Méthodes de sélection de variables appliquées en spectroscopie proche infrarouge pour l'analyse et la classification de textiles

Alexandra Durand

N° d'ordre : 4073

Université des Sciences et Technologies de Lille

École Doctorale des Sciences Pour l'Ingénieur

Thèse

pour obtenir le grade de

Docteur de l'Université des Sciences et Technologies de Lille

Instrumentation et Analyses Avancées

Présentée et soutenue publiquement par

DURAND Alexandra

le 26 novembre 2007

METHODES DE SELECTION DE VARIABLES APPLIQUEES EN SPECTROSCOPIE PROCHE INFRAROUGE POUR L'ANALYSE ET LA CLASSIFICATION DE TEXTILES

Mr Jean-Michel Roger, Chercheur, Cemagref, Montpellier.
Mr Yvan Vander Heyden, Professeur, Vrije Universiteit Brussel, Bruxelles.

Mme Mireille Bayart, Professeur, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq.
Mme Stéphanie Pessayre, Docteur-Ingénieur, Groupe SEB, Pont-Evêque.
Mr Cyril Ruckebusch, Maître de conférences, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq.
Mr Jean-Pierre Huvenne, Professeur, Université des Sciences et Technologies de Lille, Villeneuve d'Ascq.

A la mémoire de mes grand-parents

A mes parents

A tous ceux qui me sont chers,…

Remerciements

Bien qu'étant un effort personnel, un travail de thèse ne peut aboutir sans aide. Je tiens donc à remercier toutes les personnes qui ont participé à la réalisation de ce travail et qui ont rendu ces trois années de thèse très agréables. Par ailleurs, je m'excuse par avance pour les personnes que j'aurai pu oublier dans ce délicat exercice.

Ce travail a été réalisé au laboratoire de Spectrochimie Infrarouge et Raman (LASIR, CNRS UMR 8516). J'exprime ma gratitude au directeur du LASIR Monsieur Buntinx pour m'avoir accepté au sein de son laboratoire. Cette étude a été financée par le ministère de la recherche dans le cadre du projet performance n°05T11 en collaboration avec deux partenaires industriels, le groupe SEB et la société ADISON. Je remercie toutes les personnes du groupe SEB qui ont participé à l'évolution de ce projet et plus particulièrement Monsieur Dancer, Directeur des Relations Extérieures. Je remercie également Monsieur Verard, Président de la société ADISON, de m'avoir accueilli dans sa société.

Je suis très sensible à l'honneur que me font Monsieur Roger, Chercheur au Cemagref de Montpellier, et Monsieur Vander Heyden, Professeur à Vrije Universiteit Brussel de Bruxelles, en acceptant d'être rapporteurs. De même, je remercie Madame Bayart, Professeur à l'Université des Sciences et Technologies de Lille à Villeneuve d'Ascq d'avoir présidé ce jury et Madame Pessayre, Docteur-Ingénieur du groupe SEB à Pont-Evêque d'avoir accepté de faire partie de ce jury.

Je remercie vivement toutes les personnes qui m'ont aidé au cours de ce travail. Ma reconnaissance et ma sympathie vont en premier lieu à Monsieur Huvenne, Professeur à l'Université des Sciences et Technologies de Lille et à Monsieur Ruckebusch, Maître de Conférences à l'Université des Sciences et Technologies de Lille pour avoir dirigé ce travail. Je tiens à remercier chacun d'eux non seulement pour leurs apports scientifiques et techniques, mais aussi et surtout pour leur sens aigu de l'encadrement, pour leur disponibilité de chaque instant.

Je tiens à remercier les membres du LASIR pour leur sympathie, les discussions enrichissantes et leur aide au cours de mes travaux. Mes pensées s'adressent à chacun d'eux et je citerai particulièrement mes collègues de bureau Lionel et Séverine, les collègues de la pause café, Baptiste, Christophe, Myriam, Ophélie et Stéphane, Kasia et Françoise avec qui j'ai débuté mes activités d'enseignement. Je remercie tout particulièrement Emmanuelle et Julien pour leur jovialité, leur bonne humeur et leur soutien tout au long de ce travail.

Pour terminer, peut être le plus important, je voudrais remercier mes parents et ma famille entière qui, de près ou de loin, ont toujours su m'offrir leur soutien, leur compréhension, leurs encouragements, leur patience et leur affection. Merci pour l'amour et la confiance que vous m'avez toujours accordés. A eux je dédie cette thèse.

Table des matières

Chapitre 3

Chapitre 4

Liste des abréviations

AG	Algorithme(s) Génétique(s)
ANN	*Artificial Neural Networks*
CP	Composante(s) Principale(s)
DET	*Det-trending*
DER1 et DER2	Dérivée première et dérivée seconde
IM	Information Mutuelle
kNN	*k-Nearest Neighbors*
LDA	*Linear Discriminant Analysis*
MC	*Mean Center*
MLR	*MultiLinear Regression*
PA	Polyamide
PCA	*Principal Component Analysis*
PCR	*Principal Component Regression*
PES	Polyester
PIR	Proche InfraRouge
PLS	*Partial Least Squares*
RMSEC	*Root Mean Square Error of Calibration*
RMSECV	*Root Mean Square Error of Cross Validation*
RMSEP	*Root Mean Square Error of Prediction*
Rprop	*Resilient propagation*
SIMCA	*Soft Independent Modelling of Class Analogy*
SNV	*Standard Normal Variate*
SV	Vecteur support
SVM	*Support Vector Machine*
U.A.	Unités arbitraires
VC	Validation croisée

Liste des symboles

α	Multiplicateurs de Lagrange (SVM)
b	Décalage du séparateur (SVM)
C	Lot d'entraînement
C	Paramètre de régularisation (SVM)
C_d	Point de croisement simple (AG-PLS)
C_s	Point de croisement double (AG-PLS)
ξ	Variables de relâchement (SVM)
G	Paramètre lié à la largeur de la gaussienne (SVM)
h	Nombre d'individus (AG-PLS)
H	Entropie (IM)
$\mathcal{H}$	Espace de vecteurs
k	Nombre de plus proche voisins (IM)
K	Fonction noyau (SVM)
L_D et L_P	Forme duale et primale de la fonction de Lagrange (SVM)
m	Nombre de variables de la matrice $\boldsymbol{X}$
$\boldsymbol{m}$	Groupe de variables (AG-PLS)
M	Ensemble de variables (IM)
M_u	Taux de mutation (AG-PLS)
n	Nombre d'échantillons de la matrice $\boldsymbol{X}$
P	Lot de prédiction
$\boldsymbol{P}$	Matrice des vecteurs propres (*loadings*)
ϕ	Fonction non linéaire (SVM)
ρ	Marge (SVM)
t	Valeur du test de *Student*
$\boldsymbol{T}$	Matrice des coordonnées factorielles (*scores*)
τ	Fonction de coût (SVM)
ω	Vecteur normal au séparateur (SVM)
$\boldsymbol{X}$	Matrice des données spectrales
X	Variables spectroscopiques de la matrice $\boldsymbol{X}$
$\boldsymbol{Y}$	Matrice de la propriété d'intérêt

Introduction

La spectroscopie proche infrarouge (PIR) est une technique très employée pour l'analyse d'échantillons de diverses natures[1,2], en mélanges[3], bruts[4] ou pour le suivi de procédés[5]… L'intérêt de cette technique réside essentiellement dans sa rapidité et son caractère non destructif. Sa popularité est aussi due aux avancées de l'instrumentation et au développement des méthodes chimiométriques[6]. Il s'agit de méthodes mathématiques et statistiques permettant de visualiser, de traiter l'information spectrale afin d'extraire l'information analytique. En particulier, les modèles d'étalonnage multivarié offrent des alternatives intéressantes aux analyses chimiques classiques pour la détermination indirecte de la nature[7], de la composition[8] ou d'une propriété physico-chimique[9] de l'échantillon. A condition d'être généralisables à des données futures, ces modèles rendent possible l'analyse quantitative ou qualitative des échantillons sur la base de leur spectre proche infrarouge. Les problèmes rencontrés en étalonnage multivarié sont souvent liés à la dimension des matrices de données spectroscopiques. La sélection de l'information pertinente, des variables significatives ou des échantillons représentatifs est donc un domaine de recherche important. La sélection de variables[10-12] permet de réduire la dimension des données spectrales, d'améliorer les performances, la robustesse des modèles ou de tendre vers une instrumentation simplifiée rendant possibles par exemple des contrôles de routine. La spectroscopie proche infrarouge et la chimiométrie ont ainsi prouvé leur utilité dans divers domaines tels que l'agroalimentaire[13], la pharmacie[14], la plasturgie[15] et plus particulièrement dans le domaine du textile[16]en ce qui concerne l'application de cette étude.

L'industrie textile est un secteur très concurrentiel, sans cesse en évolution, avec l'apparition notamment de nouvelles fibres, permettant d'obtenir des tissus plus performants ou plus techniques. L'analyse rapide de la composition chimique des échantillons textiles est fondamentale dans certaines applications, par exemple pour le contrôle sur les lignes de production, le contrôle aux frontières, le tri sélectif[17,18]… Dans cet objectif, une étude a été réalisée en collaboration avec l'Institut Français du Textile-Habillement (IFTH) afin de déterminer précisément la teneur en coton de textiles coton/polyester ou coton/viscose par

spectroscopie proche infrarouge, sans analyse chimique. Par ailleurs, pour d'autres applications, la grandeur d'intérêt peut être plutôt une propriété physico-chimique telle que l'élasticité[19] ou la rugosité[20]. Dans ce cas, il ne s'agit plus d'une analyse quantitative[21,22], comme précédemment, mais d'une analyse qualitative[19,23,24]. Nous chercherons également dans ce travail à développer des modèles pour la classification d'échantillons textiles sur la base de leur spectre proche infrarouge en fonction d'une propriété physico-chimique d'intérêt dont la nature exacte restera confidentielle dans ce manuscrit. L'objectif de cette étude est de tendre vers une instrumentation simplifiée. Les méthodes qui seront utilisées dans le cadre de nos travaux doivent prendre en compte tous les paramètres liés aux données, en terme de nombre de classes à prédire, de distribution des lots de données ou de complexité des échantillons. Diverses méthodes ont été utilisées dans la littérature pour la classification de données textiles[19]. Ces méthodes sont fondées sur différents critères (méthodes multiclasses, linéaires, paramétriques, supervisées…) donnant ainsi accès à une large gamme d'applications. Cette étude a été financée par le ministère de la recherche dans le cadre d'un projet « Performance » regroupant deux partenaires industriels, le groupe SEB et la société ADISON spécialisée dans l'optique et l'électronique, et un laboratoire de recherche, le LASIR.

Le manuscrit s'articule en quatre chapitres. Dans un premier chapitre, nous passerons en revue les données bibliographiques sur lesquelles se fonde notre travail. Nous insisterons dans un premier temps sur les aspects de la sélection de variables pour l'analyse quantitative, puis sur les méthodes de classification dans le cadre des objectifs définis par les partenaires industriels. Viennent ensuite les chapitres consacrés à la présentation des résultats et à la discussion. Le chapitre 2 concerne l'analyse de mélanges binaires textiles coton/polyester et coton/viscose. Nous présenterons les résultats de modèles prédictifs obtenus sur les spectres complets et sur les spectres réduits par des méthodes de sélection de variables. Le chapitre 3 concerne l'analyse qualitative proche infrarouge de tissus pour la classification d'échantillons textiles dans trois classes par rapport à la propriété physico-chimique d'intérêt. Les capacités prédictives obtenues seront discutées et mises en perspective par rapport au domaine d'application. Le chapitre 4 présentera le prototype et la construction de modèles de classification pour une application grand public.

Chapitre 1

Méthodes chimiométriques

La chimiométrie est la discipline de la chimie analytique qui utilise les mathématiques et les méthodes statistiques pour extraire l'information présente dans les données de mesures expérimentales[6]. Les spectres proche infrarouge (PIR) d'échantillons naturels ou complexes contiennent des bandes d'absorption liées essentiellement aux groupements C-H, N-H, O-H, S-H qui peuvent être affectées par la nature physique de l'échantillon. L'extraction de cette information ne peut se faire qu'indirectement notamment du fait de variations de la ligne de base ou de l'enchevêtrement des bandes d'absorption. Obtenir l'information analytique nécessite donc de recourir à un modèle d'étalonnage multivarié.

Ce premier chapitre détaille les méthodes chimiométriques utilisées au cours de ce travail pour extraire et modéliser l'information présente dans les spectres. Tout d'abord, nous verrons l'étalonnage multivarié pour l'analyse qualitative et quantitative. Dans un deuxième temps, nous détaillerons deux procédures de sélection de variables afin d'améliorer en particulier la robustesse et l'interprétation des modèles. Enfin, nous présenterons deux méthodes de classification, une méthode linéaire classique et une méthode de classification non linéaire très récemment développée pour les données spectroscopiques.

