Approches parcimonieuses pour la sélection de variables et la classification : application à la spectroscopie IR de déchets de bois

Leïla Belmerhnia

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Approches parcimonieuses pour la sélection de variables et la classification : application à la spectroscopie IR de déchets de bois

THÈSE

présentée et soutenue publiquement le 2 mai 2017

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention automatique, traitement de signal et des images et génie informatique)

par

Leïla BELMERHNIA

Composition du jury

Président :	Régis Lengellé	Professeur à l'Université de Technologie de Troyes
Rapporteurs :	André Ferrari	Professeur à l'Université de Nice-Sophia Antipolis
	Alain Rakotomamonjy	Professeur à l'Université de Rouen Normandie
Examinateurs :	Cédric Carteret	Professeur à l'Université de Lorraine
	David Brie	Professeur à l'Université de Lorraine
	El-Hadi Djermoune	Maître de conférences à l'Université de Lorraine
Invités :	Antoine Bourely	Directeur R&D Pellenc ST
	Eric Masson	Responsable Pôle R&D Crittbois

Centre de Recherche en Automatique de Nancy — UMR 7039

Mis en page avec la classe thesul.

Remerciements

Mes remerciements les plus profonds et les plus sincères vont en premier à **David Brie**, Professeur à l'université de Lorraine, et à **El-Hadi Djermoune**, Maître de conférence à l'université de Lorraine, pour leur excellent encadrement et leurs conseils précieux tout au long de la thèse. Pour vous avoir côtoyés depuis plus de 4 ans, votre rigueur et votre exigence de travail m'ont toujours incitées à pousser plus loin mes réflexions. Je vous remercie également pour votre patience et surtout votre confiance en mon travail qui m'a permis d'évoluer tant sur le plan professionnel que sur le plan personnel.

Je tiens à exprimer ma gratitude à l'ensemble des membres du jury de thèse qui m'ont fait l'honneur d'examiner mon travail et de me faire part de leurs appréciations. Merci à **Monsieur Régis Lengellé**, Professeur à l'Université de Technologie de Toyes, pour avoir accepté de présider le jury de thèse de ma soutenance. Je remercie tout particulièrement **Monsieur André Ferrari** Professeur à l'Université de Nice-Sophia Antipolis, et **Monsieur Alain Rakotomamonjy**, Professeur à l'Université de Rouen Normandie pour l'intérêt et le temps qu'ils ont consacré à l'évaluation de mon travail de recherche, et surtout pour les commentaires avisées qu'il ont apportés à mon manuscrit et pendant ma soutenance de thèse.

Mes remerciements chaleureux vont à **Cédric Carteret** pour son aide précieuse lors de ma première année de thèse. Ses conseils et sont encadrement pour l'acquisition de mes données au LCPME m'ont été d'une grande utilité pour bien entamer mes travaux. Je me souviendrai toujours de sa sympathie, de son enthousiasme et de ses grandes qualités de travail.

Je remercie également nos partenaires du projet TRISPIRABOIS pour la collaboration fructueuse qui nous a unie depuis le début de ce projet. En particulier, je tiens à exprimer ma gratitude à **Monsieur Antoine Bourely**, Directeur R&D de Pellenc ST et **Monsieur Eric Masson**, Responsable Pôle R&D au Crittbois d'avoir accepté de participer au jury de thèse.

Merci aux doctorants **Marc**, **Yinyin**, **Mamadou** et **Ludivine**, qui ont partagé mon bureau pendant mes années de thèse. Merci pour votre compagnie au quotidien et l'ambiance calme et sereine dont j'ai joui pour la réalisation et la rédaction de ma thèse. Je remercie également tous les autres doctorants : **Karima**, **Magalie**, **JB**, **Simon**, **Marie**... pour leur amitié et pour les moments agréables qu'on a partagés. Merci aux permanents du laboratoire : **Sam**, **Magalie**, **Sebastien**, **Charles**... pour la qualité des discussions scientifiques et la convivialité des échanges lors de diverses occasions. Je remercie aussi l'équipe de l'IUT Brabois pour leur contribution au bon déroulement de ma thèse surtout **Brice** et **Michaël**.

Un grand merci à **Christelle**, **Monique** et **Sabine** pour leur gestion de mon dossier universitaire ainsi que leur disponibilité systématique pour m'assister dans mes démarches administratives.

Je tiens également à remercier ma famille pour leur soutien infaillible pendant ces années de thèse. Enfin, un merci tout particulier à **Michel** pour son amour, ses encouragements et sa disponibilité en tout

temps.

À la mémoire de mon très cher père,

Tables des matières

Table des figures

Liste des tableaux

Introduction générale

1 Contexte et objectifs

Ce travail a été mené au Centre de Recherche en Automatique de Nancy (CRAN), Université de Lorraine, CNRS. Il s'inscrit dans le cadre du projet TRISPIRABOIS [1] bénéficiant d'un financement FUI (Fond Unique Interministériel). Ce projet réunit l'expertise de plusieurs entités liées au secteur de tri et de recyclage de bois (Egger, Pellenc ST, Crittbois) ainsi que deux structures de recherche de l'Université de Lorraine pour le développement de méthodes d'analyse de données et d'interprétation physico-chimique (CRAN, LCPME). L'objectif principal du projet est la valorisation des déchets de bois en conditions industrielles, c'est-à-dire le recyclage des bois non pollués pour incorporation dans la fabrication de panneaux de particules et la valorisation thermique des bois pollués. La technologie retenue pour la caractérisation des déchets de bois est la spectroscopie proche infrarouge. Avec les techniques actuelles de spectroscopie IR, chaque spectre est composé d'un millier de longueurs d'ondes. Afin de réduire les coûts de calcul [2] et répondre à la contrainte « temps-réel » de l'application, il est nécessaire de sélectionner dans l'ensemble des bandes spectrales celles qui sont susceptibles de révéler la présence de polluants.

L'objectif de ce travail de thèse est double : développer des méthodes de sélection variables en utilisant des données réelles de spectroscopie IR et mettre en place un systèmes de traitement de données permettant un tri performant et efficace des déchets de bois. Le problème de tri de bois est formulé comme une classification à deux hypothèses, dans le sens où un échantillon de bois doit être récupéré s'il appartient à la catégorie des bois purs, ou rejeté s'il appartient à la catégorie des bois indésirables. Dans le cadre de du projet, ce problème est particulièrement délicat pour les raisons suivantes.

La présence de plusieurs groupes dans chaque catégorie de bois La formulation du problème initial comme un problème bi-classes, suppose que les groupes appartenants à la même catégorie ont des caractéristiques communes. Ceci, n'est pas toujours le cas, car des groupes de bois ayant des revêtements, des peintures ou d'autres types de finition, peuvent ne pas avoir des caractéristiques spectrales communes avec les groupes de bois bruts de la même catégorie ou, du moins, difficiles à discriminer. Cet aspect rend complexe la phase d'apprentissage et peut induire des erreurs de classification.

1. TRI par différentes SPetrsocopies, dont l'InfraRouge, des déchets de BOIS.
2. Le système de tri intégrant le spectromètre et l'algorithme de classification sera installé en tête de ligne de recyclage. Le temps d'acquisition et de traitement doit être de quelques dizaines de micro-secondes par spectre.

Les contraintes liées aux systèmes optiques Elles sont de nature technologiques, puisque ces systèmes n'intègrent qu'un nombre limité de bandes spectrales compte tenu des exigences en terme de vitesse d'acquisition. Ceci nous contraint à réduire la taille des données et à choisir des classifieurs robustes même lorsque le nombre de variables est faible. Cette restriction peut engendrer des dégradations des performances dues à la perte d'informations, par rapport à l'utilisation de la totalité des données.

Le choix des bandes spectrales Deux alternatives sont possibles :

1. les bandes spectrales des machines de tri sont fixées au préalable. Dans ce cas, les caractéristiques du modèle de classification doivent correspondre à ce choix,
2. les bandes spectrales ne sont pas fixées à l'avance. On peut alors réaliser une analyse spectrale ou une sélection de variables afin de déterminer les longueurs d'ondes les plus pertinentes pour la classification. Ce travail de thèse s'inscrit dans cette configuration.

La sélection de variables est une étape incontournable pour la classification lorsque le nombre de variables est grand. Cette sélection a pour but de déterminer un sous-ensemble de descripteurs permettant d'une part, de réduire la dimension du problème sans nuire aux performances de classification, et d'autre part d'éviter le phénomène de sur-apprentissage. De façon générale, les méthodes de classification basées sur l'analyse discriminante, la régression logistique et les vecteurs de support n'intègrent pas des techniques de sélection de variables dans leurs procédures. Par conséquent, il est nécessaire de réaliser une sélection de variables de façon disjointe de la classification. Cette sélection peut se faire en amont ou en aval de la classification. En effet, certaines méthodes *ad-hoc* ont été utilisées afin de supprimer les variables qui contribuent le moins à la prise de décision du classifieur. Ce sont des méthodes récursives de type *backward* [61, 113] qui consistent à éliminer les variables ayant le plus petit poids dans la prise de décision. Cependant, il a été démontré dans [114] que ces méthodes ne sont pas toujours efficaces. L'exemple donné est celui du SVM dont les performances commencent à se dégrader lorsque le nombre de variables est réduit de cette manière. À l'inverse, des méthodes de sélection en amont ont été beaucoup plus exploitées dans le cadre de la classification. Ces méthodes permettent de réduire le nombre de variables avant de réaliser la classification. Parmi les approches utilisées, on retrouve celles basées sur des représentations parcimonieuses [62, 160, 31]. Dans le contexte de sélection de variables, la représentation la plus parcimonieuse peut être obtenue en imposant une pénalité sur la cardinalité du support (via la pseudo-norme ℓ_0) du vecteur de coefficients. Cependant, cette approche nécessite la résolution d'un problème combinatoire NP-complet [101]. Une technique commune, et très utilisée pour relaxer ce problème consiste à remplacer la norme ℓ_0 par la norme ℓ_1. Cette approche est une approximation du problème initial qui conduit à la résolution d'un problème convexe dont la solution peut être calculée en utilisant des méthodes plus faciles, tout en garantissant la parcimonie du vecteur de coefficients. Dans ce travail, nous proposons des méthodes de sélection de variables basées sur les décompositions parcimonieuses simultanées. Ces approches sont ensuite mises en œuvre pour la classification des spectres proche infrarouge.

Ce mémoire est organisé en cinq chapitres.

Dans le premier chapitre, nous présentons le projet TRISPIRABOIS et quelques notions de spectroscopie IR. Enfin, nous présentons les conditions d'acquisition des spectres et la base de données réalisée.

Le deuxième chapitre constitue un état de l'art des méthodes d'analyse (ACP, ACP parcimonieuses) et de classification des données spectrales utilisées dans le domaine de la spectrométrie. Les méthodes de classification sont ensuite comparées en utilisant la base de données de déchets de bois. Cette comparaison montre en particulier que les méthodes SVM et G-SVM [46] permettent d'obtenir de bons taux de classification. Elles sont alors retenues pour la suite du travail.

Dans les derniers chapitres, nous nous intéressons aux méthodes d'approximation parcimonieuse simultanées qui peuvent être utilisées pour réaliser une sélection de variables en classification. Dans le troisième chapitre, après quelques rappels sur les méthodes d'approximation parcimonieuse, nous présentons l'extension de quelques méthodes gloutonnes pour l'approximation parcimonieuses simultanée. Des simulations multiples montrent l'avantage des versions simultanées en comparaison avec les approches vectorielles.

Le quatrième chapitre constitue la principale contribution de ce travail. Dans ce chapitre, plusieurs méthodes de sélection de variables intégrant une pénalité de constance par morceaux des lignes de la matrice des coefficients sont présentées. Les critères proposés étant convexes, ils sont minimisés en utilisant l'algorithme FISTA [6]. On montre en particulier que la minimisation est séparable (chaque ligne de la matrice des coefficients est calculée séparément des autres), ce qui permet de réduire significativement le coût de calcul et, par là même, d'utiliser ces méthodes de sélection pour des problèmes de grande dimension. Les méthodes développées sont ensuite comparées en utilisant la base des données spectrales de déchets de bois. On montre également que les approches proposées s'étendent aisément à la classification d'images hyperspectrales obtenues, par exemple, par un imageur *pushbroom*.

Le cinquième chapitre se focalise sur l'aspect industriel du projet TRISPIRABOIS conformément au cahier des charges. Le logiciel développé est présenté et différentes approches permettant de fournir des solutions adaptées aux situations les plus critiques – où la décision de classer ou non un déchet de bois dans une catégorie donnée est difficile (par exemple les déchets surfacés ou peints) – sont proposées. Notamment, on décrit une technique de ré-étiquetage des données dans le but de minimiser les pertes de bons bois et, surtout, de réduire la quantité de bois pollués dans les processus de recyclage.

2 Contributions

Les principales contributions de ce travail sont décrites ci-dessous.

Développement de méthodes gloutonnes pour l'approximation parcimonieuse simultanée

Dans le chapitre 3 nous présentons plusieurs méthodes pour l'approximation parcimonieuse. Une des premières contribution de cette thèse est l'extension de quelques méthodes gloutonnes pour l'approximation parcimonieuse simultanée. Ce problème consiste à trouver une bonne estimation de plusieurs

signaux d'entrée à la fois, en utilisant différentes combinaisons linéaires de quelques signaux élémentaires tirés d'une collection fixe. Les algorithmes gloutons pour lesquels des versions simultanées sont proposées dans le chapitre 3 sont CoSaMP [104], OLS [29] et SBR [123]. Ces approches sont comparées à l'algorithme S-OMP [134]. Nous montrons que dans le cas des signaux présentants des composantes corrélées, les versions simultanées de SBR, CoSaMP et OLS réussissent à mieux reconstruire la matrice d'observation.

Approximation parcimonieuse simultanée et régularisée

Une deuxième contribution de ce travail est la proposition de nouvelles méthodes pour la sélection de variables. Elles sont basées sur l'approximation parcimonieuse simultanée intégrant une contrainte de constance par morceaux des lignes de la matrice de coefficients. Ce type de régularisation est intéressant pour la classification lorsqu'il est appliqué à une matrice bien ordonnée dans le sens où les spectres d'une même catégorie sont regroupés dans des colonnes adjacentes dans la matrice d'observation. Dans le chapitre 4, en utilisant des normes mixtes, nous proposons différentes formes de critères pour le problème de sélection (parcimonie de la matrice de coefficients, parcimonie par groupe, positivité des coefficients). Ces critères sont ensuite minimisés en utilisant l'approche FISTA [6]. Nous montrons notamment que ces minimisations sont séparables par ligne, ce qui permet d'obtenir des algorithmes à un faible coût de calcul et donc adapté aux problèmes à grande dimension. Enfin, des tests expérimentaux de classification réalisés sur différents groupes de spectres IR et d'images hyperspectrales de déchets de bois permettent de valider ces approches et montrent, en particulier, que la régularisation sur les lignes de la matrice de coefficients permet une amélioration significative des taux de classification.

Développement d'un logiciel de traitement basé sur les méthodes proposées

Le dernier chapitre se focalise sur l'aspect industriel du projet Trispirabois conformément au cahier des charges. Le logiciel développé est présenté et différentes approches permettant de fournir des solutions adaptées aux situations les plus critiques – où la décision de classer ou non un déchet de bois dans une catégorie donnée est difficile (par exemple les déchets surfacés ou peints) – sont proposées. Pour ce faire, nous procédons à trois différents tests : le premier vise à augmenter les quantités de bois purs récupérées dans le flux des déchets de bois, ce qui représente un gain économique considérable pour les industriels. Le deuxième a pour objectif de diminuer fortement les proportions des bois pollués dans les processus de recyclage et le troisième à rejeter un type de déchet particulier, à savoir les panneaux MDF/HDF.

3 Liste des Publications

Article de revue (soumission prévue en février 2017)

[T1] L. Belmerhnia, E.-H. Djermoune, D. Brie, and C. Carteret. Regularized Simultaneous Sparse Approximation Methods for Variables Selection and Classification of NIR Spectra and Hyperspectral

Images. Research report. CRAN, Université de Lorraine, CNRS. 2017.

Actes de congrès internationaux avec comité de lecture

[C1] L. Belmerhnia, E.-H. Djermoune, and D. Brie. Greedy methods for simultaneous sparse approximation, In *Proc. EUSIPCO*, pages 1851–1855, Lisbonne, Portugal, September 2014.

[C2] L. Belmerhnia, E.-H. Djermoune, C. Carteret and D. Brie. Simultaneous regularized sparse approximation for wood wastes NIR spectra features selection. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 437–440, Cancun, Mexico, December 2015.

[C3] L. Belmerhnia, E.-H. Djermoune, D. Brie and C. Carteret. A regularized sparse approximation method for hyperspectral image classification. In *19th IEEE Workshop on Statistical Signal Processing 2016 (SSP16)*, Palma de Mallorca, Spain, June 2015.

Actes de congrès nationaux avec comité de lecture

[N1] L. Belmerhnia, E.-H. Djermoune, and D. Brie. Approximation parcimonieuse simultanée constante par morceaux. In *25ième colloque GRETSI sur le traitement du signal et de l'image*, Lyon, France, September 2015.

Communications sans actes

[P1] L. Belmerhnia, E.-H. Djermoune, and D. Brie. Approximation parcimonieuse simultanée constante par morceaux. *Forum Fédération Charles Hermite*, Nancy, Janvier 2016.

[P2] L. Belmerhnia, E.-H. Djermoune, and D. Brie. Algorithmes gloutons pour l'optimisation sous contrainte de parcimonie. Journée *Algorithmes gloutons pour l'optimisation sous contrainte de parcimonie*, GDR ISIS, Paris, Juin 2016.

Chapitre 1

Projet TRISPIRABOIS

1.1 Introduction

1.1.1 Contexte

Le projet TRISPIRABOIS [3] a pour objectif général la gestion des déchets bois et la mise en place de filières de revalorisation pertinentes. Il a pour ambition le développement d'une brique technologique, ouvrant la porte à un tri qualitatif et automatique des déchets de bois en conditions industrielles. Ce projet est le fruit d'une collaboration entre plusieurs partenaires, et regroupe des acteurs dans les domaines de valorisation de déchets (Egger, Pellenc ST), de promotion du matériau bois (Crittbois), de chercheurs (CRAN, LCPME) et des pouvoirs publics (région Lorraine, région Provence-Alpes Côte d'Azur, Conseil général des Vosges). Dans l'industrie de la collecte et du tri sur plateforme, certaines techniques de tri et de séparation existent pour le bois. La Spectroscopie Proche Infrarouge (SPIR) est aujourd'hui capable de séparer les bois des non-bois, ainsi que les bois non-traités (bois de classe A), les bois faiblement adjuvantés : mélaminés, agglomérés, peints et vernis, contreplaqués (bois de classe B) et les bois fortement adjuvantés (bois de classe C), avec une efficacité encore perfectible. Des recherches sont en cours pour améliorer l'efficacité des capteurs IR afin de séparer les bois de classe B [16, 158]. La spectroscopie est utilisée également dans la reconnaissance des matériaux (polymères entre eux, papier, bois). En complément des études sur l'utilisation de la technologie SPIR pour caractériser les déchets bois, il apparaît qu'un gain important sur la qualité du tri peut être apporté en associant cette technique à des méthodes d'analyse et de traitement de données spectroscopiques. En effet, le besoin d'incorporer des déchets de bois dans les panneaux de particules est une tendance de fond qui s'est fortement développée pendant ces dernières années. Ce projet vient donc pour répondre à un réel besoin des industriels du secteur, afin de permettre un meilleur tri des déchets de bois et l'optimisation de ceux-ci, mais aussi, l'augmentation significative du taux de déchets dans les panneaux de particules tout en conservant les propriétés mécaniques de ces derniers.

3. Financement FUI : période de réalisation 01/10/2013- 01/10/2016

1.1.2 Objectifs du projet

Un système de tri de déchets de bois devrait, idéalement, permettre de séparer quatre catégories :

1. Bois purs (faiblement adjuvantés) : Bois parfaitement propres de tous les éléments susceptibles de dégrader les couteaux de la coupeuse (verres, éléments métalliques,...). Elle peut cependant inclure des bois ou des contreplaqués avec finition.
2. Panneaux de particules purs : Les panneaux de particule bruts, peints ou avec finition sont déjà propres, il n'est donc pas nécessaire de passer par toutes les étapes de recyclage. Cette catégorie doit normalement passer directement à la phase de broyage.
3. Bois de recyclage : Bois peint ou surfacé mais ne contenant pas des polluants tels le plomb ou le créosote. Cette catégorie de bois nécessite de passer par une phase de tri et de nettoyage afin d'être incluse dans le panneau.
4. Indésirables : Cette dernière catégorie contient divers éléments, comme les MDF ainsi que les bois créosotés ou avec des sels métalliques ou du plomb. Elle inclut également des éléments tels les mousses, les déchets verts, les matières plastiques ou les textiles.

FIGURE 1.1 – Exemple de déchets de bois : de gauche à droite, bois pur, panneau de particules, MDF/HDF brut, fibre dure brut.

Quelques échantillons des quatre catégories sont présentés sur la Figure 1.1. Dans ce travail de thèse, l'objectif principal est l'extraction des bois purs et le rejet des déchets de bois indésirables. Pour atteindre cet objectif, nous utiliserons une méthodologie combinant la spectrométrie IR et le traitement statistique des données. Ces techniques doivent permettre un tri performant des bois pollués par rapport aux bois naturels, et des bois pollués entre eux. Cette nouvelle capacité de classer finement et rapidement les déchets de bois permettra de diriger les flux de déchets vers les différentes filières de valorisation (bois énergie, panneautiers, incinération), et, ouvrira la porte à de nouvelles valorisations aujourd'hui inaccessibles économiquement (bois composites, bois béton, etc.). Le pilote industriel associé au projet sera installé en tête d'une ligne de recyclage de déchets de bois chez Pellenc ST et permettra ainsi à l'usine EGGER Rambervillers d'augmenter son taux d'incorporation de déchets de bois. Concrètement pour EGGER Rambervillers, l'enjeu est d'envoyer une part des bois recyclés en coupeuse et non au broyeur : la coupeuse est une machine qui fait des copeaux plus longs avec de meilleures propriétés mécaniques que le broyeur à marteau, lequel fait des copeaux inférieurs à 2-3 mm, et accepte tous les bois recyclés. Ce besoin est illustré par le schéma de la Figure 1.2.

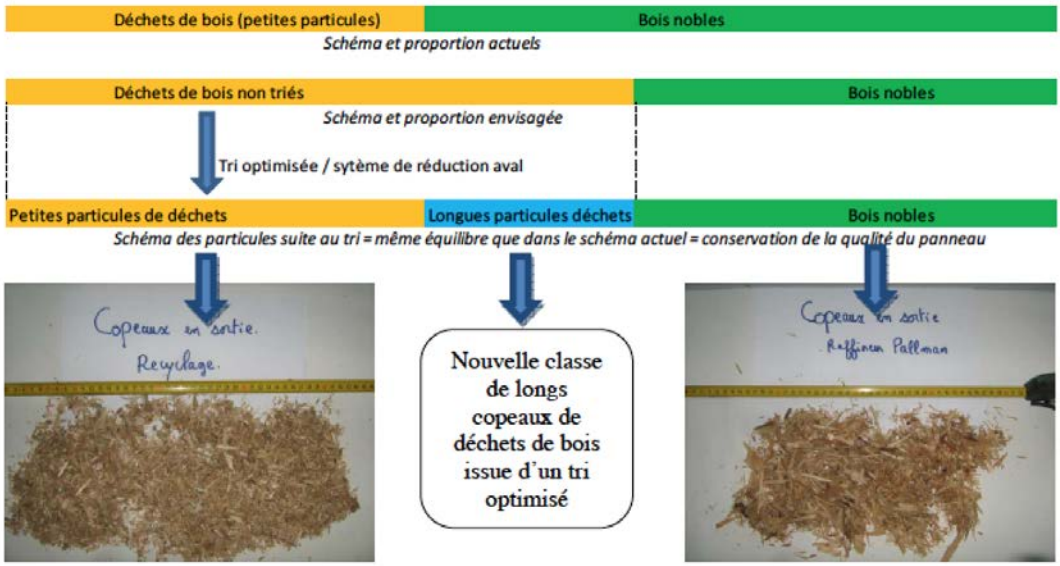


FIGURE 1.2 – Schéma représentatif des besoins de la société EGGER : nouvelle classe de copeaux pouvant être intégrés dans les panneaux de particules tout en conservant la qualité de ceux-ci.

1.2 Spectroscopie IR

1.2.1 Types de rayonnement

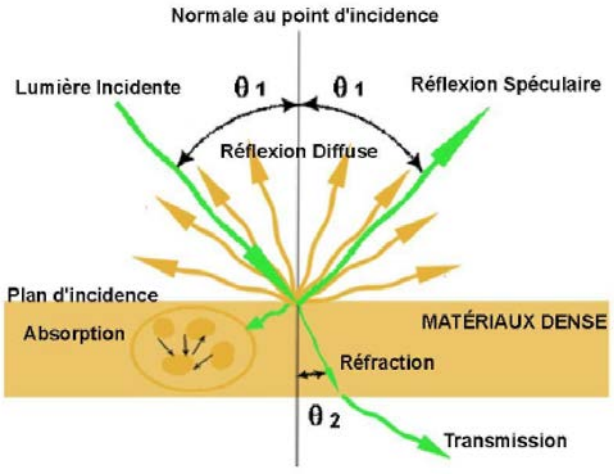


FIGURE 1.3 – Types d'effets induits par le rayonnement électromagnétique d'un objet.

Dans le cadre du projet TRISPIRABOIS, nous nous intéressons à l'analyse et la classification de données spectrales issues d'un processus d'acquisition proche/moyen infra-rouge. Cette technique consiste à envoyer sur un objet un rayonnement électromagnétique qui interagit avec l'objet, est conduit à différents effets comme présenté sur la Figure 1.3. Outre la réflexion spéculaire que nous utilisons dans le cadre de ce projet, où le rayonnement rebondi tel un miroir, on peut observer d'autres effets tels que le

réflexion diffuse (réflexion dans toutes les directions), la transmission au travers du matériau, ou encore l'absorption par le matériau de rayonnement.

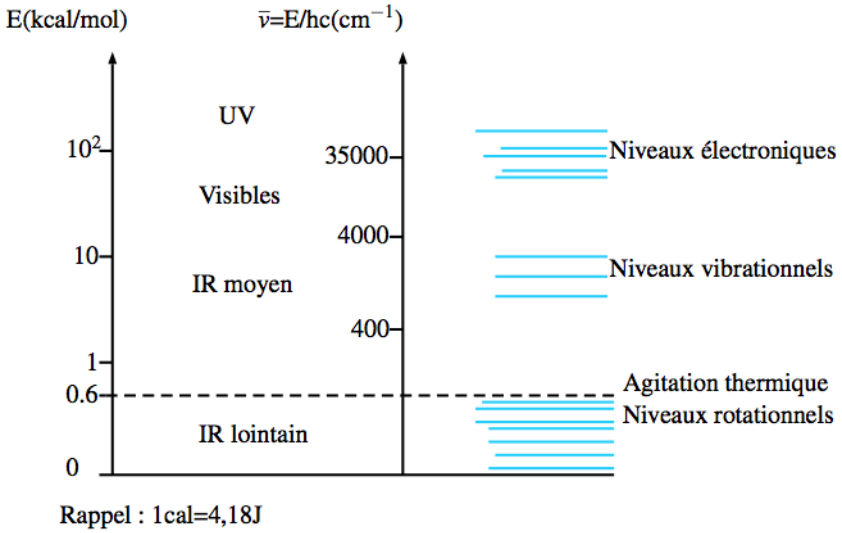


FIGURE 1.4 – Niveaux énergétiques moléculaires aux fréquences des différents domaines spectraux.

1.2.2 Spectroscopie vibrationnelle

Le potentiel de la spectroscopie vibrationnelle notamment en proche/moyen infrarouge a été étudié depuis plus d'une dizaine d'années [143, 81]. Le niveau énergétique vibrationnel se situe dans le domaine spectral IR proche/moyen comme présenté sur la Figure 1.4. Une composante spectrale peut être caractérisée par différentes mesures : longueur d'ondes, énergie, fréquence, etc. En réalité il s'agit de la même grandeur exprimée différemment.

Un rayonnement sur une matière donnée peut produire différents effets en fonction des molécules présentes. Pour le bois, les molécules peuvent être assez complexes, néanmoins, elles peuvent être composées de groupements dont les modes de vibrations peuvent être identifiés. On présente sur la Figure 1.5 quelques exemples de spectres correspondant à des essences différentes de bois. Cette illustration permet de voir le positionnement des groupements OH de la cellulose, les composés phénoliques de la lignine ainsi que l'eau qui est souvent due au taux d'humidité du bois.

1.2.3 Spectrométrie infra-rouge

La spectrométrie infra-rouge a pour but de mesurer la diminution de l'intensité du rayonnement électromagnétique, compris entre 12800 cm^{-1} et 4000 cm^{-1}, qui est réfléchi par un échantillon en fonction de la longueur d'onde. Cette technique permet d'identifier les liaisons chimiques et donc, de déterminer les molécules présentes dans la matière en identifiant les énergies vibrationnelles absorbées par la matière. Ainsi, on peut utiliser la spectroscopie infra-rouge pour réaliser différentes tâches :

1. analyse fonctionnelle pour déterminer les fonctions chimiques présentes [111, 47] ;

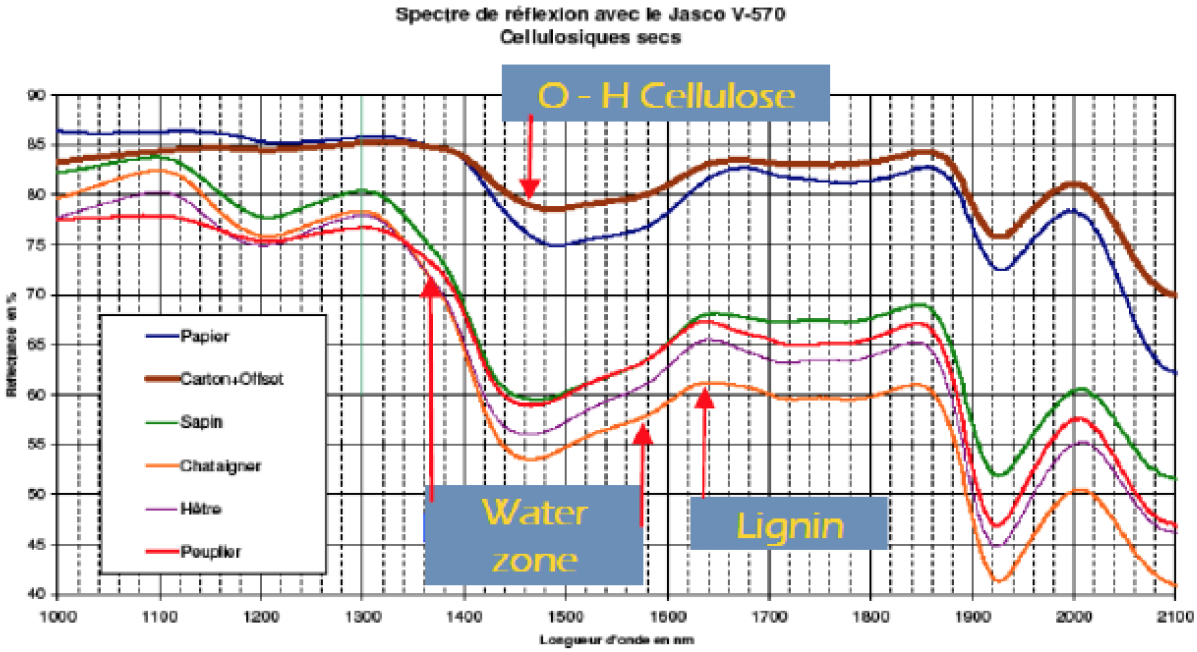


FIGURE 1.5 – Spectres de matériaux cellulosiques dont différentes essences de bois.

2. analyse structurale afin d'identifier les groupements chimiques présents, leur structures ainsi que la géométrie des molécules [120] ;

3. analyse quantitative qui determine la quantité d'une molécule présente dans la matière (H_2O par exemple pour connaître le taux d'humidité) [28, 117] ;

4. tri de matériaux, ou classification [126, 3].

1.3 Étiquetage des données

Les échantillons de déchets de bois ont été prélevés par les experts de la société EGGER et du Crittbois (Figure 1.6) dans le parc de déchets de la société EGGER . Pour chaque échantillon, trois éprouvettes ont été préparées après séchage : une destinée aux tests chez Pellenc ST avec leurs machines actuelles, une autre pour l'acquisition des données au Crittbois avec le spectromètre industriel, et la dernière éprouvette pour l'acquisition et l'analyse au sein du CRAN/LCPME. Ainsi, tous les partenaires du projet disposent des mêmes échantillons pour leurs analyses. Enfin, les experts ont pu dégagé 23 groupes d'échantillons étiquetés selon leur type et la catégorie à laquelle ils appartiennent comme présentés dans le Tableau 1.1.

FIGURE 1.6 – Exemples d'échantillons collectés et étiquetés par les experts.

FIGURE 1.7 – Machine de tri Mistral dual vision (Pellenc ST).

1.4 Caractéristiques des spectromètres

1.4.1 Spectromètre Pellenc ST

Le spectroscope intégré actuellement dans les machines Mistral de Pellenc.ST (Figure 1.7) est un spectromètre proche/moyen infra-rouge avec une gamme spectrale variant de 900 à 1900 nm (5200-11000 cm^{-1}). Ce dispositif est complété par un spectroscope dans la gamme spectrale du visible. La source utilisée consiste en des réflecteurs halogènes avec une température de couleur de 3000 Kelvin et un angle d'incidence de 10 degré. De plus, 16 voies sont utilisées pour l'acquisition des données spectrales (16 longueurs d'onde) et pouvant aller jusqu'à 20 dans le domaine du NIR. La résolution varie de 7 à 20 nm avec une fréquence d'acquisition de 16 ms. La référence utilisée est une plaque céramique blanche.

1.4.2 Spectromètre du laboratoire LCPME

Le spectromètre que nous avons utilisé pour l'acquisition des spectres sur lesquels nous avons principalement travaillés dans le cadre de cette thèse est un spectromètre Nicolet 8700 FTIR présenté sur la

Catégorie	Groupes	Destination
Bois purs	Bois brut	Coupeuse
	Bois massif avec finition	
	Bois massif peint ou surfacé	
	Contreplaqué brut	
	Contreplaqué avec finition	
	Contreplaqué surfacé	
	OSB	
	Plaquettes	
Pan. particules	Panneaux de particules bruts	Broyeur lent
	Panneaux de particules surfacés	
	Panneaux de particules avec finition	
Indésirables	Panneaux MDF/HDF bruts	Rejet
	Panneaux MDF/HDF surfacés	
	Panneaux MFD/HDF avec finition	
	Panneaux fibres dures surfacés	
	Panneaux fibres dures bruts	
	Bois massifs sels-métalliques	
	Bois massifs traité (créosote)	
	Textiles, fibres	
	Déchets verts	
	Plastiques	
	Mousses	
	Panneaux alvéolaires, rotin, bambou...	

TABLE 1.1 – Catégories et groupes de bois collectés et étiquetés par les experts

Figure 1.8. Ce spectromètre est un instrument de recherche de qualité conçu pour une flexibilité expérimentale et des performances maximales. Le Nicolet 8700 FTIR est robuste aux erreurs d'échantillonnage grâce à une suite de fonctionnalités avancées pour les applications de spectroscopie en recherche telles que les numériseurs internes à deux canaux et une plage spectrale étendue. Ce dispositif est purgé avec du N_2 ultra pur. L'appareil est également équipé d'un détecteur MCT et d'un diviseur de faisceau CaF2. La gamme spectrale varie de 3000 -10000 cm^{-1} (3.33-1 μm) avec un pas de 4 cm^{-1}. La résolution spectrale est de 16 cm^{-1}.

1.4.3 Spectromètre du Crittbois

Le spectromètre du Crittbois (Figure 1.9) est un spectromètre Brüker Matrix-F avec option Emission pour des mesures avec une sonde sans contact. Il s'agit d'un spectromètre NIR-FT (transformée de Fourier), avec une source NIR refroidie à l'air et un multiplexeur interne permettant d'installer deux sondes. La gamme spectrale varie de 4000 cm^{-1} à 12800 cm^{-1} avec une résolution allant jusqu'à 2 cm^{-1}. La vitesse d'acquisition est de 5 scans/sec. Pour ce dispositif nous utilisons une référence mobile blanche en Spectralon (Labsphere Inc.), située dans la sonde d'illumination. Les caractéristiques de la sonde d'illumination pour des analyses sans contact en réflexion sont les suivantes :

— Deux sources tungstène (5W),
— Distance de mesure : 100 mm,

FIGURE 1.8 – Spectromètre Nicolet FTIR 8700. Les spectres analysés par le CRAN/ LCPME ont été acquis par ce dispositif.

FIGURE 1.9 – Spectromètre Matrix-F Brüker utilisé pour l'acquisition des spectres par le Crittbois.

— Surface d'analyse : disque de 10 mm de diamètre,
— Fenêtre saphir,
— Référence interne,
— Fibre optique de 5 m de longueur (diamètre interne 0,6 mm).

1.5 Acquisition des données

1.5.1 Formes des spectres

L'acquisition des données par le CRAN et le LCPME a été réalisée en utilisant le spectromètre Nicolet 8700. Ce spectromètre de laboratoire ne permettant pas d'extraire les données des 24 groupes mentionnés précédemment, nous n'avons pu retenir que les spectres correspondant à 15 groupes présentés dans le Tableau 1.2. Les réflectances en proche infra-rouge des spectres, enregistrées dans la gamme

Catégories de bois	Type de bois	# d'échantillons
Bois à la coupeuse	Bois brut	123
	Bois massif finition sans peinture	139
	Contreplaqués bruts	39
	Contreplaqués finition	52
	Contreplaqués surfacés	22
Bois à valoriser	Bois massif avec peinture	143
	Pann. de particules peinture et autres	23
	Pann. de particules bruts	51
	Pann. de particules surfacés	83
Indésirables	Panneaux MDF HDF avec peinture	49
	Bois massif sels métalliques	34
	Panneaux MDF HDF bruts	27
	Panneaux MDF HDF surfacés	56
	Panneaux fibres dures surfacés	38
	Panneaux fibres dures bruts	7

TABLE 1.2 – Groupes et nombres d'échantillons de la base de données

spectrale 3000-10000 cm^{-1}, ont été réalisées avec un accessoire de réflectance spéculaire proche de la normale, fournie par Pike Technologies (angle d'incidence fixé à 10 degrés). La résolution spectrale était de 16 cm^{-1} et 100 scans ont été additionnés pour chaque spectre qui comprend 1647 longueurs d'onde. On présente sur la Figure 1.10 des exemples de spectres relatifs aux 15 groupes de bois des deux classes. Pour chaque échantillon, nous avons choisi de réaliser deux captures sur chaque face (si l'échantillon est trop petit on se contente d'une capture par face). Ainsi, on désigne un fichier de données dans un groupe donné par : le numéro de l'échantillon, la face qui lui correspond et le numéro de la capture.

Exemple :

— les deux spectres de la face 1 de l'échantillon 7 d'un bois brut sont nommés Ech7F11 et Ech7F12 ;
— les deux spectres de la face 2 de l'échantillon 7 du même bois sont nommés Ech7F21 et Ech7F22.

1.5.2 Base de données spectrales

Après l'acquisition des spectres correspondants à chaque groupe de bois, nous avons élaboré une base de données pour leur stockage et éventuellement leur modification. On dispose de 900 spectres au total avec 15 groupes différents. Il est nécéssaire de noter ici, que les échantillons sont décomposés en trois catégories : bois à la coupeuse, bois à recycler et bois pollués. Cette base peut également être enrichie avec d'autres données spectrales fournies par les autres partenaires du projet. La base est disponible à l'adresse :

http://195.220.155.5/trispirabois/trispirabois.php

1.6 Conclusion

Nous avons introduit dans ce premier chapitre, le cadre général du projet Trispirabois, dans lequel s'inscrit cette thèse. Nous avons également donné quelques éléments sur la spectroscopie proche/infrarouge afin de montrer l'intérêt de son utilisation.
Le prochain chapitre est un état d'art des méthodes de classification utilisées dans le domaine de spectrométrie. Ce chapitre nous permettra d'étudier les performances générales de ces méthodes afin de dégager les approches les plus adaptées à notre projet.

