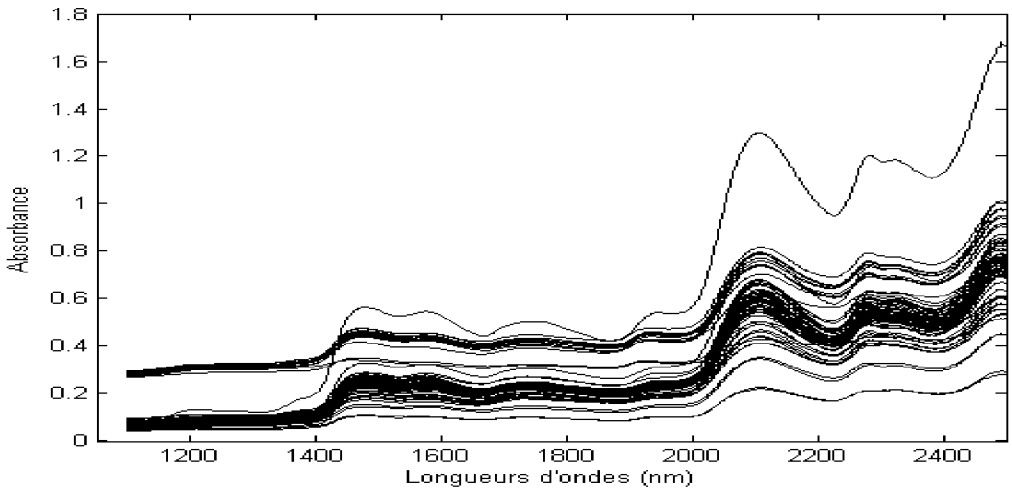
(a) Collection des spectres proche infrarouge de jus d'orange



(b) ensemble d'apprentissage



(c) ensemble de validation



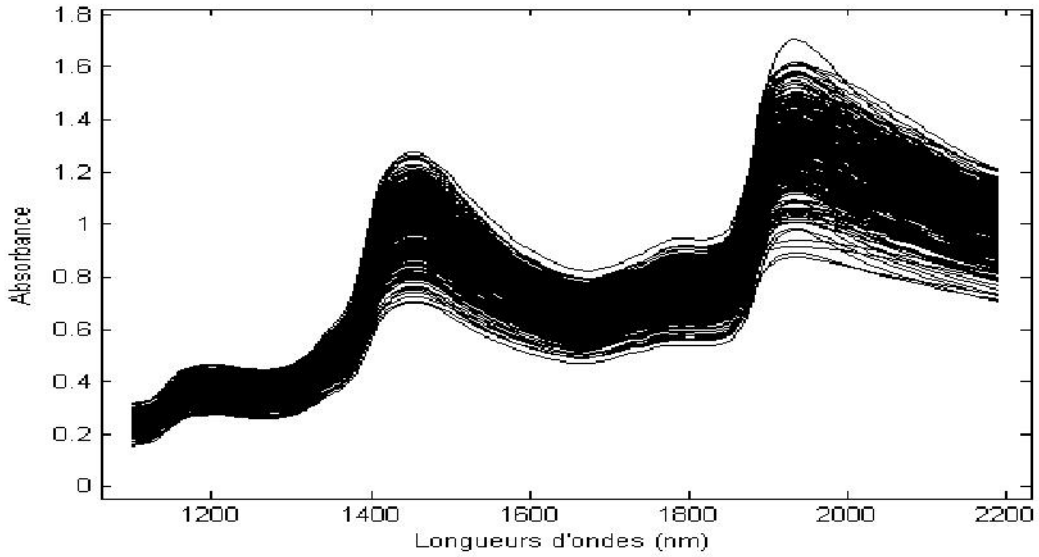
(d) ensemble de test

Figure 2.6: Spectres de réflectivité en proche infrarouge de jus d'orange

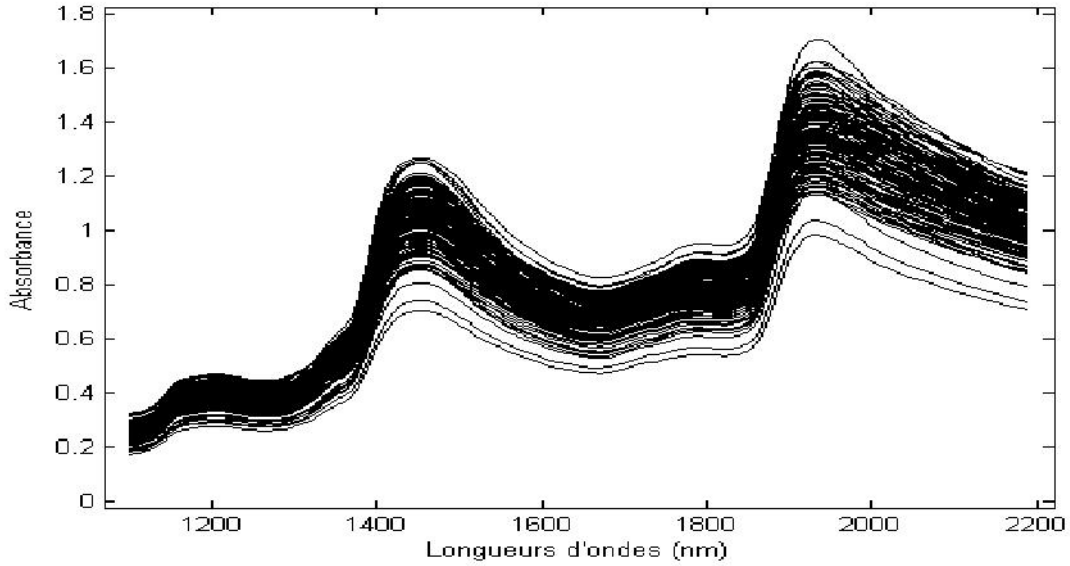
2.4.3. Base de pomme

Dans cette base de données, il s'agit de 337 spectres de réflectivité en proche infrarouge de la pomme avec 110 variables spectrales qui sont l'absorbance (log (1/R)) à 110 longueurs d'onde entre 1100 et 2190 nm (où R est le facteur de réflexion

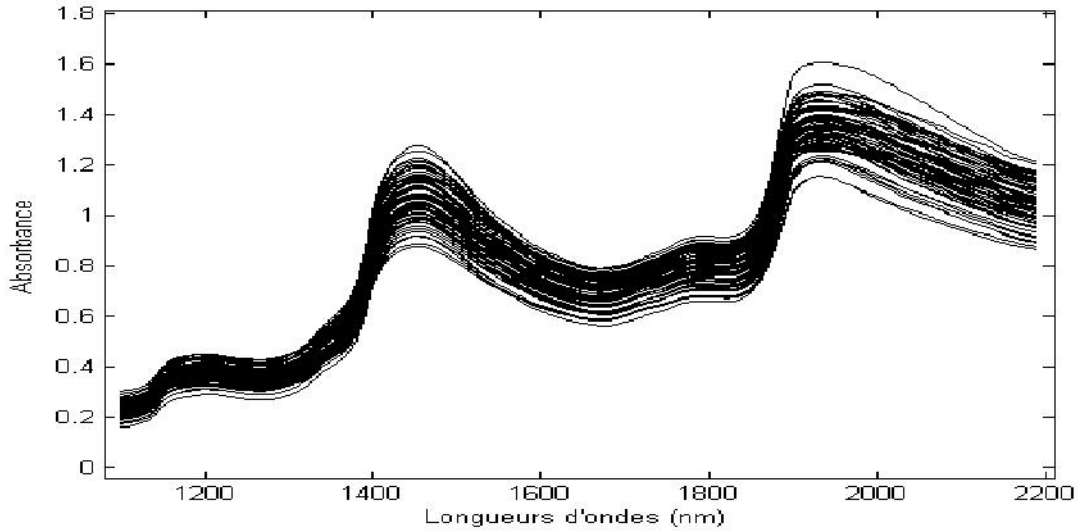
de la lumière sur la surface de l'échantillon). La variable à prédire ici est la force d'écrasement qui devrait être utilisée pour insérer un objet (la tige) dans la peau d'une pomme pour déterminer si elle est trop mûre ou non. Les données d'absorbance de cette base de données varient de 0,1622 à 1,7030 comme le montre la figure 2.7 (a). Les ensembles d'apprentissage, de validation et de test contiennent respectivement 150, 75 et 112 spectres comme indiqué dans les figures 2.7 (b, c et d). La variable de sortie (force d'écrasement) est mesurée en pourcentage en poids et varie de 8,295% à 31,105%.



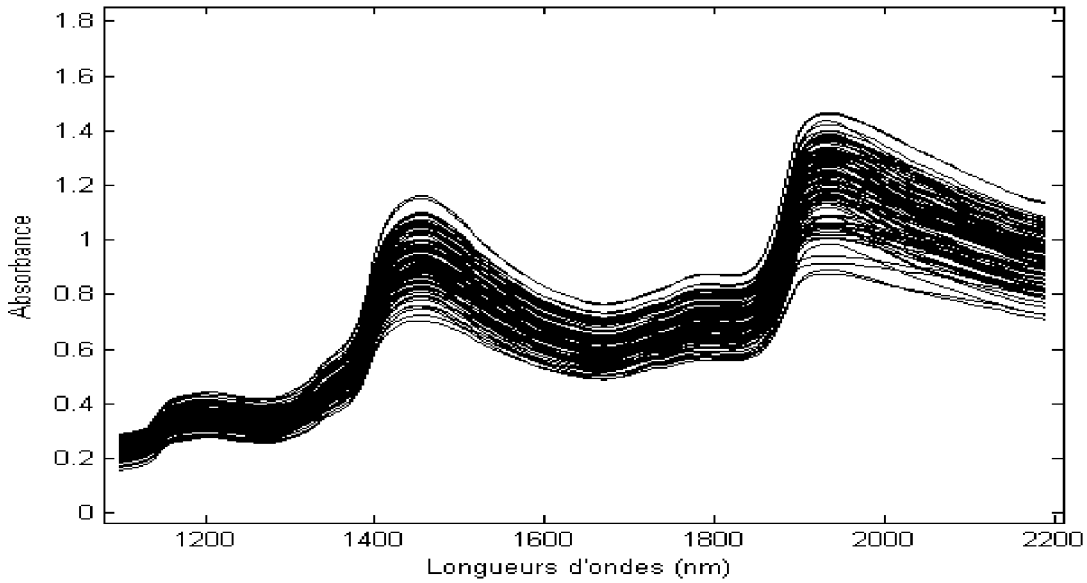
(a) Collection des spectres proche infrarouge des pommes



(b) ensemble d'apprentissage



(c) ensemble de validation



(d) ensemble de test

Figure 2.7: Spectres de reflectance en proche infrarouge des pommes

2.5. Conclusion

Dans ce chapitre, nous avons tout d'abord présenté une introduction sur les données spectrophotométriques et nous avons donné une idée sur la chimiométrie en tant qu'une science incluant les méthodes mathématiques, statistiques et informatiques des données spectrales de manière à en extraire les informations utiles. Par la suite les régions spectrales d'intérêt analytique ont été brièvement exposées. Dans la partie suivante, la description des données spectrales et le principe de la loi de Beer-Lambert ont été présentés. La dernière partie de ce chapitre a été consacrée à la description des bases de données réelles liées à l'industrie agroalimentaire qui seront utilisées dans ce travail de recherche.

Chapitre 3
La régression en spectrophotométrie: étude expérimentale

3.1. Introduction

L'objectif de ce chapitre est de présenter une application de la méthode de régression par SVM au problème d'estimation de la concentration des composants chimiques à partir des mesures spectrophotométriques. En particulier, nous présentons une investigation sur les performances de cette méthode de différents points de vue, y compris:

- L'influence du type de noyau dans la tâche de régression: en adoptant trois types différents du modèle SVM à savoir SVM linéaire (SVM-linear), SVM non linéaire mis en œuvre avec le noyau polynomial (SVM-polynomial) et SVM non linéaire basé sur la fonction de base radiale gaussienne (SVM-RBF).

- L'effet de la réduction du nombre de variables d'entrée sur les régresseurs: par la génération des sous espaces dans l'espace d'entrée de dimension décroissante.
- L'effet de la réduction du nombre d'échantillons d'apprentissage sur les régresseurs: pour cela, nous avons généré des ensembles différents sur l'ensemble d'apprentissage, en diminuant le nombre d'échantillons à partir de 100% des échantillons d'apprentissage disponibles, à 50% et jusqu'à 25%.
- La stabilité globale: en adoptant le calcul de la moyenne AVG (Average) et l'écart type STD (Standard Deviation) de l'erreur quadratique moyenne normalisée NMSE sur tous les résultats obtenus.

A titre de comparaison, les résultats obtenus par les modèles RBF, MLP et MLR sont également fournis [17].

3.2. Les résultats expérimentaux

3.2.1. évaluation d'erreur d'estimation

Les performances des méthodologies de régression explorées dans ce travail, sont évaluées en termes de l'erreur quadratique moyenne normalisée NMSE (Normalized Mean Squared Error) sur l'ensemble de test:

$$\mathrm{NMSE} = \frac{\frac{1}{N_T}\sum_{i=1}^{N_T}\left(f(\boldsymbol{x}_i) - y_i\right)^2}{\mathrm{var}(y)}\ , \qquad (3.1)$$

où N_T est le nombre d'échantillons de test et var (y) est la variance des valeurs de sortie. Cette dernière, qui joue le rôle de la constante de normalisation, est estimée sur tous les échantillons disponibles (c'est à dire sur l'ensemble d'apprentissage et l'ensemble de test) [57].

3.2.2. Résultats obtenus avec SVM, RBF, MLP et MLR

Dans cette section, pour chaque base de données, nous avons mené des études expérimentales sur l'évaluation de l'influence du type du noyau dans le processus de régression, en adoptant trois types différents du modèle SVM [43, 45, 47, 50]. À savoir SVM linéaire (SVM-linear) (équation 1.18), SVM non linéaire mis en œuvre avec le noyau polynomial (SVM-polynomial) (équation 1.19) et SVM non linéaire basé sur la fonction de base radiale gaussienne (SVM-RBF) (équation 1.20). Pour les trois types de régresseurs basés sur SVM, il était nécessaire de déterminer la valeur du paramètre de régularisation C. Au contraire avec SVM linéaire, SVM non linéaire nécessite aussi la détermination des paramètres associés au noyau, c'est à dire l'ordre du polynôme (n) et la largeur (γ) pour SVM-polynomial et SVM-RBF, respectivement.

De manière générale, pour estimer les bons paramètres de la fonction noyau utilisée il n'existe pas de critères universels permettant de choisir ces paramètres. Ce choix dépend de l'application. Ici, l'objectif est de trouver les paramètres qui donnent de meilleure précision d'estimation. Sur la base de cette étude expérimentale, nous avons fixé des intervalles de variation des valeurs de chaque paramètre pour optimiser la précision de régression sur l'ensemble de validation. La valeur du paramètre C a été varié de 0.001 à 1000 et (γ) a été varié de 0.0001à 1000.

L'algorithme SVM-polynomial a été testé avec les noyaux polynomiaux d'ordre 1,2, 3 et 4. Les études expérimentales menées sur les deux bases de données wine et jus d'orange, ont montré qu'en augmentant l'ordre du noyau polynomial, le régresseur SVM devient moins précis. D'après les résultats de simulation obtenus, le régresseur SVM-linéaire s'avère le moins efficace parmi les trois types basés sur SVM. La meilleure valeur de NMSE trouvée par SVM-linéaire (C=1) était égale à 0.007538 et 0.827646 pour les deux bases de données wine et jus d'orange respectivement. Cela confirme que la régression linéaire n'est pas adéquate pour ces applications.

Par contre le modèle SVM-RBF présente les meilleurs résultats en termes de NMSE, la meilleure valeur de γ était égale à 0,0001 et 0,01 pour la base wine et la base de jus

d'orange, respectivement. Les figures 3.1-3.2-3.3 et 3.4 montrent le comportement de l'erreur quadratique moyenne normalisée (NMSE) pour le type SVM-RBF en faisant varier l'un de ses paramètres sur l'ensemble de validation des deux bases de données.

Par la suite, nous avons utilisé un réseau de neurones artificiel perceptrons multicouches MLP à deux couches cachées (équation 1.5) [15, 30, 33, 37], le modèle est testé avec le nombre de neurones (nc1) dans la première couche cachée et le nombre de neurones (nc2) dans la deuxième couche cachée allant de 5 jusqu'à 25 neurones. Dans le cas de la base wine, le meilleur résultat est obtenu avec 15 neurones dans les deux couches cachées. Dans le cas de la base de jus d'orange, le nombre de neurones dans les couches cachées a été varié de 5 à 40 neurones. Le nombre optimal de neurones était égale à 10 neurones dans la première couche cachée et 5 neurones dans la deuxième couche cachée.