I. Etalonnage multivarié

Les spectres PIR des échantillons sont enregistrés et regroupés dans une matrice de données, notée ***X***. Cette matrice ***X*** contient les valeurs numériques d'absorbance observées pour les *m* variables (longueurs d'onde ou nombres d'onde) pour *n* échantillons. La matrice ***X*** est représentée sur la Figure 1. En parallèle, des mesures chimiques sont réalisées afin de déterminer les valeurs de la mesure de référence de la propriété à prédire. Ces valeurs sont regroupées dans le vecteur colonne, noté *Y*, qui peut être généralisable à une matrice ***Y*** contenant plusieurs colonnes dans le cas où plusieurs propriétés sont à prédire (Figure 1).

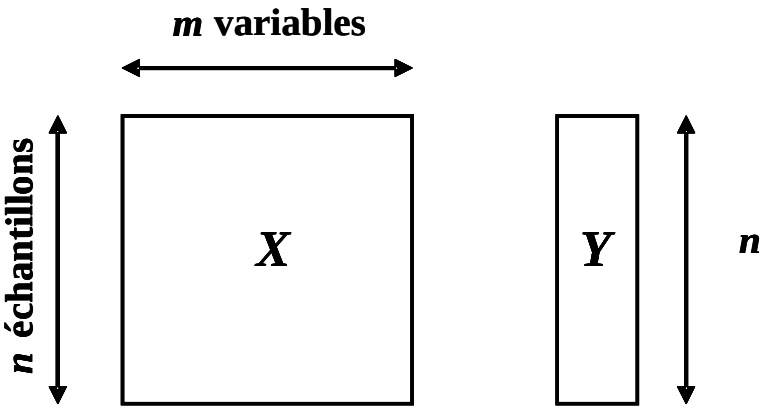


*Figure 1 : Matrice des données spectrales **X** contenant les valeurs d'absorbance pour m variables spectroscopiques et n échantillons. La matrice **Y** contient les valeurs de la mesure de référence.*

Le choix de la méthode d'analyse est fonction de l'objectif de l'étude. Il peut s'agir d'une analyse quantitative, ***Y*** contient alors des valeurs continues comme par exemple des valeurs de concentrations[25]. Il peut s'agir de la détermination d'une propriété, de l'appartenance à une classe… Dans ce cas, l'analyse est qualitative et les valeurs de ***Y*** sont codées de manières discrètes[26].

- ***Méthodes pour l'analyse quantitative***

L'étalonnage multivarié est très utilisé pour l'analyse quantitative des spectres proche infrarouge. On peut distinguer parmi les nombreuses méthodes disponibles pour l'analyse quantitative, les régressions multivariées sur les variables d'origines (méthodes non factorielles). La méthode la plus couramment utilisée en étalonnage multivarié est la régression linéaire multiple[22,27] (MLR) qui est fondée sur la méthode des moindres carrés.

Les méthodes de condensation des données sont, quant à elles, des techniques d'extraction de facteurs. L'idée de condensation des données est de caractériser des données de très grande dimension par un nombre restreint de facteurs abstraits qui sont des combinaisons linéaires des variables d'origine. On distingue parmi ces méthodes, la régression en composantes principales[6] (PCR) et la régression des moindres carrés partiels[28] (PLS). La PCR est constituée de deux étapes, un traitement des données spectrales par une analyse en composantes principales (PCA), suivi d'une régression MLR sur les coordonnées factorielles (*scores*) issues des axes principaux de la PCA.

On peut aussi distinguer parmi les méthodes employées pour l'analyse quantitative, les méthodes linéaires et les méthodes non linéaires. Par exemple, la régression PLS établit une régression linéaire entre les deux matrices, la matrice des données spectrales $\boldsymbol{X}$ et la matrice de la propriété à prédire $\boldsymbol{Y}$. Cependant les données peuvent présenter des non linéarités, c'est-à-dire notamment que la corrélation entre les variables prédictives et les valeurs de Y n'est pas linéaire. Afin de prendre en compte ces dérivations aux modèles purement linéaires diverses méthodes existent dont les réseaux de neurones artificiels[29] (ANN). Les ANN permettent d'estimer la relation entre une ou plusieurs valeurs d'absorbances et la propriété d'intérêt Y. Les procédures peuvent être considérées ici d'une certaine façon comme des méthodes standard de l'étalonnage multivarié[30]. Les détails concernant les méthodes PCA, PLS et ANN, sont présentés, respectivement, en annexe 1, 5 et 6.

- ***Méthodes pour l'analyse qualitative***

L'analyse qualitative concerne la discrimination des échantillons par une frontière et la classification, c'est-à-dire l'attribution des échantillons dans différents groupes en fonction de la valeur d'une propriété d'intérêt. L'application de méthodes de classification est importante

en chimie[31], biologie[32], agroalimentaire[33]… On distingue les méthodes dites supervisées et les méthodes non supervisées. Par définition, pour les méthodes non supervisées (*clustering*), les échantillons sont regroupés sans connaissance a priori de leur appartenance à une classe, c'est-à-dire que seule la matrice $\boldsymbol{X}$ intervient. On peut également distinguer, d'une part, les méthodes hiérarchiques qui procèdent par divisions successives du lot de données pour aboutir à la formation d'agrégats (*clusters*) représentés sous forme d'arbres[34] ou de dendrogrammes[35] et, d'autre part, les méthodes non-hiérarchiques qui procèdent par étapes itératives telles que *k-means*[35] ou les cartes de Kohonen[36]…

Dans le cas de méthodes de classification supervisées, l'attribution des classes pour les échantillons du lot de données nécessite la connaissance de la propriété de référence. L'apprentissage consiste à développer un modèle de classification sur les échantillons du lot d'entraînement. Les performances du modèle sont ensuite validées puis évaluées en comparant les valeurs prédites aux valeurs de référence sur les lots de validation et de prédiction. Dans la littérature, il existe trois critères pour distinguer les méthodes de classification.

La première distinction est basée sur la discrimination entre les classes. Les méthodes telles que l'analyse linéaire discriminante (LDA)[37] mettent en évidence les différences entre les classes alors que les méthodes comme la modélisation indépendante des analogies de classes (*Soft Independant Modelling of Class Analogy*, SIMCA)[38] construisent indépendamment pour chaque classe un modèle par classe.

La seconde différence concerne la linéarité des méthodes selon le choix des règles de discrimination entre les classes. En effet, la frontière construite par la méthode LDA est linéaire alors que la méthode des *Support Vector Machine* (SVM)[39-41] recherche une frontière non linéaire entre les classes. Les deux méthodes, LDA et SVM seront détaillées, respectivement, dans le chapitre 1, paragraphes III.1 et III.2 du manuscrit.

Enfin, une distinction peut être faite entre les méthodes paramétriques et les méthodes non paramétriques. Pour les techniques paramétriques telles que la méthode LDA, les paramètres statistiques de la distribution normale des échantillons sont utilisés dans les règles de classification. Ce n'est au contraire pas le cas pour les méthodes non paramétriques. En effet, la méthode des *k* plus proches voisins (*k*NN)[42] et la méthode des SVM par exemple ne font pas d'hypothèses sur la distribution des données.

II. Méthodes de sélection de variables

En étalonnage multivarié, la sélection de variables permet d'identifier et d'éliminer les variables qui pénalisent les performances d'un modèle dans la mesure où elles peuvent être bruitées, redondantes ou corrélées[6]. Les procédures de sélection de variables présentent un intérêt particulier en ce qui concerne les données spectroscopiques. En effet, le nombre de variables est en général très important vis-à-vis du nombre d'échantillons présents dans la matrice $\boldsymbol{X}$ ($m >> n$). Habituellement, ce problème de dimension est géré par l'utilisation de méthodes factorielles de régression. Mais les variables latentes notamment pour le modèle PLS peuvent aussi être affectées par les redondances des variables d'origine ou par la présence de variables non pertinentes[10]. Par conséquent même dans ce cas, la sélection de variables a prouvé être une solution satisfaisante pour simplifier la complexité du modèle et pour améliorer les capacités prédictives de ce dernier. De plus, la mise en évidence des variables pertinentes facilite l'interprétation et la compréhension des aspects physico-chimiques.

De nombreuses méthodes de sélection existent pour lesquelles le choix d'inclure ou non des variables est effectué à partir des erreurs de prédiction des modèles de régression construits. Parmi ces méthodes, on peut noter par exemple la régression PLS par l'élimination des variables non informatives (UVE-PLS)[43], la régression PLS par intervalles (iPLS)[44], la régression PLS par la sélection interactive de variables (IVS-PLS)[45]. Comme nous le verrons, la sélection de variables par la méthode des algorithmes génétiques (AG-PLS) entre également dans cette catégorie.

D'un point de vue plus conceptuel, une procédure de sélection de variables inclut en premier le choix d'une estimation de la pertinence d'une ou plusieurs variables et, en second, le choix d'un algorithme pour réaliser l'optimisation. Concernant la mesure de la pertinence, elle peut être basée par exemple sur l'estimation de la variance des variables X [46]. D'autres critères, comme l'information mutuelle (IM)[47], permettent de mesurer les dépendances entre les variables X et la variable à prédire Y. L'information mutuelle estime la quantité d'information contenue dans une variable X qui est utilisée pour prédire une variable Y. Cette estimation est réalisée indépendamment d'un modèle de régression, comme une étape de prétraitement. Cela présente l'avantage, ensuite, de pouvoir réaliser n'importe quel type de

régression ou modèle. Il est alors possible de construire par exemple un modèle PLS[48], ou un réseau de neurones artificiels (ANN)[49]. Concernant le choix de l'algorithme d'optimisation, des algorithmes stochastiques sont en général utilisés lorsque l'on travaille sur des données de grande dimension en particulier des données spectroscopiques. On s'intéressera notamment aux algorithmes génétiques (AG) qui sont très utilisés pour la sélection de variables en étalonnage multivarié[50-53]. Ce sont des outils d'optimisation qui réalisent une recherche aléatoire et globale dans un espace de grande dimension.

II.1. Sélection de variables par algorithmes génétiques

Les algorithmes génétiques sont des techniques d'optimisation stochastique introduites par Holland en 1975[54] qui peuvent être utilisées pour trouver une solution optimale globale dans un espace de grande dimension. Pour définir les AG, il est intéressant de faire le parallèle avec la théorie de l'évolution. En effet, ils imitent le concept de la sélection darwinienne dans une population d'organismes vivants : les individus les plus adaptés ont une plus grande chance de survie et transmettent leurs gènes par reproduction. La première application des algorithmes génétiques dans la littérature chimique remonte aux travaux de Lucius et Kateman en 1991[50]. Ces travaux ont été réalisés dans divers domaines en particulier pour l'analyse de bio-polymères[55]. En spectroscopie, les AG sont le plus souvent appliqués à la sélection d'observations ou de variables dans la matrice d'échantillons $\boldsymbol{X}$ en optimisant un critère qui correspond en général à l'erreur de la validation croisée (RMSECV)[56]. La sélection de variables par AG a été appliquée en spectroscopie proche infrarouge pour déterminer, par exemple, le nombre d'octane dans les échantillons de gasoil[57] ou le taux d'humidité dans les céréales[58].

- ***Algorithme AG-PLS pour la sélection de variables***

La première étape de l'algorithme est la construction aléatoire d'une population d'individus. Cette population est codée comme une matrice de taille h x m où h est le nombre d'individus de la population initiale et m correspond au nombre de variables (gènes) composant chaque individu (Figure 2). Au départ, pour la population initiale, la valeur des

gènes est attribuée de façon aléatoire et vaut 0 ou 1 codant l'absence ou la présence de la variable.

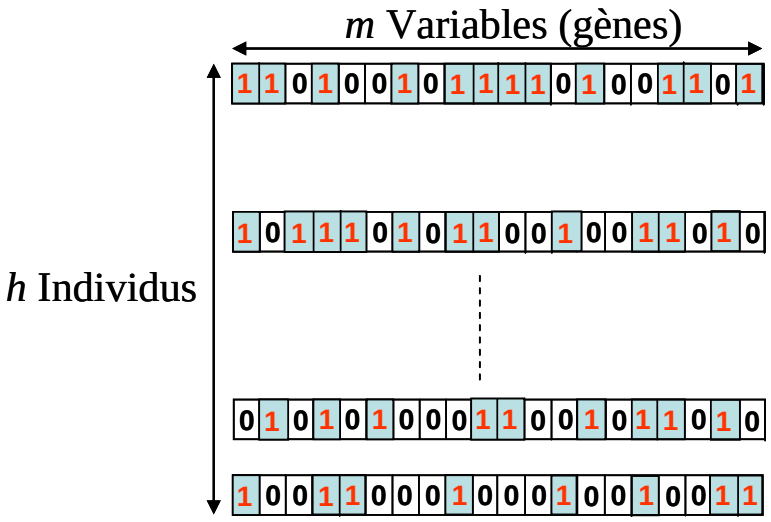


Figure 2 : Représentation de la population initiale (h individus et m gènes).