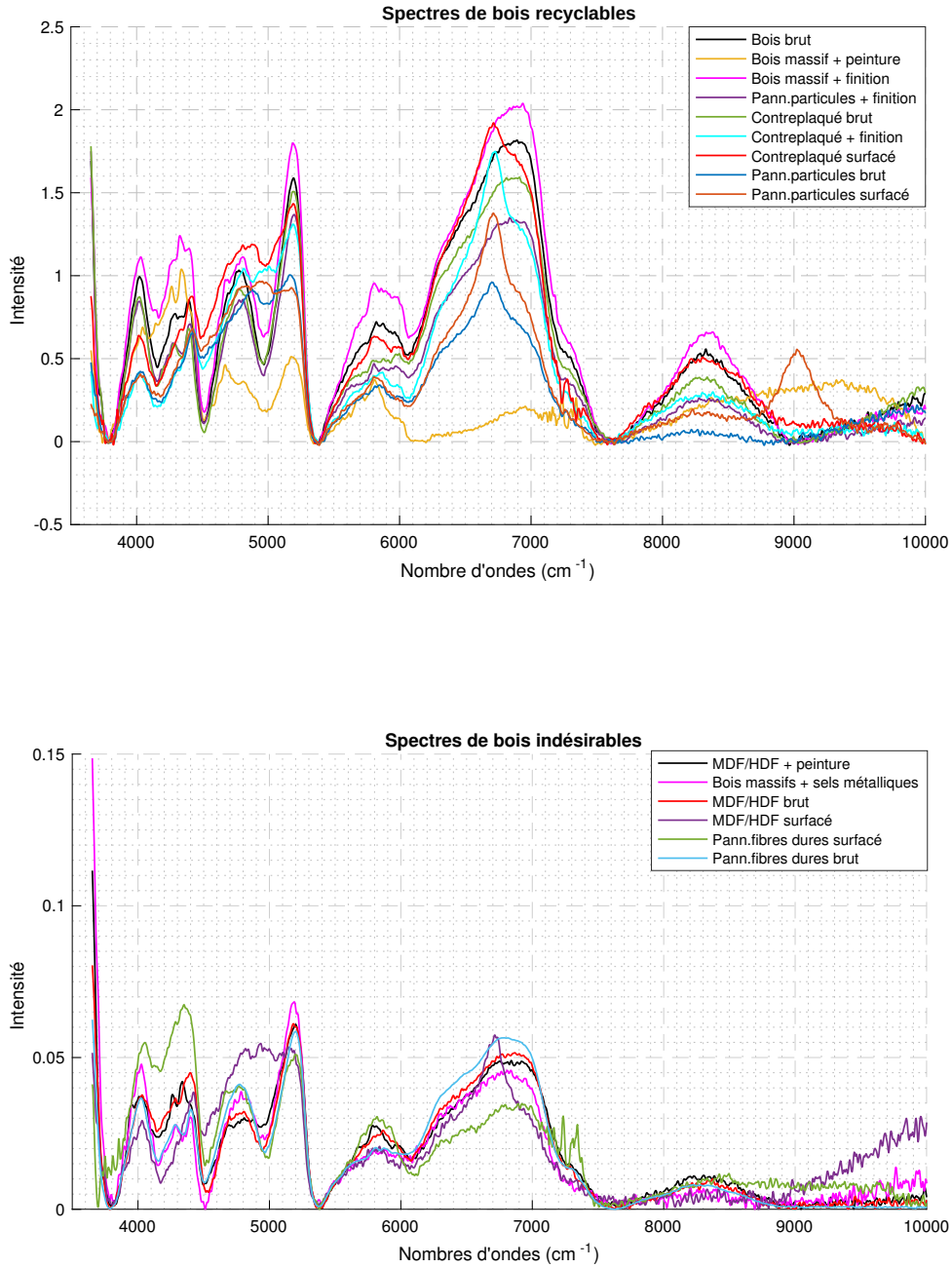


FIGURE 1.10 – Exemples de spectres de la base de données du CRAN après correction de la ligne de base : 9 groupes de bois de la première classe, et 6 groupes de la deuxième. Les spectres ne sont pas normalisés.

Chapitre 2

Analyse et classification de données spectroscopiques

2.1 Introduction

La classification de données est un problème délicat qui apparaît dans différentes disciplines telles que l'analyse d'images, les statistiques et la théorie de l'information. Elle consiste à affecter des objets à des classes, en fonction des valeurs des attributs (features) calculées. Dans ce contexte, un attribut peut être décrit soit de manière nominale quand il s'agit d'un ensemble de valeurs discrètes, caractérisant une certaine propriété (couleur, genre, type...), soit de façon continue, avec des valeurs réelles, correspondant à des mesures calculées physiquement (intensités, âge, concentration...). La classification est basée sur le postulat qu'une classe donnée contient des objets partageant des caractéristiques communes avec des valeurs identiques ou proches [54]. Comme la plupart des grandeurs physiques mesurées, les intensités spectrales obtenues par un système spectroscopique sont souvent affectées par des artefacts expérimentaux qui induisent des déviations et des erreurs. Ces imprécisions peuvent avoir comme conséquences des erreurs d'analyse et des situations de fausse classification.
Le présent chapitre décrit différentes méthodes d'exploration et de classification des données. Ces méthodes sont utilisées dans différentes applications comme la biologie [63, 130], la toxicologie [144, 57], le diagnostic de maladies [106, 115], etc. On se restreint ici aux approches appliquées principalement à des données en proche/moyen infra-rouge, ce qui nous permettra de comparer et d'évaluer les techniques les plus appropriées pour la classification des déchets de bois. Il est organisé comme suit. Dans le prochain paragraphe nous présentons brièvement les techniques de pré-traitement des données, puis nous aborderons quelques méthodes d'exploration de données. La quatrième section est consacrée aux méthodes les plus utilisées en classification non supervisée, puis les approches de classification supervisée sont traitées dans la cinquième section. Ensuite, des approches de classification supervisée sont testées sur la base de données spectrale de déchets de bois. La dernière section conclut le chapitre.

2.2 Pré-traitement des données spectrales

Un spectre IR peut être exprimé en réflectance, c'est-à-dire par l'énergie que l'échantillon observé renvoie, ou par la quantité d'énergie qu'il absorbe (absorbance) comme présenté sur la Figure 2.1a. Par ailleurs, des techniques de pré-traitement sont souvent nécessaires avant d'entamer le processus d'analyse ou de classification. Ceci a pour objectif d'éliminer ou de réduire la participation des sources de variation non désirées qui sont dues aux conditions d'acquisition. Différents pré-traitements peuvent être appliqués aux données spectrales :

— correction de la référence afin d'éviter les variations entre les spectres dues à la différence de la ligne de base. Cette correction peut être faite en utilisant la première ou la deuxième dérivée. Des travaux [102, 103] ont conclu que les meilleurs résultats sont obtenus en utilisant la première dérivée. Dans ce travail, nous utilisons une méthode plus récente, développée dans [92], pour la suppression de la ligne de base. On présente sur la Figure 2.1b un exemple de l'estimation de la ligne de base, et la forme du spectre après sa suppression,

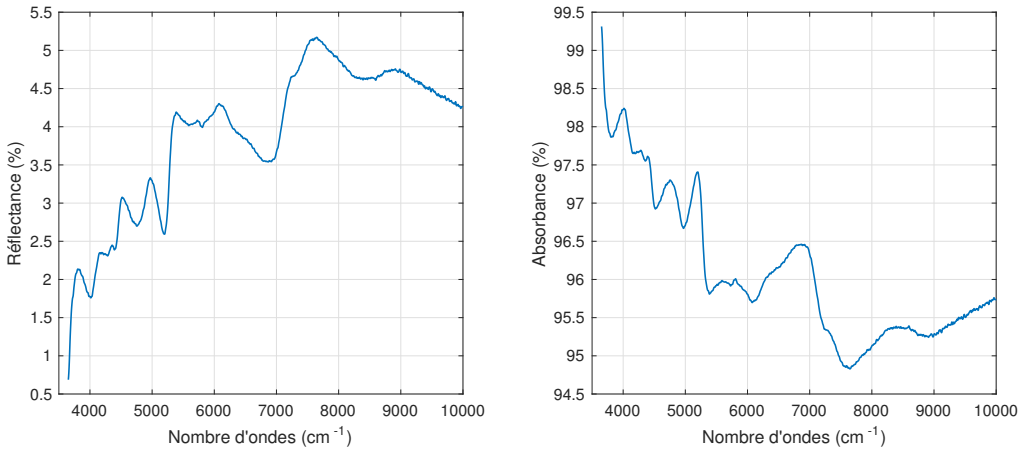
— normalisation des données, pour réduire les variations dues aux durées d'exposition et ramener les spectres à la même échelle. Cette normalisation peut être réalisée de différentes façons, la plus commune consiste à forcer une variance unité [35]. D'autres options sont parfois utilisées telles que la normalisation sur la base de l'élément ayant l'absorbance maximale ou encore la normalisation permettant d'obtenir une énergie unité (Figure 2.1c).

2.3 Exploration de données

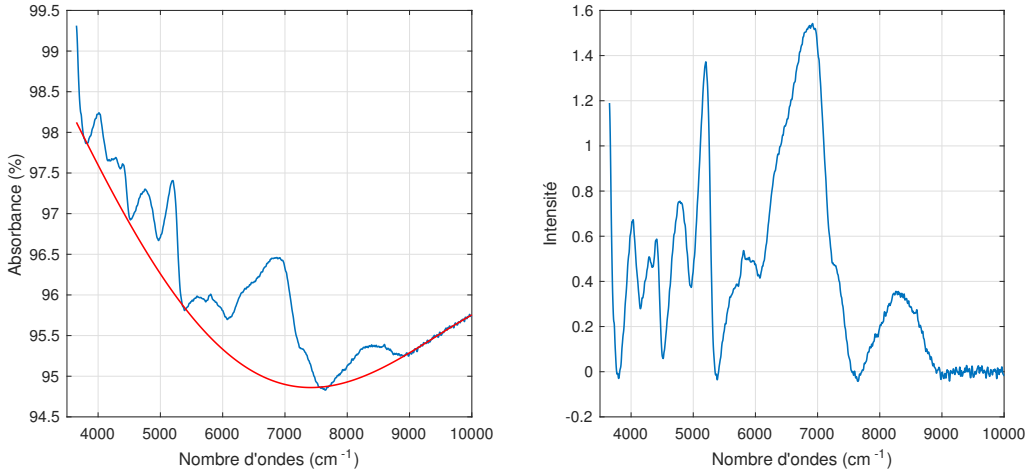
L'exploration de données peut avoir plusieurs objectifs tels que l'identification d'anomalies dans l'ensemble des mesures, la détermination des relations existantes entre certaines mesures ou encore, des relations significatives entre des groupes de mesures. En général, cette étape vise à retrouver des formes dans les observations en utilisant plusieurs outils disponibles à cet effet. L'analyse en composantes principales [76, 93], l'analyse factorielle [79], les méthodes basées sur la poursuite de projection [141] sont dans ce contexte, des techniques de réduction qui définissent un nombre de variables latentes en effectuant des combinaisons linéaires des données originales suivant un certain critère. Pour toutes ces méthodes, la projection des objets à partir de l'espace original des données dans un espace de variables latentes est désigné comme le score de ces variables. Ceci permet d'obtenir des informations sur les similarités ou les différences entre les objets. La contribution de chaque variable à ce score permet de détecter les variables responsables d'un groupement (clustering) dans les données.

2.3.1 Analyse en composantes principales

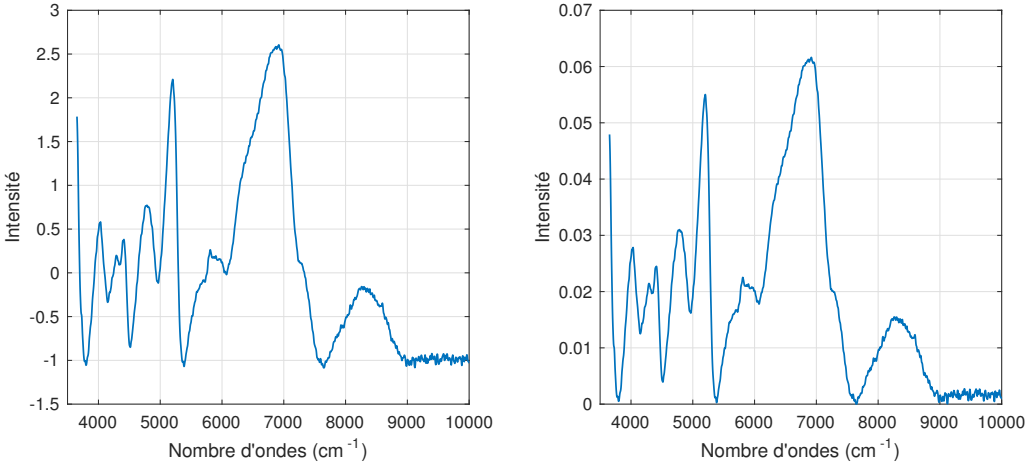
L'analyse en composantes principales (ACP) consiste à analyser la structure de la matrice de covariance de dimension $p \times p$, c'est-à-dire la dispersion des données. L'objectif de cette décomposition est de décrire à l'aide de k variables ($k \leq p$) cette dispersion en projetant le nuage de points sur les axes qui maximisent l'inertie projetée. En diagonalisant la matrice de covariance on obtient les axes portés par



(a) Exemple de spectre brut exprimé en réflectance (gauche) et en absorbance (droite).



(b) Pré-traitement : estimation de la ligne de base (gauche), suppression de la ligne de base estimée (droite).



(c) Normalisations : spectre à moyenne nulle et de variance unité (gauche), spectre rehaussé (non-négatif) avec norme unité (droite).

FIGURE 2.1 – Suppression de la ligne et base et normalisation des données spectrales.

les vecteurs propres et la longueur de chaque axe est déterminée par la racine de la valeur propre correspondante. Ainsi, le vecteur propre correspondant à la plus grande valeur propre identifie la direction de la dispersion maximale du nuage de point. On peut également calculer ces vecteurs et valeurs propres en utilisant la décomposition en valeurs singulières de la matrice de données.

En spectrométrie proche infra-rouge, l'ACP a beaucoup été utilisée dans la décennie 90, parmi ces travaux on peut citer notamment [93, 87, 91, 110].

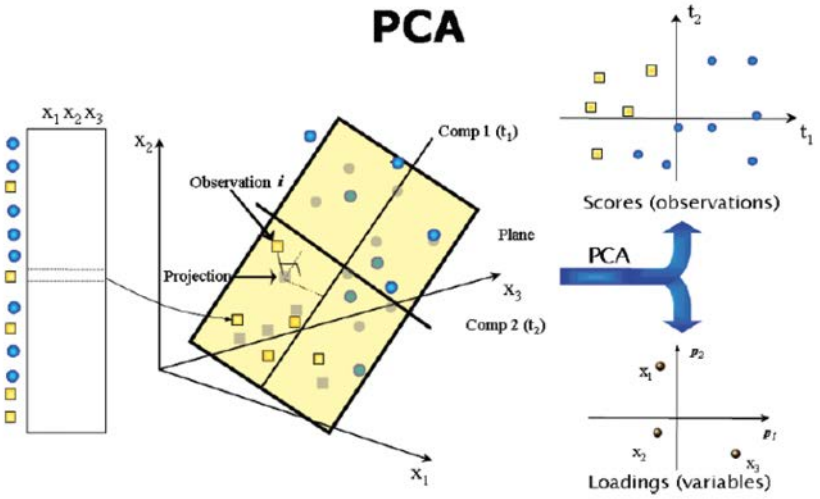


FIGURE 2.2 – Illustration du fonctionnement de l'ACP pour l'approximation de la matrice $\mathbf{X}$ dans un espace de dimension réduite [136].

On présente sur la Figure 2.2 une illustration permettant de comprendre le fonctionnement de l'ACP. On considère une matrice de données $\mathbf{X}$ avec 3 colonnes et deux composantes principales :

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \mathbf{E}. \tag{2.1}$$

où $\mathbf{E}$ représente la matrice des résidus. La première composante du modèle ACP, ($\mathbf{p}_1$) décrit la plus grande variation du nuage de points. La deuxième composante représente la deuxième plus grande variation, et ainsi de suite. Les scores ($\mathbf{T}$) permettent de revenir à un espace de dimension réduite (ici 2), en déformant le moins possible la réalité des données initiales dans $\mathbf{X}$. Un diagramme de dispersion sur le plan formé par les vecteurs $\mathbf{t}_1$ et $\mathbf{t}_2$ fournit un aperçu de toutes les observations de la matrice de données. On peut également y trouver les regroupements, les tendances ainsi que les valeurs aberrantes présentes dans les observations. La position de chaque point (correspondant à une observation) dans le plan du modèle est utilisée pour relier les points les uns aux autres. Les éléments qui sont proches les uns des autres correspondent à des observations avec un profil similaire. Les vecteurs de charge ($\mathbf{p}_1$, $\mathbf{p}_2$) définissent eux, la relation entre les variables mesurées, c'est-à-dire les colonnes de la matrice $\mathbf{X}$. Les résidus ($\mathbf{E}$) correspond à la partie de $\mathbf{X}$ qui n'est pas expliquée par le modèle.

Nous avons appliqué la décomposition ACP sur les données de déchets de bois. L'objectif est de vérifier si cette méthode nous permet de séparer les spectres des deux catégories, à savoir, le bois à ré-

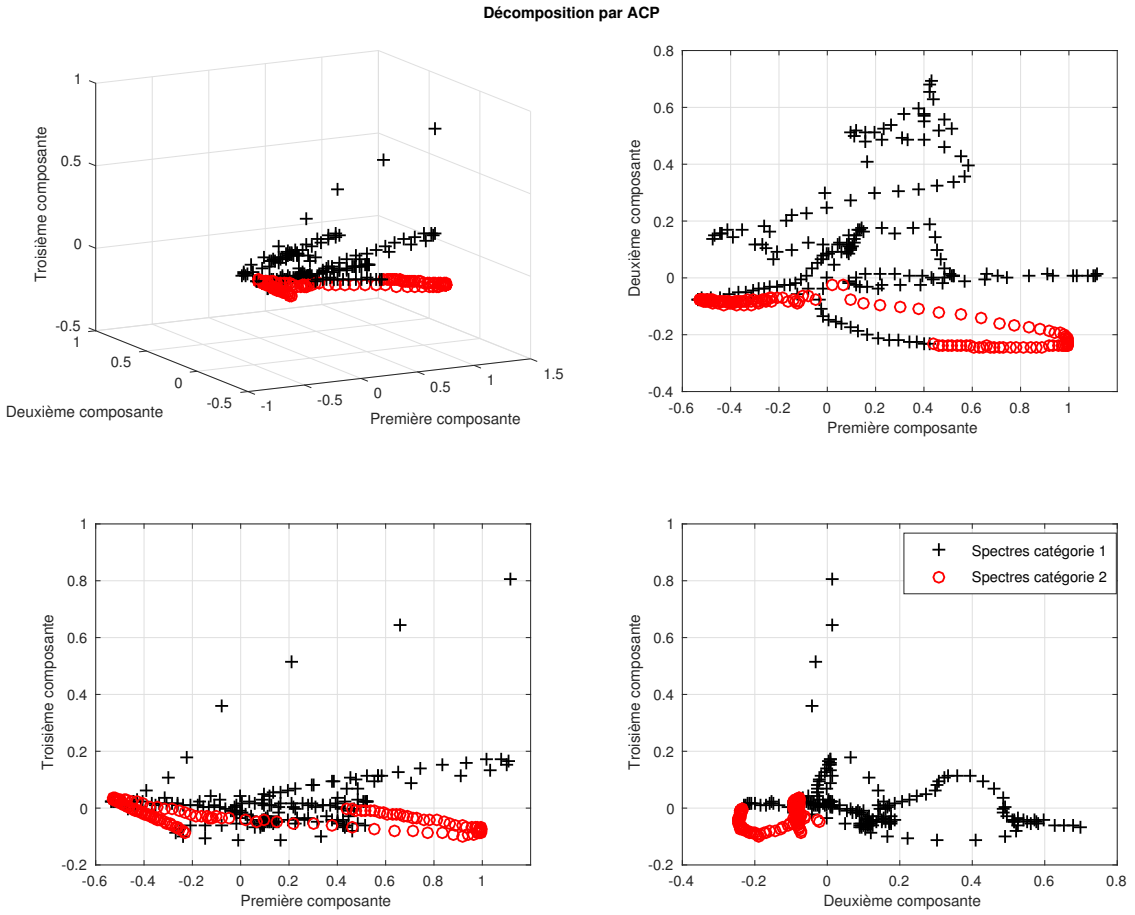


FIGURE 2.3 – Décomposition par ACP des spectres de déchets de bois.

cupérer et les indésirables. Les résultats de la décomposition avec trois axes principaux sont représentés sur la Figure 2.3. On constate que cette décomposition ne permet pas de séparer linéairement les deux catégories, puisque leurs nuages de points respectifs sont mélangés. Toutefois, la décomposition avec trois composantes seulement n'est peut être pas suffisante pour une séparation linéaire. En effet, on peut toujours essayer d'augmenter le nombre de composantes afin d'atteindre de meilleurs résultats. Malheureusement, les composantes obtenues par ACP ne sont généralement pas interprétables physiquement puisqu'il s'agit de combinaisons de plusieurs variables. Il est donc difficile d'intégrer les composantes de l'ACP sur les machines de tri puisque ces dernières ne sont pas des longueurs d'ondes utilisées par les systèmes optiques de ces machines.

Plus récemment, d'autres variantes de cette approche ont été introduites pour en corriger les limitations. En effet, l'ACP ne prenant pas en compte les dépendances non linéaires entre les variables, des alternatives telles que l'ACP à noyaux [119], l'ACP par variétés [59, 64], ou encore l'ACP pondérée [90], on été proposées pour offrir plus de flexibilité. Une autre variante consiste à imposer une parcimonie sur les vecteurs des scores. Nous présentons dans ce qui suit le principe de cette méthode.

2.3.2 Analyse en composantes principales parcimonieuses

L'ACP parcimonieuse (*Sparse Principal Components Analysis* SPCA) vise à produire des vecteurs de scores $\mathbf{t}_i$ parcimonieux et plus simple à interpréter. Joliffe *et al*. [77] ont proposé une première approche appelée *SCoTLass*, basée sur la pénalité ℓ_1 et qui résout le problème suivant :

$$\max \mathbf{t}_i'(\mathbf{X}'\mathbf{X})\mathbf{t}_i, \text{ s.c. } \sum_{j=1}^{p} |\mathbf{t}_{ij}| \leq s, \ \mathbf{t}_i'\mathbf{t}_i = 1, \ \mathbf{t}_j'\mathbf{t}_i = 0 \text{ pour } j < i, \tag{2.2}$$

pour la matrice d'observations centrée $\mathbf{X}$, de taille $N \times p$. La première contrainte induit la parcimonie des vecteurs $\mathbf{t}_i$. Néanmoins, ce problème est non convexe et les calculs peuvent s'avérer très coûteux [52]. Une autre façon de formuler ce problème a été développée dans [163]. L'idée de cette approche est de reformuler l'ACP comme un problème d'optimisation de type régression/reconstruction, en introduisant une deuxième contrainte en norme 2 sur les vecteurs $\mathbf{t}_i$:

$$\min_{\alpha,\mathbf{t}} \sum_{i=1}^{N} ||\mathbf{x}^i - \alpha\mathbf{t}'\mathbf{x}^i||_2^2 + \lambda_1||\mathbf{t}||_2^2 + \lambda_2||\mathbf{t}||_1 \text{ s.c. } \alpha'\alpha = \mathbf{I}_k. \tag{2.3}$$

où $\mathbf{t}'$ est la transposée de $\mathbf{t}$ et $\mathbf{x}^i$ est la i-ème ligne de $\mathbf{X}$. Ce problème est connu sous le nom de *Elastic net*. De la même manière, la pénalité ℓ_1 sur $\mathbf{t}$ encourage une solution parcimonieuse de ce vecteur. De plus, lorsque λ_1 et λ_2 sont nuls et $N > p$, il est démontré que $\mathbf{t} = \alpha$ et qu'il s'agit de la composante principale. Lorsque $p \gg N$ la solution n'est pas toujours unique (sauf si $\lambda_1 > 0$). Pour $\lambda_1 > 0$ et $\lambda_2 = 0$ la solution de $\mathbf{t}$ est proportionnelle à la direction de la composante principale. L'algorithme 2.4 résume les étapes cette approche.

FIGURE 2.4 – Algorithme SPCA

— **Entrée :** $\alpha = \mathbf{V}[, 1 : k]$, les poids des k premières composantes principales.

— **Pour** α fixé :
— **Pour** $j \leftarrow 1, \cdots, k$:
— Résoudre le problème suivant (problème *naive elastic net*) :

1. $\mathbf{t}_j = \arg\min_{\mathbf{t}^*} \{\mathbf{t}^{*\prime}(\mathbf{X}'\mathbf{X} + \lambda_1)\mathbf{t}^* - 2\alpha_j'\mathbf{X}'\mathbf{X}\mathbf{t}^* + \lambda_{2_j}|\mathbf{t}^*|_1\}$
2. Pour chaque $\mathbf{t}$ fixé :
3. $\mathbf{X}'\mathbf{X}\mathbf{t} = \mathbf{U}\mathbf{D}\mathbf{V}'$.
4. Mis à jour : $\alpha = \mathbf{U}\mathbf{V}'$
5. Répéter 1. et 4. jusqu'à la convergence de $\mathbf{t}$
6. $\mathbf{V}_j = \frac{\mathbf{t}_j}{|\mathbf{t}|}$ Normalisation

— **Sortie** Vecteurs de scores parcimonieux $\mathbf{t}$.

2.4 Méthodes de classification non supervisée

2.4.1 K-means

Le K-mean est une méthode de classification non supervisée qui consiste à réaliser une classification par réallocation. Dans cette approche, on fixe le nombre de clusters M avant toute analyse. Au départ, M centres de classes sont choisis parmi les individus, puis chaque individu est affecté au centre le plus proche. Une mise à jour est par la suite effectuée sur les nouveaux barycentres des classes. Ce processus est réitéré jusqu'à la stabilisation, c'est-à-dire jusqu'à ce que les centres des classes ne changent plus. Au fil des itérations, la variabilité intra-classe diminue alors que la variabilité inter-classe augmente. Soit K le nombre d'itérations complétées et m_i la moyenne de la classe c_i, $i = 1,...,M$. L'algorithme K-means fonctionne comme suit : à chaque itération, on balaye les individus en leur attribuant l'étiquette c_i si la distance à m_i est minimale. En général, il s'agit de la distance euclidienne, mais d'autres types de distances comme la mesure de corrélation peuvent être utilisés, notamment pour prendre en compte les écarts-types des classes. Le critère d'arrêt peut être relatif au nombre de clusters dont l'étiquette a changé entre deux étapes, ou sur la variation relative des paramètres m_i^k entre l'itération k et $k+1$. Cette technique est relativement efficace en termes de coût de calcul avec une complexité de l'ordre de $\mathcal{O}(pKM)$ où p est le nombre d'observations, K le nombre d'itérations et M le nombre de clusters, mais elle peut souvent se terminer dans un optimum local. Les détails de la procédure sont présentés dans l'Algorithme 2.5.

FIGURE 2.5 – Algorithme du K-means

— **Entrée :** m_i^0, M le nombre des clusters.
— **Pour :** $k \leftarrow 1, \cdots, K$:
— **Tant que :** Le critère d'arrêt dans chaque cluster c_i n'est pas vérifié :

1. $d_k \leftarrow \underset{i=1,...,M}{\arg\min}\, \text{dist}(x, m_i^k)$ $\quad$ m_i^k est le centre des clusters à l'itération k
2. $x \leftarrow c_i$. $\quad$ Affecter l'observation x à la classe c_i
3. **Pour** $i \leftarrow 1, \cdots, M$
4. $m_i^k \leftarrow$ Nouveaux barycentres $\quad$ Recalculer les barycentres m_i^k de chaque classe c_i.

— **Sortie :** les classes c_i.

2.4.2 Analyse par groupement hiérarchique

L'analyse par groupement hiérarchique (*Hierarchical Clustering Analysis* HCA),[20] est une méthode qui permet le groupement des variables selon leur similarité ou leur dissimilarité. En effet, contrairement au K-means, qui nécessite de fixer le nombre de clusters ainsi qu'une configuration initiale, cette approche ne se base que sur une mesure de similarité (distance par exemple) et n'a pas besoin d'autres spécifications. L'algorithme HCA recherche une structure sous-jacente des données à travers un processus itératif qui associe ou dissocie les éléments un par un jusqu'à ce qu'ils soient tous traités (Figure 2.6). Ces processus sont appelés *méthodes d'agglomération* et *méthodes de division* respectivement. La procédure d'agglomération démarre par les éléments d'une paire de groupes distincts dont la

différence intergroupe est minimale, et combine ces éléments de façon séquentielle, jusqu'à ce que tous les éléments appartiennent à un seul groupe. A l'inverse la procédure de division débute avec toutes les données appartenant à un seul groupe, puis réalise une partition à chaque niveau pour faire émerger un nouveau groupe. Cette partition est réalisée afin d'obtenir deux groupes dont la différence est minimale. Cela signifie que dans les deux cas, pour p objets, l'algorithme exécute $p-1$ itérations.

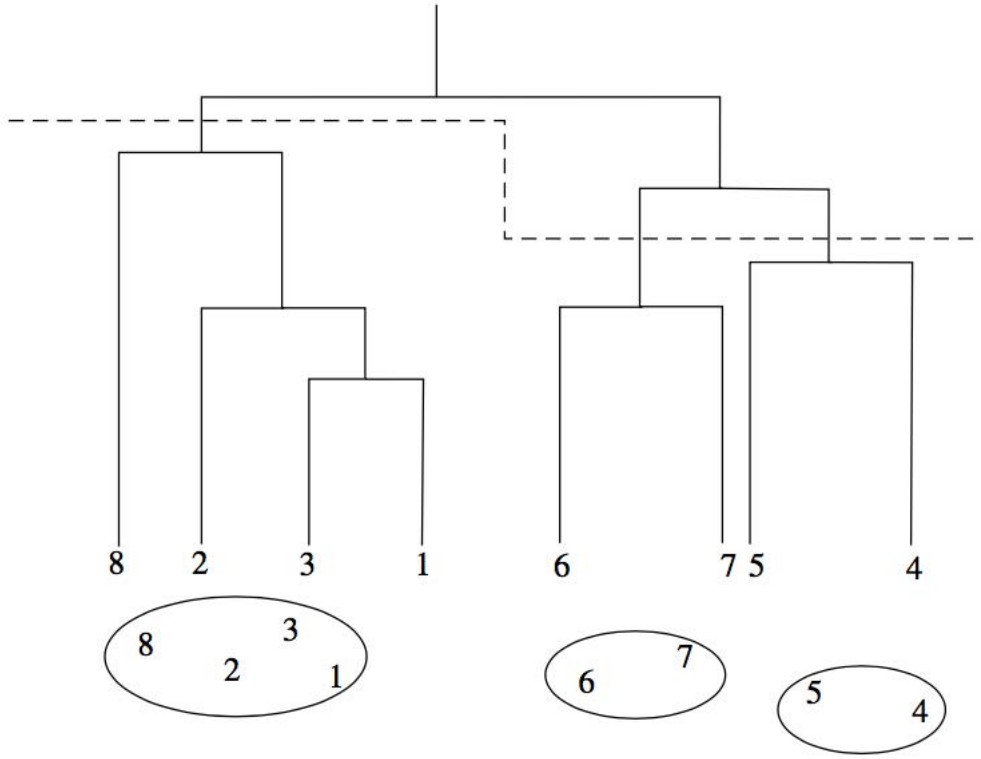


FIGURE 2.6 – Procédure du HCA. Cet exemple correspond à un processus de division qui permet d'identifier 3 groupes distincts.

Cette approche nécessite tout de même de faire deux choix importants : le type de la mesure de similarité entre les objets ainsi que la technique de liaison [18]. La mesure de similarité peut être obtenue en utilisant différentes distances, la plus naturelle étant la distance euclidienne. Néanmoins, cette distance n'est pas toujours fiable pour exprimer des profils similaires dans certaines situations. Ainsi, d'autres mesures de distance, représentant mieux la similarité entre les données ont été introduites [75, 19]. Par exemple, la mesure de corrélation a été utilisée en spectrométrie dans [100]. Concernant les méthodes de liaison, il existe trois modes de liaison entre les objets : les liaisons complètes (le voisin le plus éloigné), les liaisons uniques (les plus proches voisins) ou encore les liaisons moyennes (inter et intra groupes). Ces trois types de liaisons sont illustrés sur la Figure 2.7.

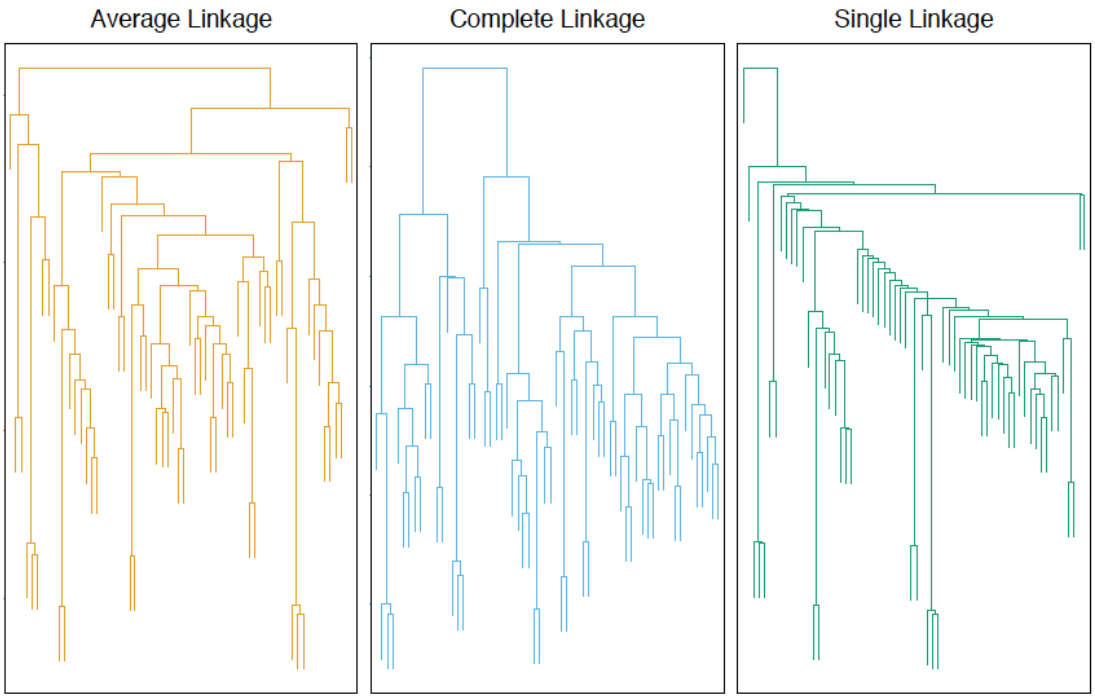


FIGURE 2.7 – Les trois types de liaison dans HCA [52]. Ce type d'affichage graphique est appelé dendrogramme et fournit une description complète et interprétable du regroupement hiérarchique dans un format graphique.

2.5 Méthodes de classification supervisée

Les techniques de classification supervisée sont basées sur une calibration, ou un ensemble d'apprentissage avec des informations connues en amont, pour construire un modèle de classification. Le modèle est par la suite testé au moyen d'un ensemble d'échantillons indépendants pour valider les propriétés prédictives du modèle, avant de l'utiliser sur des échantillons inconnus. Les méthodes de classification supervisée sont utilisées pour la classification de différents types de données. Parmi les méthodes utilisées dans le cadre la classification de données spectrales, on trouve l'analyse discriminante linéaire (*linear discriminant analysis*), la méthode des *k*-plus proches voisins (*K-Nearest Neighbours*), la modélisation indépendante d'analogie de classe (*Soft Independent Modeling of Class Analogy, SIMCA*), l'analyse discriminante des moindres carrés partiels (*Partial Least Squares Discriminant Analysis PLS-DA*) ainsi que la méthode d'analyse discriminante par projections orthogonales de structures latentes (*Orthogonal Projections to Latent Structures-Discriminant Analysis OLS-DA*). D'autres méthodes ont aussi montré un fort potentiel de classification pour ce type de données, dont les réseaux de neurones artificiels (*Artificial Neural Network ANN*) et les machines à vecteurs de support (SVM).

2.5.1 Les K-plus proches voisins (K-NN)

La méthode des K-plus proches voisins est une approche simple et intuitive pour la classification [45, 34, 67]. Un objet inconnu est classé selon l'appartenance majoritaire à une classe de ses K plus proches voisins dans l'ensemble d'apprentissage. La distance (ou proximité) est calculée moyennant une métrique appropriée au type de données traitées. La procédure du K-NN se décompose principalement en trois étapes :

1. calcul des distances entre un objet inconnu (a) et tous les objets de l'ensemble d'apprentissage ;
2. sélection des K objets de l'ensemble d'apprentissage les plus similaires à l'objet (a) selon le résultat des distances calculées ;
3. classification de l'objet (a) dans le groupe à qui appartiennent la plupart des K objets similaires.

Une valeur optimale de K peut être sélectionnée à partir de la classification des échantillons de l'ensemble de test en utilisant la méthode *leave-one-out* ou par la validation croisée. Concrètement l'algorithme attribue l'étiquette m associée à la classe c_m à l'objet (a) selon la formule suivante :

$$m = \arg \min_{i=1,\ldots,M} \text{dist}(x_{(a)}, c_i) = \arg \min_{i=1,\ldots,M} ||x_{(a)} - \mu_i||, \tag{2.4}$$

où μ_i est la moyenne de chaque classe, $x_{(a)}$ est la valeur associé à l'objet (a). L'inconvénient de cet algorithme est qu'aucune limite n'est fixée pour les différentes classes, d'autant plus que les critères d'appartenance ne sont pas définis. Par conséquent, tout point de l'espace, même correspondant à du bruit, est automatiquement rattaché au centre de la classe la plus proche.

En fait, le principe sur lequel est basé cet algorithme est assez simple, mais il n'est pas du tout réaliste, car un objet très éloigné du centre de la classe est probablement très différent de la classe à laquelle il est censé appartenir. Ceci implique la nécessité de fixer un seuil au delà duquel l'objet n'est plus classé.

2.5.2 La modélisation indépendante d'analogie de classe

Cette approche, plus connue sous le nom de SIMCA (*Soft Independant Modelling of Class Analogy*) [50, 147, 1], est une méthode de classification supervisée basée sur l'analyse en composantes principales. L'objectif de la procédure SIMCA est de définir la règle de classification pour un ensemble de m groupes en révélant des informations supplémentaires sur la structure de chaque groupe. Soit les m groupes notés $\mathbf{X}^j$, $j = 1,\ldots,m$, où j indique l'appartenance à un groupe. L'ensemble $\mathbf{X}^j$, est considéré comme l'ensemble d'apprentissage des groupes j, $j = 1,\ldots,m$. L'ensemble de validation noté $\mathbf{Y}^j$, $j = 1,\ldots,m$, contient des nouvelles observations de chaque groupe j. La première étape ici est de construire, séparément, un modèle ACP pour chaque groupe d'observations de $\mathbf{X}^j$. Ceci permet de réduire la dimension des données et fournit les matrices de scores $\mathbf{T}^j$ et de charges $\mathbf{P}^j$ pour chaque groupe. On considère alors $k_j < p$, le nombre des composantes principales retenues pour chaque groupe de données j. Les nouvelles observations sont alors classées en les projetant dans les sous-espaces engendrés par les différents modèles ACP des groupes d'apprentissage. Cette mesure est appelée distance orthogonale et est définie comme suit. On considère une nouvelle observation à classifier, $\mathbf{y}$ et soit $\widehat{\mathbf{y}}^{(n)}$ la

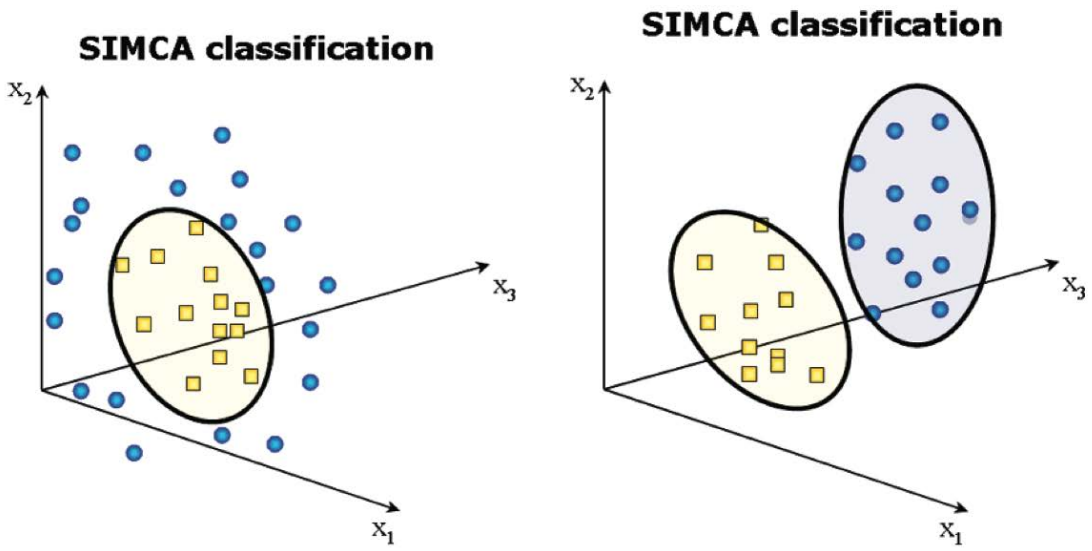


FIGURE 2.8 – Classification par la méthode SIMCA. À gauche un cas asymétrique avec une seule classe d'objets à déterminer. À droite une classification à deux classes chacune décrite par un modèle ACP [136].

projection de cette observation sur le modèle ACP du groupe n :

$$\hat{\mathbf{y}}^{(n)} = \tilde{\mathbf{x}}^{(n)} + \mathbf{P}^{(n)}\mathbf{P}^{(n)'}(\mathbf{y} - \tilde{\mathbf{x}}^{(n)}) \tag{2.5}$$

où $\tilde{\mathbf{x}}^{(n)}$ est la moyenne des observations du groupe n de l'ensemble d'apprentissage, et $\mathbf{P}^{(n)}$ est la matrice des scores de même groupe. La distance orthogonale (DO) de $\mathbf{y}$ par rapport au groupe n est définie par :

$$\mathrm{DO}^{(n)} = ||\mathbf{y} - \hat{\mathbf{y}}^{(n)}||^2. \tag{2.6}$$

L'observation $\mathbf{y}$ est classé dans le groupe n pour lequel $\mathrm{DO}^{(n)}$ est minimal. La méthode SIMCA est beaucoup plus utilisée dans le cas où il n'y a qu'une seule classe à déterminer, lorsqu'on a, par exemple, une classe d'objets bien définie avec d'autres objets non homogènes, mais aussi dans le cas d'une classification où aucune interprétation n'est exigée. En effet, les variables latentes générées par cette approche se basent sur les directions qui maximisent la variance, ce qui peut être différent des directions séparant les classes. Ces deux cas sont illustrés sur la Figure 2.8.