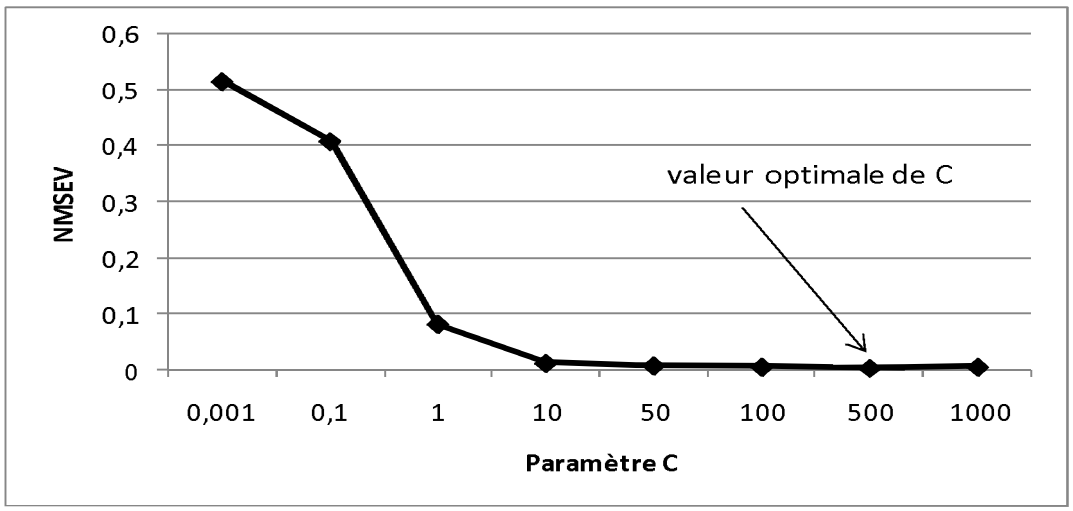


Figure 3.1: NMSE obtenue sur l'ensemble de validation de la base wine en fonction du paramètre C du régresseur SVM-RBF.

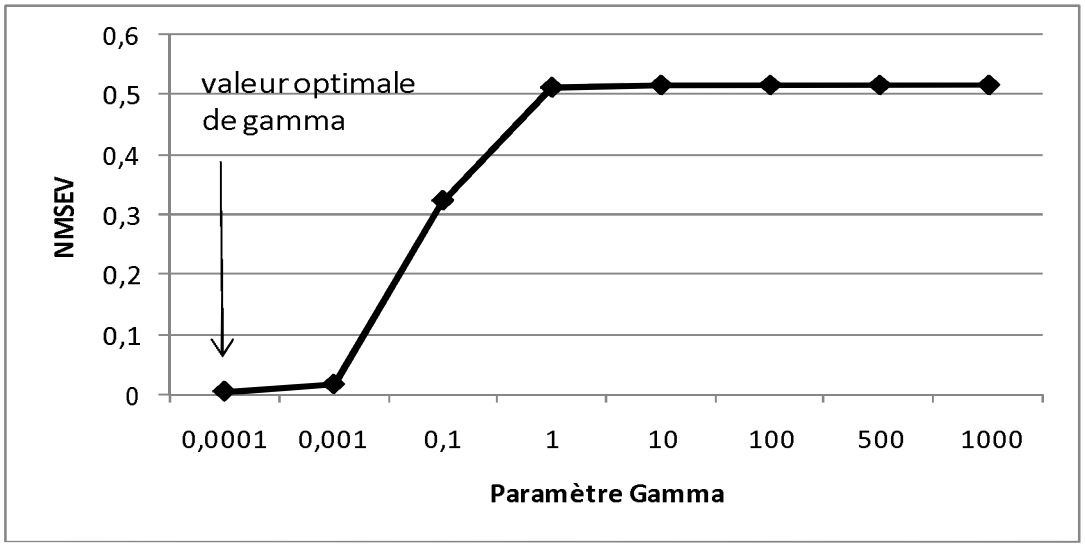


Figure 3.2: NMSE obtenue sur l'ensemble de validation de la base wine en fonction du paramètre gamma du régresseur SVM-RBF.

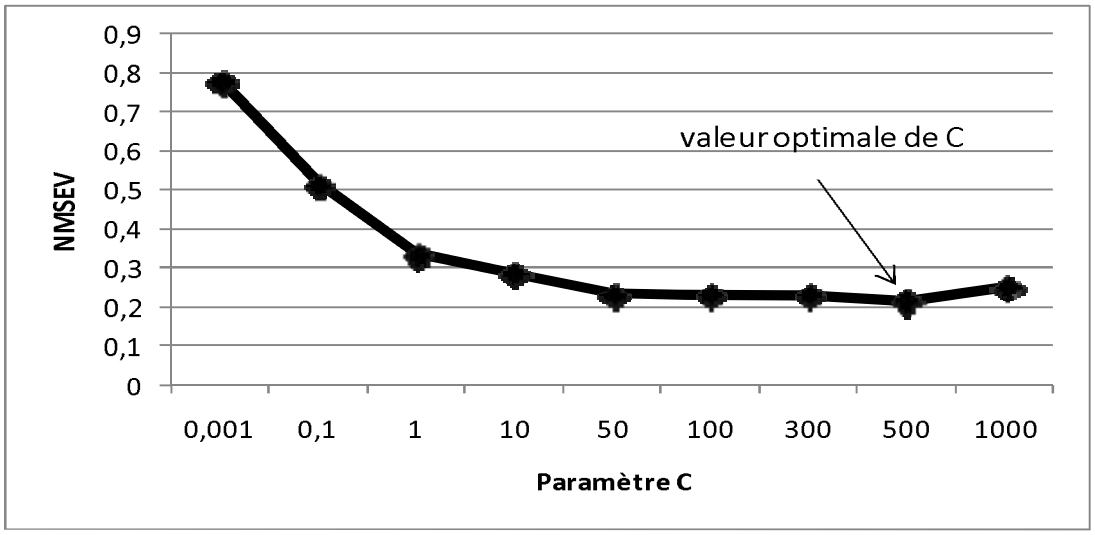


Figure 3.3: NMSE obtenue sur l'ensemble de validation de la base de jus d'orange en fonction du paramètre C du régresseur SVM-RBF.

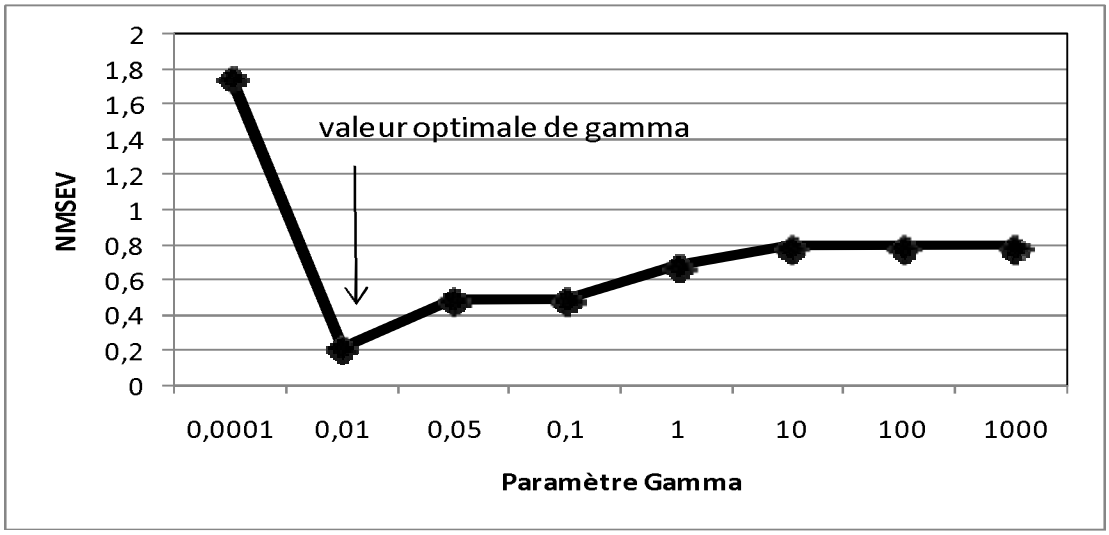


Figure 3.4: NMSE obtenue sur l'ensemble de validation de la base de jus d'orange en fonction du paramètre gamma du régresseur SVM-RBF.

Concernant le modèle RBF (équation 1.7) [30, 38, 39, 41], nous avons fait varier le nombre de neurones (ncc) dans la couche cachée de 1 à 30 neurones et la largeur de la gaussienne δ de 10^{-3} à 200 pour la base wine. En ce qui concerne la base de jus d'orange le nombre de neurones dans la couche cachée a été varié de 2 à 40 neurones et la valeur de δ a été variée de 1 à 100 dans le but également de rechercher les paramètres optimaux de manière empirique. Dans le cas de la base wine, le meilleur résultat a été obtenu avec 21 neurones cachés et avec δ égale à 2. Dans le cas de la base de jus d'orange, le meilleur résultat est obtenu avec 35 neurones dans la couche cachée et avec δ égale à 75. Les figures 3.5, 3.6, 3.7 et 3.8 montrent l'évolution de l'erreur de validation NMSEV (Normalized Mean Square Error) en fonctions des paramètres du modèle RBF (δ et ncc) sur les deux bases de données.

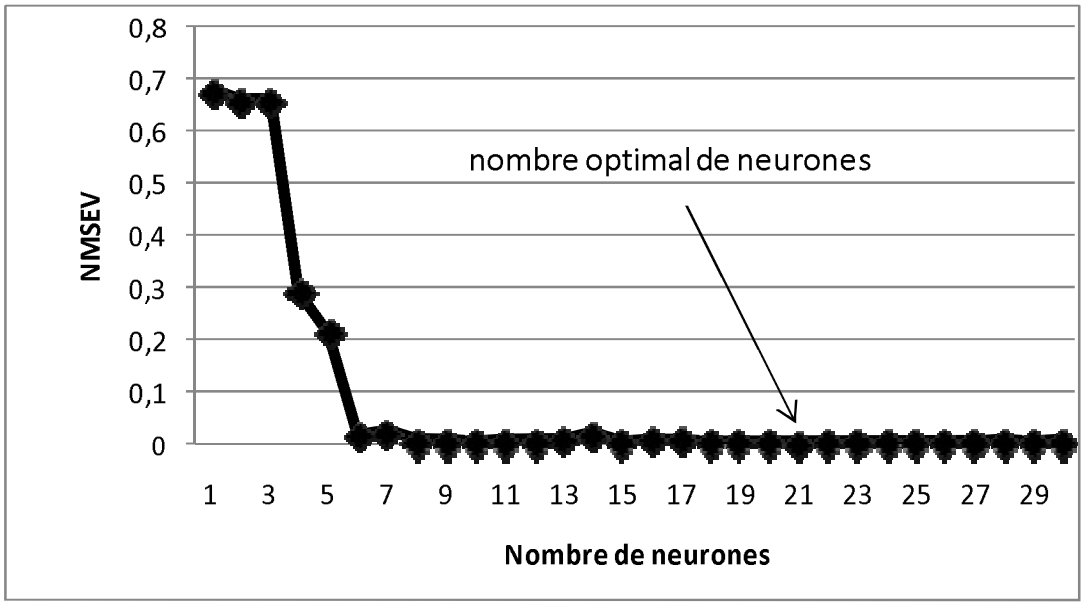


Figure 3.5: Evolution de l'erreur de validation en fonction du nombre de neurone dans la couche cachée du régresseur RBF sur la base wine.

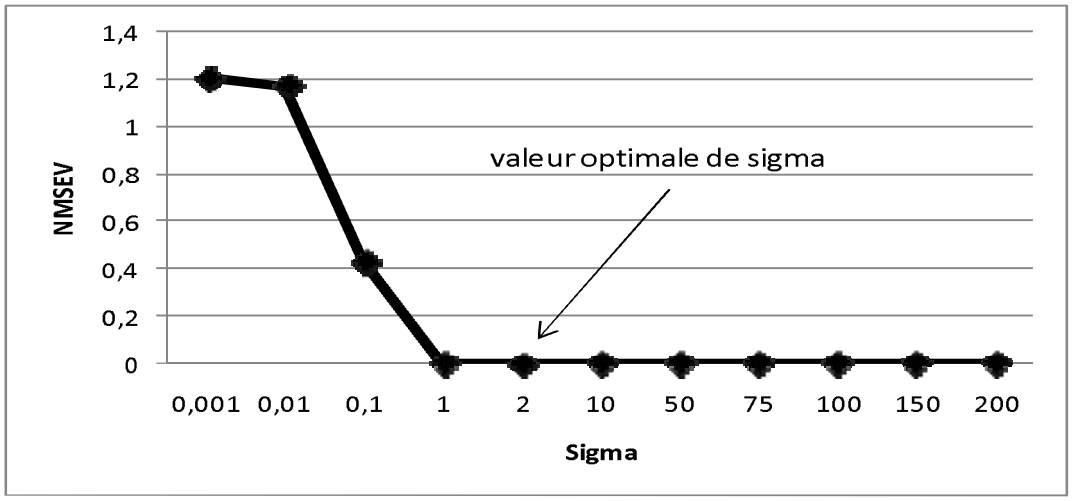


Figure 3.6: Evolution de l'erreur de validation en fonction du paramètre sigma du régresseur RBF sur la base wine.

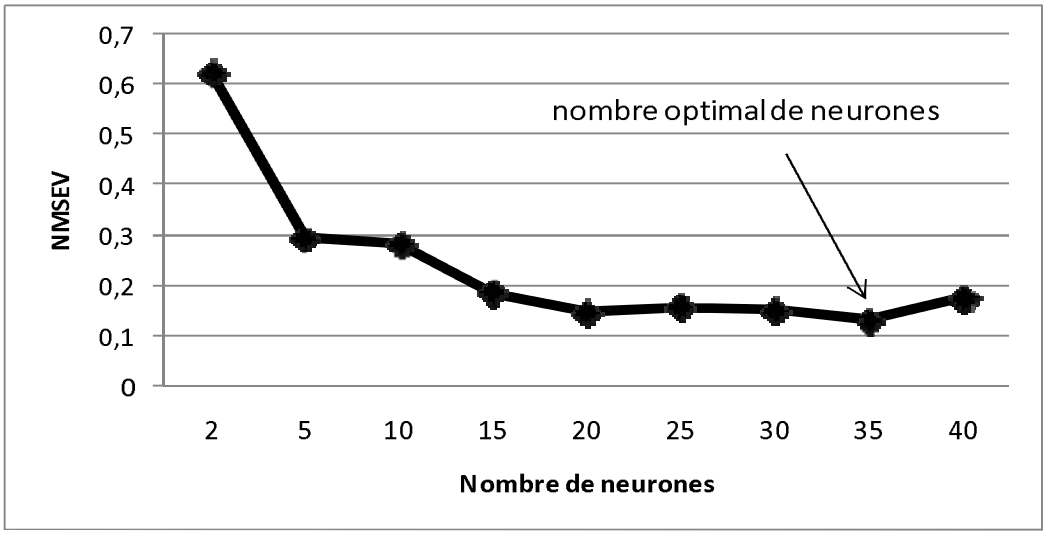


Figure 3.7: Evolution de l'erreur de validation en fonction du nombre de neurone dans la couche cachée du régresseur RBF sur la base de jus d'orange

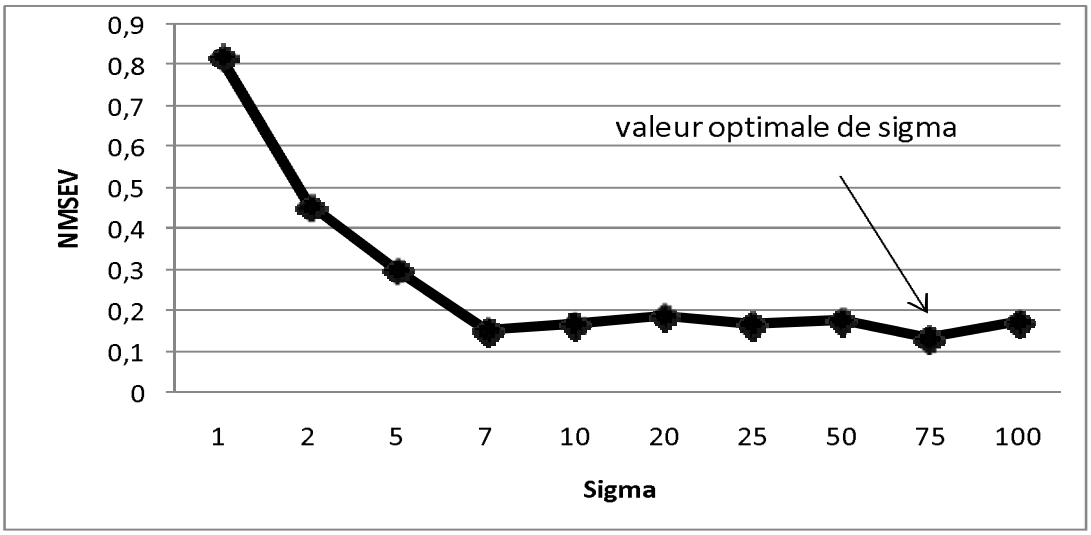


Figure 3.8: Evolution de l'erreur de validation en fonction du paramètre sigma du régresseur RBF sur la base de jus d'orange.