La seconde étape correspond à l'évaluation de cette population, c'est-à-dire à l'application de chaque individu à la matrice des données ***X*** et à la construction de modèles PLS pour décrire la relation entre les variables *X* sélectionnées et *Y*. Un modèle PLS est alors construit pour chaque individu. Les *h* individus sont évalués par une fonction de coût (*fitness*) qui correspond aux erreurs de validation croisée (RMSECV). Les individus auxquels sont associées les erreurs de RMSECV les plus petites sont conservés et seront utilisés pour générer la population future. Les individus de la population future seront ainsi plus adaptés à leur environnement que ceux de la population initiale, par principe[(59)].

La sélection est le mécanisme conduisant à l'élimination des individus de la population initiale et à la création de la population future avec pour contrainte de générer une population de même taille. Les AG utilisent pour cela des opérateurs, tels que les *points de croisement* et les *mutations* qui seront détaillés dans le paragraphe suivant.

Cette procédure est itérative et peut être résumée en 4 étapes qui sont reprises sur la Figure 3 et détaillées ci-dessous :

1. Choix d'une population initiale générée aléatoirement, par exemple 256 individus (256 modèles).

2. Evaluation par le calcul du RMSECV pour chaque individu de la population, avec élimination des individus ayant des valeurs de RMSECV élevées.
3. Création d'une population future en utilisant les opérateurs *points de croisement* et *mutations*.
4. Vérification des conditions d'arrêt de l'algorithme. Si la nouvelle population ne satisfait pas ces conditions, les étapes 2 à 4 sont répétées afin de générer une nouvelle population.

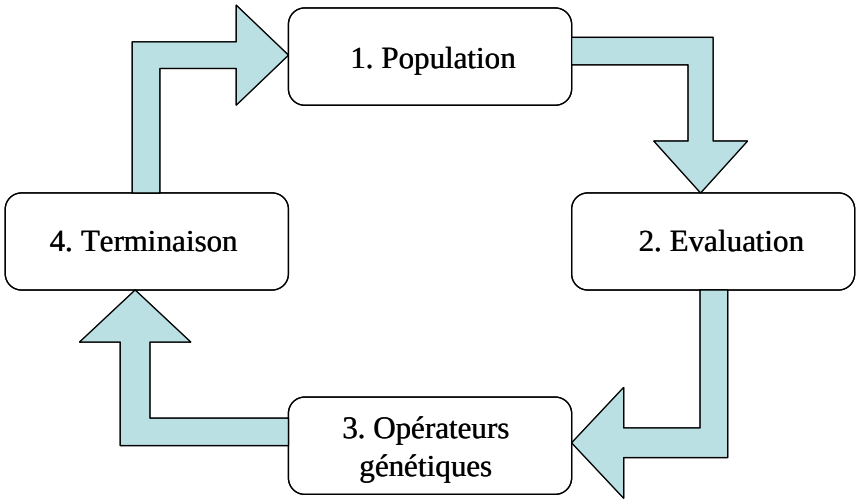


Figure 3 : Cycle de reproduction.

Les algorithmes génétiques ne sont pas applicables à tous les lots de données, la dimension de la matrice des données ***X*** étant un point important à prendre en compte dans le contexte des données spectroscopiques. En effet, comme dans la plupart des modèles quantitatifs, lorsque le nombre d'objets présentés est limité ou que les variables sont trop bruitées, le modèle va modéliser du bruit et non de l'information[(52)]. Des règles empiriques ont, donc, été proposées. Il est généralement recommandé que le nombre de variables n'excéde pas cinq fois le nombre d'échantillons[(52)]. Le risque de sur-entraînement reste malgré tout un facteur limitant pour l'utilisation des AG[(58)].

- ***Paramétrages de l'algorithme***

Le premier paramètre à définir est la taille de la population initiale c'est-à-dire le nombre d'individus (*h*). Une taille de population large permet d'avoir une représentation plus

exhaustive des différentes combinaisons des variables mais peut engendrer des temps de calculs non négligeables (de l'ordre de plusieurs heures). Une alternative à une taille de population large est l'utilisation de plusieurs itérations ou générations (étapes 2 à 4) [(10)]. De plus, si le nombre de variables (*m*) est très important, les gènes peuvent coder un groupe de variables (noté ***m***) plutôt que les variables individuelles. On définit alors un second paramètre correspondant à la largeur de cette fenêtre qui doit être adapté à la nature des données, notamment aux prétraitements utilisés. Par ailleurs, le nombre de variables à prendre en compte a priori est, en général, défini sous la forme d'un pourcentage qui correspond à la proportion des variables sélectionnées.

Une seconde série de paramètres à optimiser est composée des opérateurs génétiques, qui regroupent les *points de croisement* (C_s et C_d) et le *taux de mutation* (M_u). Dans la plupart des algorithmes, l'étape de recombinaison est réalisée soit par le *point de croisement simple* (C_s) soit par le *point de croisement double* (C_d). Concernant C_s, les gènes de deux individus pris aléatoirement sont séparés en deux. La première partie des gènes de l'individu A est échangée avec la première partie des gènes de l'individu B et les deux gènes hybrides forment deux nouveaux individus C et D[(60)]. Ceci est présenté de façon schématique sur la Figure 4. La procédure pour C_d est très similaire sauf que les gènes sont séparés en deux endroits. On notera néanmoins que le choix du *point de croisement double* produit généralement des nouveaux individus plus ressemblants aux parents[(61)].

Une fois la nouvelle population construite, l'opérateur *mutation* est appliqué (Figure 5). Cette opération permet d'éviter la sur-représentation ou la sous-représentation d'une variable dans la population en changeant arbitrairement la valeur d'un gène de 0 en 1 ou, inversement, de 1 en 0. Le taux de mutation est souvent compris entre 0,001 et 0,01[(45)]. La nouvelle population ainsi crée est de nouveau évaluée.

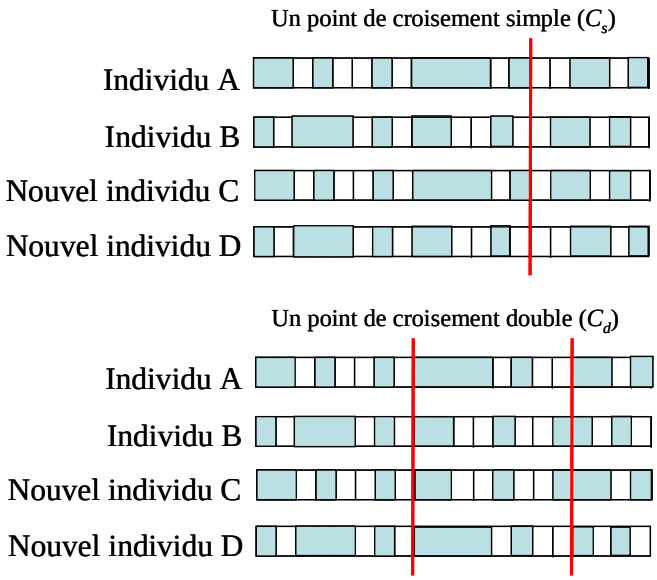


Figure 4 : Points de croisement pour la génération d'une nouvelle population.

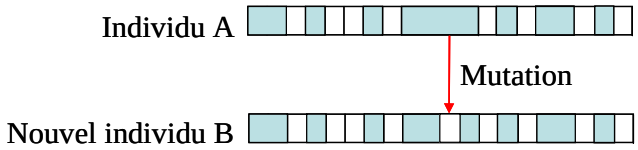


Figure 5 : Mutation.

Enfin, le critère d'arrêt de l'algorithme est défini soit par un nombre déterminé d'itérations[27] soit par l'obtention d'un certain pourcentage d'individus identiques[45]. Il faut noter que de par la nature stochastique de l'algorithme, les résultats des différentes applications des AG peuvent être légèrement différents. Les résultats sont donc moyennés en répétant la procédure plusieurs fois. Dans la littérature, le nombre d'itérations généralement proposé est souvent égal à cinq[51].

II.2. Sélection de variables par information mutuelle

L'information mutuelle (IM), qui estime la dépendance statistique entre deux variables, a été introduite récemment en spectroscopie proche infrarouge pour la sélection de

variables[62]. Par opposition aux méthodes fondées sur l'optimisation d'une fonction de coût comme par exemple l'erreur de validation croisée dans la méthode AG-PLS décrite précédemment, l'utilisation de l'information mutuelle peut se faire indépendamment des modèles prédictifs. L'information mutuelle est une méthode de sélection qui peut être appliquée avant diverses méthodes d'analyses quantitatives telles que la méthode de régression PLS ou l'application de réseaux de neurones artificiels (ANN)[63,64].

L'idée est d'utiliser une mesure de l'information mutuelle entre les variables spectroscopiques indépendantes *X*, et la variable dite dépendante *Y*, afin d'identifier la variable spectroscopique *X* ayant la plus grande information en commun avec *Y*. Une fois la première variable sélectionnée, une procédure incrémentale dite directe-indirecte (*forward-backward*) est utilisée afin de sélectionner les variables suivantes. Si l'idée d'appliquer le critère de l'information mutuelle parait pertinente, l'implémentation devient de plus en plus difficile lorsque le nombre de variables augmente. En effet, le calcul de l'information mutuelle entre un groupe de *k* variables *X* et les valeurs de *Y* nécessite l'estimation d'une fonction de densité de probabilité jointe de dimension (*k+1*), ce qui est très vite limitant du fait de la taille des lots de données analysés habituellement[62].

- ***Approche mathématique***

La théorie de l'information de Shannon et Weaver donne une définition de l'incertitude d'une variable aléatoire[65]. L'incertitude d'une variable aléatoire *Y* prenant *y* valeurs discrètes dans un ensemble *D* peut être mesurée par son entropie *H*(*Y*) décrite par l'Équation 1.

$$\text{Équation 1} \qquad H(Y) = -\sum_{y \in D} P(Y = y) \times \log P(Y = y)$$

$P(Y = y)$ représente la fonction de densité de probabilité de la variable aléatoire *Y* estimée en chaque point *y*. Afin d'illustrer le concept d'entropie, supposons un cas extrême où toutes les valeurs $y \in D$ ont une probabilité nulle exceptée l'une d'entre elles, *y**, qui a une probabilité égale à 1. Dans ce cas, l'entropie est nulle puisque la seule valeur prise par *Y* est *y**. Le *désordre* est nul et il n'y a aucune incertitude sur *Y*.

Si on s'intéresse à un couple de variables (*X*, *Y*), l'entropie conditionnelle notée $H(Y|X)$ est définie comme l'incertitude sur *Y* quand *X* est connue et peut être calculée selon l'Équation 2.

Équation 2 $$H(Y|X) = -\sum_x P(X = x)\sum_y P(Y = y|X = x) \times \log P(Y = y|X = x)$$

$P(Y = y|X = x)$ est la fonction de probabilité jointe de *X* et de *Y* estimée en chaque point *x* et *y*. Finalement, l'information mutuelle entre *Y* et *X* est la différence entre l'incertitude sur *Y* et l'incertitude sur la même variable connaissant *X*[47](Équation 3).

Équation 3 $$I(Y, X) = H(Y) - H(Y|X)$$

Cette équation peut s'interpréter comme la diminution de l'incertitude sur *Y* du fait de la connaissance sur *X*. L'information mutuelle entre *X* et *Y* est égale à zéro, si et seulement si, *X* et *Y* sont deux variables statistiquement indépendantes. De plus, on notera que l'information mutuelle ne fait pas de supposition sur la dépendance entre *X* et *Y*, relation linéaire ou non.

Les concepts d'entropie et d'entropie conditionnelle peuvent être étendus au cas des variables continues (ensemble *D* de taille infinie) et sont définis par les Équations 4 à 6.

Équation 4 $$H(u) = -\int f(u) \log f(u) du$$

Équation 5 $$H(y|x) = -\int f(x) \int f(y|x) \log f(y|x) dy dx$$

Équation 6 $$I(y, x) = \int h(x, y) \log \frac{h(x, y)}{f(x) g(y)} dx dy$$

où *f*(*u*), *f*(*x*) et *g*(*y*) sont des fonctions de densité de probabilité et *h*(*x*,*y*) est la fonction de densité de probabilité jointe de *X* et *Y*. Puisque $f(x) = \int h(x, y) dy$ et $g(y) = \int h(x, y) dx$, nous avons seulement besoin d'estimer *h*(*x*,*y*) afin d'estimer l'information mutuelle entre *X* et *Y*. Les logarithmes utilisés dans ces équations sont des logarithmes en base e[66].