2.5.3 L'analyse discriminante par moindres carrés partiels

La technique des moindres carrés partiels (*Partial-Least Squares*) proposée dans plusieurs travaux notamment [149, 148, 150] est une méthode de régression multi-variée qui calcule les axes qui maximisent la séparation entre les groupes en recherchant une relation qualitative entre deux matrices $\mathbf{X}$ et $\mathbf{Y}$. $\mathbf{X}$ est une matrice comprenant généralement des données spectrales d'un ensemble d'individus, et $\mathbf{Y}$ est la matrice des attributs contenant les valeurs quantitatives de ces échantillons (les concentrations des composantes spectrales, âge, dose...). Cette méthode peut également être exploitée dans l'analyse discri-

minante (PLS-DA) quand la matrice **Y** contient des valeurs qualitatives comme la classe d'appartenance. Le modèle PLS est exprimé par :

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}. \tag{2.7}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{C}' + \mathbf{F}. \tag{2.8}$$

Les scores **T** représentent un plan de dimension réduite des variables qui sont une bonne approximation de **X**. De façon analogue, **P** et **C** représentent les vecteurs des charges définissant les relations entre les variables mesurées, c'est-à-dire les colonnes de **X** et de **Y**, respectivement. Les parties de **X** et de **Y** qui ne sont pas expliquées par le modèle représentent les résidus **E** et **F**. L'objectif de la méthode PLS est de déterminer les composantes (variables latentes) qui seront considérées comme les nouvelles variables explicatives. Ces composantes sont des combinaisons linéaires des variables initiales $\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^p$. En fait, l'équation 2.7 est similaire à la décomposition fournie par l'ACP, cependant, le critère résolu par PLS est basé sur la maximisation de la covariance entre les éléments de **X** et de **Y**. Les détails relatifs à la procédure ainsi que les étapes de l'algorithme sont fournis dans [19]. Cette approche a, par ailleurs, prouvé son efficacité dans divers domaines dont les études de la métabonomie [78] et la transcriptomique [109]. Cependant, même si elle permet d'expliquer les différences entre les caractéristiques des classes, l'interprétation devient progressivement difficile au fur et à mesure que le nombre des classes augmente.

2.5.4 L'analyse discriminante par projections orthogonales de structures latentes

La méthode des projections orthogonales de structures latentes (*Orthogonal Projections to Latent Structures OPLS*) [137] est une approche multi-classes pour la classification supervisée. Plus récemment, cette technique a été utilisée dans plusieurs travaux en chimiométrie [56, 66, 43]. Cette méthode apporte une amélioration à l'approche PLS-DA en analysant une variation supplémentaire dans chaque composante du PLS. En effet, l'idée principale dans OPLS est la décomposition de la variation de **X** en deux parties : une composante prédictive, et une composante orthogonale à **Y**. La composante supplémentaire dans le modèle OPLS est constituée des composantes orthogonales (non-corrélées). La séparation de la composante prédictive de la composante orthogonale permet de faciliter l'interprétation de la différence entre les classes. Ainsi, OPLS comprend deux modèles de variation, la composante **Y**-prédictive ($\mathbf{T}_p\mathbf{P}'_p$) et la composante **Y**-orthogonale ($\mathbf{T}_o\mathbf{P}'_o$) :

$$\mathbf{X} = \mathbf{T}_p\mathbf{P}'_p + \mathbf{T}_o\mathbf{P}'_o + \mathbf{E} \tag{2.9}$$

$$\mathbf{Y} = \mathbf{T}_p\mathbf{C}'_p + \mathbf{F}, \tag{2.10}$$

où **E** et **F** correspondent aux matrices de résidus de **X** et **Y** respectivement. L'OPLS a également été utilisé pour l'analyse discriminante (OPLS-DA) dans les dernières études en métabonomie, notamment dans [22]. Le principal avantage de la classification avec OPLS-DA par rapport à PLS-DA réside dans la capacité de OPLS-DA à séparer la variation prédictive de la variation non-prédictive, ce qui facilite beaucoup plus l'interprétation des résultats. Cet avantage peut être illustré à l'aide d'un scénario simple d'une

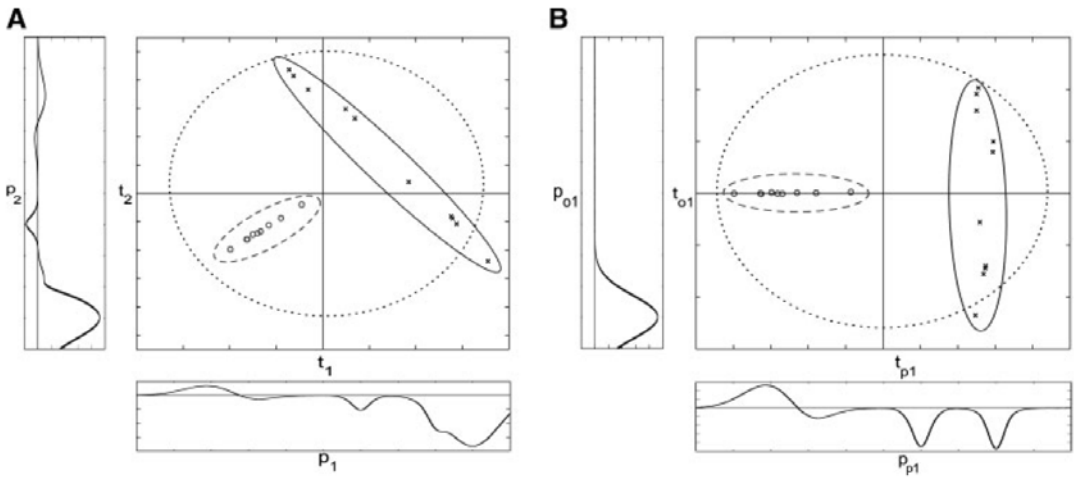


FIGURE 2.9 – La différence entre PLS-DA et OPLS-DA pour la classification de deux classes. Les observations de la première classe sont notées par des cercles et ceux de la deuxième classe par des croix. (**A**), la direction discriminatoire entre les classes pour PLS-DA est une combinaison de $\mathbf{t}_1$ et $\mathbf{t}_2$. La classe 2 a une grande variance intra-classe qui est confondue avec la direction discriminatoire ce qui rend difficile l'interprétation des vecteurs de charge $\mathbf{p}_1$ et $\mathbf{p}_2$. (**B**), OPLS-DA sépare la direction discriminatoire $\mathbf{t}_{p,1}$ de la direction orthogonale $\mathbf{t}_{o,1}$ ce qui rend les vecteurs de charges correspondants plus facile à interpréter. La différence entre OPLS-DA et PLS-DA est plus importante dans le cas multi-classes [22].

base de données spectrales avec deux classes (Figure 2.9). Pour PLS-DA, deux composantes $\mathbf{t}_1$ et $\mathbf{t}_2$ sont nécessaires pour trouver un plan parfaitement discriminant entre les deux classes comme représenté sur la Figure 2.9**A** [22]. Les vecteurs de charge correspondants $\mathbf{p}_1$ et $\mathbf{p}_2$ contiennent un mélange des propriétés discriminantes, ainsi que les propriétés non-discriminantes, qui sont principalement confondues avec la direction de $\mathbf{t}_2$. À l'inverse, OPLS-DA sépare efficacement la direction prédictive $\mathbf{t}_{p,1}$ de la direction $\mathbf{Y}$-orthogonale $\mathbf{t}_{o,1}$ ce qui permet d'interpréter beaucoup plus facilement le vecteur de charge prédictif correspondant, $\mathbf{p}_{p,1}$ (Figure 2.9**B**). Ainsi, les parties du spectre responsables de la variation résiduelle peuvent facilement être identifiées à partir du vecteur de charge de la direction $\mathbf{Y}$-orthogonale $\mathbf{p}_{o,1}$.

2.5.5 Les machines à vecteurs de support

Les machines à vecteurs de support (*Support Vectors Machines, SVM*) [21] ont été utilisées dans divers domaines, dont l'apprentissage statistique [142] et la reconnaissance de formes [145, 82]. Cette approche a également été utilisée pour la classification multispectrale [70] et hyperspectrale [60, 95]. La méthode SVM est basée sur le principe de séparation réalisée par un hyperplan, identifié à partir des échantillons d'apprentissage les plus représentatifs et qui sont situés aux extrémités des classes. Ces échantillons sont appelés vecteurs de support. Si l'ensemble d'apprentissage n'est pas linéairement séparable, une fonction noyau est introduite pour réaliser une projection non-linéaire des données dans un espace de dimension plus grande, où les classes peuvent être linéairement séparables. De plus, les vecteurs de support peuvent être retrouvés à partir d'un petit nombre d'échantillons représentatifs, sans

souffrir du phénomène de Hughes [72], qui consiste en une détérioration des taux de classification au fur et à mesure que les dimensions augmentent lorsque l'ensemble d'apprentissage n'est pas suffisamment grand [97].

Cas de deux classes

Une classification à deux classes peut être formulée de la manière suivante : On considère n échantillons d'apprentissage représentés par l'ensemble de couples $\{(y_i, \mathbf{x}_i),\ i = 1,...,n\}$, avec $\mathbf{x}_i$ un vecteur de variables à N composantes, et y_i les valeurs des étiquettes correspondantes : ± 1. Le classifieur est représenté par la fonction $f(\mathbf{x}) \longrightarrow y$. La méthode SVM consiste à trouver l'hyperplan optimal de séparation, tel que :

1. les échantillons avec des étiquettes $y = \pm 1$ sont situés sur chaque côté de l'hyperplan ;
2. la distance entre les vecteurs les plus proches de l'hyperplan est maximale de chaque côté. Ce sont ces vecteurs qui constituent les vecteurs support.

L'hyperplan est défini par : $\mathbf{w} \cdot \mathbf{x} + b = 0$, où $(\mathbf{w}, b)$ représentent les paramètres de l'hyperplan. Les vecteurs qui ne se trouvent pas sur cet hyperplan correspondent ainsi à : $\mathbf{w} \cdot \mathbf{x} + b \lessgtr 0$ et permettent de définir le classifieur par la fonction : $f(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b)$. Les vecteurs de support se trouvent sur les deux hyperplans parallèles à l'hyperplan optimal d'équation : $\mathbf{w} \cdot \mathbf{x} + b = \pm 1$ (Figure 2.10). L'hyperplan du SVM est entrainé en calculant la solution du problème quadratique :

$$\min_{\mathbf{w},\xi,b} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N} \xi_i \ \text{ s.c } \ y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i,\ i = 1,...,N \ \text{ et } \xi_i \geq 0. \tag{2.11}$$

où ξ est la variable auxiliaire qui permet d'introduire une marge d'erreur lorsque les données ne sont pas linéairement séparables et C représente la pénalité sur le terme de l'erreur. La solution du problème (2.11) est calculée en utilisant la méthode de Lagrange.

Cas multi-classes

À la base, la méthode SVM a été développée pour résoudre le cas de la classification à deux classes seulement. Néanmoins, il est possible de l'adapter à la classification à M classes de deux façons possibles :

1. la méthode un contre tous (*one versus all*) qui consiste à apprendre itérativement un classifieur permettant de discriminer chaque classe m par rapport aux $M-1$ classes restantes, puis l'étiquette la plus probable est affectée aux observations.
2. La méthode de comparaison par paires (*one versus one*) où $\frac{M(M-1)}{2}$ classifieurs sont appliqués sur chaque paire de classes. L'étiquette la plus souvent calculée est celle qui est maintenue pour chaque vecteur.

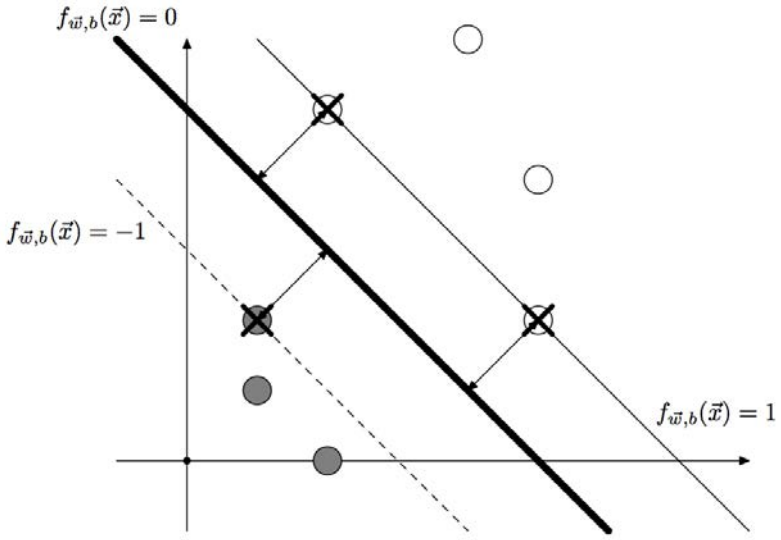


FIGURE 2.10 – Le séparateur de marge maximale définit au moins deux classes. Les vecteurs de support sont représentés par des croix. Les courbes de niveaux $f_{\vec{w},b}(\vec{x}) = \pm 1$ coupent les vecteurs de support.

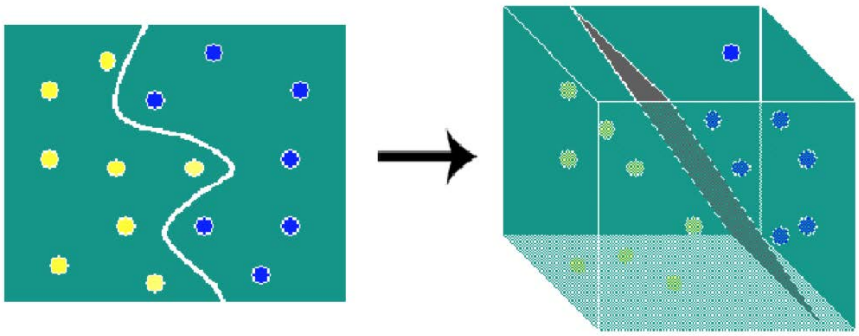


FIGURE 2.11 – Séparation linéaire par l'hyperplan dans un espace de dimension supérieure.

SVM à fonction noyau

Cette approche permet de calculer des surfaces de décision non-linéaires et consiste à faire une projection des données dans un espace de dimension supérieure où elles peuvent être séparées linéairement, comme le montre la Figure 2.11. Par conséquent, lorsque le SVM est appliqué dans cet espace il permet de déterminer les surfaces non-linéaires dans l'espace d'origine. Soit donc $\mathbf{x} \in \mathbb{R}^N$ qu'on projette sur un espace de dimension supérieure $\mathscr{H}$ par une fonction non linéaire $\Phi : \mathbb{R}^N \longrightarrow \mathscr{H}$, et le produit scalaire $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ est remplacé par $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathscr{H}}$. La fonction noyau $\ker(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathscr{H}}$ peut prendre différentes expressions à condition de satisfaire la condition de Mercer qui consiste en :

$\forall g(\cdot) \in \mathscr{L}^2(\mathbb{R}^N)$ tel que $\int g(\mathbf{x})^2 d\mathbf{x}$ est fini, alors :

$$\int \ker(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0. \tag{2.12}$$

L'hyperplan du SVM à fonction noyau est calculé cette fois en utilisant la formulation (2.11) en

remplaçant x_i par la fonction noyau $\Phi(\mathbf{x}_i)$:

$$\min_{\mathbf{w},\xi,b} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}\xi_i \text{ s.c } y_i(\mathbf{w}'\Phi(\mathbf{x}_i)+b) \geq 1-\xi_i,\ i=1,...,N \text{ et } \xi_i \geq 0. \tag{2.13}$$

Il est difficile d'expliquer les caractéristiques propres à chaque noyau mais en général, on peut définir deux types de noyau :

1. Les noyaux locaux :
   (a) Le noyau de base radiale : $\ker(\mathbf{x},\mathbf{x}_i) = \exp(-\|\mathbf{x}-\mathbf{x}_i\|^2)$,
   (b) Le noyau KMOD : $\ker(\mathbf{x},\mathbf{x}_i) = \exp\left(\frac{1}{1+\|\mathbf{x}-\mathbf{x}_i\|^2}\right) - 1$,
   (c) Le noyau multiquadratique inverse : $\ker(\mathbf{x},\mathbf{x}_i) = \frac{1}{\sqrt{\|\mathbf{x}-\mathbf{x}_i\|^2+1}}$.
2. Les noyaux globaux :
   (a) Le noyau linéaire : $\ker(\mathbf{x},\mathbf{x}_i) = \mathbf{x}\cdot\mathbf{x}_i$,
   (b) Le noyau polynomial : $\ker(\mathbf{x},\mathbf{x}_i) = (\mathbf{x}\cdot\mathbf{x}_i+1)^p$,
   (c) Le noyau sigmoid : $\ker(\mathbf{x},\mathbf{x}_i) = \tanh(\mathbf{x}\cdot\mathbf{x}_i+1)$,

Le classifieur SVM non-linéaire s'écrit alors sous la forme :

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{K_s}\lambda_i y_i \ker(\mathbf{s}_i,\mathbf{x}) + b\right) \tag{2.14}$$

où λ_i représente les multiplicateurs de Lagrange qui sont non nuls pour les vecteurs de support $\mathbf{s}_i$ et K_s le nombre de vecteurs de support.

2.5.6 SVM parcimonieux

L'idée des SVM parcimonieux est de réaliser des performances de classification similaires à celles obtenues par le SVM standard, tout en imposant une contrainte sur la cardinalité du vecteur normal à l'hyperplan. Dans le cadre des SVM, il existe deux stratégies pour la sélection de variables [25] : (1) sélection des variables conjointement à l'apprentissage du classifieur ; (2) sélection des variables après apprentissage du classifieur.

Sélection de variables conjointement à l'apprentissage du classifieur L'idée est de reformuler la fonction de coût du SVM afin d'assurer la parcimonie de la règle de décision. Les ajustements se basent en général sur la minimisation de la norme 0 (cardinalité) du plan orthogonal à l'hyperplan séparateur. Néanmoins, intégrer une contrainte en norme 0 à la fonction SVM rend la résolution difficile car ce problème est combinatoire. D'autres travaux [9, 17, 161] ont proposé une approche permettant d'obtenir des vecteurs de support parcimonieux en minimisant le critère original avec la norme 1 du vecteur normal

à l'hyperplan. Ceci se traduit en remplaçant la norme 2 par la norme 1 du vecteur du support dans la formulation du SVM standard (2.11) de la façon suivante :

$$\min_{\mathbf{w},\xi,b} ||\mathbf{w}||_1 + C\sum_{i=1}^{N}\xi_i \text{ s.c } y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1-\xi_i,\ i=1,...,N \text{ et } \xi_i \geq 0. \tag{2.15}$$

Ce critère est résolu par la programmation linéaire et est connu sous le nom de SVM par programmation linéaire (LP-SVM). De manière générale, imposer la parcimonie sur $\mathbf{w}$ revient à augmenter la fonction objectif du SVM avec un terme de pénalité sur la cardinalité (ou sa relaxation convexe) de ce vecteur. Ainsi, dans [25] une approche basée sur la résolution d'une fonction SVM intégrant une contrainte en norme 0 sur les vecteurs support, est proposée. Cette approche consiste à résoudre le problème d'optimisation suivant :

$$\min_{\mathbf{w},\xi,b} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}\xi_i \text{ s.c } y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1-\xi_i,\ i=1,...,N\ ,\ \xi_i \geq 0 \text{ et } ||\mathbf{w}||_0 \leq s. \tag{2.16}$$

La solution de ce problème n'est pas triviale à cause de la contrainte en norme 0. Une solution proposée dans [25] consiste à relaxer le problème par une norme 1. En fait, cette méthode peut être interprétée comme la combinaison du SVM standard (minimisation de la norme 2 de $\mathbf{w}$) et le LP-SVM. La solution est alors obtenue en utilisant une méthode basée sur le programme quadratique à contraintes quadratiques, *Quadratically-constrained quadratic program* (QCQP). Une approche efficace a également été développé dans [46] pour la résolution d'un critère SVM avec une régularisation en norme mixte dans le cadre des SVM à noyaux multiples. L'approche proposée impose une contrainte de parcimonie en utilisant la norme ℓ_1 ou des normes mixtes de type $\ell_1 - \ell_q$ sur les vecteurs de support de la façon suivante :

$$\min_{\mathbf{w},b}\ y_i(\mathbf{w}'\mathbf{x}_i + b) + \lambda\Omega(\mathbf{w}). \tag{2.17}$$

où $\Omega(\mathbf{w})$ est le terme de régularisation utilisée et λ est le paramètre de régularisation. Ainsi, $\Omega(\mathbf{w})$ peut prendre plusieurs formes selon la nature du problème initial. Pour le SVM standard on a :

$$\Omega_2(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||_2^2. \tag{2.18}$$

Lorsque la règle de classification est réalisée sur quelques coefficients du vecteur de support, on peut alors utiliser la norme 1 de $\mathbf{w}$ tel que :

$$\Omega_1(\mathbf{w}) = \sum_{i=1}^{N}|w_i|. \tag{2.19}$$

Enfin, si le problème de classification initial est un problème multi-tâches, c'est à dire que l'apprentissage de différentes tâches est réalisé pour plusieurs variables simultanément, la norme mixte $\ell_1 - \ell_q$ avec

$1 \leq q \leq 2$, est utilisée en considérant la régularisation suivante :

$$\Omega_{1-q}(\mathbf{w}) = \sum_{g \in G} ||\mathbf{w}_g||_q \tag{2.20}$$

où $\mathbf{w}_g$ est le vecteur de support du group g et G est une partition de l'ensemble $\{1,...,N\}$. Dans les deux derniers cas, une méthode de gradient projeté en utilisant les opérateurs proximaux adéquats permet de trouver la solution du problème.

Sélection de variables après apprentissage du classifieur Ces approches réalisent la sélection de variables après l'apprentissage du modèle. Ainsi, les variables sont classées en fonction des paramètres de l'hyperplan. Dans [61], un premier modèle SVM est entraîné, puis la variable qui influence le moins la règle de décision est éliminée. Ensuite, un nouveau modèle SVM est entraîné sur les variables restantes. Cette procédure est répétée jusqu'à obtenir le nombre de variables désiré. Ce travail a été développé dans le cadre de la sélection de gènes pour la classification du cancer. Une extension de cette approche a été proposée dans [113] en utilisant un autre critère de classement basé également sur les paramètres de l'hyperplan. Ce critère est basé sur une méthode de premier ordre dans laquelle les variables sont rangées en fonction de la valeur minimale de la valeur absolue de la dérivé de $||\mathbf{w}||^2$, permettant ainsi de sélectionner n variables ($n < N$) qui maximisent les performances du classifieur.

2.5.7 Les réseaux de neurones

Les réseaux de neurones artificiels (*Artificial Neural Network ANN*) [4] sont un outil puissant pour la modélisation des données permettant de capturer et de représenter des relations complexes entre les données d'entrée et de sortie. Cette approche a été étudiée et utilisée en chimiométrie dans les travaux de Zupan [164] et de Melssen *et.al* [96] ainsi que pour la classification de données spectrales en proche infra-rouge [153]. L'intérêt des réseaux de neurones en chimiométrie est dû au fait que les méthodes de classification classiques exigent la connaissance de certaines propriétés sur les distributions des données, alors que cette technique peut être appliquée à n'importe quel problème de classification. La seule exigence dans ce cas concerne l'ensemble d'apprentissage qui doit être suffisamment représentatif (contient suffisamment de données). Cependant, la durée de la phase d'apprentissage peut s'avérer très longue. Pendant la phase d'apprentissage aucune considération sur les propriétés physiques d'une observation n'est prise en compte. Ainsi, les paramètres du système de neurones s'adaptent progressivement aux propriétés de la signature spectrale. Par ailleurs, le modèle de réseaux de neurones le plus communément utilisé en classification est le perceptron multicouche. Il consiste en une couche d'entrée et une de sortie, avec au milieu une ou plusieurs couches intercalées appelées couches cachées. Le nombre de paramètres internes à ajuster croît avec le nombre de couches, et du nombre de neurones par couche. Cette propriété est mise en évidence dans l'exemple du réseau à trois couches. La couche d'entrée dispose de n possibilités, la couche cachée contient k neurones et celle de sortie m neurones. En connectant chaque entrée à tous les neurones et chaque neurone à tous ses suivants, le nombre de poids à ajuster est $n \times k \times m$.

2.6 Classification des spectres de déchets bois

Nous nous proposons ici d'évaluer les performances de quelques méthodes de classification supervisée sur notre base de données spectrales de déchets de bois. Nous utilisons 294 spectres : 173 spectres de bois purs (catégorie 1) et 121 spectres de bois pollués (catégorie 2). Ces spectres sont composés de 1647 longueurs d'ondes. Nous procédons d'abord à la suppression de la ligne de base et à la normalisation de la matrice de données $\mathbf{Y} \in \mathbb{R}^{1647\times294}$. Les méthodes K-means, SIMCA, K-plus proches voisins (K-NN) avec $K = 7$, la méthode G-SVM développée dans [46] en utilisant la norme 1 des vecteurs de support, et SVM sont appliquées sur l'ensemble des données. L'objectif est de comparer les résultats de classification obtenus pour chaque méthode. On considère 25% des spectres comme ensemble d'apprentissage et 75% des spectres comme l'ensemble de validation. En ce qui concerne la méthode G-SVM le paramètre qui favorise la parcimonie des vecteurs du poids est réglé de façon à obtenir 40 coefficients non nuls (afin de nous positionner par rapport à cette approche). La fonction quadratique est utilisée comme noyau du SVM standard.

Résultats	K-means	SIMCA	K-NN	G-SVM (40 coef.)	SVM
Taux de classification	66.85%	54.42%	74.55%	75.91%	73.64%
Taux de bonne détection Cat.1	49.8%	43.93%	79.07%	80.62%	63.57%
Taux de bonne détection Cat.2	79.3%	69.42%	68.13%	69.23%	87.91%

TABLE 2.1 – Résultats de la classification des spectres de déchets de bois obtenus par les méthodes K-means, SIMCA, K-NN, G-SVM et SVM.

Les résultats des cinq méthodes, présentés sur le Tableau 2.1, montrent que des bonnes performances peuvent être atteintes avec des méthodes de classification supervisée comme K-NN, G-SVM et SVM contrairement aux approches K-means et SIMCA dont les performances restent très limitées. En particulier, les taux de bonnes classification de la catégorie 1 obtenus en utilisant K-means et SIMCA sont très faibles. Les résultats obtenus avec la méthode K-NN sont satisfaisants pour la classification de la catégorie 1, cependant, cette méthode renvoie des taux plus faibles en ce qui concerne la classification de la deuxième catégorie. Les performances de classification des deux approches G-SVM et SVM sont meilleures que celles obtenues par les autres méthodes. D'une part, la méthode G-SVM permet d'atteindre des taux de classification supérieurs à 70%, en considérant uniquement 40 coefficients des vecteurs de support. Néanmoins, cette approche ne permet pas de sélectionner directement des longueurs d'ondes afin de les implémenter physiquement sur les machines de tri, puisque la sélection est réalisée sur les coefficients du vecteur de support $\mathbf{w}$. D'autre part, on constate que SVM permet d'obtenir de bons résultats de classification. En particulier, les résultats de bonnes classification de la Catégorie 2 obtenus avec SVM sont meilleurs que ceux des autres approches. Cet aspect est important dans ce type d'application car il implique un rejet considérable des bois pollués. Globalement, les résultats obtenus confirment que les méthodes basées sur SVM réussissent à bien discriminer les deux catégories.

2.7 Conclusion

Nous avons présenté dans ce chapitre un éventail des méthodes de classification qui sont principalement utilisées dans le domaine de chimiométrie et de la classification de données spectrales. De notre point de vue, le SVM est l'une des méthodes les plus performantes grâce aux fonctions noyaux qui permettent la séparation non linéaire des données, mais aussi, les plus efficaces en termes de temps de calcul (en comparaison avec les réseaux de neurones par exemple). Par ailleurs, lorsque l'application implique des données de grande dimension, il est nécessaire de réaliser une première phase de sélection de variables afin de déterminer un sous-ensemble de variables pertinentes pour la classification. En ce qui concerne la problématique du projet Trispirabois, la complexité de la phase d'apprentissage due à la présence de groupes avec des caractéristiques différentes dans une seule classe, ainsi que la contrainte liée au temps de traitement doivent être gérées efficacement. Par conséquent, une première étape de sélection de variables doit être réalisée sur les données, afin de réduire la dimension du problème et de sélectionner un sous ensemble de longueurs d'ondes qui garantit des performances de classification similaires à celles obtenues en utilisant la totalité des données. Ainsi, l'apprentissage du classifieur est réalisé sur ces variables afin de déterminer le modèle qui sera utilisé par la suite. L'objectif est que l'étape de la validation (le tri de déchets de bois) soit effectuée de façon efficace et peu coûteuse puisque l'application visée consiste en un traitement en ligne des données.
Nous présentons dans le chapitre suivant plusieurs approches pour la sélection de variables basées sur la représentation parcimonieuse.

Chapitre 3

Méthodes d'approximation parcimonieuse

3.1 Introduction

L'approximation parcimonieuse est un problème inverse largement étudié ces dernières années et appliqué dans différents domaines tels que la compression [36, 24], l'analyse spectrale [73] ou encore la régression [68] ainsi que les problèmes de classification [80]. Elle consiste à représenter un signal en utilisant le minimum d'éléments à partir d'un dictionnaire connu Φ.
Dans ce chapitre, nous commençons d'abord par formuler le problème d'approximation parcimonieuse en posant un critère qui inclut la pseudo-norme ℓ_0, puis nous présentons quelques méthodes gloutonnes permettant de trouver une solution approchée à ce problème. Ensuite, nous présentons la relaxation convexe classique du problème en substituant la norme ℓ_1 à la norme ℓ_0. D'autres types de critères sont également traités avec, entre autres, des critères en norme mixte permettant d'imposer des contraintes supplémentaires sur la nature de la solution. Par exemple, nous abordons un critère imposant une parcimonie structurée via la contrainte de groupe qui permet d'obtenir une solution parcimonieuse par groupe. Enfin, nous proposons de développer des approches gloutonnes pour résoudre le problème d'approximation parcimonieuse simultanée. Ce cas de figure consiste à estimer une matrice creuse de coefficients correspondant à des mesures multiples, à partir du dictionnaire Φ. L'approche simultanée garantit la parcimonie des lignes de la matrice de coefficients. Des simulations numériques sont réalisées pour montrer les performances de bonne reconstruction des algorithmes développés. Nous terminons le chapitre par une conclusion.

3.2 Approximation parcimonieuse

3.2.1 Méthodes gloutonnes

La notion de parcimonie consiste à reconstruire un signal en utilisant un nombre minimal d'éléments. Cette représentation a été largement étudiée ces dernières années pour résoudre de nombreux problèmes mal-posés dans plusieurs domaines tels le dé-bruitage d'images [41], le traitement audio [112], ainsi que la séparation de sources [162, 83]. Elle permet l'approximation d'un signal par une représentation

compacte et facile à manipuler qui consiste en un jeu de quelques coefficients non nuls correspondant à une collection fixe de fonctions appelée "dictionnaire". En effet, seule la connaissance des positions et des valeurs de ces coefficients non-nuls est nécessaire pour définir le signal d'entrée. On définit alors l'ensemble des coefficients non-nuls d'un vecteur, appelé le support de $\mathbf{x} \in \mathbb{R}^N$, comme étant :

$$\Gamma = \text{supp}(\mathbf{x}) = \{1 \leq i \leq N \,|\, x^i \neq 0\}, \tag{3.1}$$

et la norme 0 du vecteur parcimonieux $\mathbf{x}$ est définie par :

$$||\mathbf{x}||_0 = |\Gamma|. \tag{3.2}$$

Le problème d'approximation parcimonieuse linéaire est un problème inverse formulé dans le cas d'absence de bruit comme suit :

$$\min_{\mathbf{x}} ||\mathbf{x}||_0 \qquad \text{sous contrainte} \qquad \mathbf{y} = \Phi\mathbf{x}, \tag{3.3}$$

où $\Phi \in \mathbb{R}^{M\times N}$ est le dictionnaire et $\mathbf{y} \in \mathbb{R}^M$ est le signal d'entrée. Dans le cas d'un signal bruité, le problème (3.3) s'écrit par exemple :

$$\min_{\mathbf{x}} ||\mathbf{x}||_0 \qquad \text{sous contrainte} \qquad ||\mathbf{y} - \Phi\mathbf{x}||_2^2 \leqslant \varepsilon \tag{3.4}$$

où ε représente l'erreur d'approximation. Une autre façon d'exprimer ce problème consiste à fixer le nombre de coefficients non nuls s désiré dans la solution comme suit :

$$\min_{\mathbf{x}} ||\mathbf{y} - \Phi\mathbf{x}||_2^2 \qquad \text{sous contrainte} \qquad ||\mathbf{x}||_0 \leqslant s. \tag{3.5}$$

Le problème de trouver le vecteur parcimonieux $\mathbf{x}$ permettant l'estimation d'un vecteur d'observation donné $\mathbf{y}$ est un problème NP-complet. Par conséquent, il ne peut pas être résolu dans un temps polynomial [135]. Ainsi, plusieurs approches sous-optimales ont été explorées et ont démontré de bonnes performances de reconstruction. Ces approches sont souvent basées sur la relaxation convexe [23], l'optimisation non-convexe [49, 116], les approches bayésiennes [146, 118] ou encore les stratégies dites gloutonnes [11]. Ces dernières sont généralement plus rapides que les autres approches et possèdent aussi, sous certaines conditions, de bonnes performances [131, 24, 12, 99]. Nous proposons de présenter dans un premier temps l'algorithme OMP (*orthogonal matching pursuit*) qui est un des premiers algorithmes gloutons développé pour résoudre le problème d'approximation parcimonieuse.

Orthogonal Matching Pursuit (OMP)

Dans une implémentation typique de cet algorithme (Figure 3.1), l'étape d'identification est la partie la plus coûteuse en termes de calcul. Elle est basée sur le calcul de la corrélation maximale entre le résidu courant et le dictionnaire : $\Phi'\mathbf{r}_{k-1}$, avec un nombre de multiplication en $\mathcal{O}(MN)$. L'étape d'estimation nécessite la résolution d'un problème de moindres carrés. La technique la plus courante consiste à utiliser

une factorisation QR ou une décomposition de Cholesky de la matrice Φ_{Γ_k}, à la $k-$ème itération avec une coût de calcul en $\mathcal{O}(Mk)$.

FIGURE 3.1 – Algorithme OMP

— **Entrée** : $\mathbf{y}$, Φ, s, un critère d'arrêt.
— **Initialisation** : $\mathbf{r}_0 = \mathbf{y}$, $\Gamma_0 = \emptyset$, $\mathbf{x}_0 = 0$.
— **Pour** : $k \leftarrow 1$ jusqu'à ce que le critère d'arrêt soit satisfait **faire** :
1. $n_k = \arg\max_n |\phi_n' \mathbf{r}_{k-1}|$. Identification
2. $\Gamma_k = \Gamma_{k-1} \cup n_k$
3. $\mathbf{x}_k = \arg\min_{\mathbf{u}} ||\mathbf{y} - \Phi_{\Gamma_k}\mathbf{u}||_2$. Estimation des coefficients
4. $\mathbf{r}_k = \mathbf{y} - \Phi_{\Gamma_k}\mathbf{x}_k$. Mise à jour du résidu

— **Sortie** : Γ, $\mathbf{x}$ et $\mathbf{r}$.

Il existe plusieurs critères d'arrêt de l'algorithme :
— le nombre maximal d'itérations est atteint : $k = s$;
— la norme du résidu est en dessous d'un certain seuil : $||\mathbf{r}_k|| \leqslant \varepsilon$;
— la corrélation maximale entre un atome et le résidu passe en dessous d'un seuil τ : $||\Phi'\mathbf{r}_k||_\infty \leqslant \tau$.

Orthogonal Least Squares (OLS)

L'algorithme OLS semble être très similaire à l'algorithme OMP, de sorte que beaucoup de confusion a été faite entre les deux [13]. En effet, les deux algorithmes réalisent les mêmes étapes, la différence résidant uniquement dans la procédure de sélection. La sélection dans OMP est basée sur le calcul du produit scalaire entre le résidu et les vecteurs colonnes ϕ_n de Φ. L'élément dont la corrélation est maximale avec le résidu est sélectionné. Dans l'algorithme OLS la sélection même si elle est gloutonne, est réalisée après une étape de projection orthogonale sur les éléments sélectionnés. Ainsi, la projection orthogonale des éléments ϕ_n avant le calcul du produit scalaire caractérise la méthode OLS qui sélectionne l'élément qui garantit la corrélation maximale avec le résidu après cette projection comme représenté sur la Figure 3.2.

Les étapes de OLS sont présentées dans la Figure 3.3. Cet algorithme est en général plus coûteux en termes de temps de calcul. Néanmoins, nous adoptons ici une implémentation plus rapide de cette approche, basée sur la factorisation QR.

Par ailleurs, d'autres algorithmes gloutons pour l'approximation parcimonieuse avec une contrainte non convexe ont été développés. Par exemple, l'algorithme CoSaMP (*Compressive Sampling Matching Pursuit*) [104] est une variante de OMP permettant de sélectionner plusieurs atomes à chaque itération. Cet algorithme incorpore en plus une technique combinatoire en comparaison à l'algorithme OMP. En général, le temps de calcul de la solution par cette méthode est plus faible et l'étude de ses propriétés a conduit à des bornes d'erreur rigoureuses [10]. Des méthodes basées sur des procédures de seuillage ont également été utilisées pour résoudre le même problème, notamment, IHT (*iterative hard thresholding*) [14], NIHT (*normalized iterative hard thresholding*) [15], HTP (*hard thresholding pursuit*) et N-

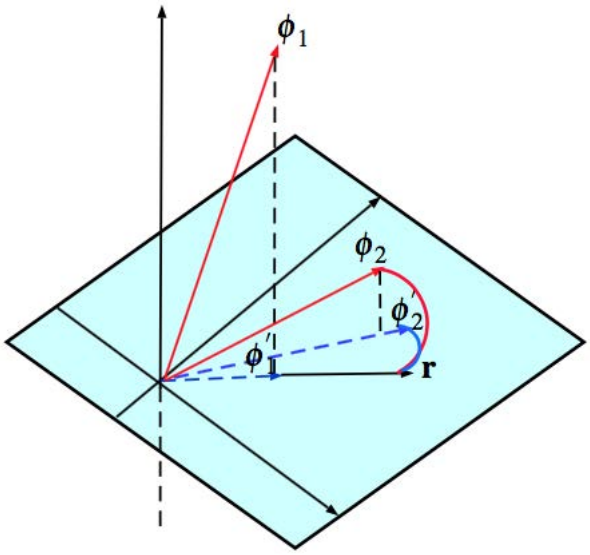


FIGURE 3.2 – La différence entre OMP et OLS avec ϕ_1 et ϕ_2 deux vecteurs non sélectionnés. l'algorithme OMP sélectionne le nouvel élément en fonction de l'angle entre le vecteur d'origine et le résidu (angle rouge) en cours. L'OLS sélectionne l'élément ayant le petit angle après sa projection sur le sous-espace orthogonal, c'est-à-dire (ϕ_1' et ϕ_2').

FIGURE 3.3 – Algorithme OLS

— **Entrée** : $\mathbf{y}$, Φ, s, critère d'arrêt.
— **Initialisation** : $\mathbf{r}_0 = \mathbf{y}$, $\Gamma_0 = \emptyset$
— **Pour** : $k \leftarrow 1$ jusqu'à ce que le critère d'arrêt soit satisfait **faire** :
  1. $n_{\max} = \arg\min_{\Gamma_k = \Gamma_{k-1} \cup n} \|\mathbf{y} - \Phi_{\Gamma_k} \Phi_{\Gamma_k}^{\dagger} \mathbf{y}\|_2$
  2. $\mathbf{x}_{\Gamma_k} = \Phi_{\Gamma_k}^{\dagger} \mathbf{y}$.
  3. $\mathbf{r}_k = \mathbf{y} - \Phi_{\Gamma_k} \mathbf{x}_{\Gamma_k}$
— **Sortie** : Γ, $\mathbf{x}_\Gamma$ et $\mathbf{r}$.