Les résultats obtenus sont répertoriés dans les tableaux 3.1 et 3.2 pour les deux bases de données wine et jus d'orange respectivement. En général, ces résultats montrent que le meilleur régresseur était le RBF tandis que le moins précis était MLR. Parmi les modèles basés sur SVM, le meilleur modèle était SVM-RBF.

Méthode de régression	**NMSE**
SVM-Linear (C=1)	0.007538
SVM-Polynomial (ordre=1, C=1)	0.003564
SVM-RBF (γ=0.0001, C=500)	0.003128
MLR	0.029056
MLP (nc1=15, nc2=15)	0.010716
RBF (δ =2, ncc=21)	**0.002929**

Tableau 3.1: Valeurs de NMSE obtenues par les méthodes SVM, MLR, MLP et RBF sur l'ensemble de test après leur application sur l'espace d'entrée original de la base de données wine (256 variables).

Méthode de régression	**NMSE**
SVM-Linear (C=1)	0.827646
SVM-Polynomial (ordre=1, C=1)	0.751848
SVM-RBF (γ=0.01, C=500)	0.213804
MLR	2.050331
MLP (nc1=10, nc2=5)	0.498961
RBF (δ =75, ncc=35)	**0.149228**

Tableau 3.2: Valeurs de NMSE obtenues par les méthodes SVM, MLR, MLP et RBF sur l'ensemble de test après leur application sur l'espace d'entrée original de la base de données jus d'orange (700 variables).

3.2.3. Effet de la réduction du nombre de variables sur les régresseurs

Dans cette section, nous nous attacherons à l'analyse de la sensibilité de chaque méthode de régression à la dimension de l'espace d'entrée. Pour ce faire, pour chaque base de données, nous avons généré des sous-espaces dans l'espace d'entrée de dimension décroissante au moyen de la technique forward-selection [57]. Les variables ont été sélectionnées en commençant avec 10 variables jusqu'à 50 variables

avec un pas de 10, puis de 75 variables jusqu'à 256 avec un pas de 25 pour la base de wine. Dans le cas de la base de jus d'orange, les variables ont été sélectionnées en commençant avec 10 variables à 50 variables avec un pas de 10, de 75 variables à 250 avec un pas de 25 et de 300 variables à 700 avec un pas de 100.

Les figures 3.9 et 3.10 représentent respectivement la variation de l'erreur quadratique moyenne normalisée NMSE obtenue sur l'ensemble de test des deux bases de données wine et jus d'orange en fonction des variables d'entrée avec les trois types de modèles basés sur SVM. Nous pouvons observer que le noyau le plus adapté aux données spectrophotométriques est le noyau Gaussien (SVM-RBF). Cela confirme ce qui a été également souligné dans d'autres domaines d'application [48, 58, 59]. Les résultats de comparaison entre SVM-RBF (c'est à dire le meilleur modèle basé sur SVM) et les autres régresseurs, à savoir MLP, RBF et MLR sont représentés dans les figures 3.11 et 3.12. Ils montrent que les mauvais résultats sont fournis par la méthode MLR, qui apparaît également très instable. Certains problèmes d'instabilité sont rencontrés aussi bien par les méthodes MLP et RBF.

Les résultats de comparaison entre les trois régresseurs SVM-RBF, MLP et RBF sont représentés dans les figures 3.13 et 3.14. Ils montrent que les deux meilleures méthodes au cours des deux bases de données sont les méthodes SVM-RBF et RBF.

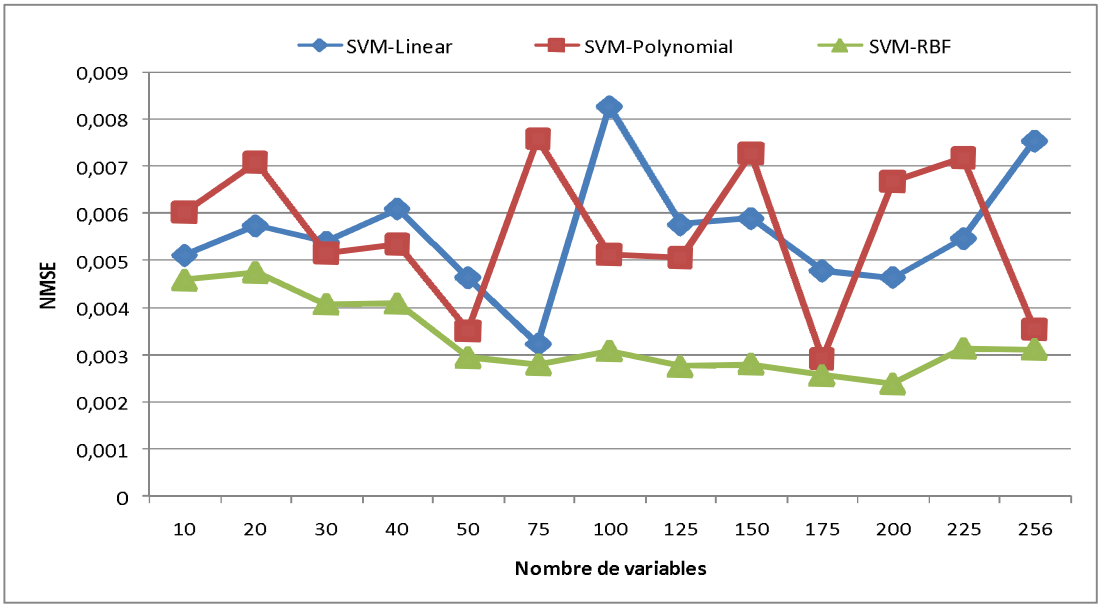


Figure 3.9: NMSE obtenue sur l'ensemble de test de la base de données wine avec les trois modèles basés sur SVM en variant la dimension de l'espace des variables d'entrée.

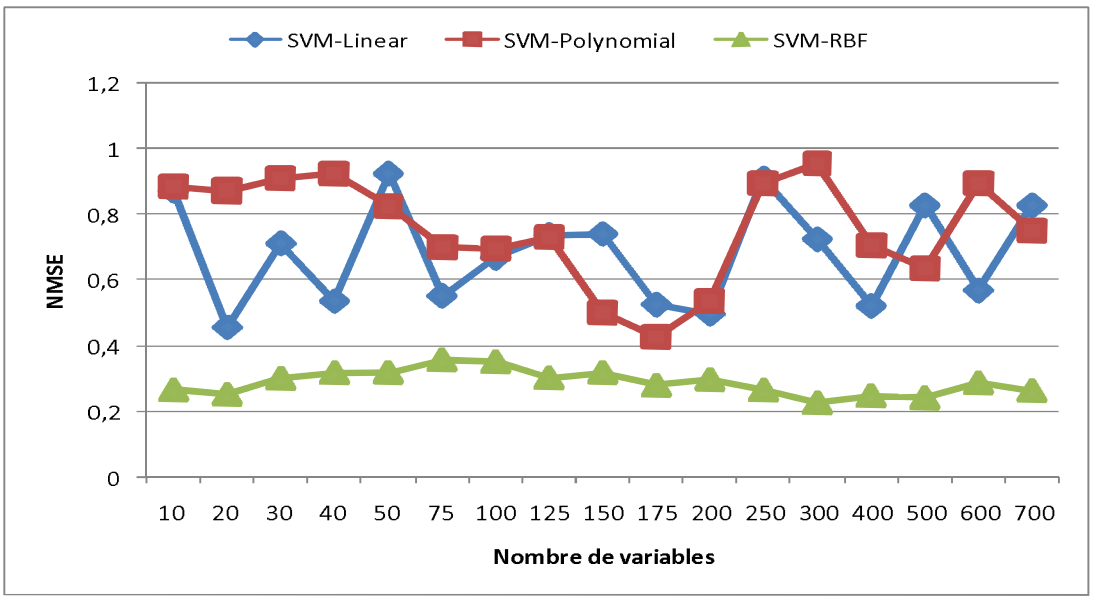


Figure 3.10: NMSE obtenue sur l'ensemble de test de la base de données de jus d'orange avec les trois modèles en variant la dimension de l'espace des variables d'entrée.

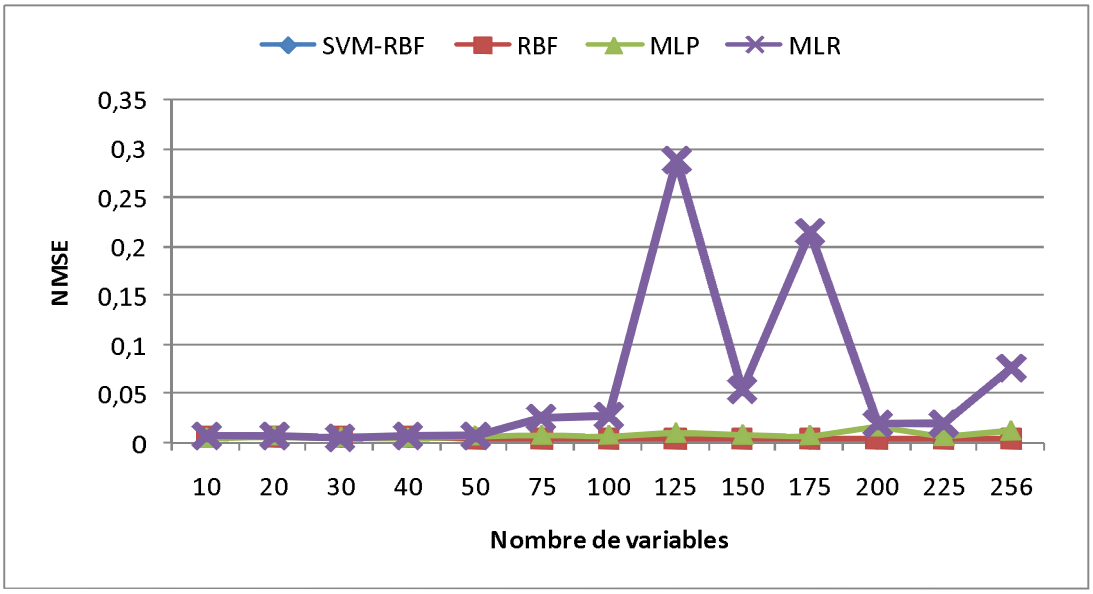


Figure 3.11: NMSE obtenue sur l'ensemble de test de la base de données wine avec les méthodes SVM-RBF, RBF, MLP et MLR en variant la dimension de l'espace des variables d'entrée.

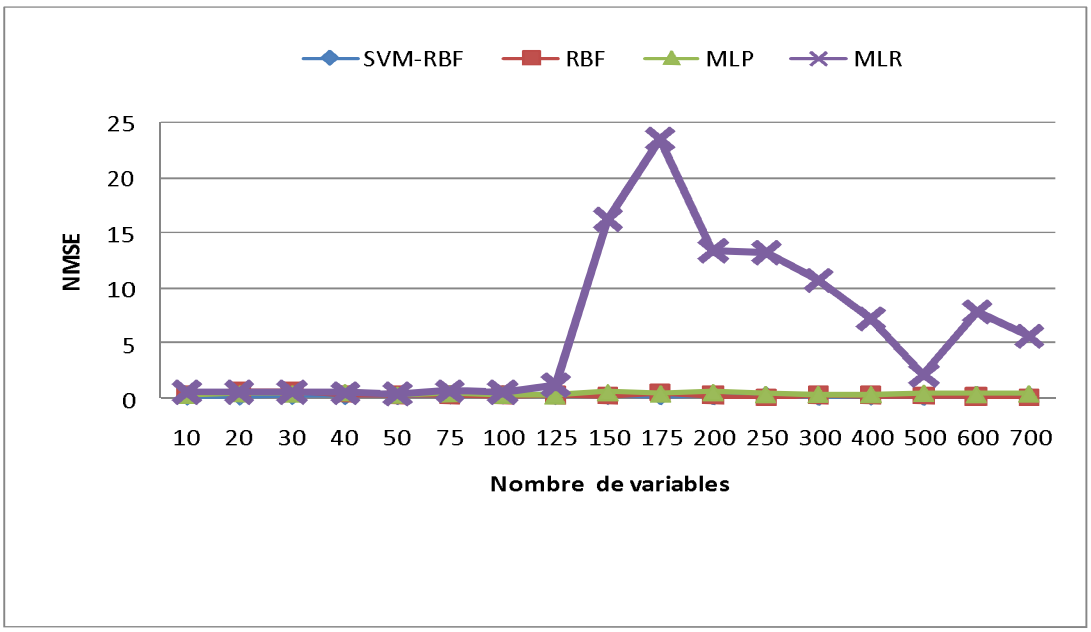


Figure 3.12: NMSE obtenue sur l'ensemble de test de la base de données de jus d'orange avec les méthodes SVM-RBF, RBF, MLP et MLR en variant la dimension de l'espace des variables d'entrée .

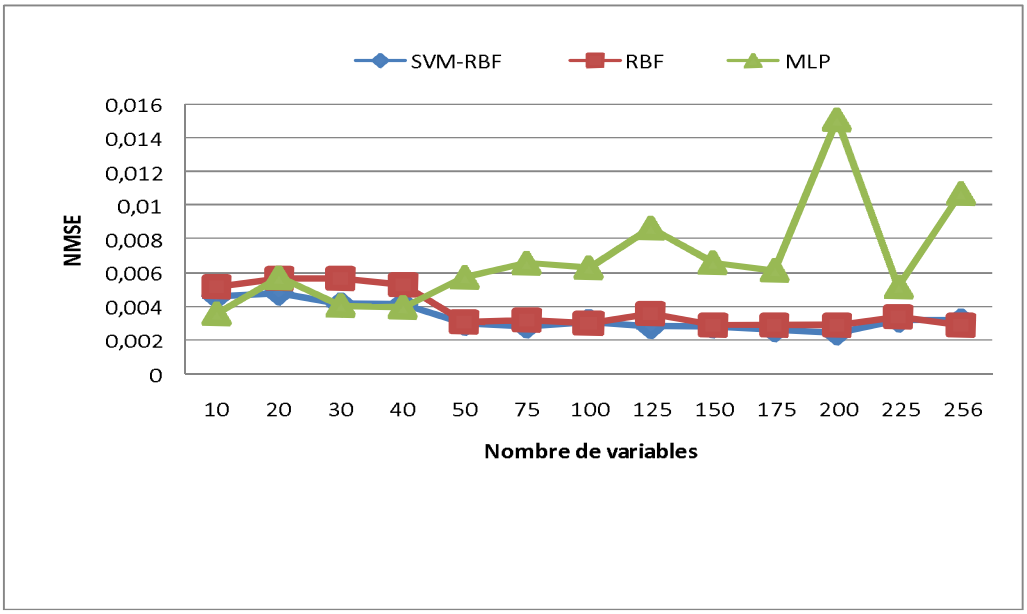


Figure 3.13: NMSE obtenue sur l'ensemble de test de la base de données wine avec les méthodes SVM-RBF, RBF et MLP en variant la dimension de l'espace des variables d'entrée.

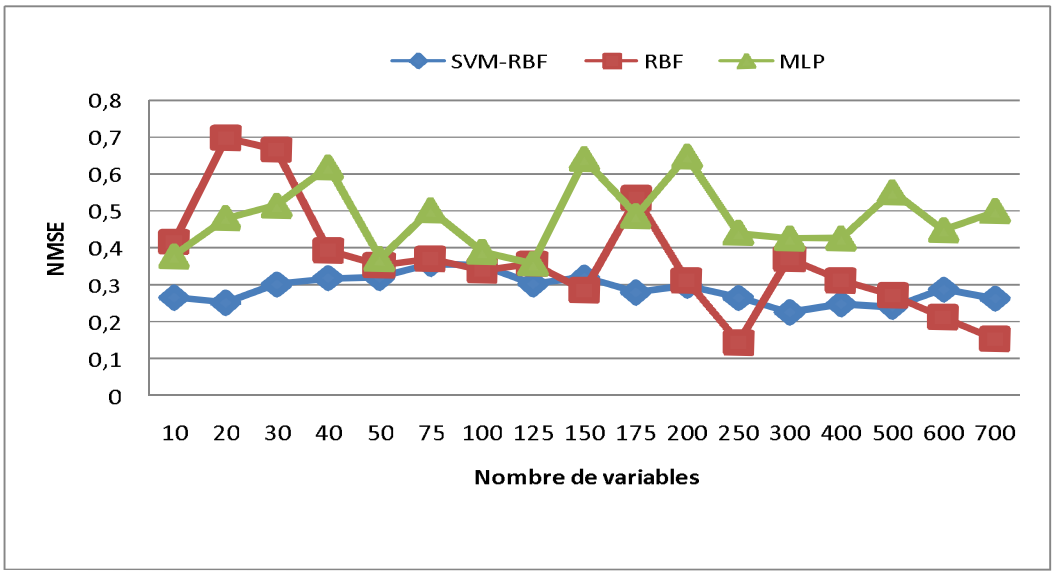


Figure 3.14: NMSE obtenue sur l'ensemble de test de la base de données de jus d'orange avec les méthodes SVM-RBF, RBF et MLP en variant la dimension de l'espace des variables d'entrée .