- ***Estimation de la fonction de densité de probabilité jointe***

Nous avons vu précédemment, que le calcul de l'information mutuelle entre *X* et *Y* requiert l'estimation de la fonction de densité de probabilité jointe de *X* et *Y*, $P(Y|X)$. Pour des fonctions de densité de probabilité unidimensionnelles ou bidimensionnelles, les estimations sont généralement basées sur des histogrammes[62] ou des noyaux [47]. Dans le contexte des données spectrales, ces estimations sont basées sur la statistique des *k* plus proches voisins[49]. La qualité de l'estimateur *I*(*Y*,*X*) est alors liée à la valeur choisie pour le nombre de plus proches voisins (*k*). Ce choix doit respecter le dilemme biais-variance[49,67]. Lorsque la valeur de *k* est trop faible, l'estimateur favorise une petite valeur du biais mais une large valeur de la variance, alors qu'une grande valeur de *k* conduit à une répartition inverse de l'erreur. La valeur de *k* est généralement comprise entre 2 et 10[66,68] pour les applications en proche infrarouge.

- ***Application pour la sélection de variables***

L'information mutuelle peut être estimée entre un sous-ensemble de variables d'entrée $\{X_1, X_2,\ldots, X_q\}$, avec $q<m$, et une variable dépendante *Y*. L'algorithme proposé initialement par Rossi *et al* est détaillé ci-dessous[49,62]. Il s'agit tout d'abord de sélectionner la première variable, notée Xs_1, qui maximise l'information mutuelle avec *Y* (Équation 7), puis de sélectionner les variables suivantes par une procédure directe-indirecte (*forward-backward*) et enfin de terminer par une étape de combinaison.

$$\text{Équation 7} \qquad Xs_1 = \arg\max_{X_j}\{I(Y, X_j)\},\ 1 \le j \le m$$

La seconde variable, Xs_2, est sélectionnée ensuite parmi l'ensemble des variables restantes $\{X_j, 1 \le j \le m,\ j \ne s_1\}$, selon l'une des deux procédures détaillées ci-dessous.

La première procédure, notée A, consiste à sélectionner la variable qui a la plus grande information mutuelle avec *Y* (Équation 8).

$$\text{Équation 8} \qquad Xs_2 = \arg\max_{X_j}\{I(Y, X_j)\},\ 1 \le j \le m,\ j \ne s_1$$

Les variables suivantes sont sélectionnées de la même façon. Néanmoins, cette procédure conduit à la sélection de variables spectroscopiques très similaires c'est-à-dire potentiellement colinéaires. Cela augmente les risques de sur-entraînement. L'option A peut être considérée comme un algorithme de classement, les variables étant ordonnées en fonction de leur information mutuelle, la procédure peut donc s'arrêter après un nombre d'étapes donné.

L'autre procédure, notée B, consiste à sélectionner la variable Xs_2 qui maximise l'information entre le lot de variables déjà sélectionnées $\{Xs_1, Xs_2\}$ et la variable à prédire *Y*, Xs_1 étant déjà sélectionnée (Équation 9).

$$\text{Équation 9} \qquad Xs_2 = \arg\max_{X_j}\left\{I\left(Y, \left\{Xs_1, X_j\right\}\right)\right\},\ 1 \le j \le m,\ j \ne s_1$$

Dans les étapes suivantes, la $r^{ième}$ variable sélectionnée Xs_r sera choisie de la même manière (Équation 10).

$$\text{Équation 10} \qquad Xs_r = \arg\max_{X_j}\left\{I\left(Y, \left\{Xs_1, Xs_2, ..., Xs_{(r-1)}, Xs_r\right\}\right)\right\},\ 1 \le j \le m,\ j \ne \left\{s_1, s_2, ..., s_{(r-1)}\right\}$$

Dans le cas de l'analyse de données spectroscopiques, les variables consécutives très corrélées ne seront ainsi pas forcément sélectionnées. Néanmoins, la procédure présente l'inconvénient d'être sensible aux minima locaux, une procédure de sélection indirecte est donc réalisée au cours de chaque étape. Supposons *r* variables sélectionnées, l'étape indirecte consiste alors à éliminer une par une toutes les variables exceptée Xs_r, et à vérifier si chaque suppression permet d'augmenter l'information mutuelle. Le cas échéant, la ou les variables corrélées sont supprimées. La sélection de variables est stoppée lorsque l'information mutuelle calculée diminue. Lorsque cette procédure est appliquée aux données spectroscopiques, on constate souvent que très peu de variables sont sélectionnées. Pour s'affranchir de cette limitation, une étape de combinaison qui consiste en une recherche exhaustive a été proposée[49]. Le sous-ensemble finalement retenu est celui qui possède la plus grande valeur de l'information mutuelle. En pratique, tous les lots possibles qui peuvent provenir d'un sous-ensemble de taille *M* sont construits et l'information mutuelle est calculée. Généralement, la valeur de *M* est inférieure à 20.

III. Des méthodes de discrimination linéaires aux méthodes non linéaires

Les méthodes décrites dans ce paragraphe sont des méthodes utilisées pour la discrimination. Nous présentons dans ce chapitre les méthodes utilisées lors de notre étude, plus particulièrement, l'analyse discriminante linéaire (LDA)[37] et la méthode des *Support Vector Machine* (SVM)[39]. La méthode des SVM est une méthode issue de la théorie de l'apprentissage et introduite très récemment en chimiométrie. D'après les premiers résultats, cette méthode présente un bon potentiel pour la discrimination d'échantillons sur la base de données spectroscopiques[19,24,69].

III.1. Analyse discriminante linéaire (LDA)

D'après la littérature, la méthode LDA est la méthode la plus étudiée pour la classification supervisée. Dans le domaine du proche infrarouge, cette méthode a été appliquée pour la classification de produits pharmaceutiques[70], alimentaires[33], mais aussi pour la classification de déchets[71]. La méthode LDA a été proposée par Fisher[37]. L'objectif de l'analyse discriminante linéaire est de classer les échantillons en établissant une fonction linéaire qui sépare les classes présentes dans le lot d'entraînement[38]. Cette méthode est fondée sur la discrimination inter-classe. Il s'agit d'une méthode linéaire et paramétrique car elle suppose que la distribution des échantillons au sein des classes est gaussienne.

- ***Approche géométrique***

A titre d'exemple, considérons deux classes, **K** et **L**, dans un espace à 2 dimensions (Figure 6a). Les distributions des probabilités normales dont sont issus les échantillons sont représentées par des ellipses sur la Figure 6b. Ces ellipses correspondent aux limites de confiance pour l'appartenance à une classe, par exemple la limite à 95%.

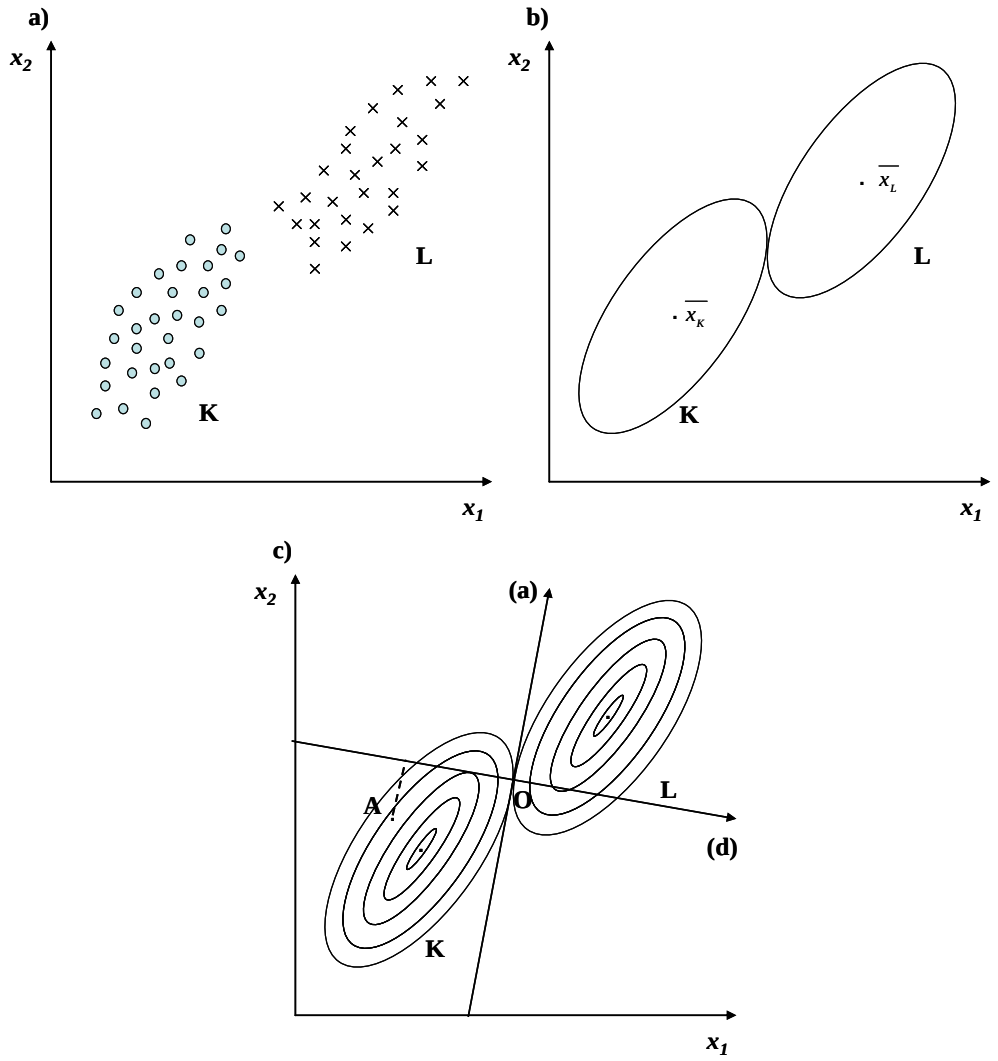


*Figure 6 : a) Distribution des échantillons dans les classes **K** et **L**, b) limites de confiance à 95% autour des centroïdes des classes **K** et **L**, c) représentation des limites de confiance à divers pourcentages, **O** est le point d'intersection de ces iso-probabilités, la droite (**a**) est la tangente à ces deux ellipses, (**d**) la direction de discrimination optimale, **A** un échantillon.*

Par principe, un objet appartient à la classe pour laquelle la probabilité calculée est la plus grande. Sur la Figure 6c, les ellipses successives représentent le même niveau de probabilité dans leurs classes respectives **K** et **L**. Les ellipses correspondantes aux limites de confiance à 95% se touchent au point **O** qui correspond, sur la figure, à la moitié de la distance entre les deux centres des classes. La droite (**a**) est la tangente à ces deux ellipses au point **O**. Les points situés à la gauche de (**a**) ont une plus grande probabilité d'appartenir à la classe **K** et réciproquement, ceux situés à droite de (**a**) ont une grande probabilité d'appartenir

à la classe **L**. D'un point de vue géométrique, la droite (**a**) peut être considérée comme une frontière séparant les deux classes. En pratique, nous préférerons une définition algébrique de la frontière. Afin d'illustrer ce propos, nous définissons une droite (**d**), perpendiculaire à la droite (**a**) et passant par le point **O** sur la Figure 6c. Chaque échantillon est ainsi projeté sur cette droite. La position du point **A** est donnée par ses coordonnées factorielles (*scores*) sur (**d**) et est définie par l'Équation 11[(35)].

$$\text{Équation 11} \quad D = w_0 + w_1 x_1 + w_2 x_2$$

Lorsque les données sont standardisées, la valeur de w_0 est nulle. Les coefficients w_1 et w_2 seront explicités dans le paragraphe ci-dessous. Ils sont tels qu'au point **O**, D=0 et que lorsque D>0 alors les échantillons appartiennent à la classe **L**, dans le cas contraire ils appartiennent à la classe **K**.

- ***Approche mathématique***

On recherche donc une fonction linéaire de variables, D, qui maximise le rapport entre les variances des deux classes **K** et **L**[(35)]. Le pouvoir discriminant des variables sera acceptable lorsque les centroïdes des deux lots d'échantillons seront suffisamment distants et que les groupes d'échantillons seront denses. D'un point de vue mathématique, cela signifie que la variance inter-classe est plus grande que la variance intra-classe (Équation 12).

$$\text{Équation 12} \quad D = \langle w, x \rangle + w_0$$

$\langle w, x \rangle$ étant le produit scalaire des deux vecteurs w et x. Les poids w sont adaptés aux caractéristiques des données pour permettre la discrimination. Pour une discrimination entre deux classes, les poids sont déterminés par l'Équation 13.

$$\text{Équation 13} \quad w^T = \left(\overline{x_L} - \overline{x_K}\right)^T S^{-1},\ w_0 = -\frac{1}{2}\left(\overline{x_L} - \overline{x_K}\right)^T S^{-1}\left(\overline{x_L} + \overline{x_K}\right)$$

$\overline{x_L}$ et $\overline{x_K}$ représentent les vecteurs des échantillons moyens qui décrivent la localisation des centroïdes dans un espace de dimension *m*. *S* est la matrice de variance-covariance combinée du lot d'entraînement des deux classes. L'utilisation d'une matrice de variance-covariance combinée implique que les matrices de variance-covariance des deux populations sont supposées être les mêmes. Par conséquent, les ellipsoïdes définies à partir des distributions des données devraient avoir en toute rigueur un volume égal (variance) et une même direction dans l'espace (covariance)[35]. Ceci est illustré sur la Figure 7.