HTP (*Normalized Hard Thresholding Pursuit*) [48]. Toutes ces approches sont des approches de type avant (*forward*), dans le mesure où, une fois qu'un atome est sélectionné, il ne peut plus être retiré de l'ensemble de la solution.

Single Best Replacement (SBR)

Plus récemment, l'algorithme SBR (*single best replacement*) a été proposé dans [123] pour une sélection avant/arrière (*forward/backward*). Il consiste à résoudre le problème pénalisé suivant :

$$\arg\min_{\mathbf{x} \in \mathbb{R}^N} \{\mathscr{J}(\mathbf{x}, \lambda) = \mathscr{E}(\mathbf{x}) + \lambda ||\mathbf{x}||_0\} \tag{3.6}$$

où $\mathscr{E}(\mathbf{x})$ est l'erreur d'approximation :

$$\mathscr{E}(\mathbf{x}) = ||\mathbf{y} - \Phi\mathbf{x}||_2^2 \tag{3.7}$$

et $\lambda > 0$ est un paramètre permettant de réaliser un compromis entre l'erreur d'approximation et la parcimonie de $\mathbf{x}$. Ainsi, à chaque itération, l'algorithme teste les N remplacements simples possibles, $\Gamma \bullet n$, $n = 1,...,N$, puis choisit l'atome qui fait décroître le plus la fonction de coût $\mathscr{J}(\mathbf{x},\lambda)$, c'est-à-dire, $\mathscr{J}_{\Gamma\bullet n}(\lambda) = \mathscr{E}_{\Gamma\bullet n} + \lambda \text{Card}(\Gamma \bullet n)$. SBR s'arrête quand il n'est plus possible de diminuer la fonction de coût (3.6). On note $\Gamma \bullet n$ une nouvelle insertion ou un retrait d'un indice de Γ :

$$\Gamma \bullet n = \begin{cases} \Gamma \cup \{n\} & \text{si } n \notin \Gamma \\ \Gamma \setminus \{n\} & \text{sinon} \end{cases} \tag{3.8}$$

Les étapes de cette méthode sont présentées dans la Figure 3.4.

FIGURE 3.4 – Algorithme SBR

— **Entrée** : $\mathbf{y}$, $\Phi \in \mathbb{R}^{M\times N}$, paramètre de régularisation λ.
— **Initialisation** : $\mathbf{r}_0 = \mathbf{y}$, $\Gamma_0 = \emptyset$.
— **Pour** $k \leftarrow 1$ jusqu'à ce que l'ensemble Γ_k n'est plus mis à jour **faire** :

1. $n_k \in \arg\min_n \{\mathscr{J}_{\Gamma_{k-1}\bullet n}(\lambda) = \mathscr{E}_{\Gamma_{k-1}\bullet n} + \lambda \text{Card}(\Gamma_{k-1} \bullet n)\}$.
2. **Si** : $\mathscr{J}_{\Gamma_{k-1}\bullet n_k}(\lambda) < \mathscr{J}_{\Gamma_{k-1}}(\lambda)$,
3. $\Gamma_k = \Gamma_{k-1} \bullet n_k$
   Fin si
4. $\mathbf{x}_{\Gamma_k} = \Phi_{\Gamma_k}^{\dagger}\mathbf{y}$.

— **Sortie** : Γ, $\mathbf{x}_\Gamma$.

3.2.2 Méthodes basées sur la relaxation ℓ_1

Une autre façon d'aborder le problème d'approximation parcimonieuse est la relaxation convexe. Cette technique consiste à remplacer la quasi-norme ℓ_0 par la norme ℓ_1 qui est une norme liée à une fonction convexe. Ceci correspond en fait, à chercher une approximation du problème original et à le résoudre par une méthode exacte. La forme convexe du problème contraint est écrite sous la forme :

$$\min_{\mathbf{x}} ||\mathbf{x}||_1 \qquad \text{sous contrainte} \qquad \Phi\mathbf{x} = \mathbf{y}. \tag{3.9}$$

Dans le cas ou le signal est bruité on écrit :

$$\min_{\mathbf{x}} ||\mathbf{x}||_1 \qquad \text{sous contrainte} \qquad ||\mathbf{y} - \Phi\mathbf{x}||_2^2 \leq \varepsilon. \tag{3.10}$$

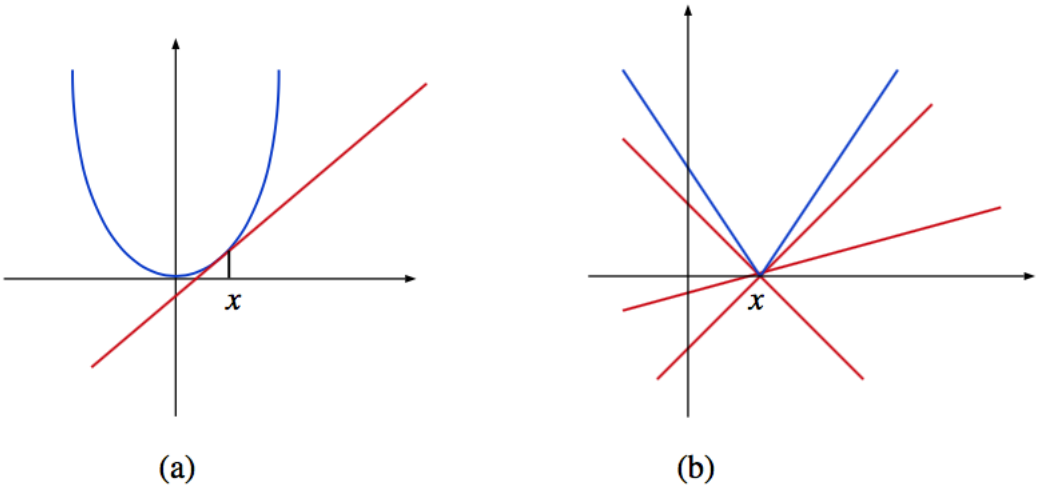


FIGURE 3.5 – Exemple de deux fonctions convexes. (a)- fonction lisse, (b)- fonction non lisse. Les droites en rouge représentent le sous-gradient des deux fonctions au point x.

où ε correspond à l'erreur de reconstruction due à la présence du bruit. Enfin, le problème intégrant le niveau de parcimonie de la solution est formulé comme suit :

$$\min_{\mathbf{x}} ||\mathbf{y} - \Phi\mathbf{x}||_2^2 \qquad \text{sous contrainte} \qquad ||\mathbf{x}||_1 \leqslant s \tag{3.11}$$

Ce problème a été étudié dans divers domaines, tel que la déconvolution [124] et la classification [53]. La forme pénalisée du problème (3.11) est donnée par :

$$\min_{\mathbf{x}} \tfrac{1}{2}||\mathbf{y} - \Phi\mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1 \tag{3.12}$$

avec λ le paramètre qui contrôle le niveau de parcimonie de la solution dans le sens où, la solution est d'autant plus parcimonieuse que λ est grand. À noter que la forme contrainte (3.11) est équivalente à la forme pénalisée dans (3.12).

3.2.3 Lasso

Le problème pénalisé (3.12) est souvent associé dans la littérature au *Lasso* (*least absolute Shrinkage and selection operator*) ou la régression *Lasso* [127]. Il a été introduit en traitement de signal sous le nom de *basis pursuit* dans [30]. Différents résultats et discussions relatifs à sa résolution sont donnés dans [138, 37, 132]. La régularisation par la norme ℓ_1 induit la parcimonie de la solution dans le sens où un certain nombre des coefficients de $\mathbf{x}$ sera nul, selon le niveau de régularisation appliqué par le paramètre λ. En revanche, si la norme 1 est remplacée par la norme 2 dans l'expression précédente, le problème correspond au *ridge regression* développé dans [69] qui est également convexe, mais dont la solution n'est pas parcimonieuse.

Le paramètre λ agit sur la parcimonie de la solution de façon indirecte, de sorte qu'il est souvent nécessaire de tester différentes valeurs pour trouver le chemin de la solution au fur et à mesure que λ

tend vers 0. En effet, des travaux se sont intéressés à la recherche du chemin de régularisation [108, 88] par des procédures d'homotopie en faisant varier λ et, ont démontré que la solution optimale du problème suit un chemin linéaire par morceaux. En fait, lorsque les colonnes du dictionnaire Φ sont orthogonales, la résolution du problème revient à trouver une solution à N problèmes de seuillage doux [38, 39]. La résolution consiste à calculer le gradient de la fonction convexe décrite par l'équation (3.12) qui correspond à une fonction affine tangente à la courbe de la fonction comme illustré sur la Figure 3.5. Lorsque le dictionnaire est orthogonal, la solution du problème (3.12) peut être trouvée en utilisant la Proposition 1.

Proposition 1. *Soit la fonction convexe* $f : \mathbb{R}^N \to \mathbb{R}$ *et* δf *son gradient,* $\mathbf{x} \in \mathbb{R}^N$ *est un minimum global de* f *si et seulement si* $0 \in \delta f(\mathbf{x})$.

Cette solution est donnée par la forme analytique du sous-différentiel δ associé à l'équation (3.12) :

$$\hat{\mathbf{x}} = \text{sign}(\tilde{\mathbf{x}}) \left(|\tilde{\mathbf{x}}| - \frac{\lambda}{2} \right)_+ \tag{3.13}$$

où $\tilde{\mathbf{x}}$ est la solution des moindres carrés et $(\kappa)_+ = \max(\kappa, 0), \forall \kappa \in \mathbb{R}$. La fonction sign est définie pour tout réel α par :

$$\text{sign}(\alpha) = \begin{cases} 1 & \text{pour } \alpha > 0 \\ -1 & \text{pour } \alpha < 0 \\ 0 & \text{pour } \alpha = 0. \end{cases} \tag{3.14}$$

Ainsi, la solution optimale est obtenue grâce à un seuillage doux des éléments sélectionnés. L'opérateur équivalent utilisé dans le cadre de l'optimisation non-convexe (norme ℓ_0) et le seuillage dur. On présente dans la Figure 3.6 la forme qui caractérise ces deux opérateurs.

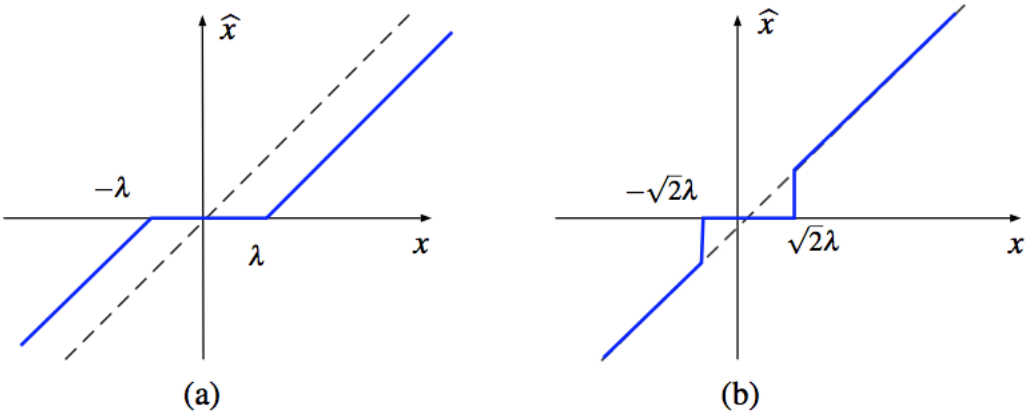


FIGURE 3.6 – Opérateurs de seuillage : (a)- seuillage doux, (b)- seuillage dur.

Par ailleurs, et plus récemment, Efron et *al.* [40] ont introduit l'algorithme d'homothopie LARS (*Least Angle Regression*), qui permet de calculer rapidement la solution du problème *Lasso*.

3.2.4 Fused sparse Lasso

Le *fused sparse Lasso* est une extension du *Lasso* standard, introduite dans [128]. Cette approche a été développée pour les problèmes dans lesquels les variables sont rangées dans un ordre significatif. En effet, le *fused sparse Lasso* incorpore un terme qui favorise la parcimonie des coefficients et de leurs différences. La fonction de coût relative au *fused sparse Lasso* est exprimée par :

$$\min_{\mathbf{x}} \tfrac{1}{2}||\mathbf{y}-\Phi\mathbf{x}||_2^2+\lambda_1||\mathbf{x}||_1+\lambda_2||\mathbf{D}\mathbf{x}'||_1 \tag{3.15}$$

où $\mathbf{D}$ est la matrice de différences finies d'ordre 1 de taille $(N-1)\times N$:

$$\mathbf{D}=\begin{bmatrix} 1 & -1 & & \mathbf{0} \\ & \ddots & \ddots & \\ \mathbf{0} & & 1 & -1 \end{bmatrix}. \tag{3.16}$$

On peut aussi ré-écrire le problème (3.15) comme suit :

$$\min_{\mathbf{x}} \tfrac{1}{2}||\mathbf{y}-\Phi\mathbf{x}||_2^2+g(\mathbf{x}), \tag{3.17}$$

où

$$g(\mathbf{x})=\lambda_1||\mathbf{x}||_1+\lambda_2\sum_{i=1}^{N-1}|x^i-x^{i+1}|, \tag{3.18}$$

avec la fonction convexe et non lisse g qui correspond aux pénalités de parcimonie et de fusion. Les paramètres $\lambda_1,\lambda_2\geq 0$ contrôlent la parcimonie de la solution et la parcimonie de la différence entre les coefficients successifs respectivement (Figure 3.7). La solution à ce problème n'est pas triviale car la pénalité de fusion est non lisse et non séparable. De plus, le critère (3.17) conduit à un problème de programmation quadratique qui, lorsque la dimension N du problème est grande, est difficile à résoudre. Cependant, des approches ont été développées pour sa résolution, la première a été proposée par Tibshirani dans [128]. Cette approche est basée sur une reformulation lisse du problème en introduisant $4N$ variables auxiliaires avec des contraintes linéaires et $4N$ contraintes de non-négativité. La résolution du problème reformulé est obtenue en utilisant l'algorithme SQOPT [4] [58]. Une deuxième approche pour résoudre le même problème dans le cadre de la régression logistique a été proposée dans [2]. Cette technique consiste à introduire $2N$ variables auxiliaires et $4N$ contraintes d'inégalité puis à résoudre le problème reformulé par le package d'optimisation CVX [5].

Néanmoins, ces méthodes basées sur la reformulation du problème original ne sont plus adaptées lorsque la dimension N est grande à cause du grand nombre de variables auxiliaires et des contraintes supplémentaires introduites. En effet, il est souligné dans [128] que la vitesse de convergence est un vrai inconvénient dans la résolution du *fused sparse Lasso* lorsque $N>2000$ et que le nombre d'échantillons $M>200$. Plus récemment, une approche basée sur la recherche du chemin de régularisation a été développée dans [68] pour le problème *Fused Lasso Signal Approximator* (FLSA), dans lequel $\Phi=I$. Cet

4. tomopt.com/tomlab/products/snopt/solvers/SQOPT.php.
5. standard.edu/ boyd/cvx.

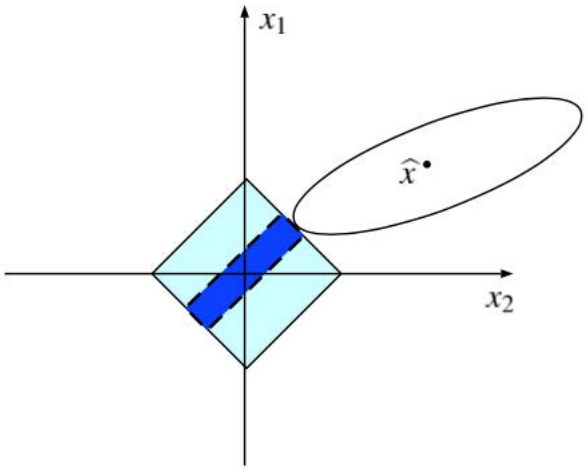


FIGURE 3.7 – Illustration géométrique de la contrainte *fused sparse Lasso*. L'ellipse correspond à une courbe de niveau de la fonction d'erreur qui satisfait la contrainte de parcimonie (losange) et la contrainte de parcimonie sur les éléments successifs (rectangle).

algorithme a aussi donné lieu à d'autres améliorations d'un point de vue temps de calcul, notamment l'algorithme *Efficient Fused Lasso Algorithm* (EFLA) [85] dont le taux de convergence est en $\mathcal{O}(1/k^2)$.

Méthodes proximales

Les algorithmes basés sur les méthodes de descente de gradient ont été utilisés plus récemment pour la résolution des problèmes du type (3.17). En effet, les méthodes proximales sont des méthodes conçues pour résoudre des critères composés d'une fonction différentiable quelconque f ayant un gradient Lipschitz continu, et une fonction non différentiable g. Ces approches ont l'avantage de pouvoir résoudre une large gamme de problèmes convexes non lisses. Soit $H(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ où g est une fonction continue convexe, et f une fonction convexe et différentiable avec un gradient Liptschitz continu $L(f)$:

$$||\nabla f(\mathbf{x}) - \nabla f(\mathbf{z})|| \leqslant L(f)||\mathbf{x} - \mathbf{z}|| \; \forall \, \mathbf{x}, \mathbf{z} \in \mathbb{R}^N. \tag{3.19}$$

On considère alors l'approximation quadratique de $H(\mathbf{x})$ au point $\mathbf{z}$ pour tout $L > 0$ comme étant :

$$A_L(\mathbf{x}, \mathbf{z}) := f(\mathbf{z}) + < \mathbf{x} - \mathbf{z}, \nabla f(\mathbf{z}) > + \frac{L}{2}||\mathbf{x} - \mathbf{z}||_2^2 + g(\mathbf{x}), \tag{3.20}$$

Ceci consiste à linéariser $f(\mathbf{x})$ autour du point $\mathbf{z}$. L'approche proximale pour l'estimation de $\mathbf{x}$ à l'itération $k+1$ consiste à trouver la solution du problème suivant :

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}_{(k)}) + \nabla f(\mathbf{x}_{(k)})'(\mathbf{x} - \mathbf{x}_{(k)}) + g(\mathbf{x}) + \frac{L}{2}||\mathbf{x} - \mathbf{x}_{(k)}||_2^2 \right\}, \tag{3.21}$$

où $L > 0$ correspond à la borne supérieure de la constante de Lipschitz de ∇f. Le terme proximal correspond au terme quadratique qui permet la mise à jour dans le voisinage de $\mathbf{x}_{(k)}$, où f est proche de son

approximation linéaire. Le problème (3.21) est réécrit comme suit :

$$\min_{\mathbf{x}} \left\{ \frac{1}{2} \left|\left| \mathbf{x} - \left(\mathbf{x}_{(k)} - \frac{1}{L} \nabla f(\mathbf{x}_{(k)}) \right) \right|\right|_2^2 + \frac{1}{L} g(\mathbf{x}) \right\}. \tag{3.22}$$

Par ailleurs, on note que lorsque le critère ne contient pas le terme non lisse g, la solution au problème (3.21) est donnée par la règle de mise à jour du gradient standard : $\mathbf{x}_{(k+1)} \leftarrow \mathbf{x}_{(k)} - \frac{1}{L}\nabla f(\mathbf{x}_{(k)})$.

Algorithme FISTA

Une technique rapide pour résoudre le problème (3.22) consiste à utiliser l'algorithme FISTA développé dans [6]. Cet algorithme a été conçu afin de résoudre des problèmes d'optimisation intégrant des termes non lisses. De plus, FISTA est une amélioration de l'algorithme ISTA développé par Nesterov dans [105], dans la mesure où il a un taux de convergence en $\mathcal{O}(\frac{1}{k^2})$ contrairement à ISTA dont le taux de convergence est en $\mathcal{O}(\frac{1}{k})$. Soit :

$$J(\mathbf{z}) = \left\{ g(\mathbf{x}) + \frac{L}{2} \left|\left| \mathbf{x} - \left(\mathbf{z} - \frac{1}{L} \nabla f(\mathbf{z}) \right) \right|\right|^2 \right\}. \tag{3.23}$$

La mise à jour de $\mathbf{x}$ à l'itération k est basée sur le calcul de l'opérateur proximal à l'étape $k-1$:

$$\mathbf{x}_{(k)} = \arg\min_{\mathbf{x}} \frac{1}{2} ||\mathbf{x} - \mathbf{v}_{(k-1)}||_2^2 + \frac{1}{L} g(\mathbf{x}), \tag{3.24}$$

avec :

$$\mathbf{v}_{(k-1)} = \mathbf{x}_{(k-1)} - \frac{1}{L} \nabla f(\mathbf{x}_{(k-1)}), \tag{3.25}$$

où f est le terme d'attache au donnés, $f(\mathbf{x}) = \frac{1}{2}||\mathbf{y} - \Phi\mathbf{x}||_2^2$. Contrairement à ISTA, l'opérateur de seuillage itératif dans l'algorithme FISTA n'est pas employé au point précédent $\mathbf{x}_{k-1}$, mais au point $\mathbf{y}_{(k)}$ qui est une combinaison linéaire spécifique des deux points précédents $\mathbf{x}_{k-1}$ et $\mathbf{x}_{k-2}$. On résume dans la Figure 3.8 les étapes de la procédure de mise à jour de $\mathbf{x}$.

FIGURE 3.8 – Procédure FISTA

- **Entrée** : L la constante de Liptschitz de ∇f, maxiter.
- **Initialisation** : $\mathbf{x}_0 = 0$, $t_1 = 1$, $\mathbf{y}_1 = \mathbf{x}_0$
- **Pour** : $k \leftarrow 1, \cdots,$ maxiter **faire** :
  1. $\mathbf{x}_{(k)} = \arg\min_{\mathbf{x}} J(\mathbf{y}_{(k)})$ Résolution par FLSA
  2. $t_{(k+1)} = \frac{1+\sqrt{1+4t_{(k)}^2}}{2}$
  3. $\mathbf{y}_{(k+1)} = \mathbf{x}_{(k)} + \left(\frac{t_{(k)}-1}{t_{(k+1)}}\right)(\mathbf{x}_{(k)} - \mathbf{x}_{(k-1)})$

L'idée d'utiliser l'opérateur proximal pour la résolution du *fused sparse Lasso* permet de simplifier

significativement les calculs. De plus, l'algorithme FLSA du package SLEP [6] est une brique dans l'algorithme FISTA permettant de résoudre le problème de la ligne 1 (estimation de $\mathbf{x}$). On montrera dans la suite qu'on peut utiliser FISTA de la même manière pour optimiser d'autres critères en modifiant simplement l'expression de l'opérateur proximal.

3.2.5 Smooth sparse Lasso

Le *smooth sparse Lasso* consiste également à combiner une pénalité *Lasso* avec un terme de fusion entre les coefficients successifs exprimé par la norme ℓ_2. Le critère à minimiser s'écrit comme suit :

$$\min_{\mathbf{x}} \frac{1}{2}||\mathbf{y}-\Phi\mathbf{x}||_2^2+\lambda_1||\mathbf{x}||_1+\frac{\lambda_2}{2}\sum_{i=1}^{N-1}(x^i-x^{i+1})^2, \tag{3.26}$$

où $\lambda_1, \lambda_2 \geq 0$ sont les paramètres qui contrôlent respectivement la parcimonie et la régularisation de la solution (*smoothness*). Cette pénalité en norme ℓ_2 a d'abord été introduite dans [65]. L'algorithme proposé se base sur une modification de l'algorithme *LARS* et les résultats théoriques relatifs aux performances de cette approche dans le cadre de la sélection de variables y sont également établis. Ici, on propose de résoudre ce problème en utilisant l'opérateur proximal et l'algorithme FISTA. D'abord on commence par regrouper les deux termes exprimés en norme 2 de l'équation (3.26) :

$$\min_{\mathbf{x}} \tfrac{1}{2}||\mathbf{y}-\Phi\mathbf{x}||_2^2+\lambda_2\sum_{i=1}^{N}(x^i-x^{i+1})^2+\lambda_1||\mathbf{x}||_1. \tag{3.27}$$

On pose

$$f(\mathbf{x})=||\mathbf{y}-\Phi\mathbf{x}||_2^2+\frac{\lambda_2}{2}\sum_{i=1}^{N}(x^i-x^{i+1})^2. \tag{3.28}$$

De la même façon que le *fused sparse Lasso*, on calcule l'estimation de $\mathbf{x}$ à l'itération k en minimisant l'équation suivante :

$$\mathbf{x}_k=\arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}_{(k-1)}||_2^2+\frac{\lambda_1}{L}||\mathbf{x}||_1, \tag{3.29}$$

où $\mathbf{v}_{(k-1)}=\mathbf{x}_{(k-1)}-\frac{1}{L}\nabla f(\mathbf{x}_{(k-1)})$, L est la constante de Lipschitz, et $\nabla f(\mathbf{x})$ est le gradient de $f(\mathbf{x})$:

$$\nabla f(\mathbf{x})=-\Phi'(\mathbf{y}-\Phi\mathbf{x})+\lambda_2\mathbf{D}'\mathbf{D}\mathbf{x}. \tag{3.30}$$

On obtient finalement :

$$\mathbf{v}_{(k)}=\mathbf{x}_{(k)}-\frac{1}{L}(\Phi'\Phi+\lambda_2\mathbf{D}'\mathbf{D})\mathbf{x}_{(k)}+\frac{1}{L}\Phi'\mathbf{y}. \tag{3.31}$$

L'algorithme FISTA avec $g(\mathbf{x})=\frac{\lambda_1}{L}||\mathbf{x}||_1$ est utilisé pour calculer itérativement $\mathbf{x}_{(k)}$. La constante de Lipschitz de la fonction $f(\mathbf{x})$ correspond à la plus grande valeur propre de la matrice $\mathbf{M}=(\Phi'\Phi+\lambda_2\mathbf{D}'\mathbf{D})$ qui est calculée en utilisant la méthode de la puissance.

6. http ://yelab.net/software/SLEP/

3.2.6 Group sparse Lasso

Le *group sparse Lasso* [157] est une autre extension de l'algorithme *Lasso* permettant de sélectionner simultanément des groupes de variables. Cet algorithme a été beaucoup exploité dans le cadre de la régression et la classification, où un groupe d'individus peut avoir des attributs communs qu'il ne partage pas avec d'autres groupes [157, 107, 44, 94]. Dans le *group sparse Lasso*, les variables sont regroupées dans des ensembles. L'idée est d'activer ou de désactiver toutes les variables d'un ensemble simultanément. Ainsi, dans le critère du *Lasso* standard, on inclut une pénalité de groupe comme suit :

$$\min_{\mathbf{x}} \tfrac{1}{2}||\mathbf{y}-\Phi\mathbf{x}||_2^2+\lambda_1||\mathbf{x}||_1+\lambda_2\sum_{i=1}^{s}||\mathbf{x}_{G_i}||, \tag{3.32}$$

où s est le nombre de groupes, G_i est l'ensemble des indices du groupe i tel que $|G_i|=n_i$ et $\mathbf{x}_{G_i}$ est le vecteur des coefficients associé au groupe i. On suppose que les groupes sont disjoints : $G_i \cap G_j = \emptyset, \forall i \neq j$. Par ailleurs, le *group sparse Lasso* peut être considéré comme une relaxation convexe de l'approximation parcimonieuse simultanée en considérant les groupes comme étant les lignes activées (non nulles) de la matrice parcimonieuse. De la même façon que précédemment, on peut utiliser l'algorithme FISTA pour la résolution de ce problème. La pénalité de groupe est résolue, en utilisant la régularisation de Moreau-Yosida décrite dans [84] pour l'apprentissage de structures d'arbre groupé (*grouped tree structure learning*). Dans cette approche, la régularisation est prise en compte à travers tous les noeuds de la structure en graphe. Ainsi, la contrainte de groupe dans le *group Lasso* peut être considérée comme un cas particulier de ces structures, où il n'existe qu'un seul niveau dans l'arbre correspondant à la racine.

Ainsi, l'opérateur proximal correspondant au *group sparse Lasso* est calculé en appliquant d'abord l'opérateur proximal associé à la norme ℓ_1 (seuillage doux) puis celui associé à la norme mixte ℓ_1/ℓ_2 (régularisation Moreau-Yosida). On présente sur la Figure 3.9 le résumé de la procédure.

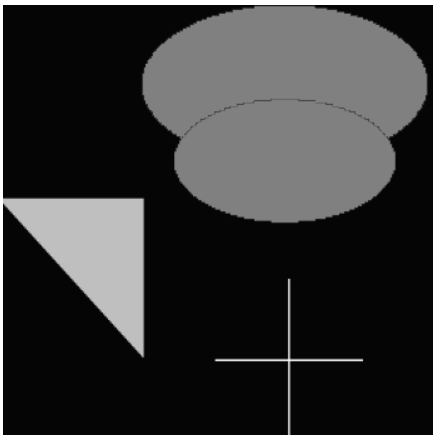
FIGURE 3.9 – Algorithme *group Lasso* : (GL)

— **Entrée** : $\mathbf{y}$, Φ, λ_1,λ_2 maxiter.
— **Pour** : $k \leftarrow 1,\cdots,$ maxiter **faire** :
1. $\mathbf{v}_{(k)} = \mathbf{x}_{(k)} - \frac{1}{L}\Phi'(\Phi\mathbf{x}_{(k)}-\mathbf{y})$.
2. $\mathbf{w}_{k+1} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}_{(k)}||_2^2+\frac{\lambda_1}{L}||\mathbf{x}||_1$. Seuillage doux
3. $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{w}_{(k)}||_2^2+\frac{\lambda_2}{L}\sum_{i=1}^{s}||\mathbf{x}_{G_i}||$. Régularisation Moreau-Yosida

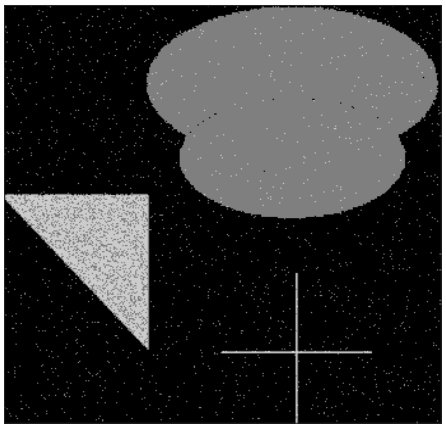
— **Sortie** : $\mathbf{x}$

3.2.7 Exemples de dé-bruitage d'image

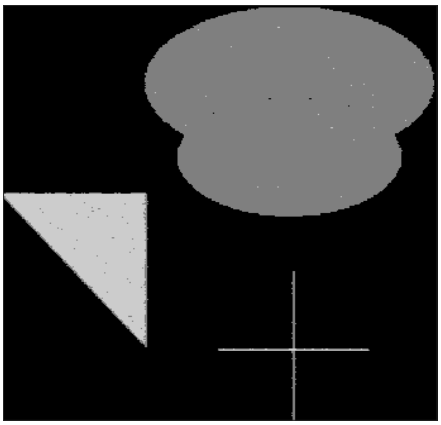
Nous nous proposons d'illustrer ici l'intérêt des critères *fused sparse Lasso* et *smooth sparse Lasso*, dans le cadre du dé-bruitage d'images. Pour ce faire, on considère deux images en niveaux de gris, la première, $\mathbf{I}_1$, représentant plusieurs formes géométriques de taille 344×344. La deuxième image, notée $\mathbf{I}_2$, est une image de deux échantillons de bois de notre base de données de taille 111×111. Ces deux images sont ensuite perturbées par un bruit Gaussien avec un RSB= 6 dB (Figures 3.10b et 3.11b). Le



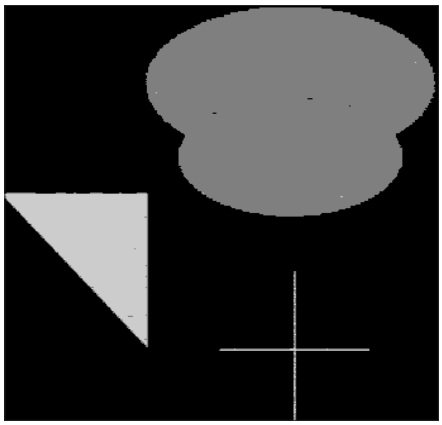
(a) Image de test $\mathbf{I}_1$ en niveaux de gris.



(b) Image $\mathbf{I}_1$ après l'ajout du bruit.



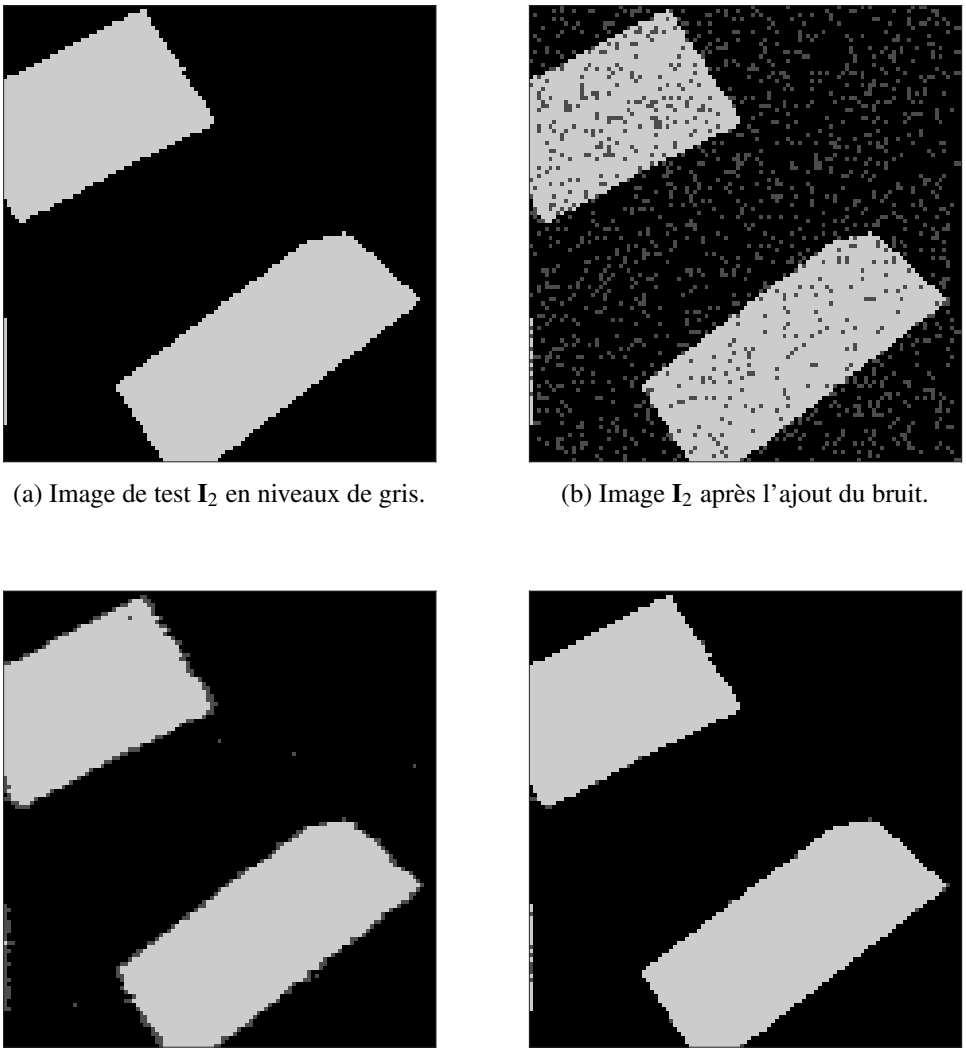
(c) *smooth sparse Lasso*



(d) *fused sparse Lasso*

FIGURE 3.10 – Dé-bruitage de l'image test $\mathbf{I}_1$.

dictionnaire Φ est une matrice identité de la même taille que les deux images. Les paramètres λ_1 et λ_2 sont réglés de façon à obtenir le meilleur résultat de dé-bruitage (en préservant les formes et les contours des formes présentes dans les images) pour chaque méthode. Les résultats du dé-bruitage des deux images, présentés respectivement sur les Figures 3.10 et 3.11 montrent bien l'intérêt de la régularisation pour la réduction du bruit. En ce qui concerne la première image nous constatons que les meilleurs résultats sont obtenus par *fused sparse Lasso*. Ces résultats sont confirmés par le deuxième exemple avec l'image $\mathbf{I}_2$ puisque le meilleur résultat est obtenu avec *fused sparse Lasso*. Ceci peut s'expliquer par l'action de la pénalité de fusion qui permet d'obtenir une image dont les lignes sont constantes par morceaux. Les critères basés sur la norme 2 sont moins efficaces et n'arrivent pas à préserver les contours. Par ailleurs, la régularisation en norme 1 sur les différences successives dans le *fused sparse Lasso* peut être étendue également aux colonnes de l'image. Ce type de régularisation est connue sous le nom de la variation totale [5] et [27]. On présente dans le Tableau 3.1 les erreurs de reconstruction de chaque méthode exprimées par la norme de Frobenius du résidu $\mathbf{R}$ entre l'image réelle et son approximation.



(a) Image de test $\mathbf{I}_2$ en niveaux de gris.

(b) Image $\mathbf{I}_2$ après l'ajout du bruit.

(c) *smooth sparse Lasso*

(d) *fused sparse Lasso*

FIGURE 3.11 – Dé-bruitage de l'image test $\mathbf{I}_2$.

Nous avons présenté jusqu'ici plusieurs approches parcimonieuses pour l'estimation d'un vecteur avec une seule mesure $\mathbf{y}$. Dans la suite, nous proposons d'étendre quelques méthodes gloutonnes pour l'approximation simultanée d'un signal à mesures multiples MMV (Multiple Measurement Vector).

$\|\|\mathbf{R}\|\|_F^2$ (dB)	*smooth sparse Lasso*	*fused sparse Lasso*
Image test 1	12.12	10.73
Image test 2	-1.57	-4.48

TABLE 3.1 – Erreur de reconstruction calculée en dB entre l'image originale et l'image dé-bruitée obtenues avec les deux méthodes.

3.3 Les approches gloutonnes pour l'approximation parcimonieuse simultanée

À la différence de l'approximation parcimonieuse standard, la version simultanée consiste à reconstruire une matrice d'observations $\mathbf{Y}$, dont les signaux élémentaires $\mathbf{y}_i$ partagent la même structure de parcimonie. Cette formulation a été exploitée pour résoudre des problèmes dans des applications telles que la localisation de source des signaux EEG/MEG [152], les problèmes d'estimation de direction d'arrivée (DOA) [32, 140] ou encore le dé-mélange hyperspectral [74]. La première approche pour l'approximation parcimonieuse simultanée a été proposée par Cotter et al. [33] qui ont développé les versions simultanées des algorithmes Matching Pursuit (MP) et FOCUSS. Ils ont également démontré que l'approximation parcimonieuse simultanée permet d'obtenir des taux de bonne reconstruction nettement supérieurs à ceux obtenus avec la version standard. Comme cela a été souligné par Tropp et al. [133], les approches précédentes correspondent à une relaxation convexe du problème d'approximation parcimonieuse simultanée exacte qui consiste à trouver la solution $\mathbf{X}$ ayant le minimum de lignes non nulles.

Formulation du problème

Le problème d'approximation parcimonieuse simultanée consiste à trouver la représentation la plus parcimonieuse du signal $\mathbf{Y}$:

$$\min_{\mathbf{X}} ||\mathbf{X}||_{0,2} \qquad \text{sous contrainte} \qquad \mathbf{\Phi X} = \mathbf{Y}, \tag{3.33}$$

où $\mathbf{Y} \in M \times K$, et $||\mathbf{X}||_{0,2}$ est la norme mixte ℓ_0/ℓ_2 de $\mathbf{X} \in N \times K$ qui correspond au nombre de lignes ayant une norme ℓ_2 non nulle. Une deuxième façon d'écrire le problème consiste à relaxer la contrainte d'égalité stricte en introduisant une tolérance d'erreur ε, dans le cas où le signal est corrompu par le bruit :

$$\min_{\mathbf{X}} ||\mathbf{X}||_{0,2} \qquad \text{sous contrainte} \qquad ||\mathbf{\Phi X} - \mathbf{Y}||_F^2 \leq \varepsilon, \tag{3.34}$$

où $||.||_F$ est la norme de Frobenius. Enfin, ce problème peut être formulé pour avoir une erreur d'approximation minimale pour un niveau de parcimonie s donné :

$$\min_{\mathbf{X}} \tfrac{1}{2}||\mathbf{Y}-\Phi\mathbf{X}||_F^2 \qquad \text{sous contrainte} \qquad ||\mathbf{X}||_{0,2} \leq s. \tag{3.35}$$

Le paramètre s contrôle le niveau de parcimonie de la solution. Ainsi, le problème d'approximation parcimonieuse simultanée consiste à estimer la matrice $\mathbf{X}$ à partir de $\mathbf{Y}$ et le dictionnaire Φ, sous l'hypothèse que tous les signaux élémentaires de $\mathbf{Y}$ ont le même profil de parcimonie. Nous définissons le support des lignes de la matrice $\mathbf{X}$, comme l'ensemble des indices de ses lignes non-nulles :

$$\text{rowsupp}(\mathbf{X}) = \{1 \leq i \leq N \,|\, \mathbf{x}^i \neq 0\} = \Gamma, \tag{3.36}$$

où $\mathbf{x}^i$ représente la i-ème ligne de $\mathbf{X}$. Ainsi, $||\mathbf{X}||_{0,2} = |\text{rowsupp}(\mathbf{X})|$.