3.2.4. Effet de la réduction du nombre d'échantillons d'apprentissage

Afin de pousser plus loin notre évaluation des performances des méthodes de régression explorées, nous avons également analysé leur sensibilité au nombre d'échantillons utilisés dans la phase d'apprentissage. Le but ici est d'étudier le comportement de chaque régresseur lorsque le nombre d'échantillons d'apprentissage varie. Pour cela, en particulier, nous avons généré des ensembles différents en diminuant le nombre d'échantillons d'apprentissage. A partir de 94 (nous avons utilisé 100% des échantillons d'apprentissage) jusqu'à 47 (nous avons exploité 50% des échantillons d'apprentissage disponibles) et à 25 (environ 25% des échantillons d'apprentissage disponibles ont été utilisés) échantillons pour la base wine, et à partir de 150 (100% d'échantillons d'apprentissage) jusqu'à 75 (50% des échantillons d'apprentissage) et à 39 (environ 25% d'échantillons d'apprentissage) échantillons pour la base de données de jus d'orange. Noter que le nombre de variables d'entrée a été varié comme dans la section précédente, de sorte d'identifier pour chaque méthode de régression appliquée sur chaque ensemble d'apprentissage le nombre optimal de variables d'entrée correspondant.

Les résultats rapportés dans le tableau 3.3 montrent que, en diminuant le nombre d'échantillons d'apprentissage l'erreur d'estimation augmente indépendamment de la méthode de régression adoptée. En général, les résultats confirment la supériorité des deux méthodes non linéaires RBF et SVM-RBF par rapport aux autres méthodes pour les deux bases de données wine et jus d'orange. Notez, cependant, que la méthode SVM-RBF se comporte mieux même lorsque quelques échantillons d'apprentissage sont disponibles.

Méthode de régression	Base de Wine			Base de Jus d'Orange		
	NMSE 100%	**NMSE 50%**	**NMSE 25%**	**NMSE 100%**	**NMSE 50%**	**NMSE 25%**
SVM-Linear	0.003222	0.005581	0.007106	0.454596	0.2644	0.319228
SVM-Polynomial	0.002927	0.004726	0.005334	0.427159	0.333411	0.344194
SVM-RBF	**0.002399**	**0.004145**	**0.003093**	0.166375	0.223859	**0.186367**
MLR	0.004929	0.007481	0.007053	0.495713	0.390706	0.494793
RBF	0.002883	0.004769	0.004765	**0.141036**	**0.16135**	0.343989
MLP	0.003581	0.00737	0.006585	0.356107	0.283466	0.401608

Tableau 3.3: Valeurs de NMSE obtenues par les méthodes SVM, MLR, MLP et RBF sur l'ensemble de test pour les bases de données wine et jus d'orange en variant la taille de l'ensemble d'apprentissage.

3.2.5. Évaluation de la stabilité globale

Finalement, afin d'évaluer la stabilité de chaque technique, nous avons adopté le calcul de la moyenne AVG (Average) et l'écart type STD (Standard Deviation) de l'erreur quadratique moyenne normalisée NMSE, sur tous les résultats obtenus par la variation simultanée du nombre de variables d'entrée et le nombre d'échantillons d'apprentissage suivant les stratégies expérimentales décrites dans les sections précédentes. Le tableau 3.4 résume les résultats de la stabilité globale réalisés sur l'ensemble de test pour les deux bases de données wine et jus d'orange.

Plus la valeur de l'écart type STD est petit, plus la stabilité est élevée. Ces résultats montrent la supériorité de la stratégie SVM-RBF en termes de la moyenne de NMSE et de la stabilité car elle fournit le plus faible écart-type. Comparée aux autres modèles, le modèle MLR apparaît le moins stable au cours des deux bases de données.

Sur la base de cette étude expérimentale nous pouvons conclure que les modèles MLR, MLP et SVM-polynomial montrent moins de stabilité et donc possèdent moins de garantie dans la réalisation d'un bon ajustement aux données.

Méthode de régression	**Base de Wine**		**Base de Jus d'orange**	
	NMSE			
	AVG	**STD**	**AVG**	**STD**
SVM-Linear	0.01062192	0.00559319	0.61179657	0.19118901
SVM-Polynomial	0.01059595	0.00572329	0.65618614	0.20631975
SVM-RBF	**0.00471319**	**0.00152188**	**0.26479676**	**0.04614403**
MLR	0.02603241	0.0186976	2.10408806	2.00869847
RBF	0.00724879	0.00377681	0.37893233	0.15450227
MLP	0.01402728	0.00922378	0.47677808	0.09338307

Tableau 3.4: Stabilité globale réalisée par les méthodes SVM, MLR, MLP et RBF sur l'ensemble de test pour les deux bases de données Wine et jus d'orange.

3.3. Conclusion

Le but de ce chapitre était de présenter une exploration expérimentale approfondie de l'efficacité de la technique SVM en tant qu'un outil de régression en spectrophotométrie. Les résultats obtenus sur les deux différentes bases de données spectrophotométriques permettent de souligner les conclusions suivantes:

- La méthode SVM apparaît comme une alternative crédible à la plupart des méthodes conventionnelles.
- La méthode SVM présente une faible sensibilité à la malédiction de la dimensionnalité.
- Elle est moins sensible au problème de la disponibilité d'un nombre limité d'échantillons d'apprentissage.

- En général, elle présente une plus grande stabilité offrant ainsi une plus grande garantie d'ajustement d'un bon modèle aux données.
- Parmi les différents noyaux étudiés, le noyau gaussien semble être le plus approprié.

Chapitre 4
Système de régression robuste

4.1. Introduction

Dans le premier chapitre, ont été présentées les méthodes linéaires et non linéaires de régression qui peuvent s'appliquer utilement aux données spectrales. Certes il existe d'autres techniques de régression potentiellement utilisables qui sont extrêmement nombreuses. L'utilisateur peut se trouver dans une situation difficile; c'est-à-dire laquelle des techniques à choisir parmi toutes les approches proposées dans la littérature (MLR, MLP, RBF, SVM etc…)? Nous proposons une solution possible pour résoudre ce problème qui est la fusion des résultats fournis par un ensemble de différentes méthodes de régression [18]. A cet égard, l'objectif de ce chapitre est double:

- Il s'agit tout d'abord d'évaluer l'intérêt de la fusion de différents types de regresseurs par rapport à l'usage d'un seul type de regresseur, via lapplication du modèle RMS (Robust Multiple System). Dans ce but, nous commençons par la formulation du problème, nous passons à la description de la

construction de l'ensemble des regresseurs, nous présentons par la suite, un ensemble de différentes stratégies de fusion. La stratégie de la moyenne simple ACS (Average Combination Strategy) et la stratégie de la moyenne pondérée WCS (Weighted Combination Strategy) basées sur la combinaison linéaire. L'autre méthode est la stratégie de fusion non linéaire NLFS (Non Linear Fusion Strategy) [18]. Les résultats expérimentaux sont ensuite présentés.

- Ensuite nous nous attacherons à analyser la précision de chaque régresseur inclut dans l'ensemble des régresseurs dans les différentes portions de l'espace des caractéristiques. Pour ce faire, nous présentons la dernière stratégie envisagée est la stratégie SBS (Selection Based Strategy). Les résultats expérimentaux sont présentés à la fin de ce chapitre [18].

4.2. Formulation du problème

Avant de décrire la méthode proposée RMS, commençons d'abord par formaliser le problème considéré. Soit $\{(x_i, y_i), i = 1...N\}$ l'ensemble de données d'apprentissage. La cible y_i (valeur réelle représentant la concentration d'un composant chimique d'intérêt dans un produit donné) est associée à chaque vecteur d'entrée x_i, qui est représenté dans l'espace des caractéristiques de dimension (d). Etant donné un ensemble formé de T régresseurs $f_i(x)$ $(i = 1,..., T)$ entraînés indépendamment sur les ensembles d'apprentissage disponibles. Après la fusion de différents régresseurs, on obtient l'estimation résultante $F(x)$ (voir figure 4.1). Le problème est de définir un mécanisme de fusion $\Phi\{.\}$ de telle sorte que $F(x)$ obtenue, est donnée par l'équation (4.1) pour un échantillon inconnu donné x.

$$F(x) = \Phi\{f_1(x), f_2(x),..., f_T(x)\} \qquad (4.1)$$

Avant de passer à la description de cette structure, nous allons d'abord présenter l'ensemble de régresseurs dans le paragraphe suivant.

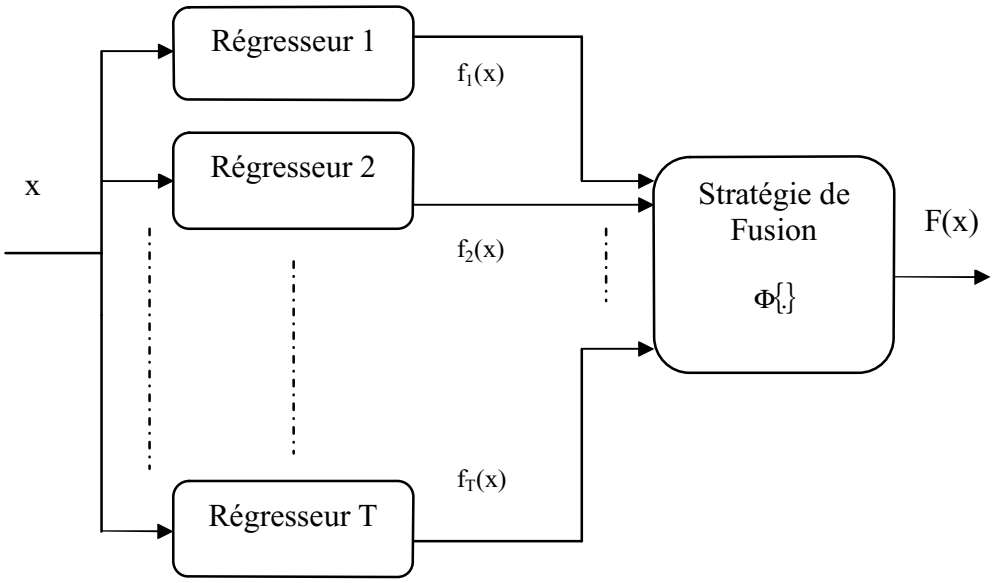


Figure 4.1: Schéma block général du modèle RMS (Robust Multiple System) [18].

4.3. Construction de l'ensemble de régresseurs

Pour construire l'ensemble de régresseurs, trois différents algorithmes de régression sont considérés. Ceux-ci sont les algorithmes RBF, MLP et SVM décrits dans le premier chapitre. Cela permet de garantir une certaine diversité dans l'ensemble, qui est nécessaire pour augmenter la probabilité d'obtenir une meilleure robustesse et précision. Une fois l'ensemble de régresseurs est disposé, la prochaine étape de l'approche proposée est la fusion. Afin d'effectuer cette tâche, nous proposons quatre stratégies différentes. La combinaison (fusion) linéaire est réalisée avec deux méthodes différentes: la stratégie de la moyenne simple ACS (Average Combination Strategy) et la stratégie de la moyenne pondérée WCS (Weighted Combination Strategy). L'autre méthode est la stratégie de fusion non linéaire NLFS (Non Linear Fusion Strategy). La dernière stratégie envisagée est la stratégie basée sur la sélection SBS (Selection Based Strategy). Toutes ces stratégies sont décrites ci-dessous.

4.4. Description du modèle RMS

4.4.1. Stratégies de fusion

Nous présentons dans ce qui suit trois stratégies différentes de fusion. La stratégie de la moyenne simple ACS et la stratégie de la moyenne pondérée WCS basées sur la combinaison (fusion) linéaire. La troisième stratégie est la stratégie de fusion non linéaire NLFS.

4.4.1.1. Stratégies basées sur fusion linéaire

A. Stratégie de la moyenne simple

L'ACS représente une stratégie non supervisée, qui est particulièrement attrayante pour sa mise en œuvre simple. Dans cette stratégie, la fusion est basée sur l'opérateur de la moyenne. L'idée derrière cette méthode est que, d'un point de vue statistique, les différents régresseurs peuvent être considérés comme des processus aléatoires différents qui modélisent la même cible. L'estimation optimale basée sur les statistiques du premier ordre peut être obtenue par une opération classique de la moyenne. En conséquence, pour un échantillon donné, l'estimation résultante $F(x)$ peut être écrite comme suit:

$$F(x) = \frac{1}{T}\sum_{i=1}^{T} f_i(x). \qquad (4.2)$$

Dans [60, 61], il a été montré que l'opérateur de la moyenne est efficace si tous les régresseurs sont non biaisés et non corrélées avec des variances similaires ou, en général, quand tous les régresseurs présentent des précisions analogues.

B. Stratégie de la moyenne pondérée

WCS est une stratégie de fusion supervisée qui vise à exploiter les informations disponibles sur les données (mesures spectroscopiques) afin d'en tirer une fusion linéaire pondérée des sorties des régresseurs. Grâce à l'affectation d'un poids à chaque

régresseur, le modèle de la fusion linéaire peut être réglé afin d'optimiser la précision d'estimation des paramètres et la robustesse du processus de régression. Il est à noter que l'approche ACS peut être considérée comme un cas particulier de WCS dans laquelle tous les régresseurs possèdent le même poids, qui peut être considéré comme un "facteur de fiabilité" associé au régresseur correspondant. Le concept du facteur de fiabilité a été introduit dans le contexte de "hybrid consensus theory" [62]. Chaque régresseur peut être considéré comme une source d'information. Les facteurs de fiabilité représentent donc un moyen de pondération de l'influence de chaque source d'information dans le processus de décision globale. L'estimation finale fournie par WCS est donnée par:

$$F(x) = \sum_{i=1}^{T} \beta_i f_i(x), \qquad (4.3)$$

où β_i représente le facteur de fiabilité (poids) affecté au $i^{ième}$ régresseur. Le problème soulevé par WCS est la détermination des valeurs des poids β_i. Ce problème peut être abordé de différentes manières. Une solution simple (qui est largement utilisée dans la littérature) est basée sur l'erreur quadratique minimale NMSE (Normaized Minimum Square Error) en utilisant la technique du pseudo inverse de la matrice des données [63]. L'avantage principal de cette technique est qu'elle permet d'obtenir une solution analytique optimale selon le critère NMSE.

C. Stratégie de fusion non linéaire

La fusion non linéaire est une stratégie supervisée dans laquelle la combinaison des trois différents régresseurs est réalisée dans ce travail par l'adoption d'un réseau de neurone artificiel (RBF) [16, 33, 38, 39, 41]. L'utilisation de ce dernier est motivée par le fait que les différents régresseurs sont utilisés dans des sous espaces différents, ce qui signifie qu'il y a une relation non linéaire attendue entre les sorties des différents régresseurs. Dans ces situations, l'utilisation des techniques de régression non linéaire peut représenter une solution efficace au problème de fusion. En plus,

l'apprentissage d'un réseau de neurones artificiel RBF est plus rapide que celui d'un réseau de neurones artificiel MLP et d'une SVM.

4.5. Résultats expérimentaux

4.5.1. Résultats obtenus avec les régresseurs simples

Dans cette étude expérimentale, pour chaque base de données, trois types différents de régresseurs ont été adoptés, à savoir RBF, MLP [15, 29, 30] et SVM non linéaire basé sur la fonction de base radiale gaussienne (SVM-RBF) [45, 47, 50]. Les trois régresseurs ont été entraînés sur leurs ensembles d'apprentissage correspondants. L'objectif ici, est de trouver les paramètres qui donnent de meilleure précision d'estimation. Dans le but de choisir empiriquement le meilleur modèle pour chaque régresseur, nous avons fixé des intervalles de variation des valeurs de chaque paramètre pour optimiser la précision de régression sur l'ensemble de validation. Concernant le modèle SVM-RBF (équation 1.20), il était nécessaire de définir deux paramètres. Un est le paramètre de régularisation C. Les simulations ont été effectuées en faisant varier la valeur de ce paramètre de 0,001 à 1000. Le second paramètre est la largeur (γ) du noyau gaussien. Ce paramètre a été varié de 0.0001 à 1000.

Par la suite, nous avons utilisé un réseau de neurones artificiel perceptrons multicouches MLP à deux couches cachées; en faisant varier le nombre de neurones nc1 et nc2 dans la première et la deuxième couche cachée respectivement , à savoir de 2 jusqu'à 16 neurones dans la première couche cachée et de 2 à 8 neurones dans la seconde couche cachée pour la base wine. Les meilleurs résultats ont été obtenus avec 6 neurones dans la première couche cachée et avec 2 neurones dans la seconde couche cachée. Les figures 4.2 et 4.3 montrent l'évolution de l'erreur de validation normalisée NMSEV (Normalized Mean Square Error) en fonction du nombre de neurones dans les couches cachées.