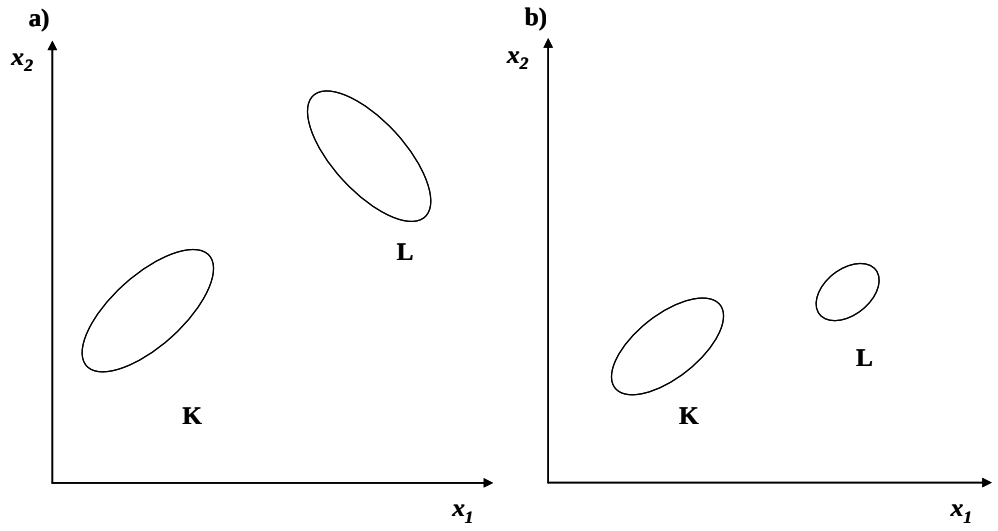


Figure 7 : a) Deux classes de même variance, b) deux classes de même covariance.

III.2. Méthode des Support Vector Machine

Les méthodes des *Support Vector Machine* (SVM) ont été proposées par Vapnik en 1995[72]. Ces méthodes ont été appliquées plus récemment à la classification d'échantillons sur la base de leurs données PIR, par exemple, en agroalimentaire[24,69] ou dans le domaine textile[19]. Afin de faire le parallèle avec la méthode LDA décrite précédemment, nous présentons tout d'abord le cas des données linéairement séparables, puis nous étendons ce cas aux données non linéairement séparables.

III.2.1 Principes de classification pour des données linéairement séparables

Supposons un espace $\mathcal{H}$ de dimension m contenant un ensemble fini de vecteurs $x_1, x_2, \ldots, x_n$. L'appartenance d'un vecteur à la classe **K** ou à la classe **L** est codée {-1} ou {1} dans la matrice des valeurs à prédire ***Y***. Le séparateur linéaire f peut être défini par l'Équation 14.

Équation 14 $$f(x) = \langle w, x \rangle + b \quad w \in \mathcal{H},\ b \in \Re$$

Avec les notations utilisées dans le chapitre 1, paragraphe III.1, $\langle w, x \rangle$ est le produit scalaire des vecteurs w et x, w est normal au séparateur linéaire et b traduit une translation des valeurs de f (décalage du séparateur). L'équation $f(x)=0$ définit la frontière de séparation entre les deux classes, qui dans le cas présent, est un hyperplan affine. Plusieurs hyperplans sont possibles pour séparer deux classes (Figure 8a). On cherche parmi ceux-ci, celui qui optimise la séparation des deux classes, intuitivement, on cherche l'hyperplan le plus « sûr ».

Lorsque $f(x)>0$, le vecteur x appartient alors à la classe des échantillons dont l'étiquette est **L** et réciproquement lorsque $f(x)<0$, le vecteur x appartient à la classe des échantillons d'étiquette **K** (Figure 8b).

Le terme Support Vector Machine (machines à vecteurs supports[41], séparateurs à vastes marges[40]) sera conservé dans ce manuscrit.

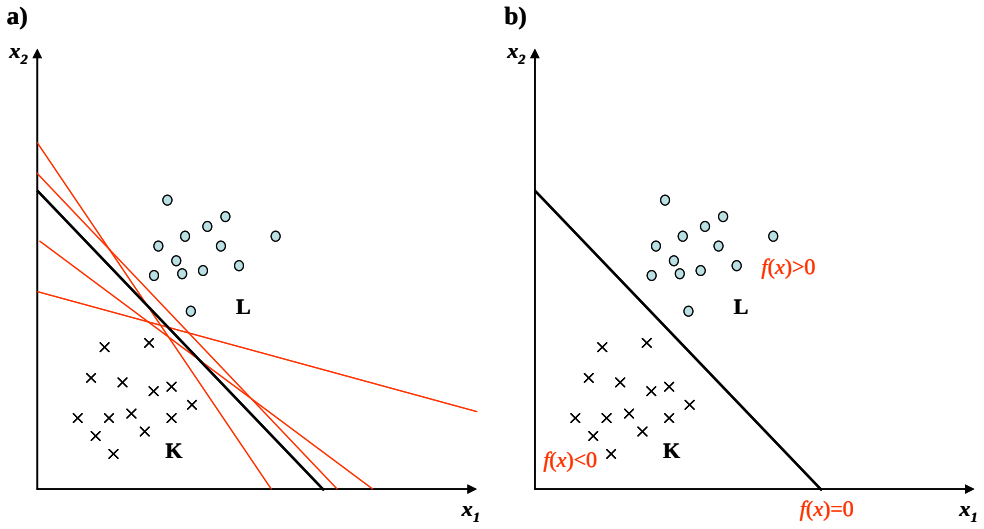


Figure 8 : Cas de données séparables par un hyperplan linéaire.

- ***Définition et optimisation de la marge***

La notion de marge sert à qualifier mathématiquement le fait qu'il existe parmi l'ensemble des solutions au problème de classification, un hyperplan qui permet de séparer les échantillons de façon optimale.

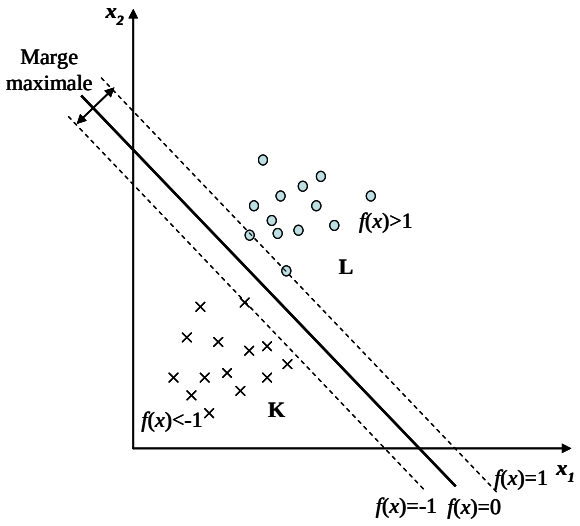


*Figure 9 : Marge entre la classe **K** et la classe **L**.*

La marge est la distance entre l'hyperplan séparateur d'équation $f(x)=0$ et un hyperplan parallèle représenté par une droite en pointillés sur la Figure 9 et qui contient les échantillons en bordure de classe. L'hyperplan parallèle est défini à une constante près, afin de simplifier la compréhension, nous avons choisi ici le cas où $f(x)=\pm 1$. La marge est définie par l'Équation 15[73] :

$$\text{Équation 15} \quad \rho(x_i, y_i) = y_i\left(\langle x_i, w\rangle + b\right)/\|w\|$$

Par conséquent, d'après l'Équation 15, la marge optimale revient donc à chercher une solution qui minimise la norme de *w*. Nous souhaitons déterminer les paramètres *w* et *b* qui minimisent une fonction de coût définie par l'Équation 16, tout en respectant la contrainte qui correspond à la règle de discrimination et qui est définie par l'Équation 17.

$$\text{Équation 16} \quad \min\left\{\frac{1}{2}\|w\|^2\right\}$$

$$\text{Équation 17} \quad \langle x_i, w\rangle + b \geq 1 \text{ si } y_i = 1$$

$$\langle x_i, w\rangle + b \leq -1 \text{ si } y_i = -1$$

pour tout $i=1, \ldots, n$.

- ***Formalisme de Lagrange et conditions de Karush-Kuhn-Tucker***

Le problème d'optimisation défini par les Équations 16 et 17 est un problème d'optimisation convexe [41], qui garantit la convergence des SVM vers la solution optimale. La résolution de ce type de problème passe par la définition d'une fonction de Lagrange donnée par l'Équation 18.

$$\text{Équation 18} \quad L_p(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i\left(\langle x_i, w\rangle + b\right) - 1\right)$$

où l'ensemble des paramètres $\{\alpha_i, i = 1, \ldots, n, \alpha_i \geq 0\}$ correspondent aux multiplicateurs de Lagrange. *Lp* est minimisée en ce qui concerne *w* et *b*. La solution au problème de

minimisation (Équations 16 et 17) est équivalente à déterminer le point neutre de la fonction *Lp* (forme primale du problème). Le point neutre correspond au point où les dérivées de *Lp* par rapport à *b* et *w* sont nulles. Cependant, *Lp* est maximisée en ce qui concerne α_i ce qui nécessite techniquement la résolution de la forme duale du Lagrangien (Équation 19).

Équation 19 $$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

L_D est maximisée pour les α_i tels que $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i y_i = 0$.

La forme duale permet d'exprimer le critère d'optimisation comme un produit scalaire des vecteurs descriptifs des données. Cependant, le formalisme de Lagrange n'est applicable que dans le cas des contraintes d'égalités. Pour résoudre les problèmes d'optimisation sous contraintes d'inégalités, il est nécessaire de tenir compte des corrections de Kuhn et Tucker [74] ainsi que de la condition complémentaire de Karush-Kuhn-Tucker[75] définies par les Équations 20 à 24.

Équation 20 $$\frac{\partial L_P}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

Équation 21 $$\frac{\partial L_P}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

Équations 22 et 23 $$y_i \left(\langle x_i, w \rangle + b \right) - 1 \geq 0, \ \alpha_i \geq 0,$$

Équation 24 $$\alpha_i \left(y_i \left(\langle x_i, w \rangle + b \right) - 1 \right) = 0$$

Les échantillons dont les valeurs des multiplicateurs de Lagrange sont strictement positives sont appelés vecteurs supports et il s'agit des échantillons les plus pertinents du lot de données d'entraînement. Pour illustrer l'importance de ces échantillons, si tous les autres échantillons sont éliminés du lot de données, la position de l'hyperplan séparateur ne sera pas modifiée. D'après l'Équation 24, ils se trouvent exactement sur la marge. Tous les échantillons restants du lot d'entraînement sont non pertinents car leurs multiplicateurs de Langrange sont nuls, leurs contraintes définies par l'Équation 22 sont néanmoins satisfaites mais ces derniers n'apparaissent plus dans l'Équation 20, ils ne participent pas à

l'optimisation de la fonction *Lp*. La classification d'un échantillon est seulement liée à l'optimisation de L_D, la fonction discriminante *f* établie par l'Équation 14 devient alors[76] :

$$\text{Équation 25} \qquad f(x) = \sum_{i \in SV} y_i \alpha_i \langle x_i, x \rangle + b$$

où *SV* est le lot de vecteurs supports.