3.3.1 Simultaneous Orthogonal Matching Pursuit

Une des premières méthodes gloutonnes qui a été développée pour l'approximation parcimonieuse simultanée est le S-OMP [134]. À l'itération k, la sélection d'un nouvel atome est basée sur le calcul de la corrélation maximale entre le dictionnaire est la matrice des résidus, $\Phi'\mathbf{R}_{k-1}$.

L'algorithme s'arrête lorsque :

— Le nombre maximal d'itérations est atteint : $k = s$;

— La norme de Frobenius du résidu est en dessous d'un certain seuil : $||\mathbf{R}_k|| \leqslant \varepsilon$;

— La corrélation maximale entre un atome et le résidu passe en dessous d'un seuil τ : $||\Phi'\mathbf{R}_k||_\infty \leqslant \tau$.

Cet algorithme est présenté dans la Figure 3.12.

FIGURE 3.12 – Simultaneous OMP

— **Entrée** : $\mathbf{Y}$, Φ ; s ; un critère d'arrêt.

— **Initialisation** : $\mathbf{R}_0 = \mathbf{Y}$, $\Gamma_0 = \emptyset$, $\mathbf{X}_0 = 0$.

— **Pour** : $k \leftarrow 1$ jusqu'à ce que le critère d'arrêt soit satisfait **faire** :

1. Trouver l'index n_k de Φ le plus corrélé avec le résidu : $n_k \in \arg\max_n ||\phi_n' \mathbf{R}_{k-1}||_2$
2. $\Gamma_k = \Gamma_{k-1} \cup n_k$
3. $\mathbf{X}_k = \arg\min_{\mathbf{U}} ||\mathbf{Y} - \Phi_{\Gamma_k}\mathbf{U}||_F^2$. Estimation des coefficients
4. $\mathbf{R}_k = \mathbf{Y} - \Phi_{\Gamma_k}\mathbf{X}_k$. Mise à jour du résidu

— **Sortie** : Γ et $\mathbf{X}$ et $\mathbf{R}$.

3.3.2 Simultaneous CoSaMP

L'algorithme CoSaMP [104] est également basé sur une recherche gloutonne tout comme OMP. Nous proposons une extension de cette approche pour résoudre le problème d'approximation parcimonieuse simultanée. L'algorithme résultant est appelé simultaneous CoSaMP (S-CoSaMP). Cet algorithme

a besoin, en entrée, d'un paramètre de parcimonie s, un paramètre de réglage μ et un critère arrêt. L'algorithme est initialisé avec une approximation triviale à savoir que le résidu initial est égal au signal d'entrée. On présente dans la Figure 3.13 les étape détaillées de S-CoSaMP.

FIGURE 3.13 – Simultaneous CoSaMP

— **Entrée** : $\mathbf{Y}$, Φ ; s ; un critère d'arrêt, paramètre de réglage $\mu < 1$.
— **Initialisation** : $\mathbf{R}_0 = \mathbf{Y}$, $\Gamma_0 = \emptyset$, $\mathbf{X}_0 = 0$.
— **Pour** : $k \leftarrow 1$ jusqu'à ce que le critère d'arrêt soit satisfait **faire** :

1. Trouver les μs colonnes de Φ les plus corrélées avec le résidu : $\Lambda \in \arg\max_{|T| \leq \mu s} \sum_{n \in T} ||\phi_n' \mathbf{R}_{k-1}||_F^2$. Identification
2. $\mathbf{U}_k = \arg\min_{\mathbf{U}} ||\mathbf{Y} - \Phi_{\Gamma_{k-1} \cup \Lambda} \mathbf{U}||_F$. Estimation
3. $\mathbf{X}_k = [\mathbf{U}_k]_s, \Gamma_k = \text{rowsupp}(\mathbf{X}_k)$ Élagation
4. $\mathbf{R}_k = \mathbf{Y} - \Phi_{\Gamma_k} \mathbf{X}_k$ Mise à jour du résidu

— **Sortie** : Γ et $\mathbf{X}$ et $\mathbf{R}$.

3.3.3 Simultaneous OLS

La version simultanée de OLS est similaire à la version standard. L'idée est d'utiliser la norme de Frobenius de la matrice de corrélation dans l'étape de sélection. La Figure 3.14 résume les étapes de cet algorithme.

FIGURE 3.14 – Simultaneous OLS

— **Entrée** : $\mathbf{Y}$, Φ ; s ; un critère d'arrêt.
—
— **Initialisation** : $\mathbf{R}_0 = \mathbf{Y}$, $\Gamma_0 = \emptyset$, $\mathbf{X}_0 = 0$.
— **Pour** : $k \leftarrow 1$ jusqu'à ce que le critère d'arrêt soit satisfait **faire** :

1. $n_{\max} = \arg\min_{\Gamma_k = \Gamma_{k-1} \cup n} \|\mathbf{Y} - \Phi_{\Gamma_k} \Phi_{\Gamma_k}^{\dagger} \mathbf{Y}\|_F$.
2. $\Gamma_k = \Gamma_{k-1} \cup n_{\max}$
3. $\mathbf{X}_{\Gamma_k} = \Phi_{\Gamma_k}^{\dagger} \mathbf{Y}$.
4. $\mathbf{R}_k = \mathbf{Y} - \Phi_{\Gamma_k} \mathbf{X}_{\Gamma_k}$

— **Sortie** : Γ et $\mathbf{X}$ et $\mathbf{R}$.

3.3.4 Simultaneous Single Best Replacement

Dans la version simultanée de l'algorithme S-SBR on cherche à minimiser la fonction de coût suivante :

$$\arg\min_{\mathbf{X} \in \mathbb{C}^{N \times K}} \{\mathscr{J}(\mathbf{X}, \lambda) = \mathscr{E}(\mathbf{X}) + \lambda ||\mathbf{X}||_{0,2}\} \tag{3.37}$$

où $\mathscr{E}(\mathbf{X}) = ||\mathbf{Y} - \Phi\mathbf{X}||_F^2$ et λ est un hyper-paramètre qui permet de contrôler le niveau de parcimonie de la solution. Dans cette version, notée S-SBR, la norme 2 dans $\mathscr{E}$ est remplacée par la norme de Frobenius. Les étapes de l'algorithme sont présentées dans la Figure 3.15.

FIGURE 3.15 – Simultaneous SBR

— **Entrée** : $\mathbf{Y}$, Φ; paramètre de régularisation λ.
— **Initialisation** : $\mathbf{R}_0 = \mathbf{Y}$, $\Gamma_0 = \emptyset$, $\mathbf{X}_0 = 0$.
— **Pour** $k \leftarrow 1$ jusqu'à ce que l'ensemble Γ_k n'est plus mis à jour **faire** :

1. $n_k \in \arg_n \min\{\mathscr{I}_{\Gamma_{k-1}\bullet n}(\lambda) = \mathscr{E}_{\Gamma_{k-1}\bullet n} + \lambda \text{Card}(\Gamma_{k-1} \bullet n)\}$.
2. **Si :** $\mathscr{I}_{\Gamma_{k-1}\bullet n_k}(\lambda) < \mathscr{I}_{\Gamma_{k-1}}(\lambda)$,
3. $\Gamma_k = \Gamma_{k-1} \bullet n_k$
   Fin si
4. $\mathbf{X}_{\Gamma_k} = \Phi_{\Gamma_k}^{\dagger}\mathbf{Y}$.

— **Sortie** : Γ et $\mathbf{X}$.

3.4 Comparaison des approches gloutonnes

3.4.1 Description

Afin d'évaluer les performances des algorithmes présentés, nous nous proposons de réaliser une simulation similaire à celle utilisée dans l'analyse de mélange spectroscopique. L'objectif est de comparer les taux de détection des composantes d'un signal obtenus avec les différentes méthodes présentées. Ainsi, nous proposons de tester les trois algorithmes développés pour résoudre le problème d'approximation parcimonieuse simultanée sur 20 vecteurs d'observations à différents niveaux de rapports de signal sur bruit (RSB) et d'évaluer les performances de chacun en comparaison avec S-OMP. La matrice de signal $\mathbf{Y} \in \mathbb{R}^{20\times 20}$ est construite comme suit : chaque colonne de $\mathbf{Y}$ est la somme de 3 fonctions gaussiennes ayant chacune un écart-type de 0.5. L'évolution de l'amplitude de chacune des gaussiennes est décrite par une fonction sinusoïdale de fréquence aléatoire et d'amplitude unitaire. Le centre des gaussiennes est fixé à $(8-d, 8, 8+2d)$, où $d \in \{0.66, 1, \ldots, 5\}$. En fait, la variable d contrôle la corrélation entre les composantes du signal. Le résultat est représenté dans la Figure 3.16 pour 3 valeurs de d. Ensuite, la matrice $\mathbf{Y}$ est perturbée par un bruit blanc additif $\mathbf{E}$, tel que le RSB défini par :

$$\text{RSB} = \frac{\|\Phi\mathbf{X}\|_F^2}{\|\mathbf{E}\|_F^2} \tag{3.38}$$

varie entre -20 dB et 20 dB. Le dictionnaire sur-dimensionné $\Phi \in \mathbb{R}^{20\times 60}$ est composé de 60 fonctions gaussiennes ayant le même écart-type (0.5) et dont les centres couvrent l'intervalle $[0, 20]$ avec un pas de 0.33. Par conséquent, le coefficient de corrélation entre deux atomes consécutifs (cohérence mutuelle) est de 0.93, ce qui signifie que le dictionnaire n'est pas orthogonal. Les paramètres de réglage sont : $s = 6$, $\mu = 0.5$ et $\lambda = 10^{-5}$. Ici, l'objectif n'est pas d'évaluer les performances des approches proposées dans le pire des cas (à l'aide de la RIP par exemple), mais de localiser les régions de reconstruction exacte,

définie par d et le RSB pour chaque méthode. Nous utilisons donc le taux de détection, lié au nombre de composantes du signal correctement identifiées, comme mesure de reconstruction exacte.

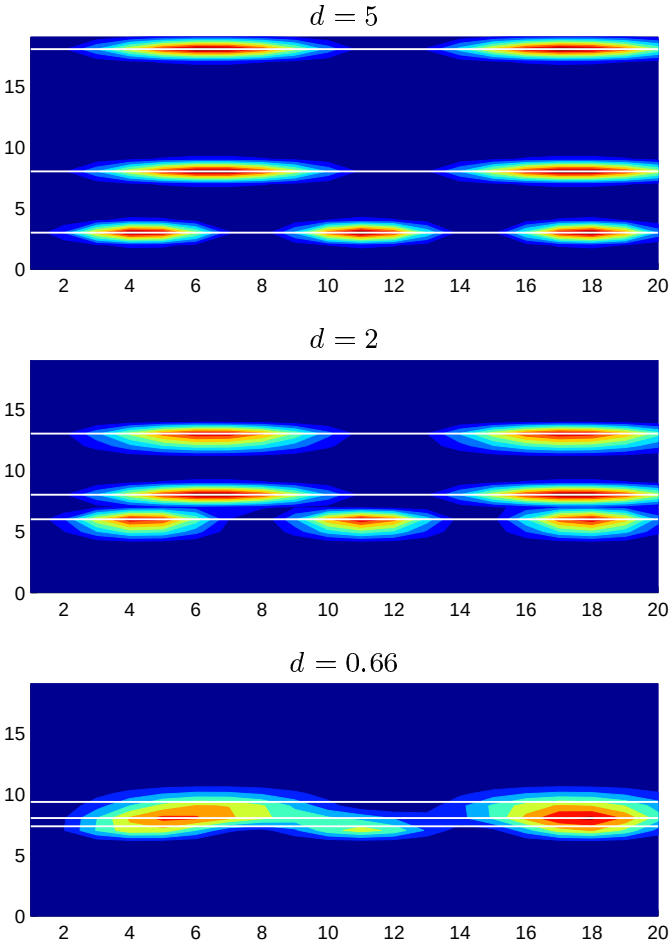


FIGURE 3.16 – Matrice de données $\mathbf{Y}$ pour $d = 5$, $d = 2$, and $d = 0.66$. Le signal est composé de 3 Gaussiennes avec des amplitudes sinusoidales le long de l'axe x.

3.4.2 Résultats des simulations

Les résultats présentés sur la Figure 3.17 montrent que toutes les méthodes réussissent à reconstruire correctement le signal pour un fort RSB et un d élevé. La reconstruction exacte (100% de taux de bonne détection) est représentée par la région rouge. On constate d'abord que la surface de cette region est plus petite dans le cas du S-OMP, en comparaison avec les autres méthodes. En effet, l'algorithme S-OLS produit de meilleurs résultats que le S-OMP, surtout pour un RSB moyen. Néanmoins, les performances du S-OLS diminuent fortement pour un faible RSB et ce, même lorsque la corrélation entre les composantes du signal est faible (d élevé). Par ailleurs, S-CoSaMP réussit à atteindre de meilleurs taux de détection que S-OMP et S-OLS pour un d élevé et un faible RSB. Toutefois, ces trois méthodes n'ont réussi à retrouver aucune composante du signal lorsque ce dernier est fortement corrompus par le bruit. Ceci est

représenté par la région blanche dans la Figure 3.17. En revanche, S-SBR a un meilleur comportement et produit de meilleurs résultats de détection même à faible SNR et/ ou d. Ainsi, nous constatons que les performances de S-SBR et S-CoSaMP sont meilleures que S-OLS et S-OMP.

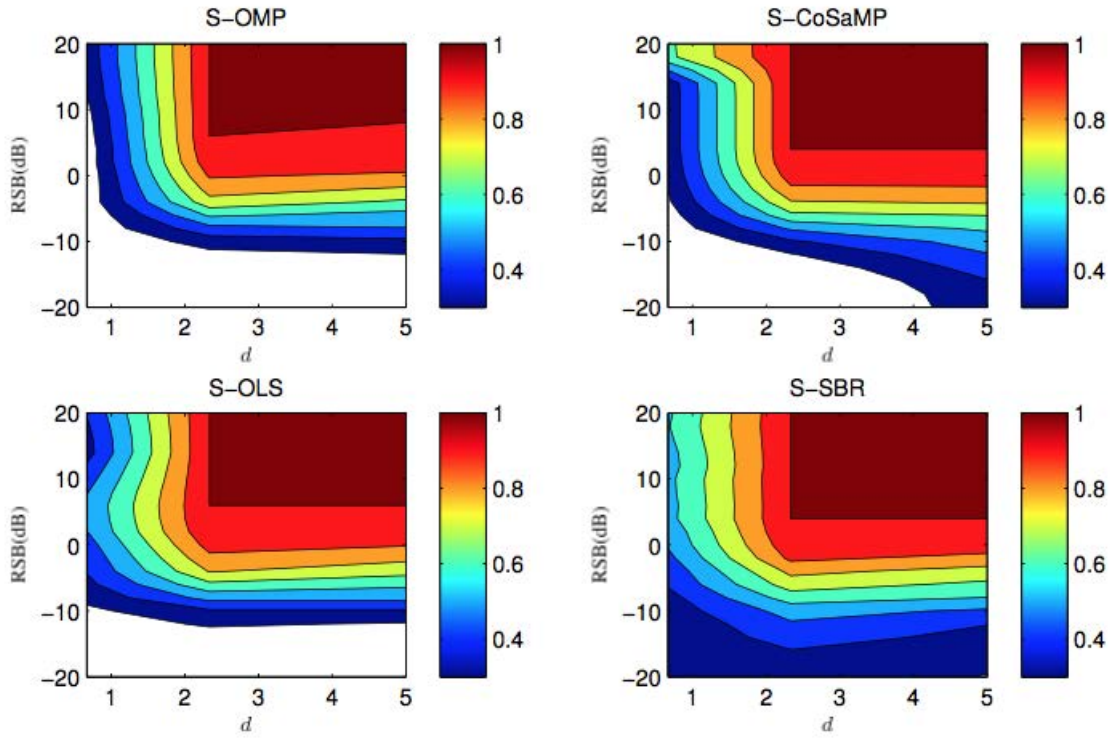
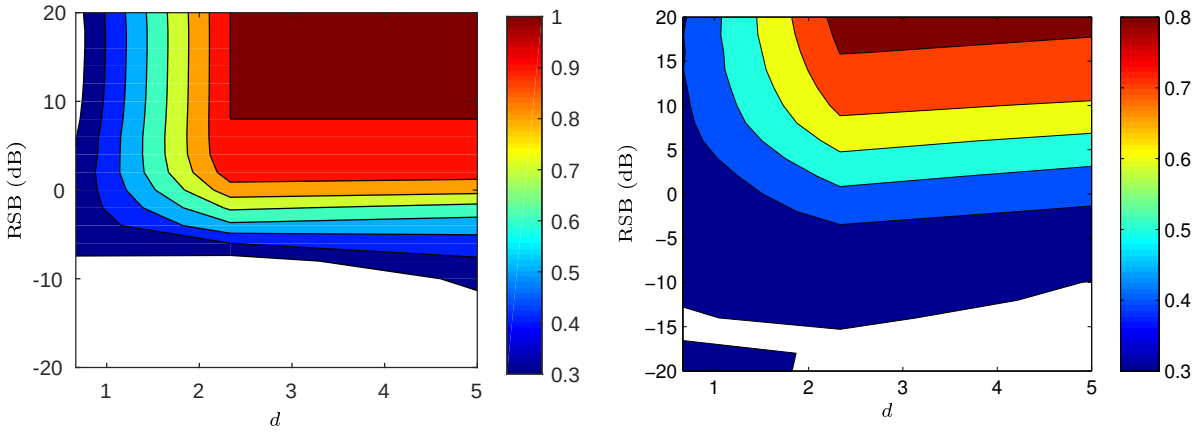


FIGURE 3.17 – Taux de détection des approches gloutonnes pour l'approximation parcimonieuse simultanée en fonction du RSB et de d pour 100 simulations Monte Carlo.

Nous présentons également sur la Figure 3.18 (a) les résultats obtenus avec la méthode *sparse group Lasso*. Ici, les paramètres de parcimonie et de groupe sont fixés à $\lambda_1 = \lambda_2 = 0.5$. Ces paramètres ont été réglés de façon à obtenir exactement 3 atomes dans le cas idéal (RSB=20 dB et $d = 5$). Néanmoins, l'inconvénient de cette méthode est de ne pas permettre le contrôle de la taille de l'ensemble des atomes actifs. En effet, la taille de cet ensemble tend à augmenter au fur et à mesure que le RSB et le niveau de la corrélation diminuent. Afin de maintenir le nombre des atomes actifs à trois, nous avons forcé l'algorithme à retourner à chaque fois les trois éléments les plus grands pour pouvoir comparer correctement les résultats avec ceux obtenus en utilisant les approches non-convexes. Ainsi, nous constatons que les performances du *group Lasso* sont similaires à celles du S-OLS. Afin de montrer l'avantage de l'approximation simultanée, nous présentons dans la Figure 3.18 (b) les résultats obtenus pour la même simulation en utilisant le SBR standard sur chaque colonne de $\mathbf{Y}$ indépendamment, et de façon successive. Nous pouvons observer que les résultats obtenus par la version standard sont moins bons que la version simultanée. En effet, cette version n'atteint jamais une taux de détection de 100%. Le meilleur résultat obtenu par cette méthode et la détection de 80% des composantes du signal. De plus la région correspondant à ce résultats est très restreinte, ce qui implique que cette approche est très sensible aux



(a) *group Lasso* (b) SBR

FIGURE 3.18 – Taux de bonne détection en fonction du RSB et d pour 100 simulations Monte Carlo. (a) : *group Lasso*, (b) : SBR.

niveaux de bruit et de corrélation entre les composantes du signal. Ceci est confirmé par les résultats obtenus pour cette méthode lorsque le RSB est faible et la corrélation forte entre les composantes du signal puisque SBR échoue complètement dans la reconstruction d'une seule composante du signal. L'évolution de la norme de Frobenius des résidus est représentée sur la Figure 3.19 dans le cas suivant : RSB$=10$ et $d = 2.33$. Il apparaît que les algorithmes S-OMP, S-CoSaMP et S-OLS atteignent un plateau après 10 itérations. En revanche, l'erreur de reconstruction de S-SBR continue de décroître. On observe également que S-CoSaMP est plus rapide car il atteint un minimum d'erreur après 8 itérations alors que les autres algorithmes nécessitent une ou deux itérations de plus.

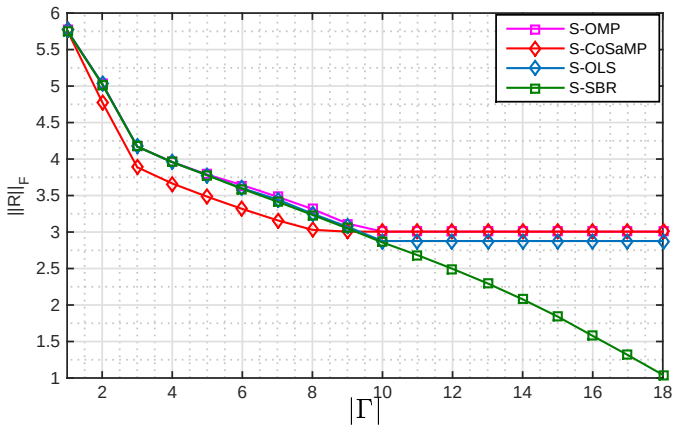


FIGURE 3.19 – Evolution de l'erreur de reconstruction $\mathbf{R} = \mathbf{Y} - \Phi_\Gamma \mathbf{X}_\Gamma$ en fonction de la cardinalité de l'ensemble des atomes actifs $|\Gamma|$ (RSB $= 10$ dB, $d = 2.33$).

3.5 Conclusion

Ce chapitre a été consacré à la présentation des méthodes d'approximation parcimonieuse. En particulier, des méthodes gloutonnes pour l'approximation parcimonieuse simultanée ont été développées. Nous avons montré que cette approche permet d'obtenir une meilleure sélection des atomes actifs d'un signal avec des observations multiples, en comparaison avec l'approche standard. Dans le chapitre suivant, cette technique sera exploitée afin de développer de nouvelles méthodes numériquement efficaces pour la sélection de variables pertinentes pour la classification des spectres de déchets de bois. Ces méthodes sont fondées sur la relaxation convexe de la norme ℓ_0.

Chapitre 4

Regularized Simultaneous Sparse Approximation Methods for Variable Selection and Classification of NIR Spectra and Hyperspectral Images

Ce chapitre est une adaption de l'article [T1].

Sommaire

4.1 Introduction

Infrared spectroscopy is a vibrational spectroscopy which provides information about the molecular composition and interactions within the studied material sample [125, 122]. As the sample spectrum is

a kind of signature characterizing the material, IR spectroscopy is used in a wide range of applications, including material identification, characterization and non destructive evaluation [1, 78]. The interest of IR spectroscopy has been reinforced with the development of fast scanning hyperspectral imaging systems allowing the development of the so called chemical imaging. In this work the targeted application is material sorting, which is envisaged as a supervised classification problem. To face the curse of dimensionality, feature selection has been recognized as a key step, especially when the number of highly correlated variables is large. The problem is to find a small subset of the observed variables which describes at best the main characteristics of the different classes. Popular features selection approaches include best subset selection [55, 98], forward and backward stepwise regression [89, 154], forward stagewise regression and sparse linear regression known as the Lasso (*Least absolute shrinkage and selection operator*) [127].

Another way to address feature selection is to state it as a dimensionality reduction problem. Basically, this consists in decomposing the data on an overcomplete dictionary (base). The feature selection consists then in selecting a small number of decomposition coefficients, also termed in statistics as explanatory variables. In fact, this is strongly connected to sparse representation which has been widely studied over the last decade, and applied to different problems such as data compression [36, 24], pattern recognition [151], classification and clustering [71, 80] and hyperspectral image unmixing [73, 31]. More recently, group sparsity was introduced to enforce certain structural constraints. Restricting our attention to simultaneous sparsity (which is a special group sparsity), we seek at finding the set of coefficients explaining at best the set of observed variables simultaneously. Solving the exact simultaneous sparse approximation problem yields to an NP-hard problem for which the greedy methods provide a good compromise between reconstruction accuracy and computational cost [134, 7]. Convex relaxations of the simultaneous sparse approximation was also proposed in [133]. Let us mention also the work of Turlach *et al.* [139] which investigated simultaneous sparse approximation in the context of variable selection of IR spectra. This work serves as a starting point of the present chapter which aim to investigate the benefit of adding a regularization term favoring piecewise constant simultaneous sparse approximation. The motivation for developing such an approach is twofold : 1) it aims at capturing the very nature of the learning dataset when it is properly ordered. 2) it allows to capture the spatial characteristics of the objects to sort when they are imaged by a pushbroom hyperspectral imaging system. In both cases the key point relies on the notion of ordering, which can be :

— a class (label) ordering for learning the classification rule,

— a spatial ordering for classifying the objects to sort.

This chapter is organized as follows : The regularized simultaneous sparse approximation problem (RSSA) is introduced in Section 4.2. The details of the convex relaxation approaches are described in Section 4.3 where it will be shown that the proposed methods have a decomposition property allowing to compute an efficient solution, suitable for large scale problems. The advantages of the proposed approaches are illustrated in Section 4.4 for two applications of material sorting : NIR spectra and hyperspectral images classification. Section 4.5 concludes the chapter.

Notations :

— $||.||_{p,q}$, denotes the mixed ℓ_p/ℓ_q norm of $\mathbf{A}$,
— $||\mathbf{A}||_F$, is the Frobenius norm of a matrix $\mathbf{A}$,
— $\mathbf{A}'$ is the transpose of $\mathbf{A}$,
— $\mathbf{x}_i$ is i-th column of $\mathbf{X}$ and $\mathbf{x}^i$ is the vector containing the elements of the i-th row,
— $\otimes$ denotes the Kronecker product and $\circ$ is the composition operator.

4.2 Regularized simultaneous sparse approximation

Let $\mathbf{Y} \in \mathbb{R}^{M\times K}$ be an observed data matrix composed of K spectra (columns) each containing M measurements. We seek to decompose the matrix $\mathbf{Y}$ such that : $\mathbf{Y} = \Phi\mathbf{X}$, as depicted in Figure 4.1. We considere here that the dictionary $\Phi \in \mathbb{R}^{M\times N}$ is composed of Gaussian-shaped functions having different locations and widths, and $\mathbf{X}$ is the sparse coefficient matrix, such that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K]$, where $\mathbf{x}_i$ is the i-th column of $\mathbf{X}$. Note that, instead of Gaussians, other kind of functions can also be used.

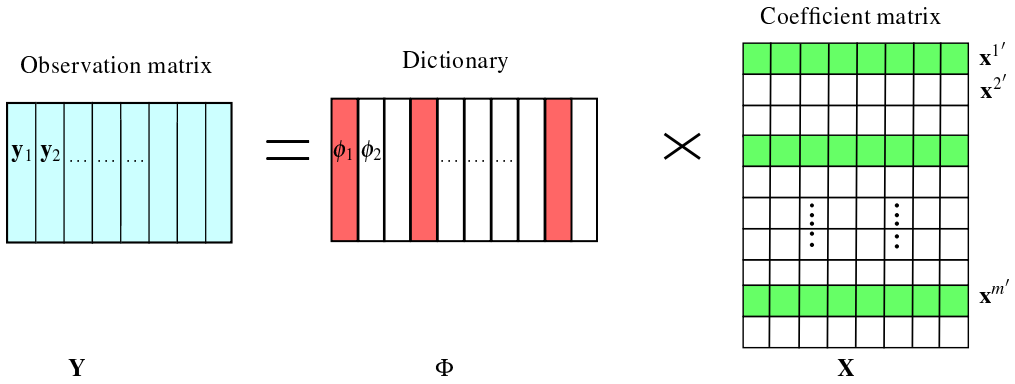


FIGURE 4.1 – Simultaneous sparse approximation. Reconstruction of matrix $\mathbf{Y}$ is a linear combination of active atoms in $\mathbf{X}$ (green rows) and the corresponding elements of the dictionary Φ.

The simultaneous sparse approximation problem [33, 26] consists in finding the solution having a limited number of active rows. The problem is then formulated as the minimization of the following cost function :

$$J_0(\mathbf{X}) = \frac{1}{2}||\mathbf{Y} - \Phi\mathbf{X}||_F^2 \text{ s.t. } ||\mathbf{X}||_{0,2} \leqslant s, \tag{4.1}$$

where $||\mathbf{X}||_{0,2}$ is the mixed ℓ_0/ℓ_2-pseudo norm of $\mathbf{X}$ (i.e. the number of rows with nonzero ℓ_2-norm) and $s \ll N$ is the sparsity parameter. We define the support of $\mathbf{X}$ as the set of active atoms :

$$\Gamma = \text{supp}(\mathbf{X}) = \{1 \leq i \leq N \,|\, \mathbf{x}^i \neq \mathbf{0}\}. \tag{4.2}$$

The regularized simultaneous sparse approximation aims at reconstructing piecewise constant sparse rows. In that respect, we propose to include a regularization term in the cost function leading to the new criterion :

$$J_1(\mathbf{X};\lambda) = \frac{1}{2}||\mathbf{Y} - \Phi\mathbf{X}||_F^2 + \lambda_2||\mathbf{D}\mathbf{X}'||_{1,1} \text{ s.t. } ||\mathbf{X}||_{0,2} \leqslant s \tag{4.3}$$

where $\mathbf{D} \in \mathbb{R}^{(K-1)\times K}$ is a matrix of finite differences of order 1 :

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & & \mathbf{0} \\ & \ddots & \ddots & \\ \mathbf{0} & & -1 & 1 \end{bmatrix}. \tag{4.4}$$

Criterion (4.3) includes an additional term enforcing the sparsity of the row derivative of $\mathbf{X}$ which in fact favors the reconstruction of piecewise constant rows. Following the terminology of the fused Lasso, this term will be referred to as fusion penalty. However, this formulation involving the mixed ℓ_0/ℓ_2-pseudo norm leads to an NP-hard problem, thus making the resolution not easy. A suboptimal but simple approach to solve (4.3) was proposed in [8]. It consists in decomposing the problem into two subproblems :

$$(SP_1) : \ \Gamma = \arg\min_{\Gamma} ||\mathbf{Y} - \Phi\mathbf{X}_\Gamma||_F^2 \ \text{ s.t. } |\Gamma| \leqslant s \tag{4.5}$$

$$(SP_2) : \ \mathbf{X}_\Gamma = \arg\min_{\mathbf{X}_\Gamma} ||\mathbf{Y} - \Phi_\Gamma\mathbf{X}_\Gamma||_F^2 + \lambda_2||\mathbf{D}\mathbf{X}'_\Gamma||_{1,1} \tag{4.6}$$

The subproblem in (4.5) may be solved using a greedy pursuit method such as S-OMP [134] or S-OLS (i.e the simultaneous version of the OLS algorithm : *Orthogonal Least Squares*) [7]. This provides the support Γ. Then, the subproblem (4.6) is reformulated as a generalized Lasso problem (for more details, see [8]). In [8] it is solved using an alternating direction method of multipliers algorithm. However, other optimization methods (such as FISTA) may be used. We will come back to this point in the next section. The corresponding algorithm will be referred as SR-SA (Simultaneous Regularized Sparse Approximation) in the experimental section.

4.3 Convex relaxation approaches of RSSA

4.3.1 Fused Sparse Lasso

The first relaxation of problem (4.3) consists in considering the following criterion :

$$J_2(\mathbf{X};\lambda_1,\lambda_2) = \frac{1}{2}||\mathbf{Y} - \Phi\mathbf{X}||_F^2 + \lambda_1||\mathbf{X}||_{1,1} + \lambda_2||\mathbf{D}\mathbf{X}'||_{1,1}. \tag{4.7}$$

where $||\mathbf{X}||_{1,1} = \sum_{i=1}^{K} ||\mathbf{x}_i||_1 = \sum_{i=1}^{N} ||\mathbf{x}^i||_1$. The parameters $\lambda_1, \lambda_2 \geq 0$ are controlling the tradeoff between data fitting, sparsity term $||\mathbf{X}||_{1,1}$, and the fusion term $||\mathbf{D}\mathbf{X}'||_{1,1}$. This criterion extends to the Multiple Measurement Vector (MMV) case the sparse fused Lasso which was studied in [128] where it is proposed to use the two-phase active set algorithm [58] designed for quadratic programming problems with linear sparsity constraints. In [129] the problem is extended to general graphs where the fusion term is promoting constant coefficients over neighboring variables. This approach does not include the additional sparsity term. It is referred to as the generalized fused Lasso (GFL). In [51] it is proposed to solve the generalized sparse fused Lasso problem (including both sparsity and fusion terms) in the special case where the dic-

tionary is $\Phi = \mathbf{I}$. This was extended in [155] to general dictionary. Here, we propose to solve this problem in the special case where the fusion term only acts on the rows of $\mathbf{X}$. According to this specific structure it is possible to have a computationally efficient implementation of the minimization problem using the proximal gradient method FISTA (Fast Iterative Shrinkage-Thresholding Algorithm) [6]. However, before going further, let us give a few comments on criterion (4.7). In fact, the sparsity term $||\mathbf{X}||_{1,1}$ does not correspond to a proper convex relaxation of $||\mathbf{X}||_{0,2}$. As will be explained in section 4.3.2, the mixed norm $||\mathbf{X}||_{1,2}$ is more appropriate. But combining $||\mathbf{X}||_{1,1}$ to $||\mathbf{D}\mathbf{X}'||_{1,1}$ yields to a kind of simultaneous sparse approximation : the simultaneity is actually enforced by the row regularization term $||\mathbf{D}\mathbf{X}'||_{1,1}$, but there is no control of the number of activated rows. Let $\mathbf{x} = \text{vec}(\mathbf{X}')$ and $\mathbf{y} = \text{vec}(\mathbf{Y}')$. Criterion (4.7) can be rewritten as :

$$J_2(\mathbf{x};\lambda_1,\lambda_2) = \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1. \tag{4.8}$$

with $\mathbf{A} = \Phi \otimes \mathbf{I}_K$, $\mathbf{F} = \mathbf{I}_N \otimes \mathbf{D}$. FISTA is an extension of the Nesterov's gradient-based method ISTA used to solve convex optimization problems including both smooth and non-smooth terms. The advantage of FISTA lies in its higher convergence rate which is $O(1/k^2)$ while the ISTA convergence rate is $O(1/k)$, where k is the number of iterations. Here, due to the block diagonal structure of $\mathbf{F}$, the problem is row decomposable. This allows the development of a fast method for solving (4.8) able to handle high dimensional problems. Let $\Omega(\mathbf{x})$ be the non-smooth term in (4.8) :

$$\Omega(\mathbf{x}) = \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1. \tag{4.9}$$

Following [155] , the update of vector $\mathbf{x}$ is :

$$\mathbf{x}_{(k+1)} = \arg\min_{\mathbf{x}} \left(\Omega(\mathbf{x}) + \frac{L}{2}||\mathbf{x}-\mathbf{v}_{(k)}||_2^2\right) \tag{4.10}$$

where

$$\mathbf{v}_{(k)} = \mathbf{x}_{(k)} - \frac{1}{L}\nabla f(\mathbf{x}_{(k)}) \tag{4.11}$$

and $\nabla f(\mathbf{x}) = \mathbf{A}'(\mathbf{A}\mathbf{x}-\mathbf{y})$ is the gradient of the smooth part of (4.8). L is the Lipschitz constant of $f(\mathbf{x})$. Note that the updating of $\mathbf{v}_{(k)}$ according to (4.11) involves the calculation and storage of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}'\mathbf{y}$. Instead, we can update the matrix $\mathbf{V}_{(k)}$ which corresponds to $\mathbf{v}_{(k)} = \text{vec}(\mathbf{V}'_{(k)})$ according to :

$$\mathbf{V}_{(k)} = \mathbf{X}_{(k)} - \frac{1}{L}\Phi'(\Phi\mathbf{X}_{(k)} - \mathbf{Y}). \tag{4.12}$$

As a consequence, only lower dimension matrices, $\Phi'\Phi$ and $\Phi'\mathbf{Y}$, need to be computed and stored. Minimizing (4.10) is similar to the 1D fused Lasso signal approximator (FLSA) [68]. However, due to the block diagonal structure of $\mathbf{F}$, it can be shown that the problem (4.10) can be solved separately for each row of $\mathbf{X}$:

$$\mathbf{x}^i_{(k+1)} = \arg\min_{\mathbf{x}^i} \frac{1}{2}||\mathbf{x}^i - \mathbf{v}^i_{(k)}||_2^2 + \frac{\lambda_1}{L}||\mathbf{x}^i||_1 + \frac{\lambda_2}{L}||\mathbf{F}\mathbf{x}^i||_1. \tag{4.13}$$

The minimizer of (4.13) is obtained by using subgradient technique. Indeed, any solution corresponding to (λ_1,λ_2) is obtained by a soft thresholding of the solution obtained for $(\lambda_1 = 0, \lambda_2)$. This is stated by

the following theorem proved in [85].

Theorem 1. *Let :*

$$\mathbf{x}^{\lambda_1}_{\lambda_2}(\mathbf{v}) = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}||_2^2 + \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1 \tag{4.14}$$

For any $\lambda_1,\ \lambda_2 \geq 0$*, we have*

$$\mathbf{x}^{\lambda_1}_{\lambda_2}(\mathbf{v}) = \text{sign}(\mathbf{x}^0_{\lambda_2}(\mathbf{v})) \odot \max(|\mathbf{x}^0_{\lambda_2}(\mathbf{v})| - \lambda_1, 0). \tag{4.15}$$

where $\odot$ *denotes the element-wise product operator.*

Hence, the solution to (4.13) is computed for each row $\mathbf{x}^i$ of the coefficient matrix $\mathbf{X}$ as following :

$$\begin{cases} \mathbf{V}_{(k)} = \mathbf{X}_{(k)} - \frac{1}{L}\Phi'(\Phi\mathbf{X}_{(k)} - \mathbf{Y}), \\ \mathbf{x}^i_{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}^i - \mathbf{v}^i_{(k)}||_2^2 + \frac{\lambda_1}{L}||\mathbf{x}^i||_1 + \frac{\lambda_2}{L}||\mathbf{D}\mathbf{x}^i||_1. \end{cases} \tag{4.16}$$

for $i = 1,\ldots,N$. In practice each problem (4.13) is solved using the FLSA routine implemented in SLEP package [7]. Finally, the main steps of the Fused Sparse Lasso algorithm are presented in Algorithm 1.

Algorithm 1 The Fused Sparse Lasso algorithm : (FSL)

Input: $\mathbf{Y} \in \mathbb{C}^{M\times K}$, $\Phi \in \mathbb{C}^{M\times N}$, $\mathbf{D} \in \mathbb{C}^{K-1\times K}$, λ_1, λ_2, maxiter, L.
Initialise: $\mathbf{X}_0 = \mathbf{0}$,
1: **for** $k \leftarrow 1$ to maxiter **do**
2: $\quad \mathbf{V}_{(k)} \leftarrow \mathbf{X}_{(k)} - \frac{1}{L}\Phi'(\Phi\mathbf{X}_{(k)} - \mathbf{Y})$.
3: $\quad$ **for** $i \leftarrow 1$ to **N do**
4: $\quad\quad \mathbf{x}^i_{(k+1)} \leftarrow \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}^i - \mathbf{v}^i_{(k)}||_2^2 + \frac{\lambda_1}{L}||\mathbf{x}^i||_1 + \frac{\lambda_2}{L}||\mathbf{D}\mathbf{x}^i||_1$. ▷ Solved using FLSA
5: $\quad$ **end for**
6: **end for**
7: **return X**.

4.3.2 Fused Sparse Group Lasso

As mentioned before, the fused sparse Lasso is not a proper relaxation of the RSSA. Indeed, the term $||\mathbf{X}||_{1,1}$ does not allow to control the number of active rows. Here, we propose to relax the ℓ_0/ℓ_2 pseudo-norm into the ℓ_1/ℓ_2 mixed norm defined by : $||\mathbf{X}||_{1,2} = \sum_{i=1}^{N} ||\mathbf{x}^i||_2$, which is a particular instance of the group Lasso penalty. So we propose the following criterion :

$$J_3(\mathbf{x};\lambda_1,\lambda_2,\lambda_3) = \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1 + \lambda_3\sum_{i=1}^{N}||\mathbf{x}^i||_2 \tag{4.17}$$

as a convex relaxation of problem (4.3). Note that the $||\mathbf{x}||_1$ penalty is maintained to eventually control the global sparsity of the solution. The proximal operator associated to the fused sparse group Lasso with

7. http://yelab.net/software/SLEP/

the composite of non-smooth penalties is defined as :

$$\text{Prox}_{FSGL}(\mathbf{v}) = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}||_2^2 + \frac{\lambda_1}{L}||\mathbf{x}||_1 + \frac{\lambda_2}{L}||\mathbf{F}\mathbf{x}||_1 + \frac{\lambda_3}{L}\sum_{i=1}^{N}||\mathbf{x}^i||_2. \tag{4.18}$$

Now, with the three non-smooth terms in the objective-function, the proximal operator may be solved as suggested in [159]. In fact, the proximal operator associated to (4.18) has a decomposition property that allows to compute it in two steps based on the following theorem :

Theorem 2. *Let us define*

$$\text{Prox}_{FSL}(\mathbf{v}) = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}||_2^2 + \frac{\lambda_1}{L}||\mathbf{x}||_1 + \frac{\lambda_2}{L}||\mathbf{D}\mathbf{x}||_1. \tag{4.19}$$

$$\text{Prox}_{GL}(\mathbf{v}) = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}||_2^2 + \frac{\lambda_3}{L}\sum_{i=1}^{N}||\mathbf{x}^i||_2. \tag{4.20}$$

Then, the following holds :

$$\text{Prox}_{FSGL}(\mathbf{v}) = (Prox_{GL} \circ \text{Prox}_{FL})(\mathbf{v}). \tag{4.21}$$

where $\circ$ is the composition operator.