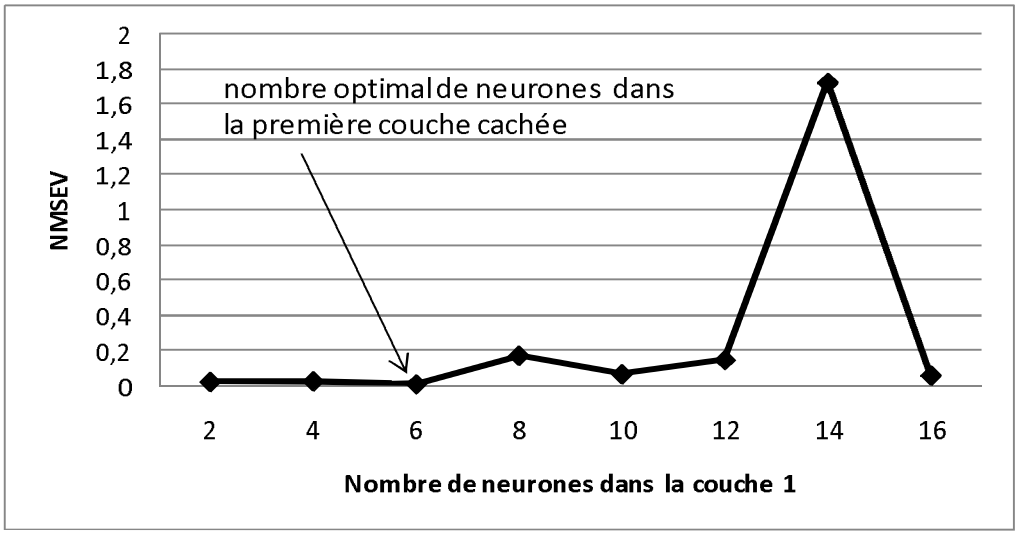


Figure 4.2: Evolution de l'erreur de validation en fonction du nombre de neurones dans la première couche cachée du régresseur MLP sur la base de données wine.

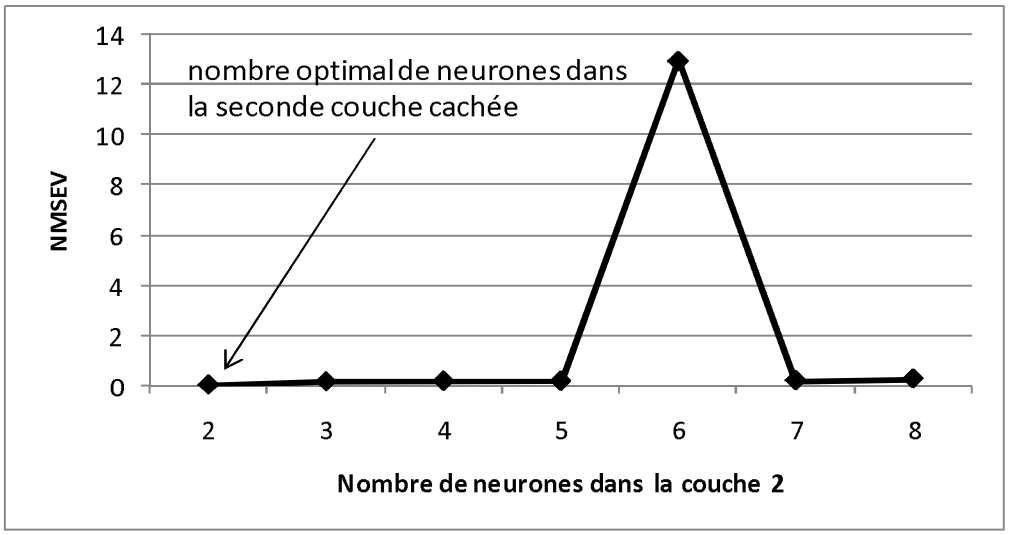


Figure 4.3: Evolution de l'erreur de validation en fonction du nombre de neurones dans la deuxième couche cachée du régresseur MLP sur la base de données wine.

Dans le cas de la base de jus d'orange, le nombre de neurones a été varié de 2 à 20 neurones dans la première couche cachée et de 2 à 6 neurones dans la seconde couche cachée. Les figures 4.4 et 4.5 montrent les meilleurs paramètres obtenus dans la première (8 neurones) et la deuxième (6 neurones) couche cachée respectivement

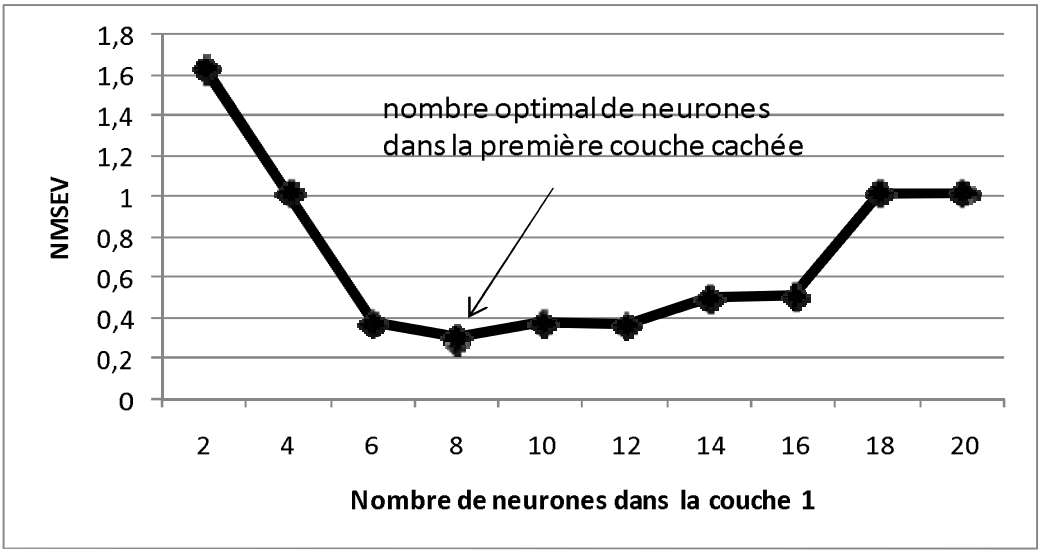


Figure 4.4: Evolution de l'erreur de validation en fonction du nombre de neurones dans la première couche cachée du régresseur MLP sur la base de jus d'orange.

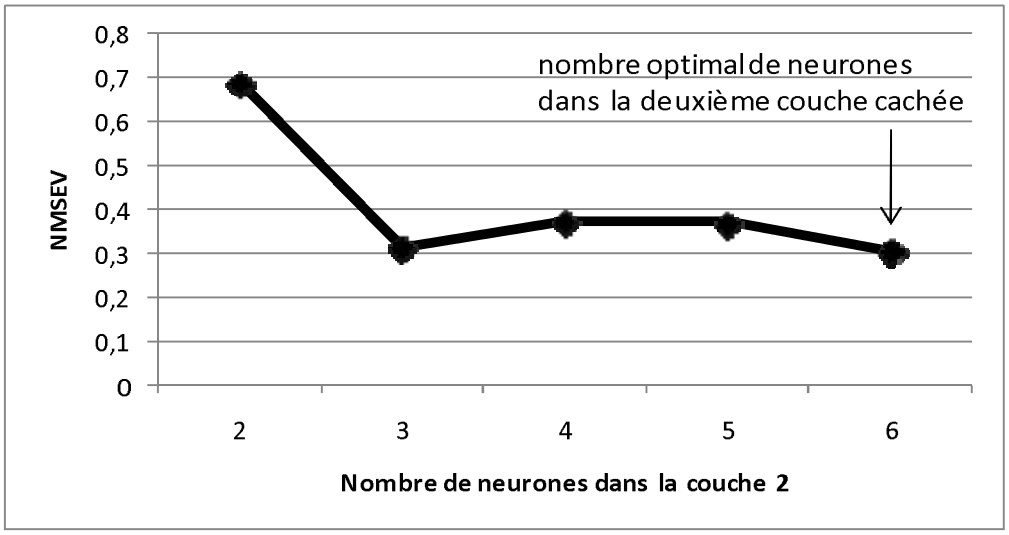


Figure 4.5: Evolution de l'erreur de validation en fonction du nombre de neurones dans la deuxième couche cachée du régresseur MLP sur la base de jus d'orange.

En ce qui concerne la base de pomme la variation du nombre de neurones dans les deux couches cachées était de 5 à 25. Le nombre optimal de neurones était 10 dans les deux couches cachées.

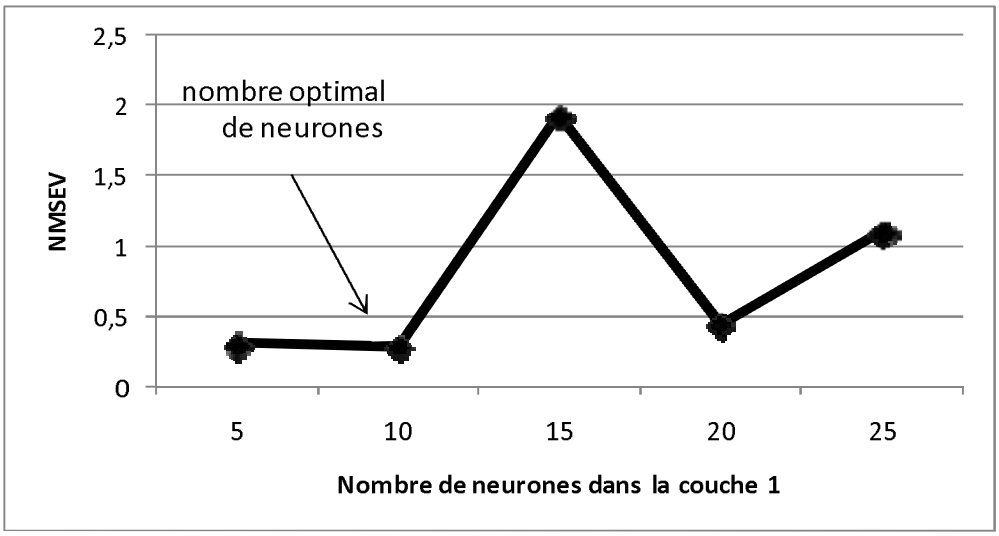


Figure 4.6: Evolution de l'erreur de validation en fonction du nombre de neurones dans la première couche cachée du régresseur MLP sur la base de pomme.

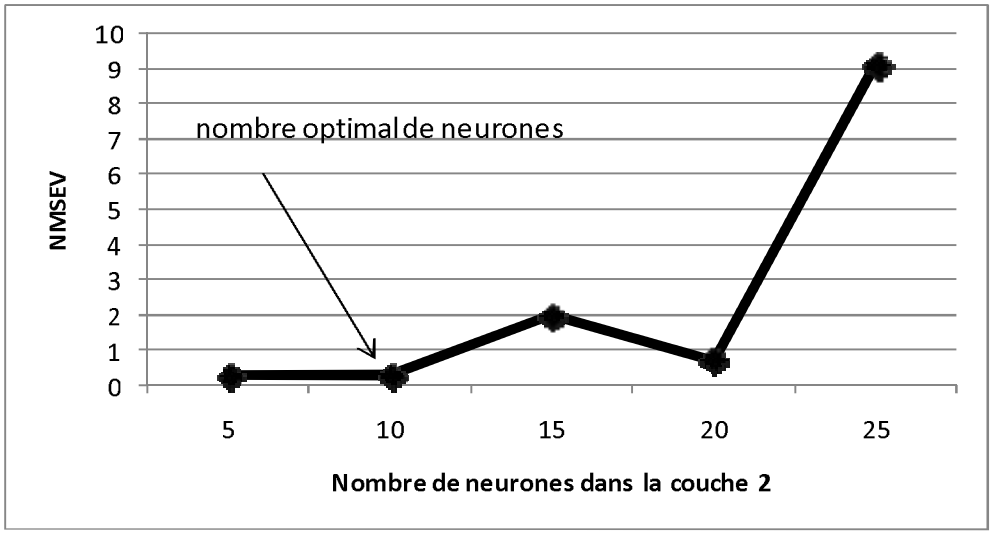


Figure 4.7: Evolution de l'erreur de validation en fonction du nombre de neurones dans la deuxième couche cachée du régresseur MLP sur la base de pomme.

Concernant le modèle RBF, le nombre de neurones (ncc) dans la couche cachée a été varié de 5 à 30 et la largeur de la gaussienne δ de 1 à 3.2 pour la base wine. Le

meilleur résultat a été obtenu avec 10 neurones cachés et δ est égale à 2.6 (voir figures 4.8 et 4.9).

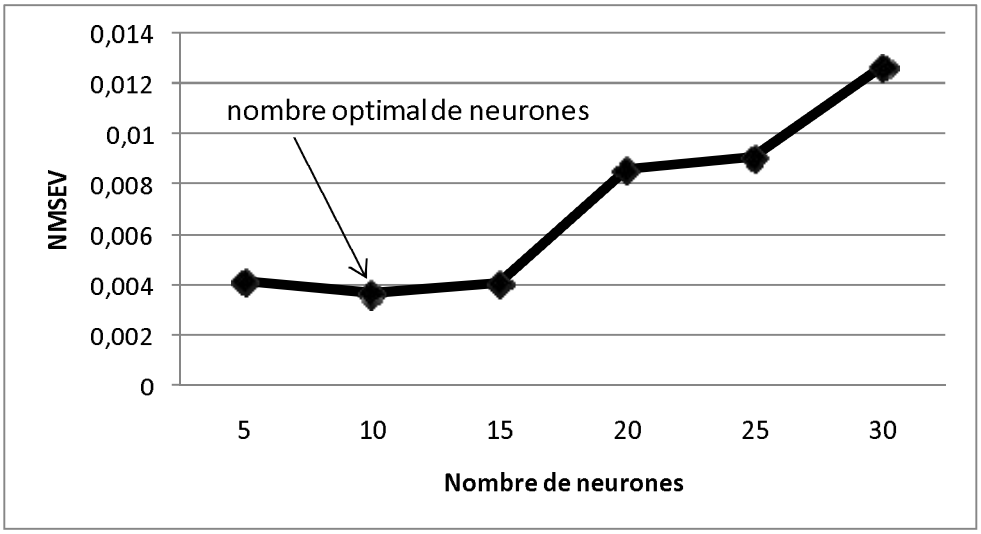


Figure 4.8: Variation de l'erreur de validation en fonction du nombre de neurones obtenue par la méthode RBF pour la base de données wine.

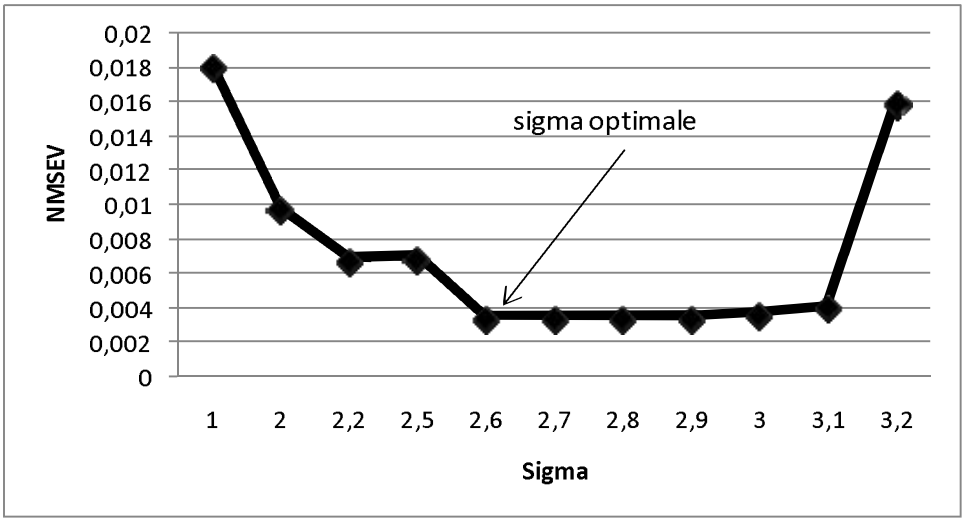


Figure 4.9: Variation de l'erreur de validation en fonction de sigma obtenue par la méthode RBF pour la base de données wine.

Dans le cas de la base de données de jus d'orange, la variation de ces paramètres était de 4 à 20 neurones dans la couche cachée et de 0.2 à 3 concernant le paramètre sigma. Les valeurs optimales de ces paramètres étaient : le nombre de neurones est égal à 16 et δ est égale à 0.7 comme indiqué dans les figures 4.10 et 4.11 respectivement.

Dans le cas de la base de pomme, le nombre de neurones dans la couche cachée a été varié de 2 à 18 et δ a été variée de 1 à 17. Les meilleurs résultats ont été obtenus avec 18 neurones cachés et δ est égale à 13 (figures 4.12, 4.13).