III.2.2 Principes de classification pour des données non linéairement séparables

L'algorithme des SVM peut être appliqué aux données non linéairement séparables. Il est alors nécessaire de trouver un compromis entre la maximisation de la marge séparatrice définie par l'Équation 15 et la minimisation du nombre d'échantillons mal prédits lors du processus d'entraînement. Dans ce cas, la solution consiste à autoriser certains échantillons à être situé de part et d'autre de la marge définie sur la Figure 9. Pour cela, on introduit des variables de relâchement, ξ_i, pour chaque échantillon *i*.
La contrainte définie par l'Équation 17 devient alors pour tout *i*=1, …, *n* :

$$\text{Équation 26} \qquad y_i\left(\langle w, x_i \rangle + b\right) \geq 1 - \xi_i \text{ avec } y_i = 1$$
$$y_i\left(\langle w, x_i \rangle + b\right) \leq -1 + \xi_i \text{ avec } y_i = -1$$

Un échantillon sera mal classé par l'hyperplan séparateur si ξ_i est strictement supérieur à 1. Un terme de pénalité est introduit dans la fonction coût définie précédemment (Équation 16). Cette fonction de coût, notée τ, est redéfinie par l'Équation 27, où *C* est le paramètre de régularisation.

$$\text{Équation 27} \qquad \tau(w, \xi) = \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

Le terme $\frac{C}{n}\sum_{i=1}^{n}\xi_i$ correspond à une mesure reliée aux nombres d'échantillons mal classés. Plus la valeur de *C* est petite, moins ces échantillons contribuent à l'erreur totale. Pour résoudre le problème de minimisation de la fonction exprimée par l'Équation 27, la méthodologie est la même que celle employée précédemment. Les formes primale et duale du Lagrangien sont données, respectivement, Équation 28 et Équation 29 :

$$\text{Équation 28} \quad L_P = \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left(y_i\left(\langle w, x_i\rangle + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{n} r_i\xi_i$$

où $\alpha_i \geq 0$ et $r_i \geq 0$. Les α_i sont les multiplicateurs de Lagrange. Le paramètre r_i est introduit afin d'assurer $\xi_i \geq 0$.

$$\text{Équation 29} \quad L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle$$

L'Équation 29 est maximisée par rapport à α_i.

$$\text{Équation 30} \quad \sum_{i=1}^{n}\alpha_i y_i = 0\,,\ 0 \leq \alpha_i \leq C$$

Les conditions de Karush-Kuhn-Tucker sont :

$$\text{Équations 31} \quad \alpha_i\left(y_i\langle w, x_i\rangle + b - 1 + \xi_i\right) = 0$$

$$\text{Équation 32} \quad r_i\xi_i = \left(C - \alpha_i\right)\xi_i = 0$$

Les échantillons pour lesquels la valeur de α_i est strictement positive sont appelés vecteurs supports. Parmi ceux-ci, les échantillons qui satisfont la condition $0<\alpha_i<C$ possèdent des valeurs des variables de relâchement (ξ_i) nulles et se trouvent donc sur l'un des hyperplans canoniques situés à une distance égale à $1/\|w\|$ de l'hyperplan séparateur.

- ***Projection des données dans un espace de plus grande dimension***

Dans le cas de données non linéairement séparables, l'intérêt des méthodes SVM est que l'espace des données d'entrée est transformé en un nouvel espace de plus grande dimension dans lequel des méthodes linéaires peuvent être appliquées (Figure 10). Cette transformation est réalisée par une fonction non linéaire, notée ϕ.

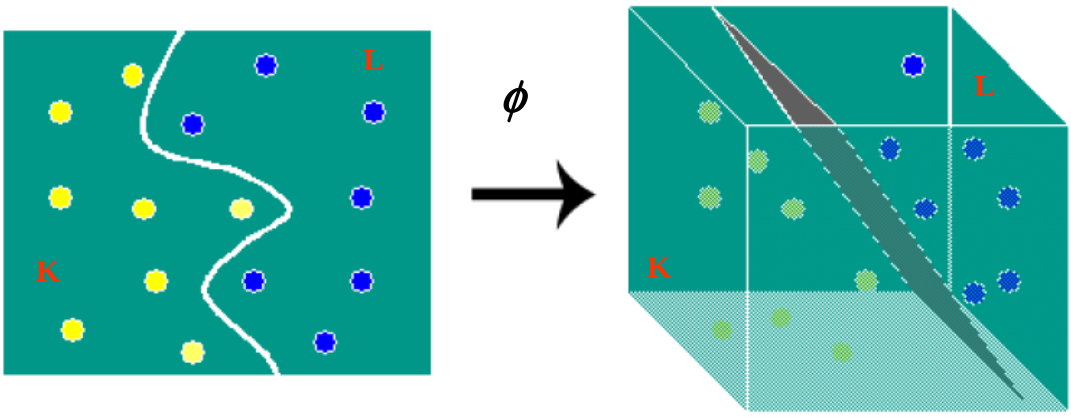


Figure 10 : Principe de la projection de données dans un espace de dimension supérieure.

Pour un problème de classification binaire, une fonction discriminante de la forme définie par l'Équation 33 est donc recherchée.

Équation 33 $$f(x) = \langle \phi(x), w \rangle + b$$

La forme duale du Lagrangien est donnée par l'Équation 34.

Équation 34 $$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$$

où α_i avec $i=1,...,n$ satisfont $0 \leq \alpha_i \leq C$ et $\sum_{i=1}^{n} \alpha_i y_i = 0$.

La solution pour *w* est $w = \sum_{i \in SV} \alpha_i y_i \phi(x_i)$ où *SV* est le lot de vecteurs supports associés aux valeurs de α_i satisfaisant $0<\alpha_i \leq C$.

La classification d'un échantillon étant liée à l'optimisation de L_D, seule la fonction noyau (*kernel K*) définie par le produit scalaire entre les vecteurs transformés (Équation 35) doit être calculée et le calcul explicite de la fonction ϕ peut être évité (*kernel trick*).

Équation 35 $$K(x, y) = \langle \phi(x), \phi(y) \rangle$$

La fonction discriminante *f* établie par l'Équation 33 devient alors :

Équation 36 $$f(x) = \sum_{i \in SV} \alpha_i y_i K(x_i, x) + b\,.$$

Les fonctions noyaux qui sont couramment utilisées dans la littérature par l'algorithme des SVM sont regroupées dans le Tableau 1. Il faut noter que les recherches dans le domaine de l'apprentissage consistent à élaborer des fonctions noyaux spécifiques des problèmes considérés, notamment en bioinformatique[76].

Non linéarité	Forme mathématique $K(x,y)$
Polynomiale	$(1 + \langle x, y \rangle)^d$
Gaussienne	$\exp\left(-\|x - y\|^2 / \sigma^2\right)$
Sigmoïde	$\tanh\left(k\langle x, y \rangle - \delta\right)$

Tableau 1 : Fonctions noyaux courantes[77].

- ***Interprétation géométrique***

Afin d'illustrer le concept du *kernel*, prenons l'exemple de données non linéairement séparables dans un espace à deux dimensions représenté sur la Figure 11. L'espace de départ ($\Re^2$) est transformé en un espace de plus grande dimension (ici $\Re^3$) par l'application d'une

fonction ϕ. Dans ce nouvel espace, les données sont linéairement séparables par un hyperplan qui correspond dans l'espace de départ à une surface de décision non linéaire. Dans ce cas, il s'agit d'un cercle (Figure 11c). Ce qui se traduit d'un point de vue mathématique de la façon suivante. On considère une fonction noyau polynomiale définie par l'équation :

Équation 37 $$K(x,x') = \langle x, x' \rangle^2 \text{ avec } x \text{ et } x' \in \Re^2$$

Ce kernel peut donc s'écrire de la façon suivante :

Équation 38
$$\begin{aligned} K(x,x') &= (x_1 x_1' + x_2 x_2')^2 \\ &= x_1^2 (x_1')^2 + 2 x_1 x_1' x_2 x_2' + x_2^2 (x_2')^2 \\ &= (x_1^2, \sqrt{2} x_1 x_2, x_2^2)((x_1')^2, \sqrt{2} x_1' x_2', (x_2')^2) \\ &= \langle \phi(x), \phi(x') \rangle \end{aligned}$$

où $\phi : \Re^2 \rightarrow \Re^3$

$x = (x_1, x_2) \rightarrow \phi(x) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)$

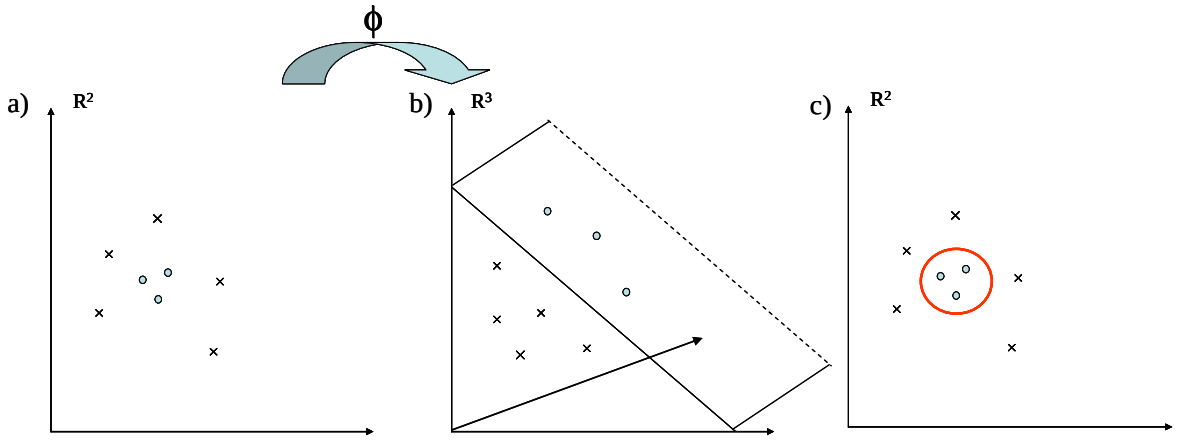


Figure 11 : a) Données dans $\Re^2$, b) Projection des données dans un espace de plus grande dimension $\Re^3$ par l'intermédiaire de la fonction noyau ϕ et calcul d'un hyperplan séparateur, c) Données reprojetées dans l'espace de départ $\Re^2$(76).

- ***Optimisation des paramètres***

Le paramètre principal à optimiser dans l'algorithme des SVM est le paramètre de régularisation *C* (Équation 27). Il exprime le compromis entre deux objectifs conflictuels : la minimisation du nombre d'échantillons mal prédits lors de la phase d'entraînement et la maximisation de la marge. Lorsque *C* devient grand, seul le nombre d'échantillons mal classés est considéré. Au contraire, lorsque *C* tend vers 0, la marge est maximisée sans tenir compte du nombre d'échantillons mal classés, ce qui peut conduire à des solutions aberrantes[19]. On cherche donc une valeur intermédiaire en faisant varier *C*. Le choix s'effectue sur la base des performances des modèles obtenus en validation croisée, par exemple.

Les paramètres définissant le *kernel* sont également à optimiser. Par exemple, si on utilise des fonctions radiales de base définies par l'Équation 39, il faut déterminer σ qui correspond à la largeur de la gaussienne.

Équation 39 $$\phi(x) = \exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right)$$

En général, les paramètres présentés précédemment sont optimisés sur des grilles d'optimisation, du type de celle proposée sur la Figure 12 définie par Hsu[78]. Cette grille permet de couvrir avec un maillage de dimension choisie, les différentes valeurs de *C* et de σ et de déterminer de façon graphique les valeurs de *C* et σ correspondantes aux pourcentages optimaux d'échantillons bien classés en validation croisée. Nous reviendrons en détails sur ces aspects dans le chapitre 3 du manuscrit.

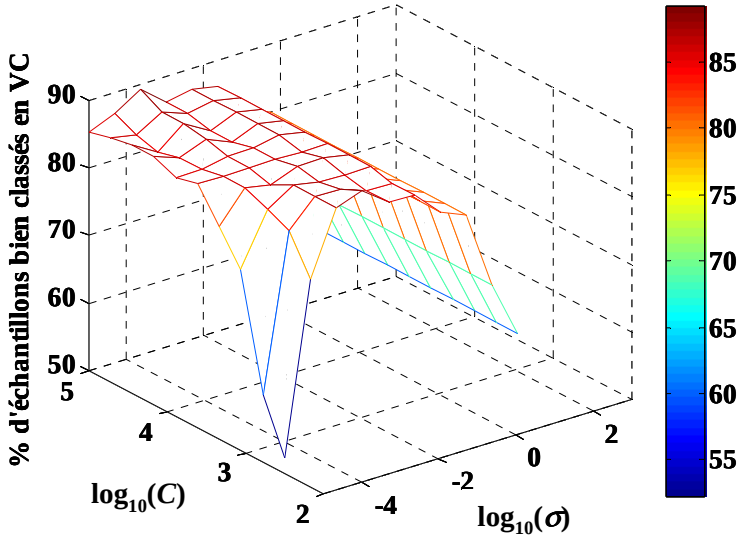


Figure 12 : Grille d'optimisation, pourcentage d'échantillons bien classés en fonction de log_{10} (C) et log_{10} (σ).

Chapitre 2

Analyse quantitative de mélanges de fibres tissées par spectroscopie proche infrarouge

---

L'objectif de ce travail est de développer des modèles prédictifs afin de déterminer la teneur en coton de deux types d'échantillons, les mélanges coton/polyester, d'une part, et les mélanges coton/viscose, d'autre part. Cette étude a été réalisée en collaboration avec l'Institut Français du Textile-Habillement (IFTH) qui nous a fourni les lots de données, c'est-à-dire les spectres PIR et les valeurs des mesures de référence (teneur en coton associée à chaque spectre). Les mélanges binaires étudiés sont représentatifs du marché de l'habillement.