The proof of the theorem above is given in [159]. This result implies that we can first compute the proximal operator associated to the fused sparse Lasso. Then, the solution obtained is plugged in the proximal operator associated to the group penalty. The implementation of proximal operator associated with the group Lasso penalty is performed using the ALTRA routine of the SLEP package to compute the solution for each row $\mathbf{x}^i$. The resulting algorithm is summarized in Algorithm 2.

Algorithm 2 The Fused Sparse Group Lasso Algorithm : (FSGL)

Input: $\mathbf{Y} \in \mathbb{C}^{M\times K}$, $\Phi \in \mathbb{C}^{M\times N}$, $\mathbf{D} \in \mathbb{C}^{K-1\times K}$, λ_1, λ_2, λ_3, maxiter.
Initialise: $\mathbf{X}_0 = \mathbf{0}$,

for $k \leftarrow 1$ to maxiter **do**
  $\mathbf{V}_{(k)} \leftarrow \mathbf{X}_{(k)} - \frac{1}{L}\Phi'(\Phi\mathbf{X}_{(k)} - \mathbf{Y})$.
  for $i \leftarrow 1$ to $\mathbf{N}$ **do**
    $\mathbf{w}^i_{(k)} \leftarrow \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}^i_{(k)}||_2^2 + \frac{\lambda_1}{L}||\mathbf{x}||_1 + \frac{\lambda_2}{L}||\mathbf{D}\mathbf{x}||_1$. ▷ Solved using FLSA algorithm
    $\mathbf{x}^i_{(k+1)} \leftarrow \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{w}^i_{(k)}||_2^2 + \frac{\lambda_3}{L}||\mathbf{x}||_2$. ▷ Solved using ALTRA routine
  end for
end for
return $\mathbf{X} \in \mathbb{C}^{N\times K}$.

4.3.3 Nonnegative Sparse Fused Group Lasso

As we deal with positive data, it is suitable to impose a nonnegativity constraint on the solution. Indeed, the solution proposed above may induce artifacts due to bad conditioning of matrices, causing the

appearance of negative values. Physically, such a solution is unacceptable and a rigorous recovery process must take into account this additional constraint. So, we propose here to minimize the nonnegative version of the fused sparse group Lasso algorithm :

$$J_4(\mathbf{x};\lambda_1,\lambda_2,\lambda_3) = \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1 + \lambda_3\sum_{i=1}^{N}||\mathbf{x}^i||_2 \text{ s.t. } \mathbf{x}\geqslant 0. \tag{4.22}$$

First, we include a slack variable $\mathbf{u}$ to the objective function which leads to :

$$J_4(\mathbf{x},\mathbf{u};\lambda_1,\lambda_2,\lambda_3) = J_3(\mathbf{x};\lambda_1,\lambda_2,\lambda_3) \text{ s.t. } \mathbf{x}-\mathbf{u}=0,\ \mathbf{u}\geqslant 0 \tag{4.23}$$

To minimize this criterion, we use the quadratic penalty method. The new objective in then :

$$J_4(\mathbf{x},\mathbf{u};\lambda_1,\lambda_2,\lambda_3) = J_3(\mathbf{x};\lambda_1,\lambda_2,\lambda_3) + \frac{\xi}{2}||\mathbf{x}-\mathbf{u}||_2^2,\ \ \mathbf{u}\geqslant 0 \tag{4.24}$$

$$= \frac{1}{2}||\mathbf{y}-\mathbf{A}\mathbf{x}||_2^2 + \frac{\xi}{2}||\mathbf{x}-\mathbf{u}||_2^2 + \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1 + \lambda_3\sum_{i=1}^{N}||\mathbf{x}^i||_2,\ \ \mathbf{u}\geqslant 0 \tag{4.25}$$

where ξ is the parameter that ensures the respect of the nonnegativity constraint, in the sens that, when $\xi\to\infty$, the entries of the vector $\mathbf{x}$ tend toward those of the vector $\mathbf{u}$ making the positivity constraint on $\mathbf{x}$ satisfied asymptotically. The surrogate problem (4.25) is unconstrained with respect to $\mathbf{x}$. Hence, by stacking the two quadratic terms of the function J_4 we have :

$$J_4(\mathbf{x},\mathbf{u};\lambda_1,\lambda_2,\lambda_3) = \frac{1}{2}||\mathbf{z}_\mathbf{u}-\mathbf{B}\mathbf{x}||_2^2 + \Omega(\mathbf{x}), \mathbf{u}\geqslant 0 \tag{4.26}$$

where $\mathbf{B} = [\mathbf{A}', \sqrt{\xi}\mathbf{I}]'$, $\mathbf{z}_\mathbf{u} = [\mathbf{y}', \sqrt{\xi}\mathbf{u}']'$ and $\Omega(\mathbf{x})$ is defined here as :

$$\Omega(\mathbf{x}) = \lambda_1||\mathbf{x}||_1 + \lambda_2||\mathbf{F}\mathbf{x}||_1 + \lambda_3\sum_{i=1}^{N}||\mathbf{x}^i||_2. \tag{4.27}$$

We should now minimize the cost function $J_4(\mathbf{x},\mathbf{u};\lambda_1,\lambda_2,\lambda_3)$ with respect to $\mathbf{x}$ (without constraint) and $\mathbf{u}$ (with the nonnegativity constraint). The minimization with respect to $\mathbf{x}$ leads to an iteration similar to that of FSGL :

$$\begin{cases} \mathbf{x}_{(k+1)} = \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x}-\mathbf{v}_{(k)}||_2^2 + \frac{1}{L_p}\Omega(\mathbf{x}). \\ \mathbf{v}_{(k)} = \mathbf{x}_{(k)} - \frac{1}{L_p}\nabla f_p(\mathbf{x}_{(k)}). \end{cases} \tag{4.28}$$

where $f_p(\mathbf{x}) = \frac{1}{2}||\mathbf{z}_\mathbf{u}-\mathbf{B}\mathbf{x}||_2^2$, $\nabla f_p(\mathbf{x}) = \mathbf{B}'(\mathbf{B}\mathbf{x}-\mathbf{z}_\mathbf{u})$ and $L_p = L+\xi$ is the Lipschitz constant of $f_p(\mathbf{x})$. Once again, the optimization is separable for each row $\mathbf{x}^i$. Replacing $\mathbf{B}$ and $\mathbf{z}_\mathbf{u}$ by their expressions gives :

$$\begin{cases} \mathbf{x}^i_{(k+1)} = \arg\min_{\mathbf{x}\in\mathbb{R}^K} \frac{1}{2}||\mathbf{x}-\mathbf{v}^i_{(k)}||_2^2 + \frac{\lambda_1}{L_p}||\mathbf{x}||_1 + \frac{\lambda_2}{L_p}||\mathbf{D}\mathbf{x}||_1 + \frac{\lambda_3}{L_p}||\mathbf{x}||_2, \text{ for } i=1,\ldots,N. \\ \mathbf{V}_{(k)} = \mathbf{X}_{(k)} - \frac{1}{L_p}\Phi'(\Phi\mathbf{X}_{(k)}-\mathbf{Y}) - \frac{\xi}{L_p}(\mathbf{X}_{(k)}-\mathbf{U}_{(\ell)}), \end{cases} \tag{4.29}$$

Algorithm 3 The Nonnegative Fused Sparse Group Lasso Algorithm (NN-FSGL)

Input: $\mathbf{Y} \in \mathbb{C}^{M\times K}$, $\Phi \in \mathbb{C}^{M\times N}$, $\mathbf{D} \in \mathbb{C}^{K-1\times K}$, λ_1, λ_2, λ_3, $\xi_1 = 1$, $\mathbf{u} = 0$, maxiter, nniter.
Initialise: $\mathbf{X}_0 = \mathbf{0}$.

for $\ell \leftarrow 1$ to nniter **do**
  for $k \leftarrow 1$ to maxiter **do**
    $\mathbf{V}_{(k)} \leftarrow \mathbf{X}_{(k)} - \frac{1}{L_p}\Phi'(\Phi\mathbf{X}_{(k)} - \mathbf{Y}) - \frac{\xi_{(\ell)}}{L_p}(\mathbf{X}_{(k)} - \mathbf{U}_{(\ell)})$.
    for $i \leftarrow 1$ to $\mathbf{N}$ **do**
      $\mathbf{w}^i_{(k)} \leftarrow \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x} - \mathbf{v}^i_{(k)}||_2^2 + \frac{\lambda_1}{L_p}||\mathbf{x}||_1 + \frac{\lambda_2}{L_p}||\mathbf{D}\mathbf{x}||_1$.
      $\mathbf{x}^i_{(k)} \leftarrow \arg\min_{\mathbf{x}} \frac{1}{2}||\mathbf{x} - \mathbf{w}^i_{(k)}||_2^2 + \frac{\lambda_3}{L_p}||\mathbf{x}||_2$.
    end for
  end for
  $\mathbf{u}_{(\ell+1)} \leftarrow \max(0, \mathbf{x})$. ▷ Update $\mathbf{u}$ with hard thresholding operator.
  $\xi_{(\ell+1)} \leftarrow \beta\xi_{(\ell)}$.
end for
return X

Hence, an external loop (ℓ) is added to update the variable $\mathbf{u}$. The minimization of J_4 with respect to the slack variable $\mathbf{u}$ is simply a hard thresholding operation :

$$\mathbf{u}_{(\ell+1)} = \max(0, \mathbf{x}). \tag{4.30}$$

where $\mathbf{x}$ is the value of $\mathbf{x}_{(k)}$ after convergence. The tuning parameter ξ is updated in the loop with the classical linear rule : $\xi_{(\ell+1)} = \beta\xi_{(\ell)}$, with $\beta > 1$ and $\xi_1 = 1$. The complete algorithm for nonnegative FSGL is summarized in Algorithm 3.

4.4 Application to wood wastes sorting

4.4.1 Motivations

One of the most promising application of spectroscopy and hyperspectral imaging in industry is material sorting [126, 156, 121] and quality control [42, 86]. In this work, we are interested in sorting wood wastes which have to be separated into two broad categories : recyclable and non recyclable. Each category includes a number of wood wastes types, termed as "groups" as given in Table 4.1.

The wood wastes sorting is addressed as a binary classification problem of NIR spectra. A single spectrum is acquired for each wood sample and the classifier has to decide whether it is recyclable or not recyclable. This is a much more laboratory oriented approach. In fact, as the goal is to develop an industrial system allowing a fast scanning of the wood pieces, hyperspectral imaging is the final objective since it allows the simultaneous inspection of many wood pieces placed on the convoyer as sketched on Figure 4.2. Such a technology is already used for example by Pellenc.ST which is a leading company in

TABLE 4.1 – Composition of the two categories of wood wastes

Class 1 : to Recycle			Class 2 : to Reject		
Group	Name	Samples	Group	Name	Samples
1.1	raw wood	32	2.1	MDF-HDF	28
1.2	painted solid wood	36	2.2	painted MDF-HDF	50
1.3	vanished solid wood	35	2.6	CCA preserved wood	35
1.4	vanished plywood	18	2.4	raw fiber board	8
1.5	raw particle board	28	-	-	-
1.7	painted particle board	6	-	-	-
1.8	raw plywood	18	-	-	-

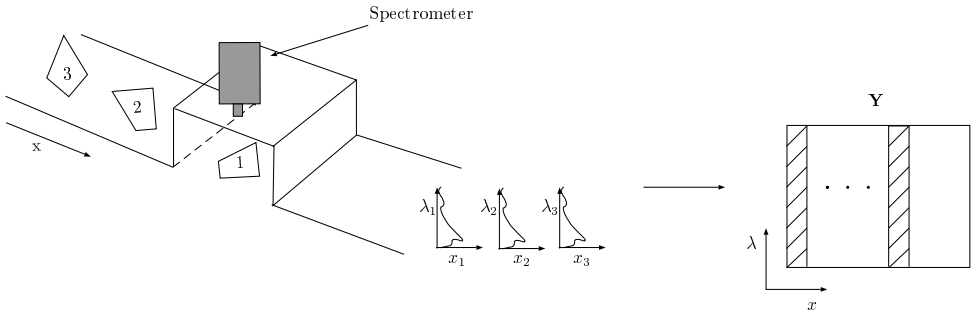
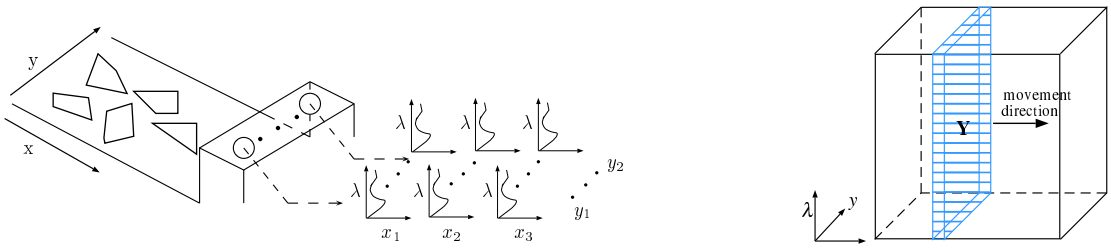


FIGURE 4.2 – Data matrix build from spectra of samples acquired using a spectrometer.

optical sorting solutions for waste management and recycling [8]. Such a system makes use of pushbroom hyperspectral imaging system which yields the hyperspectral data cube slice by slice as depicted in Figure 4.3.



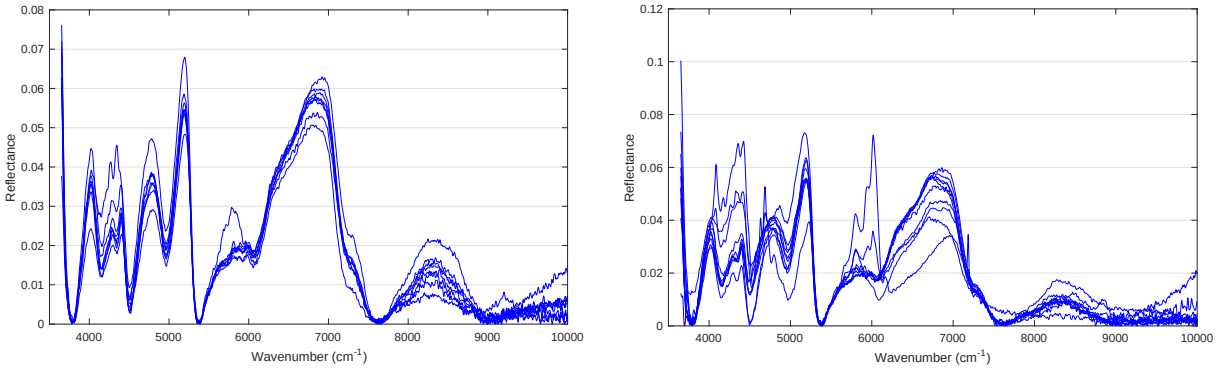
(a) Imaging system

(b) Data cube

FIGURE 4.3 – Hyperspectral images acquired by an industrial NIR spectro-imager.

The goal of this section is to show the advantages of RSSA algorithms developed in the previous section in a binary classification problem. RSSA algorithms are primarily intended at selecting the explanatory variables used in classifiers. Here we restrict our attention to kernel SVM classifiers which proved to be among the most effective for the considered problem, and the question at hand is : it is possible to improve the classification rates and decrease the computational burden by performing a proper variable selection ? This question will be addressed for both NIR spectra classification and NIR hyper-

8. http://www.pellencst.com/fr/



(a) Class 1 : to recycle

(b) Class 2 : to reject

FIGURE 4.4 – Some spectra from the two classes.

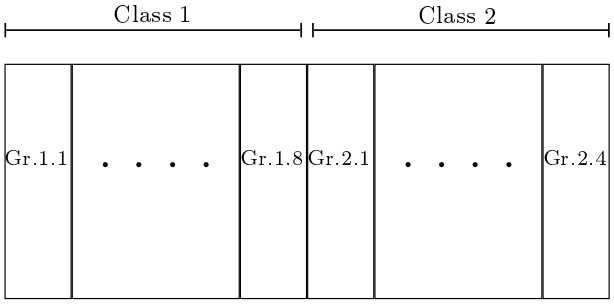


FIGURE 4.5 – Spectra in data matrix $\mathbf{Y}$ ordered according to their group.

spectral classification

4.4.2 Data acquisition and pre-processing

A large number of wood wastes (294 samples) collected on a wood wastes park were gathered by experts into 11 labeled groups (see Table 4.1). The data acquisition was carried on a Nicolet 8700 FTIR spectrometer continuously purged with ultrapure liquid nitrogen N2 and equipped with a MCT detector and a CaF2 beam splitter. Near infrared reflectance spectra covering the range 3562–10000 cm^{-1} (2.8-1 μm), were obtained with near-normal specular reflectance accessory (with a fixed 10 degree angle of incidence) provided by Pike Technologies. The spectral resolution is 16 cm^{-1}. The spectral sampling step is 4 cm^{-1} yielding a number of 1647 spectral bands. Each spectrum is obtained by averaging 100 scans resulting in a high SNR. The data pre-processing includes baseline removal (using the method proposed in [92]), offset correction ensuring zero lower bound, and unit energy normalization. Some spectra from these different groups are shown in Figure 4.4. It appears that the discriminant features cannot be determined by a simple visual examination. After pre-processing, the data are gathered into the data matrix $\mathbf{Y} \in \mathbb{R}^{1647\times 294}$. Note that the spectra are ordered according to the group they belong to as shown in Figure 4.5. This is a very important point since it is this ordering which enforces the piecewise constant characteristic of the approximation coefficient $\mathbf{X}$. In fact, the underlying implicit assumption is that the spectral characteristics are almost constant within each group.

TABLE 4.2 – The values of the parameters of the chosen configuration.

Method	Parameters						
	s / λ_1	λ_2	λ_3	maxiter	ξ_1	β	nniter
SR-SA	32	0.9	-	200	-	-	
FSL	0.18	0.24	-	200	-	-	
FSGL	0.12	0.12	0.12	200	-	-	
NN-FSGL	0.13	0.13	0.2	50	1	1.1	100

4.4.3 Variable selection

Four sparse methods for variable selection are tested : SR-SA, FSL, FSGL, and NN-FSGL. The latter is initialized with the unconstrained version of FSGL. The dictionary Φ is composed of normalized Gaussian functions whose means $m_i \in [3000, 1000]$ cm^{-1} and widths $\sigma_j \in [30, 600]$ cm^{-1} are covering uniformly their respective intervals. The discretization leads to 20 different values for σ_j. For each σ_j, the interval $[3000, 1000]$ cm^{-1} is discretized such that two adjacent m_i's are separated by σ_j. As a consequence, the number of atoms in the dictionary is 848. The set of parameters for each algorithm is given in the Table 4.2.

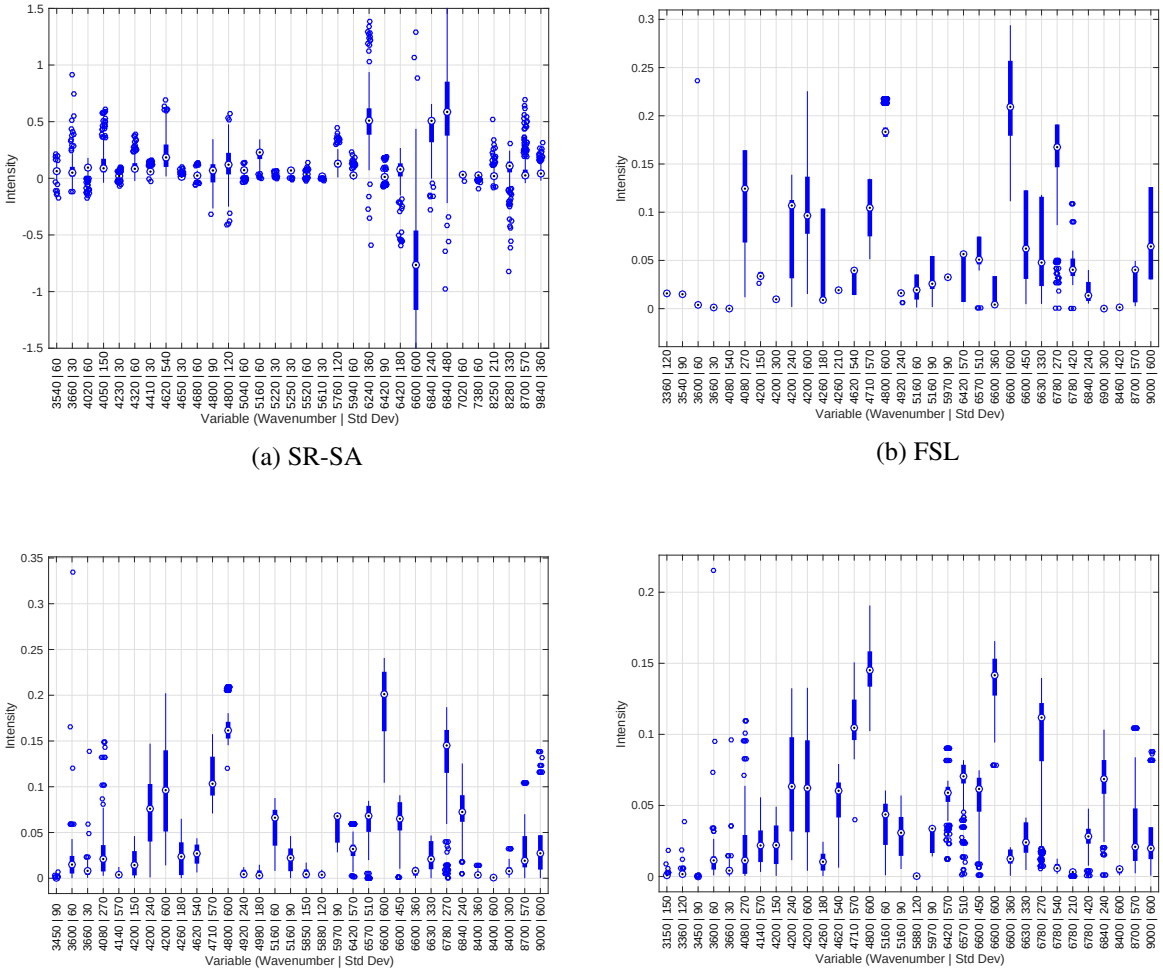
We present in Figure 4.6 the variables selected by the four algorithms. We observe that those obtained using SR-SA exhibit low dispersion of relative intensity characterized by tiny boxes. As the selection and regularization steps in this method are performed separately, the activated variables tend to cover the whole spectral range. In contrast, variables found using ℓ_1-based methods often display boxes with larger heights meaning that they are common to several spectra in the database. In particular, the variables selected using FSGL and NN-FSGL algorithms exhibit larger boxes thanks to the group penalty which reveals some coherent groups in the set of variables. It is worth to mention that the results obtained by FSGL and NN-FSGL are very similar. This is because the nonnegativity constraint in not very active in our configuration due to the pre-processing step. Notice that the three ℓ_1-based methods share 23 common variables over 32, and sometimes the same wavenumber appears with different standard deviations, e.g. at 6600 cm^{-1}. These methods share only 8–9 common variables with SR-SA.

4.4.4 Classification of wood wastes using NIR spectra

Once the 32 representative variables are selected, we perform the classification using SVM with quadratic kernel function using the least-squares estimates of the coefficients. We present in Figure 4.7 a view of the process steps diagram for this experiment. The results obtained using our methods are compared to :

— SR-SA : this is the sub-optimal approach used to solve equation (4.3) (this method is detailed in [8]).

— SVS + SVM : this is the approach proposed in [139] for simultaneous variable selection. The sparsity parameter t (as referred to in the original paper) in set to 0.2. This tuning allows to obtain 37 variables. We failed to find exactly 32 variables due to the sensitivity of SVS on this parameter. The SVM algorithm is performed on the set of selected variables.



(a) SR-SA

(b) FSL

(c) FSGL

(d) NN-FSGL.

FIGURE 4.6 – Intensity dispersion of the selected variables obtained by four approaches.

— G-SVM : this algorithm was proposed in [46]. It consists in solving an augmented SVM criterion where the sparsity constraint is imposed on the support vectors. The solution is computed using a projected gradient method. The sparsity of the support vectors is controlled by the parameter[9] C. Here C is set to 1280 making the decision rule made on 32 coefficients of the support vector.

— SVM : standard SVM with quadratic kernel function. This means that no variables selection is made and the classification is performed directly on the original data.

The classification results for 10 cross validation runs are reported in Table 4.3. We present for each method the total classification rate, the classification rate of class 1 (true positives), and the classification rate of class 2 (true negatives). In terms of total accuracy, the proposed methods clearly outperform the approaches developed in [139] and [46]. Moreover, the classification performances obtained on the subset of selected variables using the proposed approaches are better than those of SVM applied on the original data. The best result is about 89% obtained with FSGL + SVM. As in our application it is also

9. http ://remi.flamary.com/soft/soft-gsvm.html

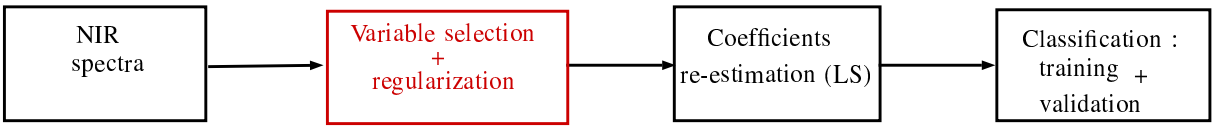


FIGURE 4.7 – Steps diagram of the NIR spectra classification process.

TABLE 4.3 – Comparison of the accuracy of wood wastes classification

Algorithm	Parameters			$\lvert\Gamma\rvert$	Total Accuracy ± Std. Dev.	Class 1 ± Std. Dev.	Class 2 ± Std. Dev.
	Sparsity	Fusion	Group				
SR-SA + SVM	32	0.9	-	32	87.6 ± 0.9%	87.1 ± 1.3%	88.2 ± 0.7%
FSL + SVM	0.18	0.24	-	32	87.6 ± 0.8%	84.6 ± 1.2%	91.9 ± 1.3%
FSGL + SVM	0	0.2	1.2	32	**88.9** ±0.8%	**87.1** ±1.3%	**91.3**±1.6%
	0.12	0.12	0.12	32	**89.0** ±1.1%	**86.3** ±1.7%	**90.4**±1.1%
NN-FSGL + SVM	0.13	0.13	0.2	32	86.2 ±1.3%	85.3 ±1.9%	87.4 ±1.7%
SVS + SVM	0.2	-	-	37	81.5 ± 1.2%	79.9 ± 1.3%	83.8 ± 1.9%
G-SVM	1280	-	-	32	72.2 ± 1.5%	70.5 ± 1.8%	74.5 ± 1.7%
SVM	-	-	-	-	83.0 ± 1.3%	80.4 ± 2.1%	86.6 ± 1.2%

important to reject the maximum polluted samples from the recycling process, the optimal parameters for FSGL are $\lambda_1 = 0$, $\lambda_2 = 0.2$ and $\lambda_3 = 1.2$.

The computational time required by each algorithm to perform variable selection is reported on Table 4.4. The results are obtained using a 2.4 Ghz Intel Core i5 processor with 8 Gigabytes of RAM. We note that the FSL algorithm is generally faster than all other approaches. As it is based on a greedy algorithm, the computational time of SR-SA increases with the sparsity parameter (i.e. the number of selected variables). The FSGL algorithm is a bit slower than FSL. The additional loop with hard thresholding operator makes the nonnegative FSGL algorithm about ten times slower than its unconstrained version. For the SVS algorithm, we did not try all the configurations because we found that this approach is much more slower and needs about four hours to select 37 variables. Finally, the proposed FSL and FSGL not only lead to good classification rates, but are also numerically efficient.

The performances of all the algorithms considered here depend on the choice of some tuning parameters. For instance, the set of selected variables (and thus, the overall classification performance) with

TABLE 4.4 – Computational time (in seconds) of the different approaches for variable selection .

$\lvert\Gamma\rvert$	SR-SA	FSL	FSGL	NN-FSGL	G-SVM
25	8	9	17	110	27
30	12	9	17	101	27
40	14	10	15	112	29
50	21	9	12	105	31

FSGL depend on λ_1, λ_2 and λ_3. Our aim now is to study the impact of each parameter on the number of selected variables and classification rates using 10-fold cross validation. For $\lambda_2 = 0.2$ and without the group penalty ($\lambda_3 = 0$), the results in terms of total classification error and cardinality of the support are reported in Figure 4.8a, for several values of λ_1. We note that the classification error rate decreases from 35% ($\lambda_1 = 10^{-2}$) to 14% ($\lambda_1 \in [0.18, 0.28]$). Naturally, the performances degrade drastically for values of λ_1 beyond 0.3 which correspond to less than 20 variables. For $\lambda_2 = 0.5$ and without the sparsity term ($\lambda_1 = 0$), the results are shown on Figure 4.8b, for different values of the grouping parameter λ_3. We observe that the total classification error rate is under 15% for $\lambda_3 \in [0.1, 1.5]$. In particular the value $\lambda_3 = 0.5$ provides the lowest classification error rate in this case (12.4%). This configuration (i.e $\lambda_3 = 0.5$) corresponds to a set of 39 active variables. It is worth to notice from these two tests that both the sparsity and grouping parameters act directly on the number of selected variables but not with same intensity. The sparsity parameter has greater influence on the cardinality of the support Γ. For instance, to obtain less than 80 variables, the grouping parameter should be set to 0.2 (and $\lambda_1 = 0$) while the same number of variables is obtained for $\lambda_1 \approx 0.06$ (and $\lambda_3 = 0$). To analyse the impact of the fusion parameter on the general classification performances, we set the sparsity parameter λ_1 to 0 and vary both the grouping and the fusion parameters such that 40 variables are retained. The results are reported on Figure 4.8c. The error rate is less than 15% in the range $\lambda_2 \in [0.1, 0.6]$. The minimum values of the classification error rate are between 13% and 11.6% ; they correspond to $\lambda_2 = 0.45$ and $\lambda_2 = 0.55$, respectively.

4.4.5 Classification of hyperspectral images of wood wastes

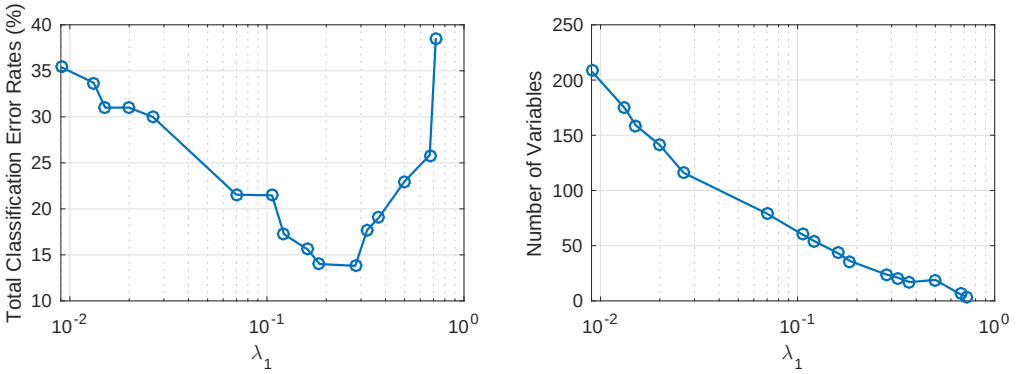
The acquisition process of hyperspectral images is composed of a conveyor on which the wood wastes samples progress at a speed of 3 m/s, a focused illumination system and an acquisition cabin which extracts the NIR spectra of samples row by row at a resolution of 12 mm along each spatial dimension. The experiments presented here consist in simulating this type of applications using *real-world* wood spectra. We start with an image of resolution 28×225 pixels representing wood samples, as depicted in Figure 4.9. The three-dimensional hyperspectral data cube $\mathscr{S}$ is then generated by affecting to each active pixel (black or red) a real spectrum from our database. As a result, the size of the data cube is $28 \times 225 \times 412$. Finally, $\mathscr{S}$ is perturbed by an additive white Gaussian noise $\mathscr{N}$ such that the SNR defined by :

$$\text{SNR} := \frac{\|\mathscr{S}\|_F^2}{\|\mathscr{N}\|_F^2}, \tag{4.31}$$

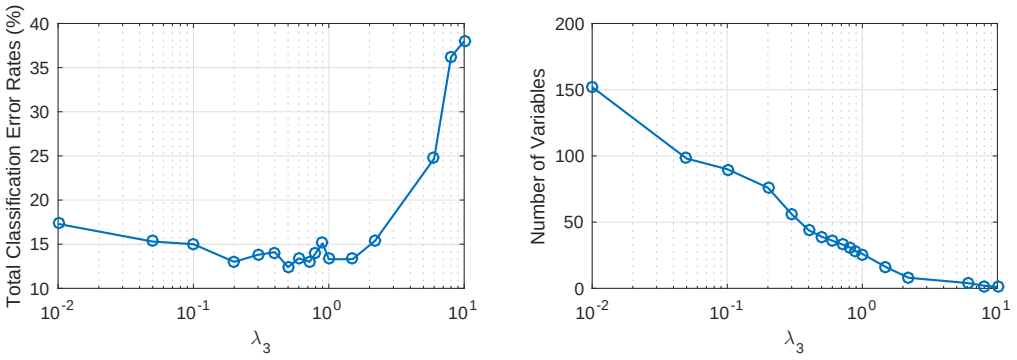
is varying between -5 and 10 dB.

The diagram sketched on Figure 4.10 presents the steps followed in this experiment. The training model used here is the same one used in the first experiment with the selected 40 variables. The proposed algorithms are applied to the noisy hyperspectral cube slice by slice [10]. The aim is to study how the spatial regularization influences the classifier performances. Thus, the coefficients are computed using the dictionary Φ_Γ with parameters $\lambda_1 = \lambda_3 = 0$ (the variables are already selected) and λ_2 varying from 0 to 0.008 for each slice successively. Thereafter, the classification is carried out on the resulting decom-

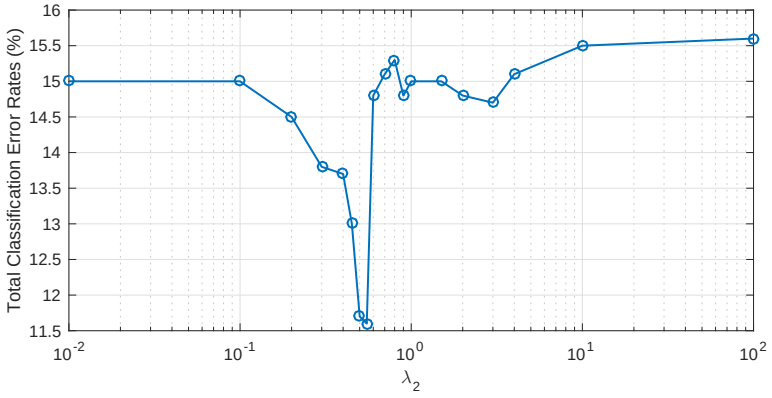
10. The size of each slice is 412×225. The first dimension is a spectral dimension and the second one is the spatial dimension.



(a) Sparsity parameter ($\lambda_2 = 0.2, \lambda_3 = 0$)



(b) Grouping parameter ($\lambda_1 = 0, \lambda_2 = 0.5$)



(c) Fusion parameter ($\lambda_1 = 0$, λ_3 fixed such that $|\Gamma| = 40$)

FIGURE 4.8 – Evolution of the total classification error rates as a function of the regularization parameters.

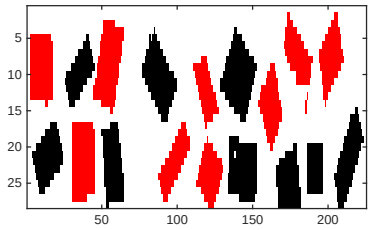


FIGURE 4.9 – Ground truth : black and red colors for category 1 and 2, respectively.

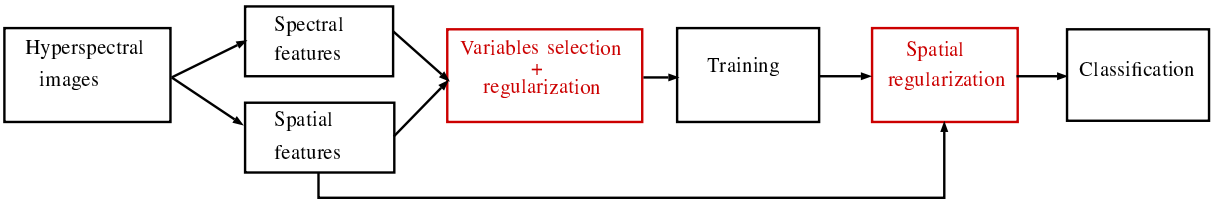
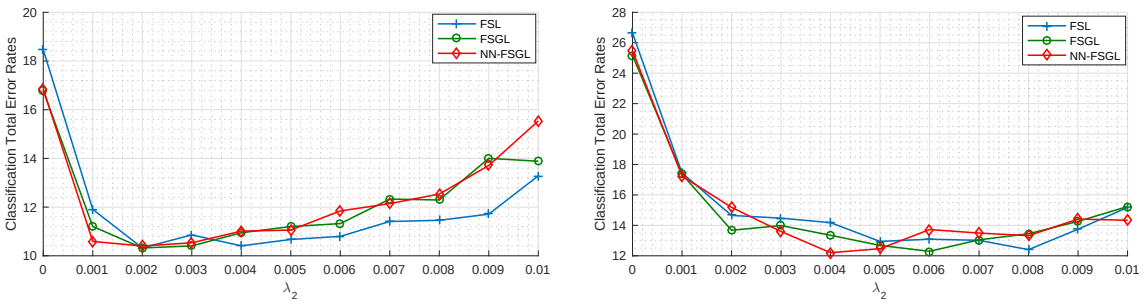


FIGURE 4.10 – General view of classification of the hyperspectral images including spectral and spatial regularization.

position using the training model. Classification rates are reported in Table 4.5 for four values of SNR. For each test we compute the total classification accuracy, the true positive rates (Class 1 accuracy) and the true negative rates (Class 2 accuracy).

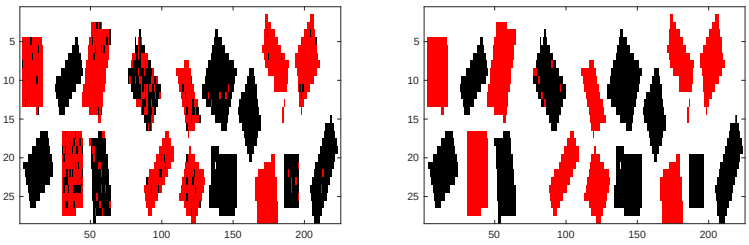
The classification rates obtained confirm the importance of using joint spectral and spatial regularization. Indeed, for 0 dB $\leq$ SNR $\leq$10 dB, which is a reasonable range of SNR in most hyperpectral applications, the classification performances increase when the regularization is applied and the three algorithms achieve about 80% of classification for $\lambda = 0.008$ even for a low SNR (-5 dB). In particular, we found that the best results are obtained using FSGL and NN-FSGL for 5 and 10 dB SNR values achieving about 90% of classification accuracy. Moreover, we succeed each time to increase the classification rates of Category 2 (true negative rates) significantly using the fusion constraint. We illustrate in Figure 4.11 the evolution of the classification total error rates as the value of the regularization parameter λ_2 increases for the proposed approaches. For SNR = 5 dB (Figure 4.11a), the minimum classification error is reached for $0.001 \leq \lambda_2 \leq 0.003$. At low SNR (Figure 4.11b) the value of the regularization parameter λ_2 has



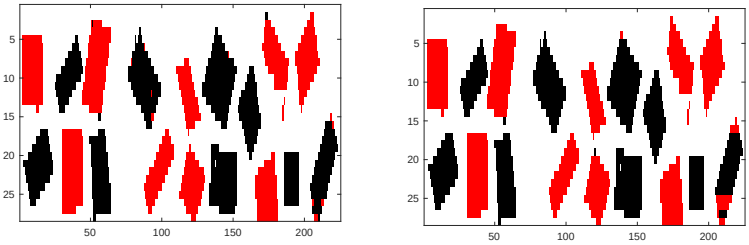
(a) SNR = 5 dB

(b) SNR = 0 dB.

FIGURE 4.11 – Total classification error rate (in %) versus λ_2 for the three proposed algorithms.