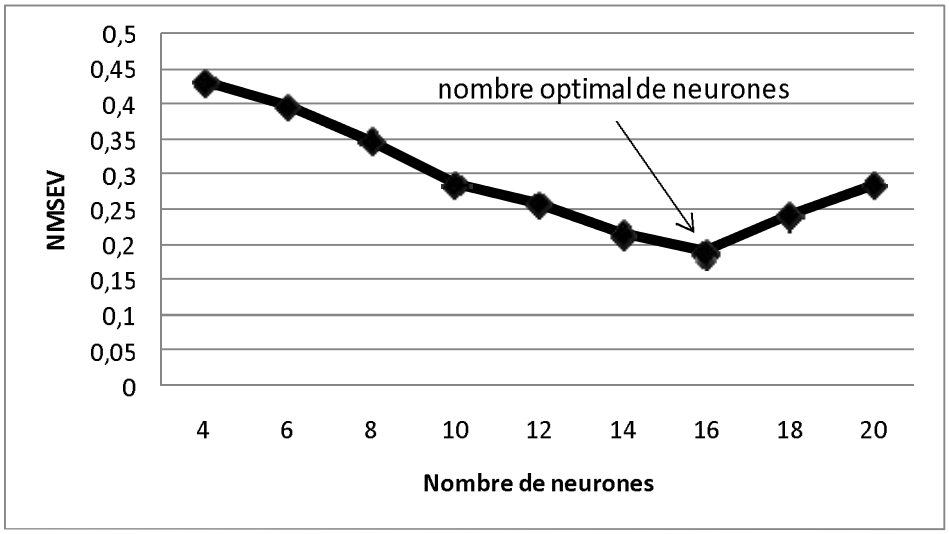


Figure 4.10: Variation de l'erreur de validation en fonction du nombre de neurones obtenue par la méthode RBF pour la base de données de jus d'orange.

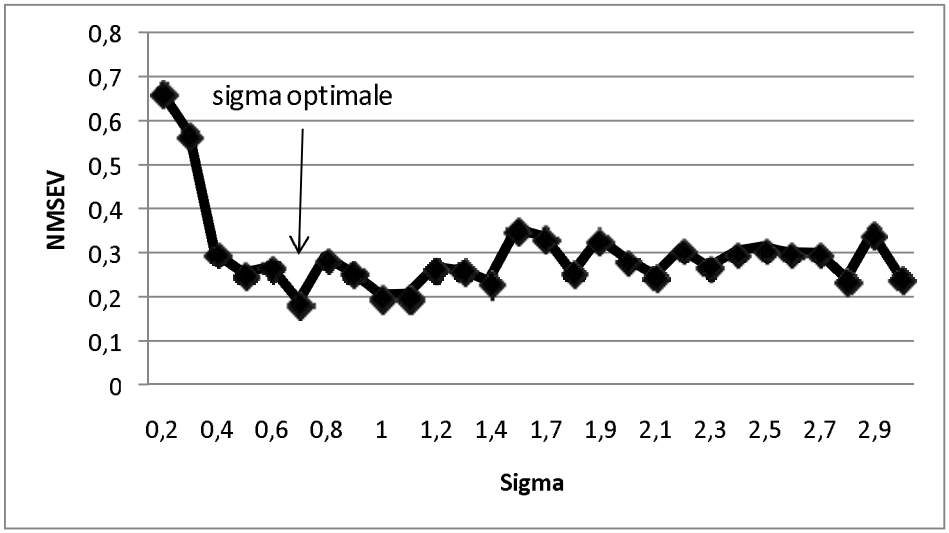


Figure 4.11: Variation de l'erreur de validation en fonction de sigma obtenue par la méthode RBF pour la base de données de jus d'orange.

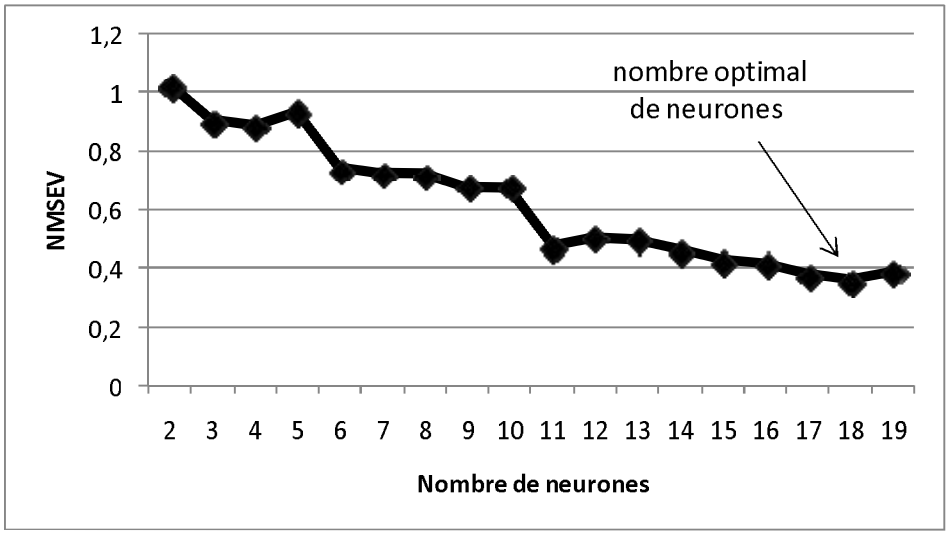


Figure 4.12: Variation de l'erreur de validation en fonction du nombre de neurones obtenue par la méthode RBF pour la base de données de pomme.

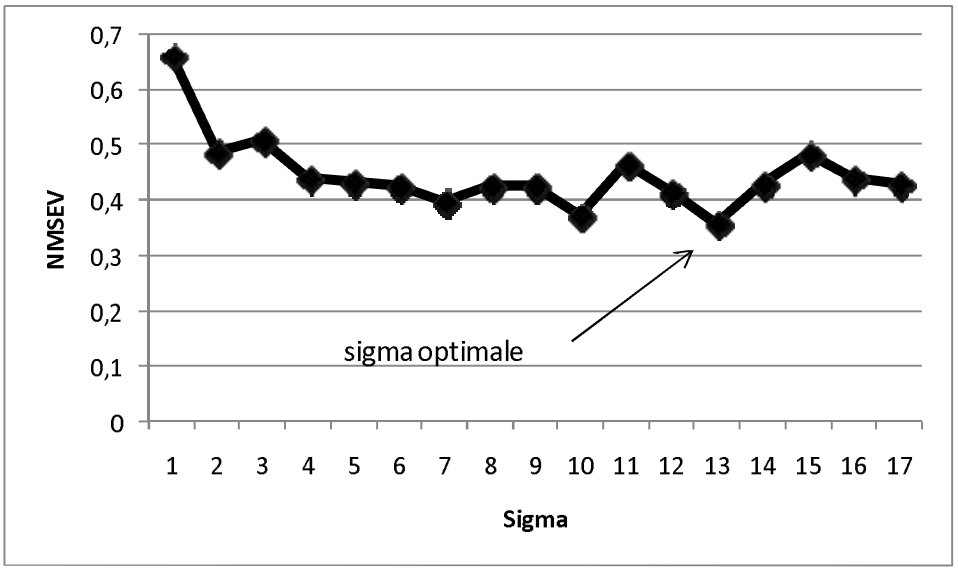


Figure 4.13: Variation de l'erreur de validation en fonction de sigma obtenue par la méthode RBF pour la base de données de pomme.

L'évaluation des régresseurs entraînés en termes de l'erreur quadratique moyenne normalisée NMSE sur l'ensemble de test a donné les valeurs indiquées dans les tableaux 4.1, 4.2 et 4.3 pour les trois bases de données wine, jus d'orange et pomme respectivement. Notons que le temps d'apprentissage de chaque régresseur est indiqué

en seconde par T(s). Nous observons empiriquement que le meilleur régresseur est le MLP pour la base de données wine, tandis que le moins précis est le régresseur RBF. En revanche, pour la base de données de jus d'orange, parmi les trois régresseurs utilisés, le modèle SVM était le meilleur régresseur alors que le modèle MLP a donné de mauvaises performances en termes de NMSE. Le modèle SVM, qui s'est avéré très précis pour la base de données de jus d'orange, a complètement échoué pour la base de données de pomme. Le modèle RBF s'avère le plus efficace entre les trois régresseurs considérés. Afin d'atténuer les risques possibles de sur-apprentissage (over-fitting), la régression a été appliquée sur un sous-ensemble de variables obtenu à partir de l'ensemble des variables (données spectrales) de chaque spectre. En particulier, les meilleures trente variables ont été sélectionnées en utilisant la procédure forward-sélection décrite dans [57] et ont été adopté pour toutes les méthodes de régression utilisées. Noter que d'autres techniques de sélection de variables rapportées dans la littérature peuvent être adoptées également.

Méthodes	**NMSE**	**T(s)**
RBF (δ=2,6, ncc=10)	0.0051	30.5
MLP (nc1=6, nc2=2)	**0.0035**	1104
SVM (C=200,γ=0.001)	0.0047	3.6

Tableau 4.1: Résultats obtenus sur l'ensemble de test de la base de données wine par les méthodes de régression RBF, MLP et SVM-RBF.

Méthodes	**NMSE**	**T(s)**
RBF (δ=0,7, ncc=16)	0.3750	55.6
MLP (nc1=8, nc2=6)	0.3867	1461
SVM (γ=0.1, C=58)	**0.3147**	4.2

Tableau 4.2: Résultats obtenus sur l'ensemble de test de la base de données de jus d'orange par les méthodes de régression RBF, MLP et SVM-RBF.

Méthodes	NMSE	T(s)
RBF (δ=2.6, ncc=10)	**0.4845**	42.5
MLP (nc1=4, nc2=3)	0.5832	462
SVM (γ=0.001, C=600)	0.7918	2.5

Tableau 4.3: Résultats obtenus sur l'ensemble de test de la base de données de pomme par les méthodes de régression RBF, MLP et SVM-RBF.

En général, les résultats expérimentaux obtenus dans cette section confirment que le choix d'une technique de régression dépend fortement de la base de données considérée. Pour une base de données, une approche peut apparaître la meilleure, alors qu'elle peut échouer complètement pour une autre.

4.5.2. Résultats obtenus par les stratégies de fusion

Nous évaluons dans cette phase l'intérêt de la fusion de différents types de regresseurs par rapport à l'usage d'un seul type de regresseur, via l'application du modèle RMS. Pour construire le modèle, trois régresseurs différents ont été envisagés. Ceux-ci sont les algorithmes RBF, MLP et SVM appliqués dans la section précédente. Les sorties des régresseurs ont été exploitées pour améliorer la robustesse et la précision du processus d'estimation. Nous avons considéré trois méthodes différentes pour la conception du modèle RMS à savoir ACS, WCS et NLFS comme décrites dans les sections (4.4.1.1-A, B, et C). L'évaluation des performances des différentes stratégies de fusion a été réalisée avec trente variables. Le tableau 4.4 présente les résultats pour les trois bases de données (wine, jus d'orange et pomme). Pour la base wine, la meilleure précision a été obtenue avec NLFS était égal à (0,0038). Concernant la base de jus d'orange, les résultats présentés par les deux stratégies de fusion (ACS et WCS) sont relativement proches les uns des autres (voir tableau 4.4). Malgré sa simplicité, ACS s'est avérée efficace. Sa précision a été trouvée égale à (0,3030). L'approche RMS mis en œuvre avec la stratégie de fusion NLFS a montré une robustesse et a même été capable d'améliorer l'erreur d'estimation du meilleur régresseur unique de

l'ensemble (0,2964 contre 0,3147). En ce qui concerne la base de pomme; malgré la présence d'un régresseur qui est le moins précis dans l'ensemble (régresseur SVM), toutes les stratégies de fusion ont prouvé leur efficacité car elles ont permis d'obtenir une meilleure précision finale que celle obtenue par le meilleur régresseur simple de l'ensemble.

Stratégies de fusion	**Base de Wine**		**Base de jus'Orange**		**Base de Pomme**	
	NMSE	**TIME(s)**	**NMSE**	**TIME(s)**	**NMSE**	**TIME(s)**
ACS	0.0044	0.06	0.3030	0.16	**0.3865**	0.05
WCS	0.0047	0.09	0.3350	0.36	0.4234	0.07
NLFS	**0.0038**	39.27	**0.2964**	32.1	0.41946	35.32

Tableau 4.4: Résultats obtenus sur les bases de données wine, jus d'orange et pomme par l'approche RMS mise en œuvre avec un ensemble de 3 régresseurs non linéaires.

4.6. Stratégie basée sur la classification

4.6.1. Stratégie de sélection SBS

4.6.1.1. Description

Dans cette approche, le modèle RMS analyse la précision de chaque régresseur inclut dans l'ensemble dans les différentes portions de l'espace des caractéristiques de dimension "d". Cette analyse se traduit en termes d'une partition de l'espace des caractéristiques indiquant quel est le meilleur régresseur en termes d'erreur minimale pour n'importe quelle position dans l'espace des caractéristiques (voir figure 4.14) [18].

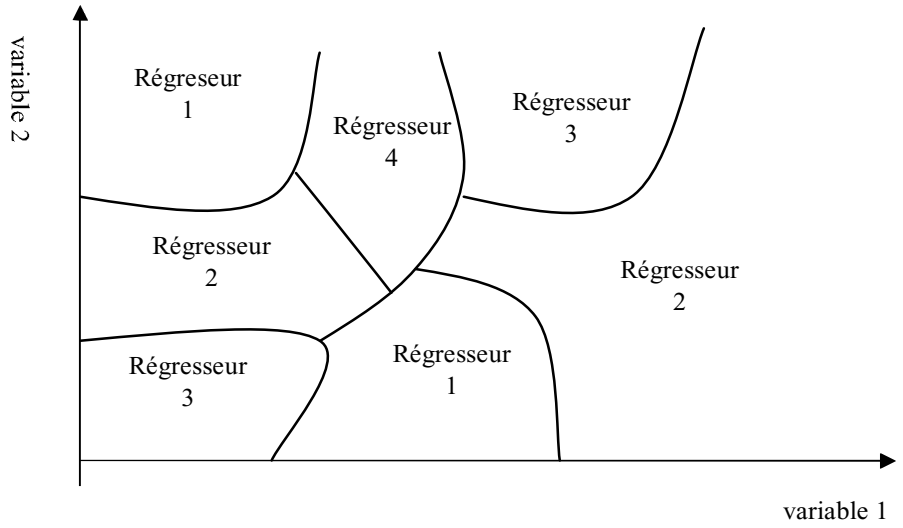


Figure 4.14: Exemple de partition de l'espace de deux entrées (variables) pour un ensemble composé de quatre régresseurs

En d'autres termes, le modèle RMS se comporte comme un sélecteur idéal de la meilleure (la plus précise) estimation trouvée réalisée par l'ensemble des régresseurs disponibles. De cette façon, il est possible d'exploiter au mieux les particularités des différents régresseurs en vue d'augmenter la robustesse et, éventuellement, la précision du processus d'estimation dans l'espace des caractéristiques entier. Une telle idée peut être mise en œuvre par l'utilisation d'un classificateur qui va apprendre d'une manière supervisée la partition optimale de l'espace des caractéristiques de dimension "d" selon un critère donné prédéfini. En fait, l'efficacité de la méthode dépend de sa capacité à exploiter correctement les différentes précisions des régresseurs inclus dans l'ensemble dans différentes parties de l'espace des caractéristiques d'entrée. L'approche de sélection se compose de deux phases: la phase d'apprentissage et la phase d'estimation. La phase d'apprentissage comprend l'identification (parmi l'ensemble des régresseurs uniques disponibles) du meilleur régresseur unique $\hat{E}(x_i)$ pour chaque point x_i de l'espace des caractéristiques (voir figure 4.15).

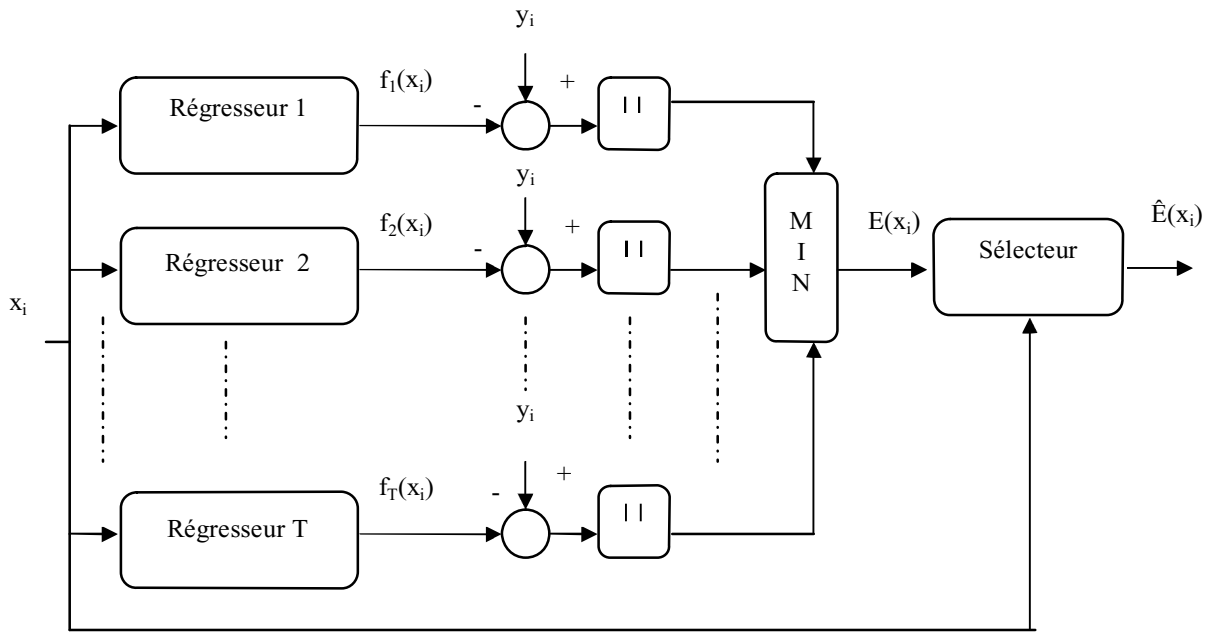


Figure 4.15: Schéma bloc représentant l'approche proposée basée sur la sélection (SBS) [18].