Une des difficultés de cette étude est que l'analyse est réalisée directement sur les produits textiles et que l'on n'autorise pas la dégradation du tissu. Dans la littérature, des études ont montré que la spectroscopie proche infrarouge associée aux traitements chimiométriques peut être une alternative aux méthodes chimiques qui sont polluantes et relativement longues (de l'ordre de plusieurs heures). Par exemple, les travaux de Cleve *et al*[79] ont présenté la caractérisation des mélanges coton-polyester, ceux de Sohn *et al*[3] pour l'analyse quantitative de mélanges coton-lin, ou ceux de Carillo *et al*[80] pour la caractéristique thermale de différentes viscoses. La difficulté principale de cette étude est de respecter la directive européenne CE 96/74[81], qui définit la teneur d'un mélange de fibres à ±3% en masse.

Nous verrons dans un premier temps, une description des deux lots de données à travers l'attribution des bandes d'absorption caractéristiques. Nous discuterons des variabilités physiques des textiles et présenterons la méthode de référence. Dans un second

temps, l'analyse quantitative sur le lot de données coton/polyester sera présentée pour des modèles prédictifs établits sur le domaine spectral complet, puis sur les variables sélectionnées par la procédure des algorithmes génétiques (GA-PLS) et enfin, sur les variables sélectionnées par le calcul de l'information mutuelle (IM). La troisième partie de ce chapitre sera consacrée à l'analyse quantitative sur le lot de données coton/viscose. Une approche similaire sera appliquée. Les résultats seront discutés et les difficultés propres à ce type d'échantillons seront explicitées.

I. Description des données

Les spectres du polyester, du coton et de la viscose ont des bandes d'absorption caractéristiques dans le domaine proche infrarouge (Figure 14). Dans la littérature[3,79,80], ces bandes d'absorption ont été partiellement attribuées et nous en faisons le bilan.

- ***Polyester***

La fibre de polyester (PES) est une fibre synthétique dont la formule semi-développée est présentée sur la Figure 13. Les contributions spécifiques de la liaison C=O (Figure 14a) sont attribuées entre 5100 cm^{-1} et 5300 cm^{-1} à la seconde harmonique de la fonction ester RCOOR' (c'est-à-dire, à trois fois la fréquence fondamentale d'élongation de la liaison C=O située à ~ 1750 cm^{-1}). Les régions comprises entre 5600 et 6000 cm^{-1} et comprises entre 8000 et 9000 cm^{-1} correspondent, respectivement, à la première et à la seconde harmonique de la liaison C-H (à deux fois et trois fois, respectivement, la fréquence fondamentale d'élongation de la liaison C-H située à ~ 3010 cm^{-1}).

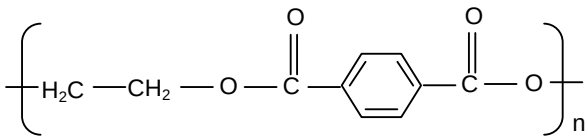


Figure 13 : Formule semi-développée du polyester.

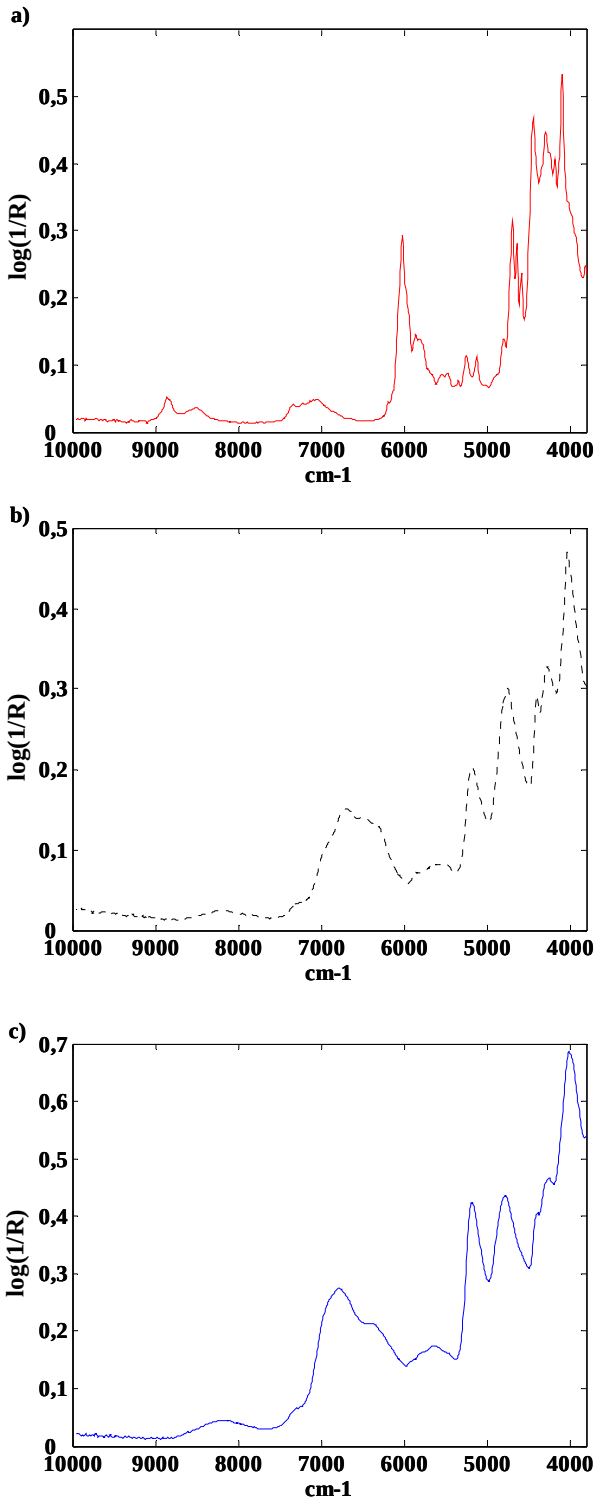


Figure 14 : Spectres bruts de tissus a) 100% polyester, b) 100% coton, c) 100% viscose.

- ***Coton***

La fibre de coton est une fibre naturelle constituée à 95% de cellulose[82] dont la formule semi-développée est présentée sur la Figure 15. Les autres constituants de cette matière sont des protéines, des acides organiques, des sucres, de la cire et d'autres minoritaires.

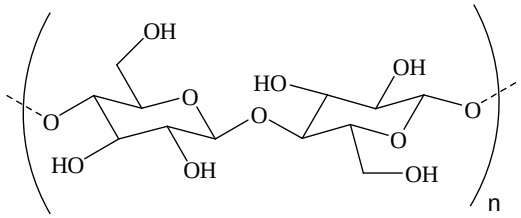


Figure 15 : Formule semi-développée de la cellulose.

Les principales bandes d'absorption du spectre proche infrarouge du coton sont attribuées à la cellulose :

– la bande d'absorption située proche de 4500 cm^{-1} correspond aux bandes de combinaisons des liaisons C-H, ce qui correspond à la somme des fréquences fondamentales d'élongation située à ~ 3050 cm^{-1} et de déformation située à ~ 1450 cm^{-1} ;
– la bande d'absorption située proche de 4850 cm^{-1} correspond aux bandes de combinaisons des liaisons O-H des fonctions alcool (R-OH), ce qui correspond à la somme des fréquences fondamentales d'élongation située à ~ 3500 cm^{-1} et de déformation située à ~ 1350 cm^{-1} ;
– la bande d'absorption située proche de 5100 cm^{-1} est attribuée aux bandes de combinaison des liaisons O-H, ce qui correspond à la somme des fréquences fondamentales d'élongation située à ~ 3500 cm^{-1} et de déformation située à ~ 1645 cm^{-1} ;
– la région comprise entre 6000 et 7500 cm^{-1} correspond à la région des premières harmoniques des liaisons C-H. La bande fondamentale d'élongation des liaisons C-H est située à 3010-3040 cm^{-1}.

De plus, on constate que les spectres du coton et de la viscose présentent une courbure de la ligne de base souvent attribuée à un effet de diffusion lié à la rugosité de la surface des fibres cellulosiques[83].

- ***Viscose***

On observe sur la Figure 16a, que les spectres du coton et de la viscose sont très similaires. En effet, ces fibres sont toutes les deux des fibres cellulosiques puisque la viscose est obtenue à partir de fibres cellulosiques chimiquement modifiées (Figure 17).

Les principales différences spectrales sont observées essentiellement dans la large bande d'absorption comprise entre 6000 et 7500 cm^{-1}. Afin de mettre en évidence ces différences, le spectre d'un échantillon 100% coton et le spectre d'un échantillon 100% viscose sont dérivés et présentés sur la Figure 16b. La dérivée utilisée est basée sur la technique de convolution de Savitzky et Golay[84]. Nous nous se focalisons sur la région comprise entre 6000 et 7500 cm^{-1} qui correspond aux premières harmoniques des liaisons C-H. Ce constat est en accord avec la formule semi-développée de la viscose (Figure 17) puisque la différence majeure entre le coton et la viscose est le groupement ($-C_3H_6OH$).

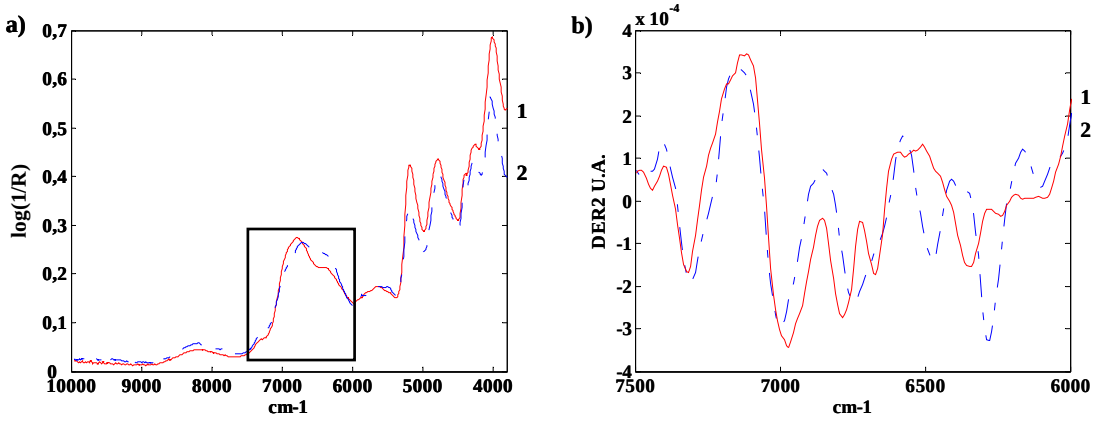


*Figure 16 : Spectres d'un échantillon (**1**) 100% viscose et d'un échantillon (**2**) 100% coton. a) spectres bruts sur le domaine 3800-10000 cm^{-1}. b) spectres prétraités par une dérivée seconde de Savitzky-Golay (DER2, 2, 2, 15) sur le domaine 6000-7500 cm^{-1}.*

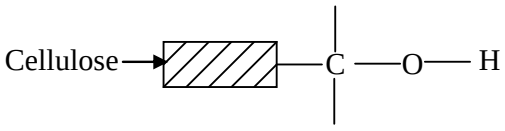


Figure 17 : Formule semi-développée de la viscose.

- ***Mélanges coton/PES et coton/viscose***

Dans cette partie de l'étude, les fibres de polyester, coton et viscose sont rencontrées dans des mélanges binaires coton/PES (lot $\boldsymbol{X_1}$) ou coton/viscose (lot $\boldsymbol{X_2}$). La teneur en coton dans ces mélanges varie de 0 à 100%. La répartition du nombre d'échantillons en fonction de la teneur en coton est présentée pour le lot $\boldsymbol{X_1}$ sur la Figure 18a et pour le lot $\boldsymbol{X_2}$ sur la Figure 18b.

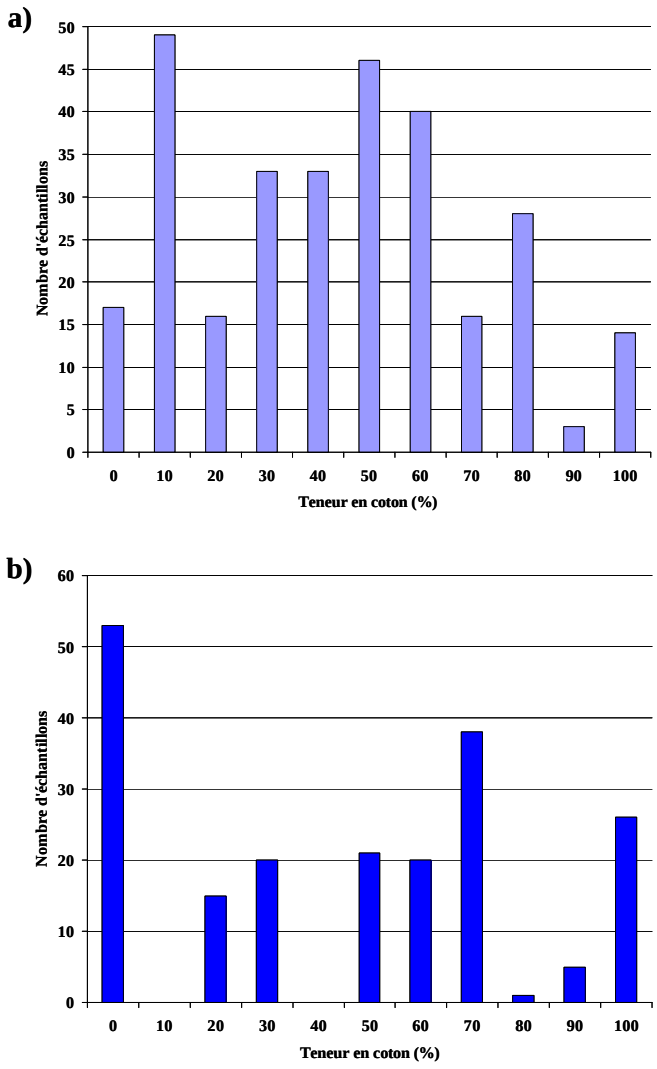


Figure 18 : Répartition des échantillons en fonction de la teneur en coton (%), a) pour le lot coton/PES ($\boldsymbol{X_1}$), b) pour le lot coton/viscose ($\boldsymbol{X_2}$).