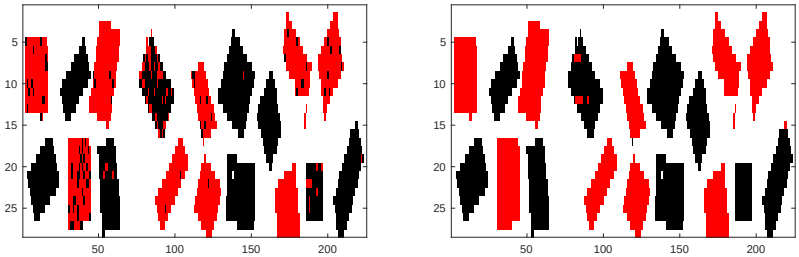


(a) $\lambda_2 = 0.$ (b) $\lambda_2 = 0.002.$

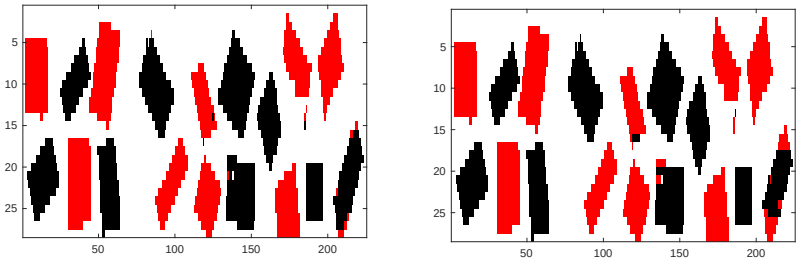


(c) $\lambda_2 = 0.005.$ (d) $\lambda_2 = 0.008.$

FIGURE 4.12 – Image reconstruction results using FSL algorithm for SNR = 5 dB.



(a) $\lambda_2 = 0.$ (b) $\lambda_2 = 0.002.$



(c) $\lambda_2 = 0.005.$ (d) $\lambda_2 = 0.008.$

FIGURE 4.13 – Image reconstruction results using FSGL algorithm for SNR = 5 dB.

TABLE 4.5 – Classification accuracy rates for four different noise levels

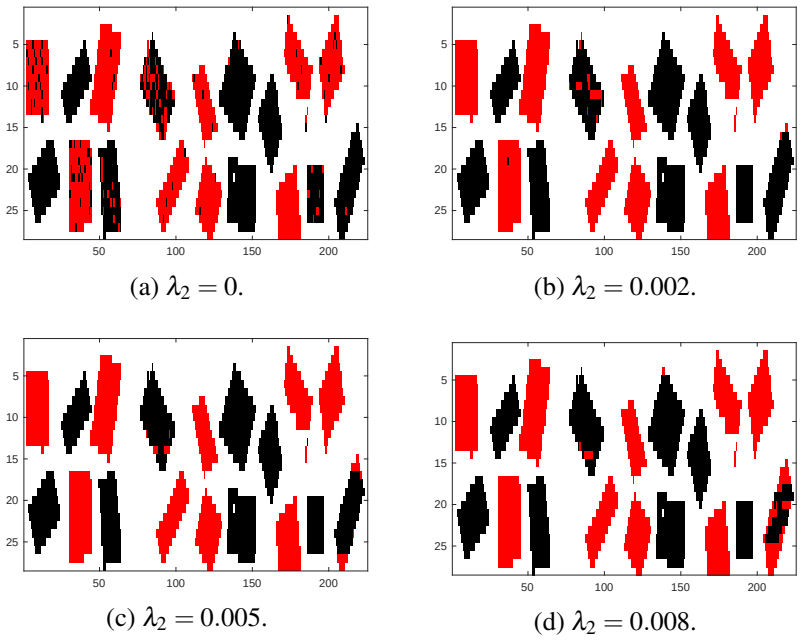
Algorithm	λ_2	SNR = -5 dB			SNR = 0 dB		
		Total	Class 1	Class 2	Total	Class 1	Class 2
FSL	0	64.8%	69.9%	59.6%	73.8%	76.3%	69.4%
	0.002	75.1%	76.5%	73.7%	85.9%	85.4%	86.3%
	0.005	77.7%	79.1%	76.2%	**88.0%**	**87.1%**	**88.8%**
	0.008	**80.4%**	**80.1%**	**80.6%**	85.3%	83.3%	87.3%
FSGL	0	64.14%	71.1%	57.2%	74.7%	78.2%	71.2%
	0.002	75.9%	76.6%	75.2%	85.6%	85.0%	86.1%
	0.005	78.0%	77.2%	79.1%	**87.3%**	**86.4%**	**88.2%**
	0.008	**79.2%**	**76.3%**	**82.0%**	86.0%	83.5%	88.5%
NN-FSGL	0	66.3%	70.7%	62.0%	75.1%	77.7%	72.4%
	0.002	74.4%	75.2%	73.6%	86.1%	84.8%	87.5%
	0.005	76.8%	75.4%	78.2%	**87.3%**	**85.5%**	**89.1%**
	0.008	**81.7%**	**78.0%**	**85.5%**	85.7%	81.8%	89.5%

Algorithm	λ_2	SNR = 5 dB			SNR = 10 dB		
		Total	Class 1	Class 2	Total	Class 1	Class 2
FSL	0	81.6%	83.0%	80.3%	87.0%	86.8%	87.2%
	0.002	89.2%	88.6%	89.7%	**89.9%**	**89.8%**	**89.9%**
	0.005	**89.5%**	**89.2%**	**89.7%**	89.5%	89.5%	89.6%
	0.008	88.5%	87.5%	89.5%	88.4%	87.1%	89.6%
FSGL	0	85.2%	84.2%	88.4%	88.7%	88.6%	88.7%
	0.002	**89.6%**	**89.4%**	**89.9%**	**89.9%**	**90.0%**	**89.8%**
	0.005	89.2%	88.6%	89.9%	89.3%	88.7%	89.9%
	0.008	87.3%	85.0%	89.6%	87.4%	85.3%	89.4%
NN-FSGL	0	83.2%	83.5%	83.0%	86.5%	83.2%	89.9%
	0.002	**89.4%**	**88.9%**	**89.9%**	**89.9%**	**90.0%**	**90.0%**
	0.005	88.9%	87.9%	89.9%	88.7%	87.5%	89.9%
	0.008	86.5%	83.2%	89.9%	87.2%	84.4%	89.8%

to be increased (i.e, $0.004 \leq \lambda_2 \leq 0.007$) in order to achieve a minimum classification error rates. This confirm that the classification errors caused by a high level of noise may be managed more efficiently by increasing the value of the regularization parameter. The images obtained from the labeled coefficients for four values of λ_2 in the case of SNR = 5 dB using FSL, FSGL and NN-FSGL algorithms are presented in Figures 4.12, 4.13 and 4.14, respectively. For each method we observe the salt-and-pepper appearance when no regularization is applied ($\lambda_2 = 0$), resulting from classification errors, while the reconstruction accuracy increases as the regularization become slightly stronger. In particular, we found that the best results are obtained for $\lambda_2 = 0.002$ and $\lambda_2 = 0.005$ for the three methods.

4.5 Conclusion

In this chapter, simultaneous regularized sparse approximation methods for feature selection are proposed. The idea is to reduce data dimensionality of NIR spectra using sparse decomposition. Also, to improve the classification performances, we incorporate a regularization constraint along the coefficient matrix rows to enforce a piecewise constant form. This is done by applying a ℓ_1-norm penalty on the derivative of successive coefficients differences. The corresponding algorithm is the sparse fused Lasso.



(a) $\lambda_2 = 0$. (b) $\lambda_2 = 0.002$.

(c) $\lambda_2 = 0.005$. (d) $\lambda_2 = 0.008$.

FIGURE 4.14 – Image reconstruction results using NN-FSGL algorithm for SNR = 5 dB.

Using a FISTA iteration, we have shown that the criterion may be solved efficiently thanks to the fused Lasso signal approximator (FLSA), applied on each row of the coefficient matrix. Additional constraints has been also incorporated to the criterion to preserve the simultaneous character of the problem and the non-negativity of the solution using respectively the mixed norm ℓ_1/ℓ_2 and the positivity constraint. The resulting algorithms has a low computational cost suitable for large-scale problems. Experiments conducted on NIR spectra of wood wastes validate the advantages of the proposed approaches for variables selection to increase the classification performances. Finally, we test the proposed algorithms in simulations of sorting material application where spatial information of hyperspectral images can be exploited. Results in terms of classification accuracy show that including a spatial regularization in the validation stage is suitable for such applications.

Chapitre 5

Automatisation du tri de bois : aspects industriels

5.1 Introduction

Dans ce dernier chapitre, nous nous proposons d'utiliser les méthodes développées afin de proposer des solutions à l'automatisation du tri de bois. L'objectif principal est de répondre aux besoins exprimés par EGGER concernant l'optimisation de son flux entrant de déchets de bois. Idéalement la société souhaite séparer le flux de déchets de bois entrant de la façon suivante [11] :

1. Bois purs,
2. Panneaux de particules purs,
3. Bois de recyclage,
4. Indésirables.

Les bois purs sont une catégorie qui doit être valorisée, les quantités de bois de cette catégorie doivent donc être augmentées. Les panneaux de particules sont séparés pour les faire passer directement au broyeur sans passer par le circuit de recyclage. Les autres groupes de bois sont orientés vers la zone de recyclage pour être traités. Il s'agit des bois qui n'ont pas étaient bien triés et qui doivent passer par des étapes de tri et de nettoyage avant d'être incorporés dans les panneaux de particules. Cette catégorie ne sera pas gérée dans le système actuel et continuera à suivre le même cheminement mis en place dans l'usine. Les indésirables sont tous les MDF/HDF ou bois pollués qui ne peuvent pas rentrer dans la filière de recyclage (voir Tableau 5.1). Nous proposons de mettre en place un schéma séquentiel de classification afin de séparer au fur et mesure les catégories des bois purs, les panneaux de particules et les indésirables. D'abord nous allons nous intéresser à la séparation des bois purs dans l'objectif d'en récupérer le maximum. Ensuite, les différents groupes de panneaux de particules sont séparés des indésirables qui doivent être rejetés.

Les tests sont réalisés au moyen du logiciel mis en place au CRAN. Cet outil possède différentes fonctionnalités, notamment pour le pré-traitement des données, la sélection de variables, et la classification

11. Source : Rapport intermédiaire du Crittbois pour le projet Trispirabois : Besoin d'EGGER, pages 23–25.

Catégorie	Groupes	# Ech.	Destination
Bois purs	Bois brut	96	Coupeuse
	Bois massif avec finition	79	
	Bois massif peint ou surfacé	140	
	Contreplaqué brut	120	
	Contreplaqué avec finition	129	
	Contreplaqué surfacé	26	
	OSB	119	
	Plaquettes	38	
Pan. particules	Panneaux de particules bruts	66	Broyeur lent
	Panneaux de particules surfacés	78	
	Panneaux de particules avec finition	12	
Indésirables	Panneaux MDF/HDF bruts	50	Rejet
	Panneaux MDF/HDF surfacés	129	
	Panneaux MFD/HDF avec finition	76	
	Panneaux fibres dures surfacés	66	
	Panneaux fibres dures bruts	14	
	Bois massifs sels-métalliques	46	
	Bois massifs traité (créosote)	33	

TABLE 5.1 – La base de données du Crittbois avec les différents groupes de déchets de bois, leur catégorie et leur destination.

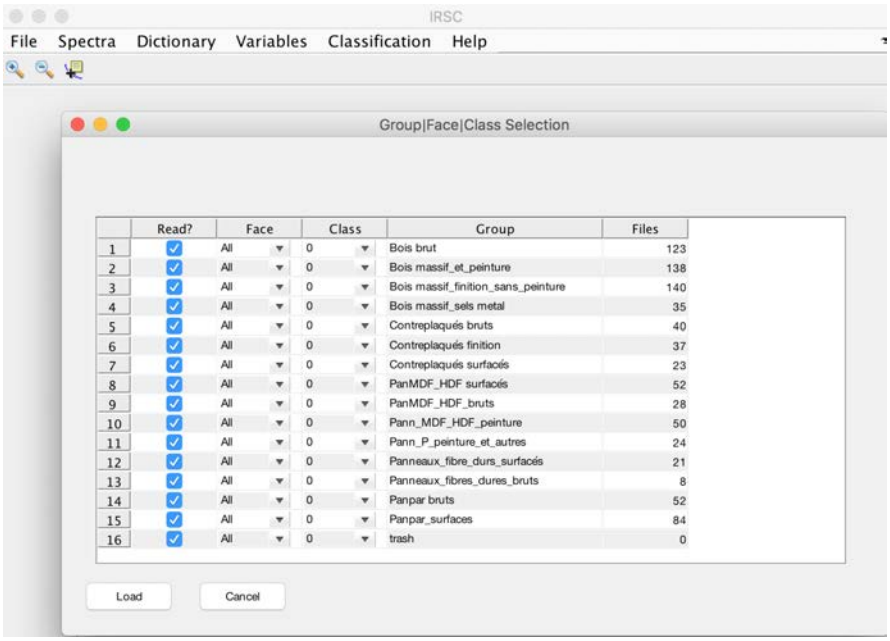
des échantillons de bois. Par ailleurs, l'évaluation des performances des méthodes proposées est faite ici sur les spectres acquis au Crittbois par le spectromètre industriel. La base de données spectrale obtenue et présentée dans le Tableau 5.1 est plus complète que celle du CRAN puisqu'elle contient 18 groupes de bois. Cependant les spectres de cette base ne se composent que de 1037 longueurs d'ondes contrairement à ceux du CRAN qui sont constitués de 1647 longueurs d'onde. Les tests réalisés sur cette base de données vont permettre de valider les méthodes de traitement développées et testées jusqu'ici sur les données du CRAN.

5.2 Description de l'outil de traitement

Le logiciel de traitement, développé sous Matlab dans le cadre de ce projet, est un outil permettant essentiellement la sélection d'un sous-ensemble de longueurs d'ondes pertinentes, et la classification des données spectrales. De plus, cet outil dispose de plusieurs fonctionnalités pour faciliter le traitement et l'interprétation des résultats. D'abord, la base de données que l'on souhaite traiter est chargée, en affectant une étiquette à chaque groupe de bois comme présenté sur la Figure 5.1. Une fois cette base chargée, l'utilisateur peut consulter les informations concernant le nombre de spectres dans chaque groupe et dans chaque classe. De plus, des possibilités de sauvegarde ou de chargement d'un modèle préalablement établit sont également disponibles dans ce premier onglet (*File*).

Ensuite, différents pré-traitements sont mis à disposition tels la suppression de la ligne de base, la normalisation ou encore le sous-échantillonnage des données. Cette deuxième rubrique (*Spectra*) permet également l'affichage des spectres en fonction de l'absorbance ou de la réflectance (Figure 5.2).

La construction du dictionnaire est proposée dans le troisième onglet (*Dictionary*), avec des paramètres fixés par défaut dans le logiciel. Ainsi, ce dictionnaire est formé par des fonctions gaussiennes



	Read?	Face	Class	Group	Files
1	☑	All	0	Bois brut	123
2	☑	All	0	Bois massif_et_peinture	138
3	☑	All	0	Bois massif_finition_sans_peinture	140
4	☑	All	0	Bois massif_sels metal	35
5	☑	All	0	Contreplaqués bruts	40
6	☑	All	0	Contreplaqués finition	37
7	☑	All	0	Contreplaqués surfacés	23
8	☑	All	0	PanMDF_HDF surfacés	52
9	☑	All	0	PanMDF_HDF_bruts	28
10	☑	All	0	Pann_MDF_HDF_peinture	50
11	☑	All	0	Pann_P_peinture_et_autres	24
12	☑	All	0	Panneaux_fibre_durs_surfacés	21
13	☑	All	0	Panneaux_fibres_dures_bruts	8
14	☑	All	0	Panpar bruts	52
15	☑	All	0	Panpar_surfaces	84
16	☑	All	0	trash	0

FIGURE 5.1 – Chargement des données spectrales et affectation des étiquettes aux différents groupes de bois analysés.

dont les centres varient de 3000 cm^{-1} à 10000 cm^{-1} et les largeurs de 30 à 600 cm^{-1} avec un pas de 30 cm^{-1}. Ces paramètres peuvent être modifiés en fonction des données traitées (une autre gamme spectrale par exemple).

La rubrique dédiée à la sélection de variables (*Variables*) permet de choisir entre quatre méthodes de sélection : *fused sparse Lasso*, *fused sparse group Lasso*, *smooth sparse Lasso* et *smooth sparse group Lasso*. Les différentes valeurs des paramètres de parcimonie et de régularisation ainsi que le nombre d'itérations sont fixées en fonction du nombre de variables que l'on souhaite retenir. Les versions non-négatives de ces algorithmes sont également proposées. Par ailleurs, différents types d'affichage des longueurs d'ondes sélectionnées peuvent être consultés. Enfin, la rubrique consacrée à la classification, permet de choisir différentes configurations pour le choix des ensembles d'apprentissage et de validation : on peut ainsi découper la base en deux ensembles : l'ensemble d'apprentissage et l'ensemble de validation ou choisir la technique de validation croisée avec la méthode K-fold. Les paramètres du modèle SVM sont également disponibles et peuvent éventuellement être modifiés. Le modèle proposé par défaut est le SVM à noyau quadratique.

5.3 Tests réalisés

Nous proposons de réaliser une série de tests permettant de valider les techniques d'analyse et de traitement développées tout au long de ce travail de thèse. Ainsi, nous pouvons vérifier jusqu'à quel point les objectifs posés initialement par les industriels ont été atteints. Notons que l'acquisition des spectres est faite une seule pour tous les tests. Pour chaque test, les étapes suivantes sont réalisées : une fois la base de données chargée, on réalise les pré-traitements suivant sur les spectres :

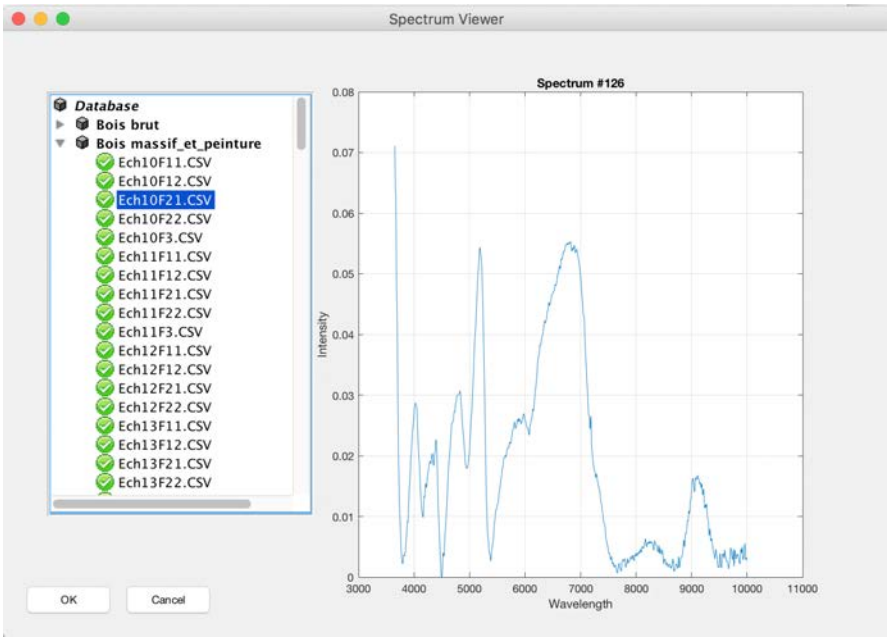


FIGURE 5.2 – Rubrique pour le pré-traitement des données spectrales. Ici, les spectres sont affichés en fonction de la réflectance.

— Suppression de la ligne de base,
— Normalisation (spectres non-négatifs de norme unité).

Puis le dictionnaire gaussien est construit avec les paramètres fixés par défaut. Le même dictionnaire avec les mêmes paramètres sera utilisé pour les différents tests. La configuration de la sélection de variables et de l'apprentissage du modèle pour les différents tests est mise en place en amont et de façon définitive. Elle se présente comme suit :

— La sélection de variables est effectuée par la méthode *fused sparse group Lasso* et 32 longueurs d'ondes sont retenues à chaque fois. Le choix du nombre des longueurs d'onde est justifié par les machines de tri de Pellenc ST. En effet, ces machines intégrant 16 voies, la possibilité de mettre en place deux scanners peut être envisagée afin d'assurer une meilleure efficacité du traitement.
— L'apprentissage du modèle de classification est réalisé sur ces 32 variables en utilisant le SVM à noyau quadratique en adoptant deux schémas différents :

1. **Schéma 1** : en considérant 20% de l'ensemble de la base comme ensemble d'apprentissage. La validation est réalisée sur les 80% des données restantes.

2. **Schéma 2** : en réalisant une validation croisée avec la méthode K-fold. Cette technique consiste à estimer les performances de classification à partir d'échantillons n'ayant pas servi à la conception du modèle. Pour ce faire, on scinde la base d'apprentissage en K parties de taille égale. L'apprentissage du modèle est réalisé avec $K-1$ sous-ensembles et la validation sur le dernier sous-ensemble. Ce processus est répété K fois. Pour chaque partie utilisée à la validation on calcule l'erreur de classification moyenne (e_k). Le score de la validation croisée correspond à la moyenne des erreurs e_k. Ici, le nombre de partitions est fixé à $K = 10$.

Nous proposons de mettre en place trois modèles. Chaque modèle retenu pourra être chargé et utilisé en temps réel dans une chaîne de classification des bois.

5.3.1 Récupération des bois purs

Un des premiers objectifs du projet concerne l'augmentation des taux de bois purs dans les flux de déchets de bois entrants. Le schéma présenté sur la Figure 5.3 décrit le déroulement de l'essai qui correspondant à cet objectif. Ce dernier consiste à séparer deux catégories de bois : les bois purs et les bois à traiter, avec comme objectif principal la récupération de quantités maximales de bois purs. Ceci doit se traduire par un taux de bonne classification maximale de la première classe tout en sachant qu'actuellement, la société EGGER récupère environ 20% de bois purs dans ses flux de déchets de bois entrants. On commence par charger les différents groupes de bois de la base de données en attribuant

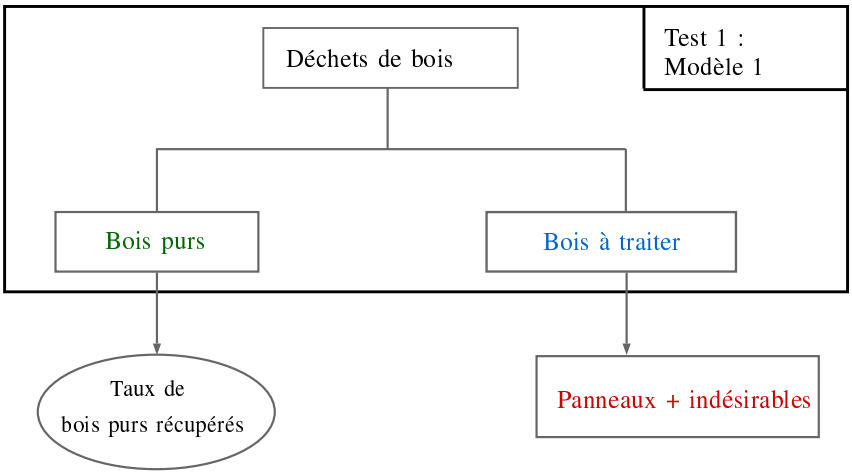


FIGURE 5.3 – Schéma descriptif du test 1 qui consiste à récupérer les bois purs.

l'étiquette 0 à tous les groupes de bois purs, et l'étiquette 1 aux autres groupes de bois à savoir les panneaux de particules et les indésirables. La matrice de données $\mathbf{Y}$ est constituée de 1317 spectres au total, dont 747 spectres appartenant à la classe 1 et 570 appartenant à la classe 2 ($\mathbf{Y} \in \mathbb{R}^{1037\times1317}$). La matrice de coefficient est obtenue en utilisant la méthode *fused sparse group Lasso* avec $\lambda_1 = \lambda_2 = \lambda_3 = 0.14$ telle que $\mathbf{X} \in \mathbb{R}^{32\times1317}$. Les résultats obtenus pour ce premier test sont présentés sur le Tableau 5.2.

	Taux de bonne classification	Résultats
Ensembles apprentissage- validation	Totale	80.25%
	Classe 1	85.26%
	Classe 2	73.68%
Validation croisée	Totale	88.69%
	Classe 1	92.11%
	Classe 2	84.21%

TABLE 5.2 – Résultats de classification du test 1.

Nous obtenons environ 80% et 88% de taux de bonne classification totale respectivement pour les deux schémas de classification. Aussi, nous pouvons constater que les groupes de bois purs sont bien classifiés à plus de 85%. Les résultats de la validation croisée permettent de valider les performances de classification sur la totalité de la base et de confirmer la robustesse du modèle mis en place puisque le taux de bonne classification des bois purs est supérieur à 90%. Ainsi, on peut estimer que le gain apporté en utilisant le traitement proposé est d'environ 60 à 70% de bois purs récupérés en plus par rapport aux taux actuels. Afin d'analyser quels sont les groupes de bois qui induisent le plus d'erreur, nous détaillons

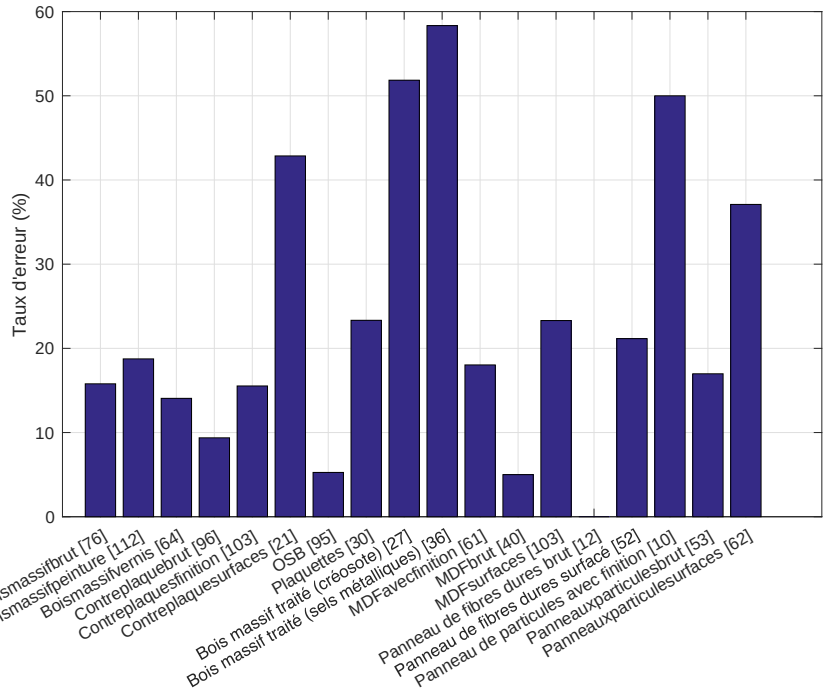


FIGURE 5.4 – Histogramme des taux d'erreur de chaque groupe de bois pour le test 1.

sur la Figure 5.4 les taux d'erreur obtenus pour chaque groupe de bois. Cet histogramme concerne les résultats obtenus pour le schéma 1. D'abord, on constate que les groupes de bois les plus difficiles à classer sont les bois traités tels les bois massifs avec du créosote ou des sels métalliques. Ceci peut s'expliquer par le fait que ce type de polluant n'est pas assez visible en proche-infra rouge. Aussi, des groupes de bois peints ou surfacés induisent des erreurs de classification dues au fait que le revêtement ne concerne qu'une face des échantillons en général ce qui peut parfois induire une confusion lors de l'apprentissage lorsqu'un même échantillon possède une face non revêtue (un bois ou un panneau brut) et l'autre face surfacée (tissu, plastique, etc).

Par ailleurs, les taux d'erreur sur les autres groupes varient entre 0 et 25% à l'exception des panneaux de particules avec finition dont le taux d'erreur est de 50%. Ceci peut s'expliquer par la faible représentativité de cette catégorie (10 individus). Par ailleurs, il est important d'évaluer l'impact des groupes de bois surfacés et les massifs avec du créosote ou des sels métalliques. Pour ce faire, nous procédons à deux tests supplémentaires afin d'étudier l'influence de ces types de bois sur les performances globales. Dans le premier test nous allons supposer que les massifs avec la créosote ou les sels métalliques sont

	Taux de bonne classification	Résultats
Ensembles apprentissage- validation	Totale	84.75%
	Classe 1	88.44%
	Classe 2	79.13%
Validation croisée	Totale	92.56%
	Classe 1	94.95%
	Classe 2	89.61%

TABLE 5.3 – Résultats de classification du test 1.1.

gérés en amont. Ainsi, notre analyse se fera sur tous les autres groupes de la même manière que présentée sur la Figure 5.3 mais sans considérer ces deux groupes. Dans un deuxième temps, nous proposons un autre test pour les bois surfacés qui consiste à considérer uniquement la face revêtue lors du chargement des données afin de mieux caractériser ces groupes de bois. Ce test nous permettra de vérifier si on peut obtenir de meilleurs résultats de classification, principalement pour la classe 1.

5.3.2 Traitement sans les bois massifs avec du créosote ou des sels métalliques

Nous avons constaté lors du test précédent que l'erreur de classification sur les groupes de bois massifs avec de la créosote ou des sels métalliques est très importante (plus de 50%). Ces groupes ont peut-être des caractéristiques spectrales qui ne peuvent pas être identifiées par la spectrométrie proche infra-rouge, ce qui expliquerait qu'ils soient souvent confondus avec des bois massifs bruts ou avec finition. Nous proposons dans ce test de ne pas considérer ces deux groupes de bois lors du chargement de la base de données et de procéder à la même analyse que le test 1. La matrice de données est constituée de 1238 spectres dont 747 spectres appartenant à la classe 1 et 491 à la classe 2 telle que $\mathbf{Y} \in \mathbb{R}^{1037\times1238}$. La matrice de coefficient est obtenue par la méthode de décomposition parcimonieuse *fused sparse group Lasso* avec $\lambda_1 = \lambda_2 = \lambda_3 = 0.14$ telle que $\mathbf{X} \in \mathbb{R}^{32\times1238}$.

Les résultats présentés dans le Tableau 5.3 montrent que nous pouvons atteindre de meilleurs taux de classification. D'abord, pour la classe 2, on a réussi à augmenter d'environ 6% les taux de rejet des bois pollués pour les deux schémas de classification. Ce résultat implique que la classe des bois purs contient moins de bois indésirables et est donc plus propre. Cet amélioration concerne également les résultats de la classe 1 puisque les taux de classification pour cette classe ont augmenté par rapport au test précédent (entre 2% et 4% de gain en plus).

En visualisant les taux d'erreur commis sur chaque groupes de bois présentés sur la Figure 5.5 on constate que plusieurs groupes sont parfaitement classés. Une remarque intéressante à faire ici, concerne les taux d'erreur sur les groupes de bois massifs bruts, avec finition et avec peinture qui ont tous diminué en comparaison avec le test 1 comme présenté sur le Tableau 5.4.

Globalement plusieurs groupes de bois de la classe 1 sont mieux classés. Ces résultats confirme l'hypothèse que les 3 groupes de bois massifs sont souvent confondus avec les bois massifs avec la créosote et les sels métalliques, puisqu'ils partagent les mêmes caractéristiques spectrales relatives à l'essence de bois mais que celles qui pourraient les discriminer (relatives aux polluants) sont imperceptibles, ce qui

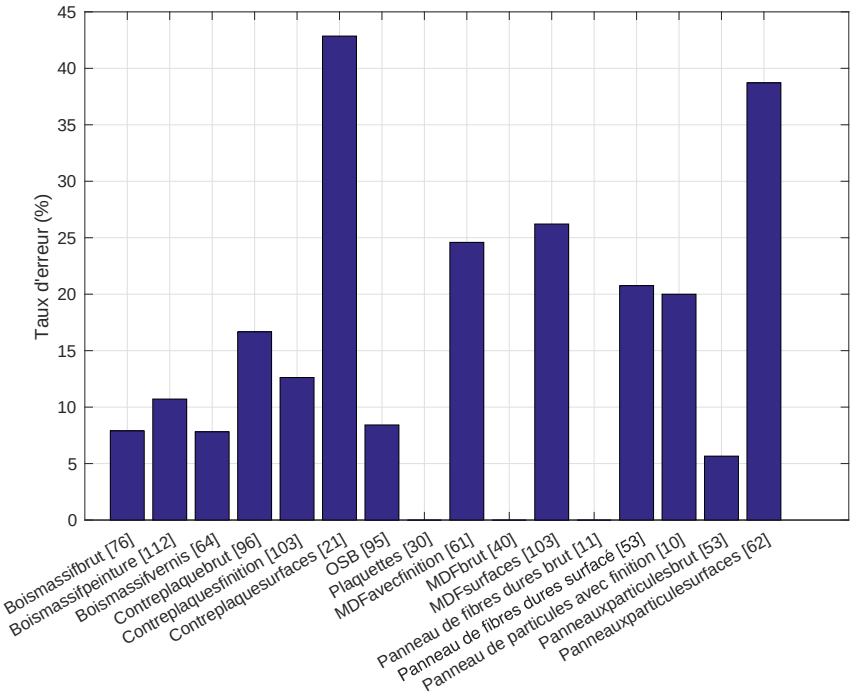


FIGURE 5.5 – Histogramme des taux d'erreur de chaque groupe de bois pour le test 1.1.

	Nombre d'éch.	Taux d'erreur Test 1	Taux d'erreur Test 1.1
Bois massifs bruts	76	16%	7%
Bois massifs vernis	64	14%	11%
Bois massifs peints	112	19%	7%
Erreur moyenne	-	17%	8%

TABLE 5.4 – Comparaison des taux d'erreur de classification sur les groupes de bois massifs dans le test 1 et le test 1.1.

conduit à une difficulté à bien les séparer. Une autre hypothèse possible concerne la gamme spectrale utilisée par le spectromètre du Crittbois. En effet, ce spectromètre couvre une gamme dans l'intervalle [4000 cm^{-1} 10000 cm^{-1}] avec seulement 1037 points par spectre à la différence des spectres obtenus en utilisant le spectromètre de LCPME qui couvre l'intervalle [3556 cm^{-1} 10000 cm^{-1}] avec 1647 points par spectre. Nous avons réalisé le même test que le test 1 (récupération des bois purs) sur la base de données du CRAN, et nous avons trouvé que le taux d'erreur commis sur les bois massifs avec des sels métalliques est de 32% seulement alors que sur celle du Crittbois l'erreur est de plus de 70%. Il est donc possible que les longueurs d'ondes en dessous de 4000 cm^{-1}, mais aussi d'autres longueurs d'ondes qui n'ont pas été acquises par le dispositif du Crittbois apportent plus d'informations sur les caractéristiques spectrales de ces types de polluants. Par conséquent, il est nécessaire d'envisager une gamme spectrale plus grande ou encore une autre technique d'acquisition pour analyser ce type de déchets de bois afin de les rejeter avant d'entamer la récupération des bois purs.

	Taux de bonne classification	Résultats
Ensembles apprentissage- validation	Totale	87.97%
	Classe 1	93.02%
	Classe 2	76.63%
Validation croisée	Totale	92.84%
	Classe 1	95.09%
	Classe 2	87.79%

TABLE 5.5 – Résultats de classification du test 1.2.

5.3.3 Traitement des bois surfacés

Une autre remarque que nous avons faite lors des tests précédents concerne les groupes de bois surfacés. Les revêtements de ces types de déchets induisent une plus grande erreur de classification car un type de revêtement donné peut être le même pour deux ou plusieurs groupes de déchets de bois. De plus, les échantillons de bois de ces groupes ne sont souvent pas complètement surfacés dans la mesure ou le revêtement n'existe que sur une face et pas toutes les faces. Cette remarque a été faite lors de l'acquisition des échantillons, et le Crittbois a fait le choix de faire correspondre la face A à la face revêtue comme présenté sur la Figure 5.6. Nous proposons ici de réaliser un test qui consiste à considérer

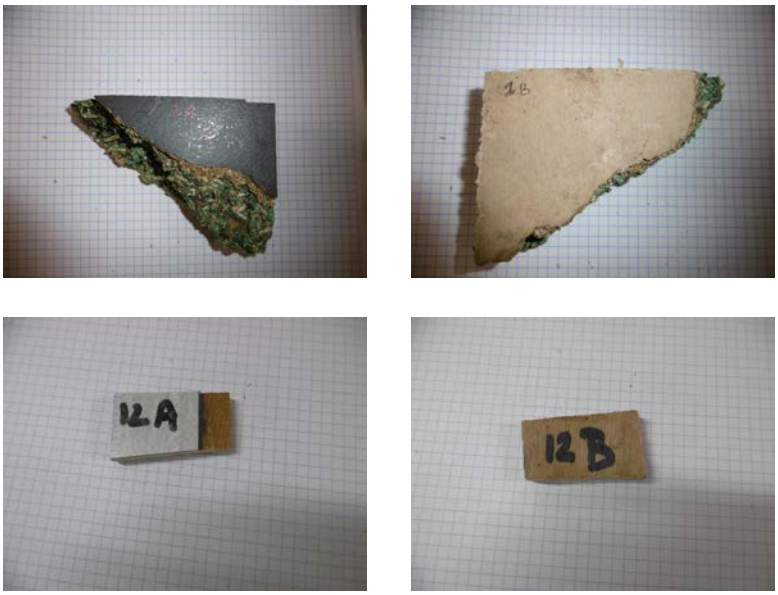


FIGURE 5.6 – Exemple d'échantillons de bois surfacés. La face A est celle avec un revêtement.

uniquement la première face (face A) des groupes surfacés. Lors de ce test, la matrice de données est constituée de 1061 spectres dont 721 spectres appartenant à la classe 1 et 327 à la classe 2 telle que $\mathbf{Y} \in \mathbb{R}^{1037\times1238}$. La matrice de coefficient est obtenue par la méthode de décomposition parcimonieuse *fused sparse group Lasso* avec $\lambda_1 = \lambda_2 = \lambda_3 = 0.13$ telle que $\mathbf{X} \in \mathbb{R}^{32\times1061}$. L'objectif est de vérifier si on arrive à diminuer les taux d'erreur de classification sur ce type de groupes. Les résultats de classification de ce test sont présentés dans le Tableau 5.5.

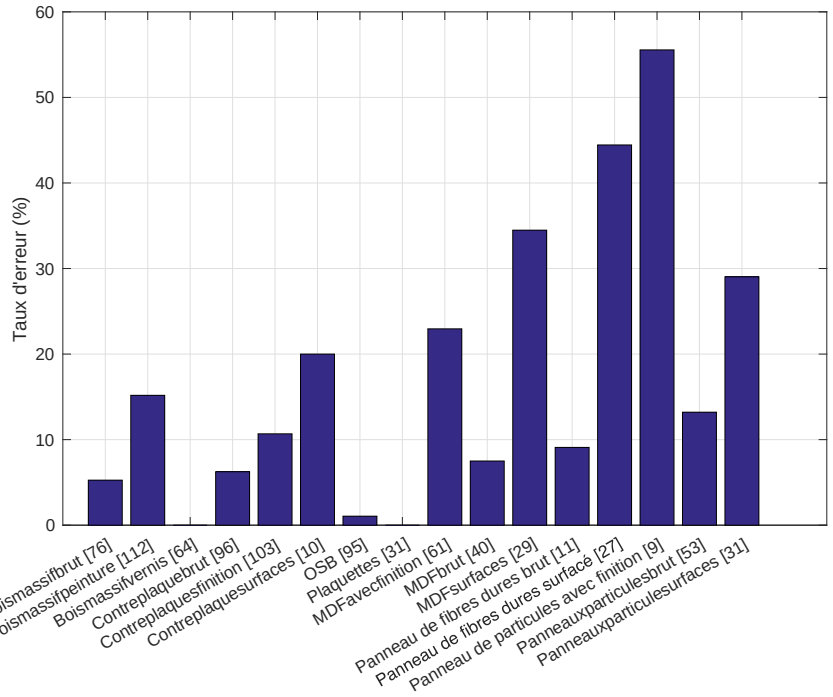


FIGURE 5.7 – Histogramme des taux d'erreur de chaque groupe de bois pour le test 1.2.

	Nombre d'éch.	Taux d'erreur Test 1.1	Taux d'erreur Test 1.2
Panneaux de particules surfacés	21	38%	29%
Panneaux de fibres dures surfacés	103	21%	43%
Panneaux MDF/HDF surfacés	53	26%	33%
Contreplaqués surfacés	62	43%	20%
Erreur moyenne	-	29%	33%

TABLE 5.6 – Comparaison des taux d'erreur de classification sur les groupes de bois surfacés dans le test 1.1 et le test 1.2.

D'abord, on constate que les taux de bonne classification sont supérieurs à ceux obtenus dans le test 1. En particulier, les taux de bonne classification de la classe 1 sont supérieurs d'environ 8% pour le schéma 1. Par ailleurs, les taux en validation croisée sont nettement meilleurs que ceux obtenus lors du schéma 1 pour la classe 2 pour laquelle nous avons obtenue environ 88% de taux de bonne classification, alors que ce taux est de 76% seulement dans le schéma 1. En fait, en examinant les taux d'erreur commis dans chaque groupe de bois (Figure 5.7), on remarque que ce dernier est le plus élevé dans le groupe des panneaux de particules avec finition (60%). Ce résultat est justifié par le faible nombre d'échantillons dans ce groupe (10 échantillons seulement). Globalement, les résultats de ce test sont meilleurs que ceux obtenus lors du test 1, néanmoins ils sont assez comparable à ceux du test 1.1. Nous présentons dans le Tableau 5.6 une comparaison entre les taux d'erreur obtenus pour ces bois dans le test 1.1 et ceux obtenus dans le test 1.2 (en considérant uniquement la face A).