La partition optimale de l'espace des caractéristiques de dimensions "d" dans un ensemble de régions est obtenue conformément à l'analyse de l'ensemble d'apprentissage $x_i (i = 1,2,..., N)$, chacune étant affectée à un régresseur unique donné. Dans notre cas, la notion d'optimalité est exprimée en termes d'erreur minimale absolue MAE (Minimum Absolute Error).

$$\hat{E}(x_i) = \arg\min_{k=1,2,...,T} \left\{ \left| f_k(x_i) - y_i \right| \right\}. \qquad (4.4)$$

La tâche du classificateur est de modéliser aussi bien que possible une telle partition optimale. Au cours de la phase d'estimation, chaque échantillon inconnu $x \in \Re^d$ (pour lequel la valeur réelle d'un paramètre n'est pas connue à priori) est donné en entrée du sélecteur, qui fournit en sortie l'estimation $\hat{E}(x) \in \{1,2,...,T\}$ du régresseur le plus

précis pour l'échantillon considéré. L'estimation $F(x)$ de RMS peut être écrite comme suit:

$$F(x) = f_{\hat{E}(x)}(x). \qquad (4.5)$$

Autrement dit, pour un échantillon donné en entrée du sélecteur, les réponses $f_k(x_i)$ (k=1,2,…T) des différents régresseurs sont comparées avec la sortie désirée y_i par le critère de la valeur absolue comme indiqué dans la figure 4.15. Parmi les erreurs absolues obtenues on choisit celle qui est minimale. Le régresseur le plus précis parmi l'ensemble des régresseurs correspond à cette erreur. Donc, le meilleur régresseur pour l'échantillon considéré correspond à l'erreur minimale trouvée.

Le sélecteur qui agit comme un classificateur a été mis en œuvre par un classificateur SVM parce que ce genre de classificateur a souvent été prouvé pour fournir une meilleure précision de classification que d'autres techniques couramment utilisées dans la reconnaissance de formes, telles que les classificateurs basés sur les réseaux de neurones artificiels MLP et RBF.

4.7. Résultats expérimentaux

4.7.1. Résultats obtenus par la stratégie SBS

Dans cette phase, nous avons appliqué le modèle RMS comme dans la section précédente mais cette fois en considérant la stratégie de fusion SBS pour la conception du modèle comme décrite dans la section 4.6.1. L'évaluation des performances de la stratégie de fusion a été réalisée avec trente variables. Les résultats sont décrits dans le tableau 4.5 pour les trois bases de données.

Stratégie de fusion	**Base de Wine**		**Base de jus d'Orange**		**Base de Pomme**	
	NMSE	**TIME(s)**	**NMSE**	**TIME(s)**	**NMSE**	**TIME(s)**
SBS	0.0040	34.8	0.3109	27.66	0.3849	29.21

Tableau 4.5: Résultats obtenus sur les bases de données wine, jus d'orange et pomme par la stratégie SBS mise en œuvre avec un ensemble de 3 régresseurs non linéaires.

Les histogrammes représentés dans les figures 4.16, 4.17 et 4.18 montrent les résultats de comparaison entre les trois régresseurs (RBF, MLP et SVM) avec la stratégie de fusion SBS obtenus sur l'ensemble de test. On voit clairement que SBS apporte une amélioration importante en termes de NMSE pour les trois bases de données wine, jus d'orange et pomme. Pour la base de jus d'orange, une amélioration d'erreur de 0.3147 du meilleur régresseur unique (SVM) à 0.3109 de la stratégie SBS. En particulier, dans le cas de la base de pomme l'amélioration est plus importante, à savoir une diminution d'erreur de 0.4845 du meilleur régresseur unique (RBF) à 0.3849 de l'approche RMS appliquée par la stratégie SBS. Pour la base de données wine, la stratégie SBS a permis d'obtenir une précision (0.004) comparable à celle atteinte par le meilleur régresseur (MLP) était égale à 0.0038.

Comme on peut le voir, au cours des trois bases de données les stratégies de fusion permettent d'obtenir des performances plus robustes par rapport à ceux réalisés par les régresseurs simples (seuls), avec un temps de calcul moins, comme indiqué dans les tableaux 4.4 et 4.5. Notez que les temps de calcul présentés dans ces tableaux sont ceux des stratégies de fusion seules et donc ne comprennent pas les temps d'apprentissage des régresseurs uniques.

Figure 4.16 : L'erreur quadratique moyenne normalisée (NMSE) obtenue sur l'ensemble de test de la base de données wine par les trois régresseurs uniques et la stratégie de fusion SBS.

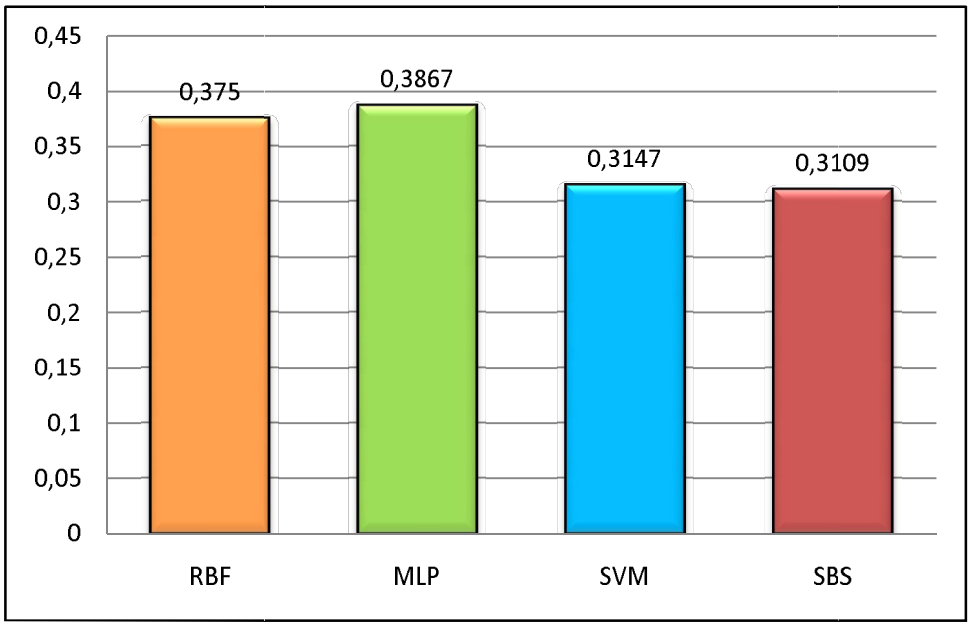


Figure 4.17 : L'erreur quadratique moyenne normalisée (NMSE) obtenue sur l'ensemble de test de la base de données de jus d'orange par les trois régresseurs uniques et la stratégie de fusion SBS.

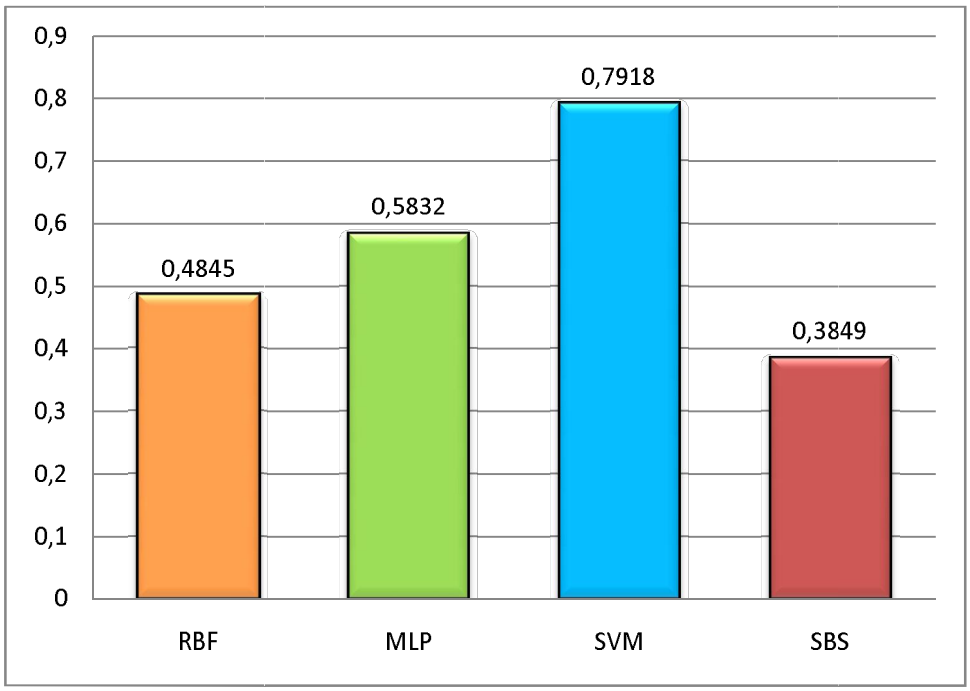


Figure 4.18 : L'erreur quadratique moyenne normalisée (NMSE) obtenue sur l'ensemble de test de la base de données de pomme par les trois régresseurs uniques et la stratégie de fusion SBS.

Les secteurs des figures 4.19, 4.20 et 4.21 présentent le taux de sélection de chaque régresseur (RBF, MLP et SVM) estimé par le classificateur SVM pour les trois bases de données : wine, jus d'orange et pomme respectivement.

Dans le cas de la base wine, 63.3333% des échantillons de la base de données ont été utilisés avec le régresseur SVM, 23.3333% des échantillons de la base de données ont été utilisés avec le régresseur RBF et 13.3333% des échantillons de la base de données ont été utilisés avec le régresseur MLP. Dans le cas de la base de données de jus d'orange, le régreseur SVM a également pris le taux de sélection le plus grand à savoir 91.1765% des échantillons de la base de données doivent être utilisés avec le régresseur SVM, 5.8824% et 2.9412% des échantillons de la base de données doivent être utilisés avec les régresseurs RBF et MLP respectivement. Au contraire avec la base de données de pomme, 81.25% des échantillons de la base de données ont été utilisé avec le régresseur MLP, 12.5% des échantillons de la base de données ont été utilisés avec le régresseur RBF et 6.25% des échantillons de la base de données ont été utilisés avec le régresseur SVM.

D'après les résultats de simulation obtenus, nous pouvons conclure que les échantillons d'une base de données utilisée peuvent donner leurs meilleurs résultats avec l'un des régresseurs, par contre les autres échantillons de la même base de données donnent leurs meilleurs résultats avec un autre régresseur.

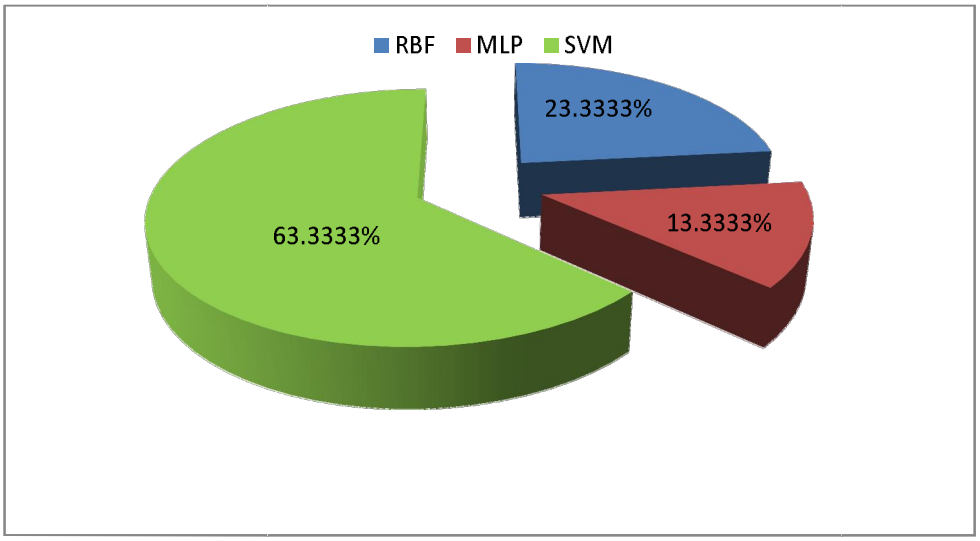


Figure 4.19 : Taux de sélection de chaque régresseur pour la base wine.

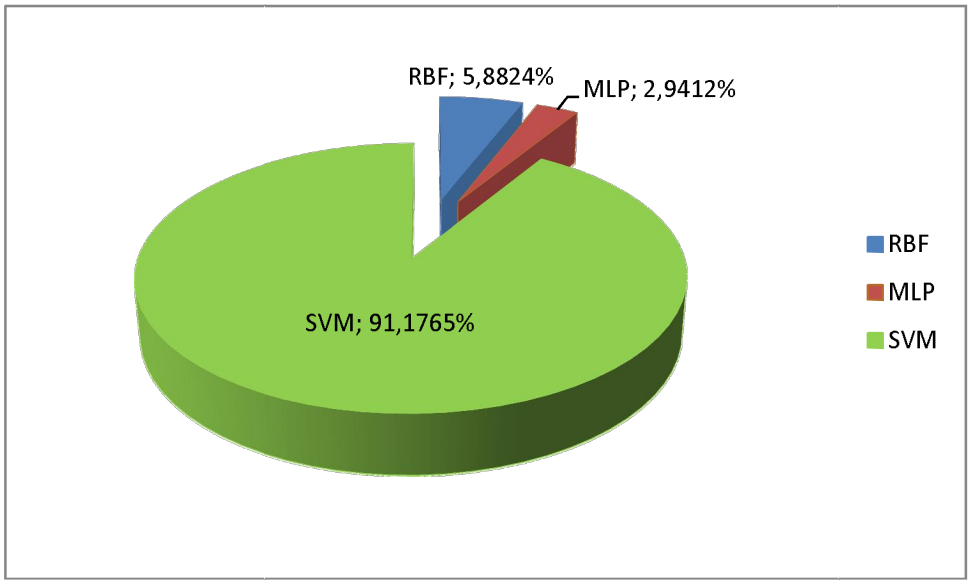


Figure 4.20 : Taux de sélection de chaque régresseur pour la base de jus d'orange.

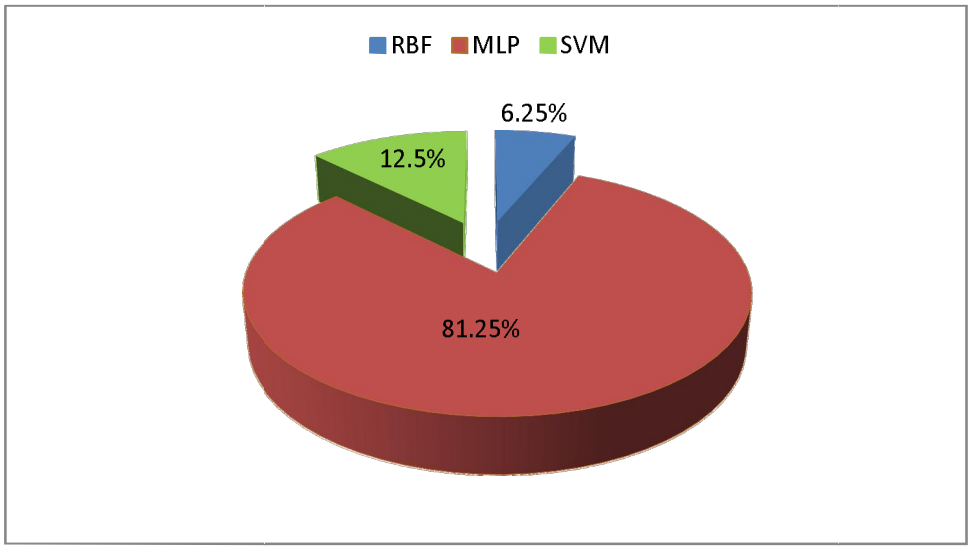


Figure 4.21 : Taux de sélection de chaque régresseur pour la base de pomme

4.8. Conclusion

Dans ce chapitre, une nouvelle approche pour l'estimation de la concentration d'un composant chimique à partir des mesures spectrophotométriques basée sur un système robuste (RMS) a été présentée. L'approche RMS vise à exploiter les particularités d'un ensemble de différents régresseurs pour améliorer la robustesse et la précision du processus de régression. Quatre stratégies différentes pour mettre en œuvre l'approche RMS ont été décrites. Parmi ces méthodes, SBS représente une nouvelle méthode d'estimation de la concentration des composants chimiques. Elle diffère des autres méthodes de fusion classiques par le fait qu'elle n'est pas basée sur une combinaison directe des estimations produites par tous les régresseurs, mais sur un mécanisme de sélection qui identifie la meilleure estimation prévue disponible. A partir des résultats expérimentaux obtenus sur les trois bases de données considérées dans ce travail, il est possible de tirer les conclusions suivantes:

- Le choix d'une approche de régression dépend fortement de la base de données considérée. Une approche peut être la meilleure solution pour une base de données, alors qu'elle peut échouer complètement pour une autre.