5.3.4 Rejet des bois indésirables

Un deuxième besoin exprimé par la société EGGER, et illustré sur le schéma de la Figure 5.8, concerne la séparation des panneaux de particules des indésirables. Ce besoin implique le rejet d'une quantité maximale de bois indésirables dans les flux de bois qui sont acheminés vers la zone de recyclage. En effet, ces types de bois présentent plusieurs inconvénients. Premièrement, en considérant que les déchets de bois dans la zone de recyclage contiennent des bois créosotés ou avec des sels métalliques, ces derniers ne peuvent être tolérés que très peu dans le flux des bois recyclés puisqu'ils sont considérés comme des bois pollués. Par conséquent, il est nécessaire d'en rejeter le maximum. Aussi, des groupes de bois dans la catégorie des indésirables comme les panneaux MDF/HDF, ont des mauvaises propriétés mécaniques pour le recyclage. Ils doivent également être rejetés. En effet, le panneau MDF/HDF est fabriqué à partir de fibre de bois. Ainsi, lorsqu'il est broyé, on en obtient en partie des morceaux de panneaux et de la fibre. Ce type de bois produit d'autant plus de fibres lorsqu'il est finement broyé. Les fibres de bois peuvent polluer les aspirations, créent des bouchons et provoquent des bourrages dans le tamis dédié à la récupération des poussières d'aspirations. Par conséquent, un arrêt de la production est inévitable.

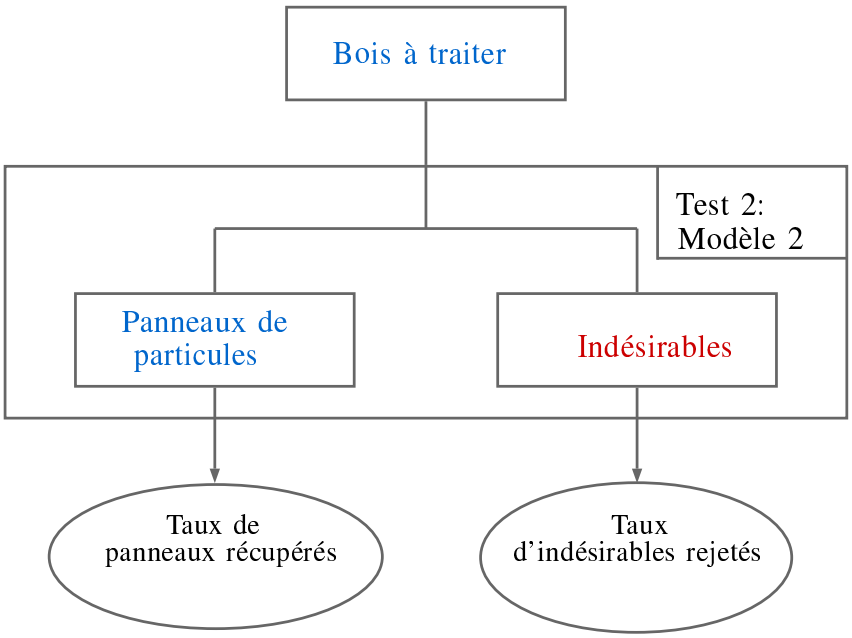


FIGURE 5.8 – Schéma du deuxième test qui vise à rejeter les bois indésirables.

L'autre aspect géré lors du test 2, et qui a également été exprimé pas les industriels, concerne les panneaux de particules. Ces groupes de bois ont une composition (petites particules de bois) qui produit énormément de poussière. De plus, les panneaux de particules ont tendance à user les couteaux des broyeurs. Par conséquent, ce type de bois doit être séparé pour être orienté vers un système de broyeur lent.

Pour réaliser ce test, nous commençons par charger les bois qui doivent être traiter, à savoir les groupes

	Taux de bonne classification	Résultats
Ensembles apprentissage- validation	Totale	82.89%
	Classe 1	64.52%
	Classe 2	89.76%
Validation croisée	Totale	92.81%
	Classe 1	85.33%
	Classe 2	95.66%

TABLE 5.7 – Résultats de classification du test 2.

de panneaux de particules et les groupes de bois indésirables, puis on procède à un nouvel étiquetage des données en fonction des catégories que l'on souhaite séparer. Ainsi, on affecte l'étiquette 0 aux groupes de la catégorie des panneaux de particules, et l'étiquette 1 aux groupes de la catégorie des bois indésirables. La matrice de données résultante est constituée de 570 spectres au total, dont 156 spectres appartenant à la classe 1 (panneaux de particules) et 414 appartenant à la classe 2 (indésirables). La matrice de coefficient est obtenue par l'algorithme *fused sparse group Lasso* avec $\lambda_1 = \lambda_2 = \lambda_3 = 0.11$ telle que $\mathbf{X} \in \mathbb{R}^{32\times570}$. Les résultats du test 2 sont reportés dans le Tableau 5.7. On constate que ce deuxième tri nous permet de rejeter environ 90% et 96% des bois indésirables, pour les schémas 1 et 2 respectivement. De plus, on a trouvé que les panneaux de particules sont bien séparés également en particulier pour en validation croisée (85% de bonne classification).

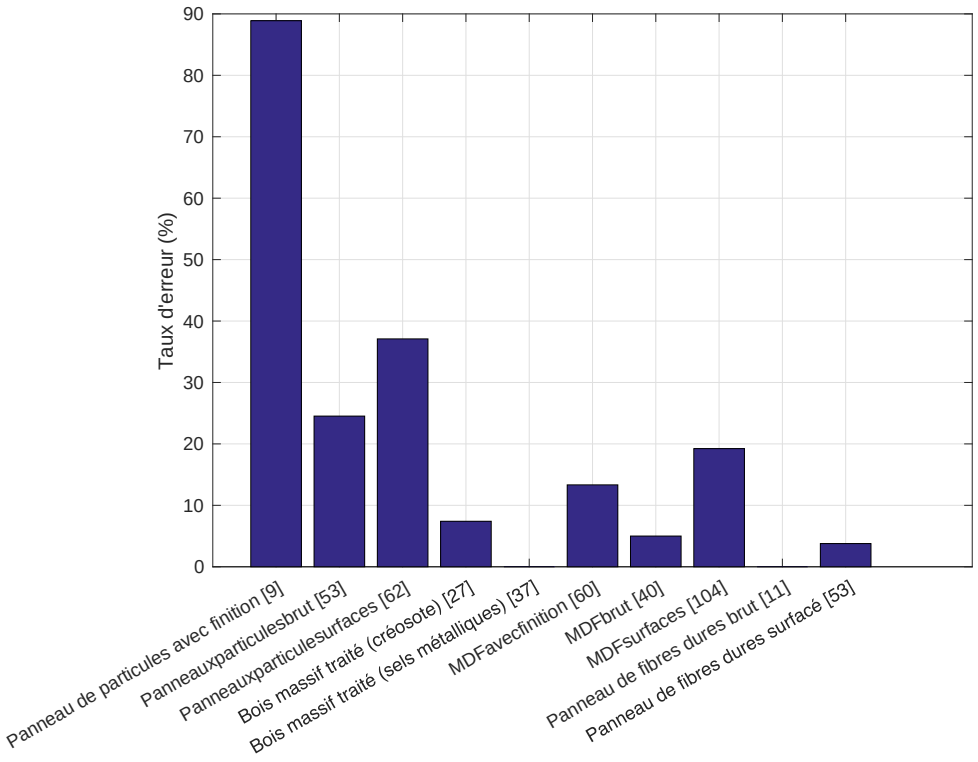


FIGURE 5.9 – Histogramme des taux d'erreur de chaque groupe de bois pour le test 2.

On présente sur la Figure 5.9 les taux d'erreur commis sur chaque groupe de bois. On constate que les taux de rejet des panneaux MDF/HDF varient entre 5% et 20% . Ce résultat confirme que ce type de bois peut être traités efficacement. Aussi, les groupes de panneaux de fibre dures sont très bien classés avec 0% et 4% d'erreur de classification respectivement pour les bruts et les surfacés. Par ailleurs, on note sur la la même figure que le groupe qui induit le plus d'erreur est le groupe des panneaux de particules avec finition, mais cela est dû au faible nombre d'échantillons dont on dispose (9 échantillons). Lorsque nous avons réalisé le même test sur la base de données du CRAN qui contient 24 échantillons dans le groupe de panneaux de particules avec finition, nous avons obtenu de meilleurs résultats que ceux obtenu ici, avec environ 32% de taux d'erreur seulement sur ce groupe de bois. Enfin, nous remarquons qu'en l'absence des bois massifs de la classe 1 (bois purs), les groupes de bois massifs avec la créosote et les sels métalliques sont très bien classés avec 5% d'erreur de classification pour le premier groupe et 0% pour le deuxième.

5.4 Conclusion

Nous avons présenté dans ce dernier chapitre des techniques de traitement pour l'automatisation et l'optimisation du tri de bois. En effet, on a montré qu'il est possible d'apporter des améliorations significatives au systèmes de tri de bois qui existent actuellement grâce à l'outil de traitement permettant la sélection des variables pertinentes pour la classification. Ces améliorations interviennent sur deux niveaux. Premièrement, la sélection de variables par les approches proposées permet de réduire considérablement le temps de traitement. Cet aspect est essentiel pour le tri en ligne des déchets de bois. De plus, nous avons également réussi à augmenter considérablement les taux de bois qui doivent être valorisés dans les systèmes de tri tout en diminuant fortement les déchets de bois indésirables. Néanmoins, nous constatons que les solutions proposées jusque là doivent encore répondre à plusieurs contraintes que nous résumons comme suit :

1. Le nombre des longueurs d'ondes est plus important que celui proposé par les industriels des machines de tri. Cette contrainte ne peut être dépassée actuellement avec les méthodes de traitement que nous avons développées sans perdre en efficacité. Une solution éventuelle consiste à aligner deux scanners afin d'atteindre 32 voies. Cependant cette solution engendre nécessairement des coûts supplémentaires ;
2. Le stockage des bois pour une certaine durée et dans certaines conditions peut provoquer une augmentation des taux d'humidité dans les bois. Ce processus peut engendrer des changements des propriétés spectrales des bois stockés et par conséquent affecter les résultats du traitement ;
3. L'apprentissage du modèle doit être mis à jour avec une certaine fréquence. Cette fréquence peut être déterminée lors des futurs tests qui vont être réalisés sur le terrain.

De plus, les tests que nous avons menés nous montrent que les méthodes de traitement mises en place dans le cadre ce projet continuent de souffrir de quelques limitations, à savoir :

1. Les déchets verts, les mousses et les plastiques ne peuvent être analysés par les méthodes développées. Il doivent continuer à être traités de la même manière que dans le système actuel ;

2. Les polluants du type peinture en plomb, la créosote ainsi que les sels métalliques doivent éventuellement être traités avec une autre technologie que la spectrométrie IR (par exemple machine à rayon X) car leur propriétés spectrales sont imperceptibles dans le proche-infra rouge ;
3. Les bois surfacés provoquent inévitablement plus d'erreurs dans le traitement à cause de la nature et de l'épaisseur du revêtement.

Conclusion générale

1 Bilan

Nous avons traité dans ce mémoire des méthodes de traitement de données pour la classification de spectres infra-rouges de déchets de bois afin de répondre aux besoins d'automatisation et d'optimisation de la gestion des déchets de bois.
Ainsi, nous avons commencé ce document par décrire le contexte de l'étude, ses objectifs ainsi que le type de données qui nous nous proposons d'exploiter. Aussi, dans le cadre du projet Trispirabois nous nous somme intéressés aux données issues d'un processus d'acquisition proche infra-rouge. Ce type d'acquisition fournit une bonne description de la composition d'un matériau donné. Enfin, nous avons présenté toute la partie qui concerne la collecte et la préparation des données. Cette première tâche nous a permis de construire une base de données des spectres IR de déchets de bois.
Le problème initial est ainsi formulé comme un problème de classification permettant de déterminer si un spectre d'un échantillon donné doit être récupéré ou rejeté. Par conséquent, nous présentons dans le deuxième chapitre un éventail de méthodes de classification utilisées dans divers domaines liés à la spectrométrie. Ces méthodes consistent essentiellement en des techniques de clustering et d'exploration de données, des méthodes de classification non supervisée et supervisée. Cette étude nous a permis de tester et de comparer plusieurs approches de classification, et à terme, de déterminer la méthode qui est la mieux adaptée pour résoudre notre problématique.
Dans le troisième chapitre nous avons abordé plusieurs approches pour l'approximation parcimonieuse. Ce chapitre est en réalité un point de départ, qui nous permis de présenter plusieurs méthodes basées sur la représentation parcimonieuse dans le but de les exploiter afin de faire de la sélection de variables. Ainsi, nous avons pu développer des méthodes gloutonnes pour l'approximation parcimonieuse simultanée, notamment les algorithmes S-CoSaMP, S-OLS et S-SBR. L'intérêt de la simultanéité a été démontré dans le cadre de simulations que nous avons réalisées pour la reconstruction d'une matrice de signal.
Le quatrième chapitre du document propose des contributions algorithmiques dans le cadre de la sélection de variables pour la classification des spectres proches infra-rouges et d'images hyperspectrales de déchets de bois. L'idée principale de ce travail consiste à développer des méthodes pour l'approximation parcimonieuse simultanée avec des critères qui incluent un terme de régularisation qui permet d'imposer une forme constante par morceaux le long des lignes de la matrice de coefficients. Cette formulation est utilisée pour décrire les spectres appartenant à la même classe et qui doivent partager des caractéristiques spectrales similaires (constants). Les algorithmes développés permettent d'aboutir à une résolution ef-

ficace du problème de décomposition en utilisant une technique d'optimisation de descente de gradient du type FISTA. Les performances de ces méthodes pour la sélection de variables sont évaluées dans le cadre de la classification des données spectrales et des images hyperspectrales de déchets de bois.
Dans le dernier chapitre nous nous sommes focalisés sur les améliorations qui peuvent être apportées aux systèmes de tri utilisés actuellement en industrie. Ces avantages sont illustrés au moyen d'une solution logicielle que nous avons développée dans le cadre de cette thèse. L'outil mis en place permet d'exploiter les méthodes de traitement proposées pour, d'une part, augmenter les taux de bois purs dans les processus de recyclage et d'autre part, rejeter une plus grande quantité de bois indésirables, présents dans les flux de déchets de bois à traiter.

2 Perspectives

D'un point de vue efficacité, les performances des méthodes de traitement développées peuvent être améliorées en ajoutant la dimension spatiale à la dimension spectrale. En effet, il apparaît que la seule signature spectrale ne suffit pas à totalement discriminer les différents types de déchets de bois. D'une part, une solution peut être de combiner cette signature spectrale et l'information de texture afin d'augmenter les taux de bonne classification.
Par ailleurs, nous avons montré dans le quatrième chapitre, que les méthodes conçues dans le cadre de ce projet peuvent parfaitement être exploitées pour imposer une régularisation spatiale sur les images hyperspectrales d'autant plus que le système d'imagerie hyperspectrale industriel développé par Pellenc-ST peut parfaitement intégrer ces méthodes. D'autre part, une amélioration des performances basée sur des méthodes de déconvolution adaptative est également en cours d'élaboration au sein du CRAN. Les techniques développées permettraient d'augmenter la résolution des images hyperspectrales tout en maintenant un temps d'acquisition plus faible.

Bibliographie

[1] M. J. Adams. *Chemometrics in analytical spectroscopy*. Royal Society of Chemistry, Cambridge (UK), 1995.

[2] A. Ahmed and E. P. Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29) :11878–11883, 2009.

[3] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina. Gasoline classification using near infrared (nir) spectroscopy data : comparison of multivariate techniques. *Analytica Chimica Acta*, 671(1) :27–35, 2010.

[4] R. Beale and T. Jackson. *Neural Computing : An introduction*. Adam Hilger, Bristol, 1990.

[5] A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11) :2419–2434, 2009.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1) :183–202, 2009.

[7] L. Belmerhnia, E.-H. Djermoune, and D. Brie. Greedy methods for simultaneous sparse approximation. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 1851–1855, 2014.

[8] L. Belmerhnia, E.-H. Djermoune, C. Carteret, and D. Brie. Simultaneous regularized sparse approximation for wood wastes NIR spectra features selection. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 437–440. IEEE, 2015.

[9] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1) :23–34, 1992.

[10] J. D. Blanchard, C. Cartis, J. Tanner, and A. Thompson. Phase transitions for greedy sparse approximation algorithms. *Applied and Computational Harmonic Analysis*, 30(2) :188–203, 2011.

[11] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jing. Greedy algorithms for joint sparse recovery. *IEEE Transactions on Signal Processing*, 62 :1694–1704, 2014.

[12] J. D. Blanchard and A. Thompson. On support sizes of restricted isometry constants. *Applied and Computational Harmonic Analysis*, 29(3) :382–390, 2010.

[13] T. Blumensath and M. Davies. On the difference between orthogonal matching pursuit and orthogonal least squares. Technical report, [Online] Available at : http://www.personal.soton.ac.uk/tb1m08/papers/BDOMPvsOLS07.pdf, 2007.

[14] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3) :265–274, 2009.

[15] T. Blumensath and M. E. Davies. Normalized iterative hard thresholding : Guaranteed stability and performance. *IEEE Journal of Selected Topics in Signal Processing*, 4(2) :298–309, 2010.

[16] M. A. Bouslamti. *Identification et évaluation des différents types et niveaux des contaminants chimiques dans les bois recyclés*. PhD thesis, École Supérieur de Bois, Nantes, 2012.

[17] P. S. Bradley and O. L. Mangasarian. Massive data discrimination via linear support vector machines. *Optimization Methods and Software*, 13(1) :1–10, 2000.

[18] N. Bratchell. Cluster analysis. *Chemometrics and Intelligent Laboratory Systems*, 6(2) :105–125, 1989.

[19] R. Brereton. *Chemometrics : data analysis for the laboratory and chemical plant*. John Wiley & Sons, Chichester, 2003.

[20] C. C. Bridges. Hierarchical cluster analysis. *Psychological Reports*, 18(3) :851–854, 1966.

[21] C. J. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2) :121–167, 1998.

[22] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg. OPLS discriminant analysis : combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 20(8-10) :341–351, 2006.

[23] T. T. Cai, G. Xu, and J. Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55(7) :3388–3397, 2009.

[24] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery for incomplete and anaccurate measurements. *Communication on Pure and Applied Mathematics*, 59 :1207–1223, May 2006.

[25] A. B. Chan, N. Vasconcelos, and G. R. Lanckriet. Direct convex relaxations of sparse SVM. In *Proceedings of the 24th International Conference on Machine Learning*, pages 145–153. ACM, 2007.

[26] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Transactions on Signal Processing*, 54(12) :4634–4643, 2006.

[27] Q. Chen, P. Montesinos, Q. S. Sun, and P. A. Heng. Adaptive total variation denoising based on difference curvature. *Image and Vision Computing*, 28(3) :298–306, 2010.

[28] Q. Chen, J. Zhao, H. Zhang, and X. Wang. Feasibility study on qualitative and quantitative analysis in tea by near infrared spectroscopy with multivariate calibration. *Analytica Chimica Acta*, 572(1) :77–84, 2006.

[29] S. Chen, S. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5) :1873–1896, 1989.

[30] S. Chen and D. Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

[31] Y. Chen, N. M. Nasrabadi, and T. D. Tran. Hyperspectral image classification using dictionary-based sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 49(10) :3973–3985, 2011.

[32] S. F. Cotter. Multiple snapshot matching pursuit for direction of arrival (DOA) estimation. In *Signal Processing Conference, 2007 15th European*, pages 247–251. IEEE, 2007.

[33] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Transactions on Signal Processing*, 53(7) :2477–2488, 2005.

[34] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1) :21–27, 1967.

[35] A. Craig, O. Cloarec, E. Holmes, J. K. Nicholson, and J. C. Lindon. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7) :2262–2267, 2006.

[36] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52 :1289–1306, 2006.

[37] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1) :6–18, 2006.

[38] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432) :1200–1224, 1995.

[39] D. L. Donoho and J. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455, 1994.

[40] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2) :407–499, 2004.

[41] M. Elad and M. Aharon. Image denoising via learned dictionaries and sparse representation. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 895–900. IEEE, 2006.

[42] G. Elmasry, M. Kamruzzaman, D. W. Sun, and P. Allen. Principles and applications of hyperspectral imaging in quality evaluation of agro-food products : a review. *Critical Reviews in Food Science and Nutrition*, 52(11) :999–1023, 2012.

[43] L. Eriksson, J. Rosén, E. Johansson, and J. Trygg. Orthogonal PLS (OPLS) modeling for improved analysis and interpretation in drug design. *Molecular Informatics*, 31(6-7) :414–419, 2012.

[44] A. Evgeniou and M. Pontil. Multi-task feature learning. *Advances in Neural Information Processing Systems*, 19 :41, 2007.

[45] E. Fix and J. L. Hodges Jr. Discriminatory analysis-nonparametric discrimination : consistency properties. Technical report, California University, Berkeley, 1951.

[46] R. Flamary, N. Jrad, R. Phlypo, M. Congedo, and A. Rakotomamonjy. Mixed-norm regularization for brain decoding. *Computational and Mathematical Methods in Medicine*, 2014, 2014.

[47] W. J. Foley, A. McIlwee, I. Lawler, L. Aragones, A. P. Woolnough, and N. Berding. Ecological applications of near infrared reflectance spectroscopy–a tool for rapid, cost-effective prediction of the composition of plant and animal tissues and aspects of animal performance. *Oecologia*, 116(3) :293–305, 1998.

[48] S. Foucart. Hard thresholding pursuit : an algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6) :2543–2563, 2011.

[49] S. Foucart and M. J. Lai. Sparsest solutions of underdetermined linear systems via ℓ_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3) :395–407, 2009.

[50] I. E. Frank and S. Lanteri. Classification models : discriminant analysis, SIMCA, CART. *Chemometrics and Intelligent Laboratory Systems*, 5(3) :247–256, 1989.

[51] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2) :302–332, 2007.

[52] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer Series in Statistics Springer, Berlin, 2001.

[53] J. H. Friedman. Greedy function approximation : a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[54] J. Friedrich. *Spatial Modeling in Natural Sciences and Engineering : Software Development and Implementation*. Springer, Berlin, ISBN 3-540-20877-1, 2004.

[55] G. M. Furnival and R. W. Wilson. Regressions by leaps and bounds. *Technometrics*, 42(1) :69–79, 2000.

[56] J. Gabrielsson, H. Jonsson, C. Airiau, B. Schmidt, R. Escott, and J. Trygg. OPLS methodology for analysis of pre-processing effects on spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 84(1) :153–158, 2006.

[57] A. Giacomino, O. Abollino, M. Malandrino, M. Karthik, and V. Murugesan. Determination and assessment of the contents of essential and potentially toxic elements in Ayurvedic medicine formulations by inductively coupled plasma-optical emission spectrometry. *Microchemical Journal*, 99(1) :2–6, 2011.

[58] P. E. Gill, W. Murray, and M. A. Saunders. *User's guide for SNOPT 5.3 : A Fortran package for large-scale nonlinear programming*. Department of Mathematics, University of California, San Diego, USA, 1998.

[59] S. Girard. A nonlinear PCA based on manifold approximation. *Computational Statistics*, 15(2) :145–167, 2000.

[60] J. A. Gualtieri and R. F. Cromp. Support vector machines for hyperspectral remote sensing classification. In *The 27th AIPR Workshop : Advances in Computer-Assisted Recognition*, pages 221–232. International Society for Optics and Photonics, 1999.

[61] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3) :389–422, 2002.

[62] X. Hang and F. X. Wu. Sparse representation for classification of tumors using gene expression data. *BioMed Research International*, 2009, 2009.

[63] M. Harz, P. Rösch, K. D. Peschke, O. Ronneberger, H. Burkhardt, and J. Popp. Micro-Raman spectroscopic identification of bacterial cells of the genus Staphylococcus and dependence on their cultivation conditions. *Analyst*, 130(11) :1543–1550, 2005.

[64] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406) :502–516, 1989.

[65] M. Hebiri. Regularization with the smooth-lasso procedure. *arXiv preprint arXiv :0803.0668*, http://prunel.ccsd.cnrs.fr/docs/00/33/09/11/PDF/Smooth_Lasso2.pdf, 2008.

[66] M. Hedenström, S. Wiklund, B. Sundberg, and U. Edlund. Visualization and interpretation of OPLS models based on 2D NMR data. *Chemometrics and Intelligent Laboratory Systems*, 92(2) :110–117, 2008.

[67] M. E. Hellman. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6(3) :179–185, 1970.

[68] H. Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of Computational and Graphical Statistics*, 19(4) :984–1006, 2010.

[69] A. E. Hoerl and R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1) :55–67, 1970.

[70] C. Huang, L. S. Davis, and J. R. G. Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4) :725–749, 2002.

[71] K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems*, pages 609–616, 2006.

[72] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1) :55–63, 1968.

[73] M. D. Iordache, J. Bioucas-Dias, and A. Plaza. Sparse unmixing of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 49 :2014–2039, June 2011.

[74] M. D. Iordache, J. M. Bioucas-Dias, A. Plaza, and B. Somers. MUSIC-CSR : Hyperspectral unmixing via multiple signal classification and collaborative sparse regression. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7) :4364–4382, 2014.

[75] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Prentice Hall Upper Saddle River, New Jersey, 2002.

[76] I. Jolliffe. *Principal component analysis*. Wiley Online Library, Berlin, 2005.

[77] I. T. Jolliffe, N. T. Trendafilov, and M. Uddin. A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, 12(3) :531–547, 2003.

[78] P. Jonsson, S. J. Bruce, T. Moritz, J. Trygg, M. Sjöström, R. Plumb, J. Granger, E. Maibaum, J. K. Nicholson, and E. Holmes. Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. *Analyst*, 130(5) :701–707, 2005.

[79] H. R. Keller and D. L. Massart. Evolving factor analysis. *Chemometrics and Intelligent Laboratory Systems*, 12(3) :209–224, 1991.

[80] J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology, 2008.

[81] M. Kokalj, J. Kolar, T. Trafela, and S. Kreft. Differences among epilobium and hypericum species revealed by four IR spectroscopy modes : transmission, KBr tablet, diffuse reflectance and ATR. *Phytochemical Analysis*, 22(6) :541–546, 2011.

[82] P. Li and S. Xu. Support vector machine and kernel function characteristic analysis in pattern recognition. *Computer Engineering and Design*, 26(2) :302–304, 2005.

[83] Y. Li, A. Cichocki, and S. Amari. Analysis of sparse representation and blind source separation. *Neural Computation*, 16(6) :1193–1234, 2004.

[84] J. Liu and J. Ye. Moreau-Yosida regularization for grouped tree structure learning. In *Advances in Neural Information Processing Systems*, pages 1459–1467, 2010.

[85] J. Liu, L. Yuan, and J. Ye. An efficient algorithm for a class of fused lasso problems. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 323–332. ACM, 2010.

[86] D. Lorente, N. Aleixos, J. Gómez-Sanchis, S. Cubero, O. L. García-Navarrete, and J. Blasco. Recent advances and applications of hyperspectral imaging for fruit and vegetable quality assessment. *Food and Bioprocess Technology*, 5(4) :1121–1142, 2012.

[87] J. F. MacGregor and T. Kourti. Statistical process control of multivariate processes. *Control Engineering Practice*, 3(3) :403–414, 1995.

[88] D. M. Malioutov, M. Cetin, and A. S. Willsky. Homotopy continuation for sparse signal representation. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–733. IEEE, 2005.

[89] T. Marill and D. Green. On the effectiveness of receptors in recognition systems. *IEEE transactions on Information Theory*, 9(1) :11–17, 1963.

[90] H. Martens and M. Martens. *Multivariate analysis of quality : An introduction*. John Wiley and Sons, Chichester, 2001.

[91] E. B. Martin and A. J. Morris. An overview of multivariate statistical process control in continuous and batch process performance monitoring. *Transactions of the Institute of Measurement and Control*, 18(1) :51–60, 1996.

[92] V. Mazet, C. Carteret, D. Brie, J. Idier, and B. Humbert. Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems*, 76(2) :121–133, 2005.

[93] R. R. Meglen. Examining large databases : a chemometric approach using principal component analysis. *Marine Chemistry*, 39(1-3) :217–237, 1992.

[94] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70(1) :53–71, 2008.

[95] F. Melgani and L. Bruzzone. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8) :1778–1790, 2004.

[96] W. J. Melssen, J. R. M. Smits, L. M. C. Buydens, and G. Kateman. Using artificial neural networks for solving chemical problems : Part II. Kohonen self-organising feature maps and Hopfield networks. *Chemometrics and Intelligent Laboratory Systems*, 23(2) :267–291, 1994.

[97] G. Mercier and M. Lennon. Support vector machines for hyperspectral image classification with spectral-based kernels. In *IEEE Geoscience and Remote Sensing Symposium, 2003. IGARSS'03.*, volume 1, pages 288–290. IEEE, 2003.

[98] A. J. Miller. *Subset selection in regression*. Chapman and Hall, London, 1990.

[99] Q. Mo and Y. Shen. A remark on the restricted isometry property in orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 58(6) :3654–3656, 2012.

[100] M. Mohammadi, M. Nezamabadi, R. S. Berns, and L. A. Taplin. Spectral imaging target development based on hierarchical cluster analysis. In *Color and Imaging Conference*, pages 59–64. Society for Imaging Science and Technology, 2004.

[101] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2) :227–234, 1995.

[102] D. Naumann, V. Fijala, H. Labischinski, and P. Giesbrecht. The rapid differentiation and identification of pathogenic bacteria using Fourier transform infrared spectroscopic and multivariate statistical analysis. *Journal of Molecular Structure*, 174 :165–170, 1988.

[103] D. Naumann, D. Helm, and H. Labischinski. Microbiological characterizations by FT-IR spectroscopy. *Nature*, 351(6321) :81–82, 1991.

[104] D. Needell and J. A. Tropp. CoSaMP : Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26 :301–321, May 2009.

[105] Y. Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Soviet Mathematics Doklady*, 27(2) :372–376, 1983.

[106] Y. Ni, Y. Peng, and S. Kokot. Fingerprinting of complex mixtures with the use of high performance liquid chromatography, inductively coupled plasma atomic emission spectroscopy and chemometrics. *Analytica Chimica Acta*, 616(1) :19–27, 2008.

[107] G. Obozinski, B. Taskar, and M. Jordan. Joint covariate selection for grouped classification. Technical report, Technical Report 743, Dept. of Statistics, University of California Berkeley, 2007.

[108] M. R. Osborne, B. Presnell, and B. A. Turlach. A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20(3) :389–403, 2000.

[109] M. Pérez-Enciso and M. Tenenhaus. Prediction of clinical outcome with microarray data : a partial least squares discriminant analysis (PLS-DA) approach. *Human Genetics*, 112(5-6) :581–592, 2003.

[110] M. J. Piovoso and K. A. Kosanovich. Applications of multivariate statistical methods to process monitoring and controller design. *International Journal of Control*, 59(3) :743–765, 1994.

[111] M. M. Plichta, S. Heinzel, A. C. Ehlis, P. Pauli, and A. J. Fallgatter. Model-based analysis of rapid event-related functional near-infrared spectroscopy (NIRS) data : a parametric validation study. *Neuroimage*, 35(2) :625–634, 2007.

[112] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies. Sparse representations in audio and music : from coding to source separation. *Proceedings of the IEEE*, 98(6) :995–1005, 2010.

[113] A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3(Mar) :1357–1370, 2003.

[114] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, and J. P. Mesirov. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26) :15149–15154, 2001.

[115] M. Rantalainen, O. Cloarec, O. Beckonert, I. D. Wilson, D. Jackson, R. Tonge, R. Rowlinson, S. Rayner, J. Nickson, and R. Wilkinson. Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *Journal of Proteome Research*, 5(10) :2642–2655, 2006.

[116] R. Saab and Ö. Yılmaz. Sparse recovery by non-convex optimization–instance optimality. *Applied and Computational Harmonic Analysis*, 29(1) :30–48, 2010.

[117] S. Šašic and Y. Ozaki. Short-wave near-infrared spectroscopy of biological fluids. 1. Quantitative analysis of fat, protein, and lactose in raw milk by partial least-squares regression and band assignment. *Analytical Chemistry*, 73(1) :64–71, 2001.

[118] P. Schniter, L. C. Potter, and J. Ziniel. Fast bayesian matching pursuit. In *Information Theory and Applications Workshop, 2008*, pages 326–333. IEEE, 2008.

[119] B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5) :1299–1319, 1998.

[120] V. H. Segtnan, Š. Šašic, T. Isaksson, and Y. Ozaki. Studies on the structure of water using two-dimensional near-infrared correlation spectroscopy and principal component analysis. *Analytical chemistry*, 73(13) :3153–3161, 2001.

[121] S. Serranti, A. Gargiulo, and G. Bonifazi. Classification of polyolefins from building and construction waste using NIR hyperspectral imaging system. *Resources, Conservation and Recycling*, 61 :52–58, 2012.

[122] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise. *Near-infrared spectroscopy : principles, instruments, applications*. John Wiley & Sons, Siesler, 2008.

[123] C. Soussen, J. Idier, D. Brie, and J. Duan. From Bernoulli–Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10) :4572–4584, 2011.

[124] J. L. Starck, F. Murtagh, and J. M. Fadili. *Sparse image and signal processing : wavelets, curvelets, morphological diversity*. Cambridge University Press, Cambridge (UK), 2010.

[125] B. Stuart. *Infrared spectroscopy*. Wiley Online Library, New York, 2005.

[126] P. Tatzer, M. Wolf, and T. Panner. Industrial application for inline material sorting using hyperspectral imaging in the NIR range. *Real-Time Imaging*, 11(2) :99–107, 2005.

[127] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[128] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(1) :91–108, 2005.

[129] R. J. Tibshirani, J. E. Taylor, E. J. Candès, and T. Hastie. *The solution path of the generalized lasso*. PhD thesis, Stanford University, 2011.

[130] D. F. Toh, L. S. New, H. L. Koh, and E. C. Y. Chan. Ultra-high performance liquid chromatography/time-of-flight mass spectrometry (UHPLC/TOFMS) for time-dependent profiling of raw and steamed Panax notoginseng. *Journal of Pharmaceutical and Biomedical Analysis*, 52(1) :43–50, 2010.

[131] J. A. Tropp. Greed is good : Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50 :2231–2242, Oct 2004.

[132] J. A. Tropp. Just relax : Convex programming methods for subset selection and sparse approximation. Technical report, The University of Texas, Austin, 2004, 2004.

[133] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II : Convex relaxation. *Signal Processing*, 86 :589–602, March 2006.

[134] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I : Greedy pursuit. *Signal Processing*, 86 :572–588, 2006.

[135] J. A. Tropp and S. J. Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98 :948–958, June 2010.

[136] J. Trygg, E. Holmes, and T. Lundstedt. Chemometrics in metabonomics. *Journal of Proteome Research*, 6(2) :469–479, 2007.

[137] J. Trygg and S. Wold. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3) :119–128, 2002.

[138] B. A. Turlach. On algorithms for solving least squares problems under an ℓ_1 penalty or an ℓ_1 constraint. In *2004 Proceedings of the American Statistical Association ; Statistical Computing Section [CD-ROM]*, pages 2572–2577, 2005.

[139] B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3) :349–363, 2005.

[140] G. Tzagkarakis, D. Milioris, and P. Tsakalides. Multiple-measurement bayesian compressed sensing using GSM priors for DOA estimation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2610–2613. IEEE, 2010.

[141] J. A. Van Leeuwen, R. J. Jonker, and R. Gill. Octane number prediction based on gas chromatographic analysis with non-linear regression techniques. *Chemometrics and Intelligent Laboratory Systems*, 25(2) :325–340, 1994.

[142] V. N. Vapnik and V. Vapnik. *Statistical learning theory*, volume 2. Wiley, New York, 1998.

[143] N. Vlachos, Y. Skopelitis, M. Psaroudaki, V. Konstantinidou, A. Chatzilazarou, and E. Tegou. Applications of Fourier transform-infrared spectroscopy to edible oils. *Analytica Chimica Acta*, 573 :459–465, 2006.

[144] H. W. Wang and Y. Q. Liu. Evaluation of trace and toxic element concentrations in Paris polyphylla from China with empirical and chemometric approaches. *Food Chemistry*, 121(3) :887–892, 2010.

[145] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *7th European Symposium on Artificial Neural Networks*, volume 99, pages 219–224, 1999.

[146] D. P. Wipf and B. D. Rao. Sparse bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, 52(8) :2153–2164, 2004.

[147] S. Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, 8(3) :127–139, 1976.

[148] S. Wold, H. Martens, and H. Wold. The multivariate calibration problem in chemistry solved by the PLS method. In *Proceedings of the Conference on Matrix Pencils*, pages 286–293. Springer, Heidelberg, 1983.

[149] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn. The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3) :735–743, 1984.

[150] S. Wold, M. Sjöström, and L. Eriksson. PLS-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2) :109–130, 2001.

[151] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2) :210–227, 2009.

[152] S. C. Wu and A. L. Swindlehurst. Matching pursuit and source deflation for sparse EEG/MEG dipole moment estimation. *IEEE Transactions on Biomedical Engineering*, 60(8) :2280–2288, 2013.

[153] W. Wu, B. Walczak, D. L. Massart, S. Heuerding, F. Erni, I. R. Last, and K. A. Prebble. Artificial neural networks in classification of NIR spectral data : design of the training set. *Chemometrics and Intelligent Laboratory Systems*, 33(1) :35–46, 1996.

[154] Z. Xiaobo, Z. Jiewen, M. J. Povey, M. Holmes, and M. Hanpin. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta*, 667(1) :14–32, 2010.

[155] B. Xin, Y. Kawahara, Y. Wang, and W. Gao. Efficient generalized fused lasso and its application to the diagnosis of Alzheimer's disease. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2163–2169, 2014.

[156] S. C. Yoon, B. Park, K. C. Lawrence, W. R. Windham, and G. W. Heitschmidt. Line-scan hyperspectral imaging system for real-time inspection of poultry carcasses with fecal material and ingesta. *Computers and Electronics in Agriculture*, 79(2) :159–168, 2011.

[157] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1) :49–67, 2006.

[158] S. Zahri. *Analyse quantitative et qualitative des substances chimiques responsables des durabilités naturelle et conférée des bois de chêne européen et de pin maritime par la spectroscopie dans le proche infrarouge*. PhD thesis, Université de Pau, 2007.

[159] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1095–1103. ACM, 2012.

[160] Y. Zhou, T. Liu, J. Li, and Z. Chen. Rapid identification of edible oil and swill-cooked dirty oil by using near-infrared spectroscopy and sparse representation classification. *Analytical Methods*, 7(6) :2367–2372, 2015.

[161] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems*, 16(1) :49–56, 2004.

[162] M. Zibulevsky, P. Kisilev, Y. Y. Zeevi, and B. A. Pearlmutter. Blind source separation via multinode sparse representation. *Advances in Neural Information Processing Systems 14*, pages 1049–1056, 2002.

[163] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2) :265–286, 2006.

[164] J. Zupan, M. Novič, and I. Ruisánchez. Kohonen and counterpropagation artificial neural networks in analytical chemistry. *Chemometrics and Intelligent Laboratory Systems*, 38(1) :1–23, 1997.

Résumé

Le présent travail de thèse se propose de développer des techniques innovantes pour l'automatisation de tri de déchets de bois. L'idée est de combiner les techniques de spectrométrie proche-infra-rouge à des méthodes robustes de traitement de données pour la classification. Après avoir exposé le contexte du travail dans le premier chapitre, un état de l'art sur la classification de données spectrales est présenté dans le chapitre 2.
Le troisième chapitre traite du problème de sélection de variables par des approches parcimonieuses. En particulier nous proposons d'étendre quelques méthodes gloutonnes pour l'approximation parcimonieuse simultanée. Les simulations réalisées pour l'approximation d'une matrice d'observations montrent l'intérêt des approches proposées. Dans le quatrième chapitre, nous développons des méthodes de sélection de variables basées sur la représentation parcimonieuse simultanée et régularisée, afin d'augmenter les performances du classifieur SVM pour la classification des spectres IR ainsi que des images hyperspectrales de déchets de bois. Enfin, nous présentons dans le dernier chapitre les améliorations apportées aux systèmes de tri de bois existants. Les résultats des tests réalisés avec logiciel de traitement mis en place, montrent qu'un gain considérable peut être atteint en termes de quantités de bois recyclées.

Mots-clés: spectres IR, classification, SVM, sélection de variables, approches parcimonieuses, images hyperspectrales.

Abstract

In this thesis innovative techniques for sorting wood wastes are developed. The idea is to combine infrared spectrometry techniques with robust data processing methods for classification task. After exposing the context of the work in the first chapter, a state of the art on the spectral data classification is presented in the chapter 2.
The third chapter deals with variable selection problem using sparse approaches. In particular we propose to extend some greedy methods for the simultaneous sparse approximation. The simulations performed for the approximation of an observation matrix validate the advantages of the proposed approaches.
In the fourth chapter, we develop variable selection methods based on simultaneous sparse and regularized representation, to increase the performances of SVM classifier for the classification of NIR spectra and hyperspectral images of wood wastes. In the final chapter we present the improvements made to the existing sorting systems. The results of the conducted tests using the processing software confirm that significant benefits can be achieved in terms of recycled wood quantities.

Keywords: infrared spectra, classification, SVM, variable selection, sparse approaches, hyperspectral images