- Une solution possible à la question mentionnée dans l'introduction est représentée par une stratégie de régression basée sur la fusion d'un ensemble d'algorithmes de régression différents. En général, les résultats obtenus montrent que par la combinaison d'un ensemble d'algorithmes de régression différents, il est possible de capturer les meilleures particularités pour atteindre un régresseur robuste. Ceci est illustré par le fait qu'elle conduit à des résultats souvent proches de ceux du meilleur algorithme de régression unique de l'ensemble.
- Les résultats obtenus dans ce chapitre confirment que la fusion d'un ensemble de régresseurs est un moyen efficace pour améliorer non seulement la robustesse du processus de régression, mais aussi la précision finale de régression.

Conclusion générale

Travailler avec des données en grande dimension n'est pas une question mathématique ou théorique sans conséquence dans des situations concrètes. Au contraire, la plupart des problèmes d'analyse de données rencontrés dans les applications du monde réel traitent explicitement des données en grande dimension.

Les centres d'intérêt du travail présenté dans le cadre de cette thèse s'articulent autour des problèmes liés à la régression des données en grande dimension. Nous nous sommes intéressés à des applications dans le domaine de la chimiométrie pour traiter les données spectrophotométriques qui sont souvent représentées en grande dimension.

Dans un premier temps, nous avions un objectif principal: tester l'effet de la dimensionnalité sur différentes méthodes de régression.

Les réseaux de neurones artificiels ont été inventés pour résoudre les problèmes là ou d'autres outils traditionnels d'analyse de données échouent. Concernant la régression

dans les espaces de grande dimension, les résultats rapportés dans le troisième chapitre ont montré l'avantage des réseaux de neurones artificiels RBF par rapport aux perceptrons multicouches MLP. Les RBF présentent l'avantage d'être rapides et faciles à utiliser, leurs performances sont moins sensibles aux minima locaux. Cependant, les réseaux de neurones artificiels qui se révèlent efficaces dans certains cas, ne parviennent pas dans d'autres situations en raison de la malédiction de la dimensionnalité.

Une orientation possible pour résoudre le problème de la dimensionnalité a été présentée dans le troisième chapitre; la méthode des machines à vecteurs supports (SVM). Elle représente une méthode d'apprentissage statistique caractérisée par un background théorique solide qui constitue l'un de ses atouts. Elle est marquée par une grande capacité de généralisation et une convergence assurée. Le fait que cette méthode présente une faible sensibilité à la malédiction de la dimensionnalité, rend SVM un bon candidat pour le problème de la régression dans les espaces de grande dimension. Sa comparaison avec d'autres méthodes de la littérature, et notamment avec les méthodes présentées dans le troisième chapitre, a montré de bons résultats, prouvant l'efficacité et la robustesse de cette méthode en tant qu'un outil de régression en spectrophotométrie. D'après les résultats de simulation rapportés dans le troisième chapitre, la méthode SVM a montré une robustesse en termes de précision lors de la réduction du nombre d'échantillons utilisés dans la phase d'apprentissage. La méthode a présenté aussi une grande stabilité offrant ainsi une grande garantie d'ajustement d'un bon modèle aux données. Parmi les différents noyaux étudiés au cours du troisième chapitre, le noyau gaussien était le plus approprié. Les résultats obtenus par simulation nous ont permis de conclure que SVM est une méthode bien adaptée pour les problèmes de grande dimension. L'avantage présenté par l'algorithme SVM est que la solution produite correspond à l'optimum d'une fonction convexe. Elle ne possède donc pas plusieurs optima locaux comme pour les réseaux de neurones artificiels, mais un optimum global.

Dans un deuxième temps, nous avions deux objectifs principaux: 1) combiner plusieurs types de régresseurs afin d'améliorer la qualité de la prédiction. De cette

façon, il était possible d'exploiter au mieux les particularités des différents régresseurs pour améliorer la robustesse et, éventuellement, la précision du processus de régression. La diversité dans l'ensemble des régresseurs et un choix approprié de la stratégie de fusion sont deux points essentiels que nous avons pris en compte, afin d'obtenir une meilleure robustesse et précision, 2) puis, analyser la précision de chaque régresseur inclut dans l'ensemble des régresseurs dans les différentes portions de l'espace des caractéristiques pour atteindre un régresseur robuste. Nous pouvons dire que ces objectifs ont été atteints. En effet, nous avons tout d'abord présenté trois stratégies de fusion: ACS et WCS. Ces stratégies sont basées sur le principe de la combinaison linéaire. L'autre stratégie est la stratégie de fusion non linéaire NLFS. Les trois stratégies diffèrent les unes des autres à la fois dans la procédure de fusion et le fait qu'elles sont supervisées ou non supervisées. Ces trois méthodes ont prouvé leur efficacité sur trois bases de données réelles dans le domaine agroalimentaire, et ont montré l'amélioration apportée sur les performances du processus de régression en termes de précision. Ensuite, nous avons proposé une nouvelle approche pour l'estimation de la concentration d'un composant chimique à partir des mesures spectrophotométriques, nommée SBS (Selection Based Strategy). Elle diffère des autres méthodes de fusion proposées par le fait qu'elle n'est pas basée sur une combinaison directe des estimations produites par tous les régresseurs, mais sur un mécanisme de sélection qui identifie la meilleure estimation prévue disponible. Les résultats de simulation présentés dans le quatrième chapitre nous ont permis de conclure que: le choix d'une approche de régression dépend fortement de la base de données considérée. C'est-à-dire qu'une approche peut être la meilleure solution pour une base de données, alors qu'elle peut échouer complètement pour une autre, les méthodes de fusion proposées surpassent en termes de précision les méthodes traditionnelles de régression. Les résultats obtenus confirmaient aussi que la fusion d'un ensemble de régresseurs est un moyen efficace pour améliorer non seulement la robustesse du processus de régression, mais aussi la précision finale de régression.

Dans ce dernier paragraphe, nous allons présenter brièvement les perspectives de recherche qui s'offrent à nous. Nous pouvons dégager de cette thèse les points d'extension suivants:

- Prolonger le test des méthodes proposées à d'autres applications de l'estimation des paramètres.
- Améliorer les résultats de régression par l'inclusion d'autres modèles dans le système régresseur multiple.
- La formulation standard de l'algorithme SVR présenté dans le premier chapitre ne considère que le problème à sortie unique. Nous proposons d'étendre la régression à vecteurs de support à des objectifs multidimensionnels. Cela permet d'introduire une généralisation de SVR pour résoudre le problème de régression pour plusieurs variables en prenant en compte les relations non linéaires entre les variables de sortie.

Bibliographie

[1] Y. Ozaki, R. Cho, K. Ikegaya, S. Muraishi, K. Kawauchi, Potential of near-infrared Fourier transform Raman spectroscopy in food analysis, Appl. Spectrosc., 46:1503-1507, 1992.

[2] S. Sekulic, H. W. Ward, D. Brannegan, E. Stanley, C. Evans, S. Sciavolino, P. Hailey, P. Aldridge, On-line monitoring of powder blend homogeneity by near-infrared spectroscopy, Anal. Chem., 68: 509-513, 1996.

[3] M. Blanco, J. Coello, J. M. Garcia Fraga, H. Iturriaga, S. Maspoch, J. Pagès, Determination of finishing oils in acrylic fibers by Near-Infrared Reflectance Spectroscopy, Analyst., 122: 777-781, 1999.

[4] P. Geladi and E. Dabakk. An overview of chemometrics applications in NIR spectrometry. J. NIR Spectrosc., 3:119–132, 1995.

[5] D. Bertrand, E. Dufour, La spectroscopie infrarouge et ses applications analytiques, Collection sciences et techniques agroalimentaires, 2e édition, 2006.

[6] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, NY, 1991.

[7] N. Draper and H. Smith. Applied regression analysis. Wiley, New York, 1981.

[8] R. Gunst and R. Mason. Regression analysis and its applications. M. Dekker, New York, 1980.

[9] L. Massart, B. G. M. Vandeginste, L.M. C. Buydens, S. De Jong, P. J. Lewi, and J. Smeyers-Verbeke. Handbook of Chemometrics and Qualimetrics: Part A. Elsevier Science, Amsterdam, first edition, 1997.

[10] P. Geladi. Some recent trends in the calibration literature. Chemometrics and Intelligent Laboratory Systems, 60:211–224, 2002.

[11] P. Geladi and B. R. Kowalski. Partial least squares regression: A tutorial. Analytica Chimica Acta, 185:1–17, 1986.

[12] B. Mevik, V. Segtnan, T. Naes. Ensemble methods and partial least squares regression. J. Chemom., 18:498–507,2004.

[13] M. Meurens, Acquisition et traitement du signal spectrophotométrique. dans La spectroscopie infrarouge et ses applications analytiques,D. Bertrand et E. Dufour, pages 233–245. Collection sciences et techniques agroalimentaires, 2006.

[14] C.E. Miller. NIR News, 4(6):3–5, 1993.

[15] K. Hornik, M. Stinchcombe, and H. White, Multilayer feedforward networks are universal approximators, Neural Networks, vol. 2, pp. 359–366, 1989.

[16] R. J. Howlett and L. C. Jain. Radial Basis Function Networks 2: New Avances in Design. Physica-Verlag Heidelberg Printed in Germany, 1st edition, 2001.

[17] L. Douha, N. Benoudjit, F. Douak and F. Melgani, Support Vector Regression in Spectrophotometry: An Experimental Study, Critical Reviews in Analytical Chemistry, 42:214–219, 2012.

[18] L. Douha, N. Benoudjit and F. Melgani, A robust regression approach for spectrophotometric signal analysis, Journal of Chemometrics, 26:400–405, 2012.

[19] M. Verleysen, Learning high-dimensional data, Limitations and Future Trends in Neural Computation, S. Ablameyko et al. (Eds.), ISO Press, pp.141-162, 2003.

[20] M. Verleysen and D. François, The Curse of Dimensionality in Data Mining and Time Series Prediction, IWANN 2005, LNCS 3512, pp. 758-770.

[21] R Bellman. Adaptive Control Processes: A Guided Tour. Princeton University Press, New Jersey, 1961.

[22] M. Verleysen, Machine learning of high-dimensional data: Local artificial neural networks and the curse of dimensionality, Thèse présentée en vue de l'obtention du grade d'agrégé de l'enseignement supérieur, Université catholique de Louvain, Belgium, décembre 2000.

[23] Jacques Goupy, La régression PLS 1, cas particulier de la régression linéaire séquentielle orthogonale (RLSO), ReConFor, 24 avenue Perrichont. 75016 Paris. France.

[24] N. Draper and H. Smith. Applied regression analysis. Wiley, New York, 1981.

[25] R. Gunst and R. Mason. Regression analysis and its applications. M. Dekker, New York, 1980.

[26] N. Benoudjit, Variable selection and neural networks: Application in infrared spectroscopy and chemometrics, VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG, Dudweiler Landstr. Germany, 2008.

[27] Hoskuldsson. PLS regression methods. J. Chemometrics, 2:211–228, 1988.

[28] M. Tenenhaus. La régression PLS théorie et pratique. Editions Technip., Paris, 1998.

[29] E. Davalo, P. Naim, Des réseaux de neurones, Eyrolles, 1993.

[30] F.Blayo et M.Verlysen, Les réseaux de neurones artificiels, Presses universitaires de France, Paris, 1996.

[31] Marc Parizeau, Réseaux de neurons GIF-21140 et GIF-64326, université Laval, 2004.

[32] Probabiltés, statistiques et modélisation, librairie AI ACCESS, 91940 Les Ulis, France.

[33] Simon Haykin, Neural Networks: A Comprehensive Foundation, second edition, Prentice Hall International, Inc. 1999.

[34] G. Dreyfus J.-M. Martinez M. Samuelides M. B. Gordon F. Badran S. Thiria L. Hérault, Réseaux de neurones: méthodologie et applications, Eyrolles 2002.

[35] T. Poggio, F. Girosi. Networks for approximation and learning, Proceedings of the IEEE, 78:1481–1497, 1990.

[36] C. M.Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[37] L. Robert Harvey, Neural Networks Principles, Prentice Hall International Editions, 1994.

[38] Young-Sup Hwang and Sung-Yang Bang. An Efficient Method to Construct a Radial Basis Function Neural Network Classifier. Neural Networks, 10(8):1495–503, 1997.

[39] J. Park and I. W. Sandberg. Universal Approximation Using Radial-Basis-Functions Networks. Neural Comput, 3:246–257, 1991.

[40] L. Faussett, Fundaments of Neural Networks Architectures Algorithmes and Applications, Prentice Hall International Editions, 1994.

[41] N. Benoudjit, M. Verleysen, On the .Kernel Widths in Radial-Basis Function Networks, Neural Processing Letters 18:139-154, 2003.

[42] J. Moody and C.J.Darken, Fast learning in networks of locally-tuned processing units, Neural Computation, 1:281-294, 1989.

[43] R. Steve Gunn, Support Vector Machines for Classification and Regression, Technical Report, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, University of Southampton, 1998.

[44] F. Melgani L. Bruzzone, Classification of Hyperspectral Remote Sensing Images With Support Vector Machines IEEE Transactions. Geoscience and Remote Sensing, 42:1778–1790, 2004.

[45] H. Frezza-Buet, Machines à Vecteurs Supports Didacticiel, Ecole Supérieure D'électricité Supélec, 2012.

[46] http://www.imtech.res.in/raghava/rbpred/svm.jpg.

[47] V. Vapnik. Statistical Learning Theory. Wiley: New York, 1998.

[48] G. Camps-valls, L.Bruzzone, J.L.Rojo-'Alvarez and F.Melgani, Robust Support Vector Regression for Biophysical Variable Estimation From Remontly Sensed Images, IEEE Geoscience and Remote Sensing Letters, Vol. 3, N°. 3, July 2006.

[49] Vapnik, V. N. The Nature of Statistical Learning Theory, 2nd ed.;Springer: New York, 2000.

[50] Smola and B. Schölkopf, A tutorial on support vector regression, Royal Holloway College, Univ. London, London, U.K., NeuroCOLT Tech. Rep. NC-TR-98-030, 1998.

[51] N. Cristianini and J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, first edition, 2000.

[52] J. C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining Knowl. Discov., 2:121–167, 1998.

[53] Set of tutorials on SVM's and kernel methods [Online]. Available: http://www.kernel-machines.org/tutorial.html.

[54] D. Bertrand et E.M. Qannari, Introduction à la chimiométrie, dans La spectroscopie infrarouge et ses applications analytiques,D. Bertrand et E. Dufour, pages 309–315. Collection sciences et techniques agroalimentaires, 2006.

[55] Dataset provided by Prof. Marc Meurens, Université catholique de Louvain, BNUT, meurens@bnut.ucl.ac.be. Dataset available from www.ucl.ac.be/mlg/.

[56] T. Eklove, P. Martenson, and I. Lundstrom. Selection of variables for interpreting multivariate gas sensor data. Analytica Chimica Acta, 381:221–232, 1999.

[57] N. Benoudjit, E. Cools, M. Meurens, M. Verleysen. Chemometric calibration of infrared spectrometers: selection and validation of variables by non-linear models. Chemom. Intell. Lab. Syst., 70:47–53, 2004.

[58] L. Bruzzone, F. Melgani, Robust Multiple Estimator Systems for the Analysis of Biophysical Parameters From Remotely Sensed Data, IEEE Transactions on Geoscience and Remote Sensing, Vol. 43, N°. 1, January 2005.

[59] Y. Bazi and F. Melgani, Semisupervised PSO-SVM Regression for Biophysical Parameter Estimation, IEEE Transactions on Geoscience and Remote Sensing, 45: 1887-1895, 2007.

[60] H. Wolpert, Stacked generalization, Neural Networks, 5:241–259, 1992.

[61] A. Jacobs, Methods for combining expert's probability assessment, Neural Comput., 7: 867–888, 1995.

[62] J. A. Benediktsson and P. H. Swain, Consensus theoretic classification methods, IEEE Trans. Syst., 22:688-704, 1992.

[63] R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edn.,Wiley, New York, 2001.