Comme mentionné en introduction, les mélanges textiles des lots $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$ sont représentatifs du marché de l'habillement. Le mélange de fibres textiles le plus communément utilisé dans le domaine de l'habillement est le mélange coton/PES[83]. Nous constatons Figure 18, que la distribution des échantillons du lot $\boldsymbol{X_1}$ est relativement unimodale à l'exception des échantillons avec une teneur en coton de 10% qui sont sur-représentés. Au contraire, les échantillons du lot $\boldsymbol{X_2}$ ne couvrent pas la gamme des 0-100% coton. En effet, les mélanges coton/viscose sont moins répandus et certains mélanges de fibres sont d'ailleurs inexistants pour notre lot de données comme par exemple pour des teneurs en coton de 10% et 40%.

Dans le cadre de nos travaux, le domaine spectral s'étend de 3800 cm^{-1} à 10000 cm^{-1} avec une résolution spectrale de 7,7 cm^{-1}. Chaque spectre contient 806 nombres d'onde. Le lot $\boldsymbol{X_1}$ (Figure 19 a) comporte 318 échantillons. La matrice $\boldsymbol{X_1}$ a pour dimensions 318x806. Le lot $\boldsymbol{X_2}$ (Figure 19 b) contient 216 échantillons. Les dimensions de la matrice $\boldsymbol{X_2}$ sont 216x806. Une analyse exploratoire des données par une PCA a permis de révéler la présence d'échantillons aberrants (annexes 1 et 2). Les échantillons aberrants sont identifiés en réalisant des tests statistiques tels que le test du T^2 de Hotelling[85] et le test Q[86]. Nous constatons, suite à l'application de ces tests, que le lot $\boldsymbol{X_1}$ comporte 23 échantillons aberrants et que le lot $\boldsymbol{X_2}$ comporte 17 échantillons aberrants. La répartition des échantillons des matrices $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$ en un lot d'entraînement et en un lot de prédiction est réalisée par une procédure de distribution aléatoire. La distribution des échantillons du lot $\boldsymbol{X_1}$ est 186 échantillons en entraînement et 109 échantillons en prédiction. Pour le lot $\boldsymbol{X_2}$, les lots d'entraînement et de prédiction contiennent, respectivement, 146 et 53 échantillons. Diverses combinaisons de prétraitements sont testées.

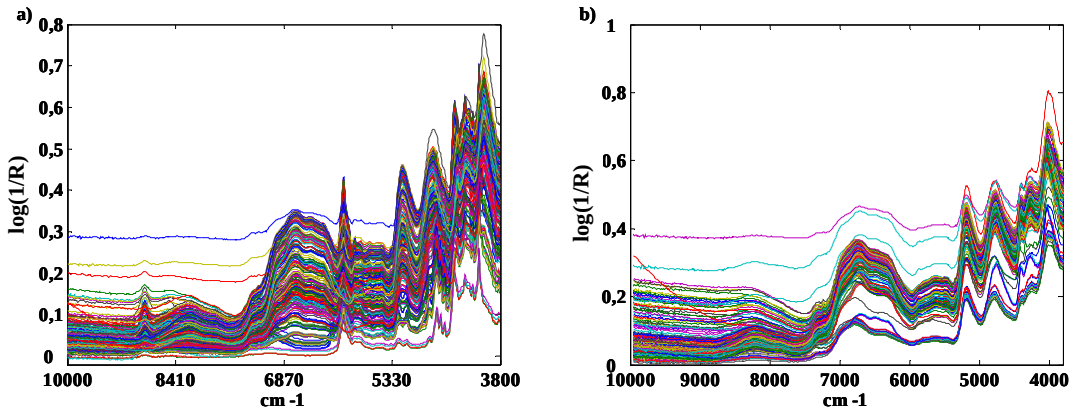


Figure 19 : a) Spectres bruts du lot $\boldsymbol{X_1}$ et b) du lot $\boldsymbol{X_2}$.

I.1. Variabilités physico-chimiques des textiles

Les variabilités physico-chimiques des échantillons textiles ont des origines diverses. On distingue celles liées à la fabrication des fils et des tissus (filature, tissage et tricotage) pour les fibres naturelles et chimiques et celles dues aux procédés d'élaboration des fibres dans le cas des fibres chimiques uniquement (la viscose et le PES).

Les échantillons étudiés sont des produits finis qui se présentent principalement soit sous la forme tissée, soit sous la forme tricotée. Ainsi, pour une composition chimique identique, les spectres de ces échantillons peuvent présenter des différences qui sont dues aux mélanges de fibres, aux aspects de surfaces (mat ou brillant)… Par ailleurs, les mélanges de fibres peuvent être réalisés soit au cours du filage, dans ce cas on parle de mélange intime, soit lors du tissage ou du tricotage de fils. Pour illustrer cette variabilité incontrôlable à notre niveau d'analyse, la Figure 20 est une représentation de 12 spectres d'échantillons de composition identique (30%coton/70%PES) extraits du lot $\boldsymbol{X_1}$.

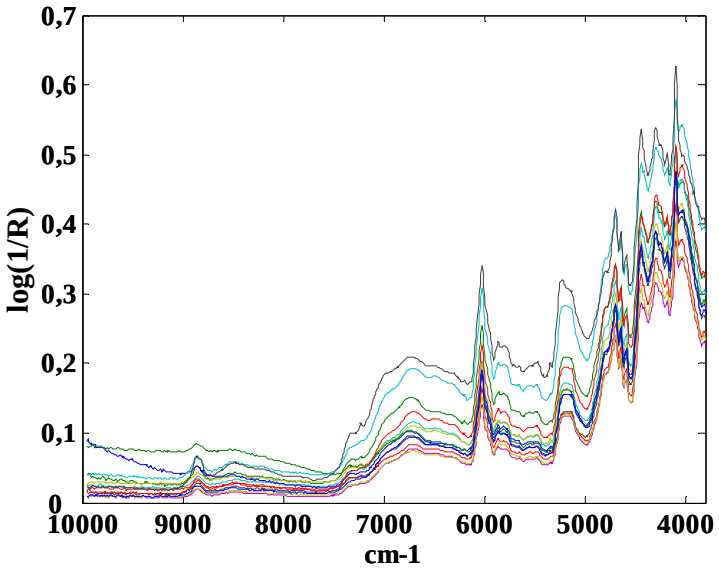


Figure 20 : Spectres bruts d'un ensemble d'échantillons de même nature.

Dans un second temps, nous revenons sur les variabilités dues aux procédés d'élaboration des fibres chimiques, en particulier pour la viscose. Dans le domaine du textile,

il existe trois principaux types de viscose (Viscose®, Modal® et Lyocell®) qui correspondent aux différents procédés de fabrication[80]. En fonction du solvant utilisé pour modifier la fibre cellulosique, la structure physique des fibres n'est pas exactement la même en ce qui concerne par exemple le taux de cristallinité ou l'orientation moléculaire[80]. Afin de mettre en évidence les différences entre les trois viscoses, nous avons représenté sur la Figure 21, les spectres dérivés (DER2) d'un échantillon 100% viscose de chacun des trois fabricants. Les différences majeures entre ces trois viscoses sont marquées dans la région 7000-7500 cm^{-1} qui correspond à la région des premières harmoniques des liaisons C-H.

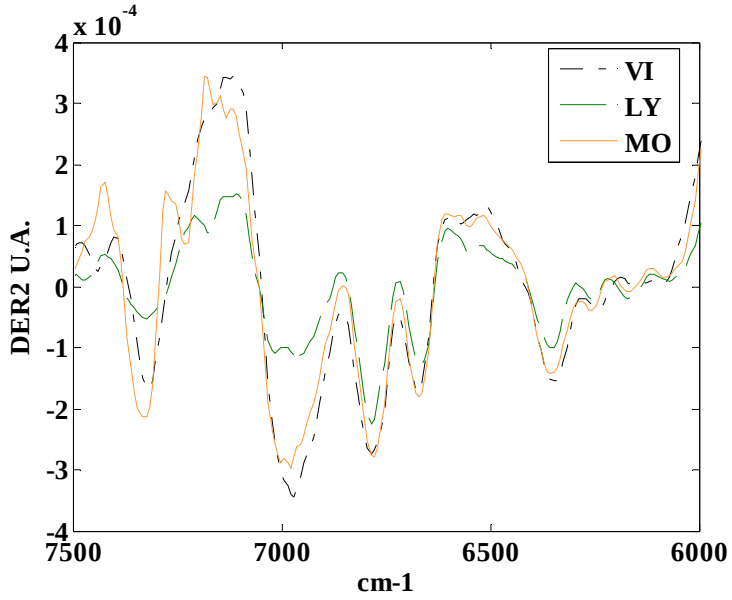


Figure 21 : Spectres dérivés (DER2) de 3 échantillons 100% viscose des différents fabricants.

I.2. Méthode de référence

La directive européenne CE 96/74[81] requiert que la composition d'un mélange de fibres pour un vêtement soit définie à ± 3% en masse. La méthode analytique de référence pour déterminer la composition des mélanges textiles dépend du type d'échantillon analysé. Dans le cas des mélanges où les fibres sont séparables (mélanges de fibres lors du tissage ou tricotage), la teneur des fibres est obtenue par une analyse microscopique[3]. Dans le cas de mélanges intimes, la séparation manuelle des différentes fibres n'est pas possible. La méthode

conventionnelle est alors une procédure gravimétrique au cours de laquelle la partie coton du mélange est dissoute. Pour les mélanges coton/PES, la dissolution est effectuée dans de l'acide sulfurique (H_2SO_4) concentré à 70%. Pour les mélanges coton/viscose, la dissolution est réalisée dans une solution de sulfate de zinc ($ZnSO_4$). Dans tous les cas, la teneur en coton du mélange est calculée à partir de la différence de masse entre le tissu d'origine et le tissu traité. La procédure complète nécessite 4 à 8 heures. Les procédures de référence réalisées par des dissolutions sont définies par des normes établies par l'organisation internationale de normalisation (ISO)[87,88]. La précision de ces méthodes est donnée de l'ordre du pourcent (±1%) pour un intervalle de confiance à 95%.

II. Analyse quantitative de matières textiles coton/PES

II.1. Régressions PLS sur spectres complets

Une analyse quantitative est réalisée sur les spectres complets afin de déterminer la teneur en coton dans les mélanges coton/PES. Les modèles PLS sont construits sur les échantillons du lot d'entraînement. Pour chaque modèle, la dimension est choisie en fonction des résultats de la validation croisée complète (*full cross-validation*) de type échantillon par échantillon (*leave-one-out*). On rappellera que les échantillons du lot de prédiction sont indépendants du lot d'entraînement afin que l'erreur de prédiction (RMSEP) soit non biaisée. Différentes combinaisons de prétraitements ont été testées. Nous présentons dans le Tableau 2 les résultats obtenus pour deux prétraitements particuliers, la SNV et la SNVDETDER2.

	SNV			SNVDETDER2		
	C	VC	P	C	VC	P
Nombre d'échantillons	186		109	186		109
Variables latentes (PLS)	15			13		
Nombre de variables	806			806		
R^2	0,995		0,998	0,997		0,996
RMSE(%)	1,61	2,20	2,93	1,71	2,54	2,62

Tableau 2 : Résultats obtenus sur les spectres complets prétraités SNV et SNVDETDER2. C : lot d'entraînement, VC : lot de validation croisée, P : lot de prédiction.

Les erreurs de prédiction (RMSEP) obtenues sur les spectres complets sont de 2,93% et 2,62%[89] pour les spectres traités SNV et SNVDETDER2, respectivement. D'une manière générale, ce niveau de performance est très appréciable dans le contexte industriel[81] car les résultats satisfont la directive CE 96/74 de la commission européenne qui requiert une tolérance à ± 3% en masse. Les modèles présentés requièrent, respectivement, 15 variables latentes pour les spectres prétraités SNV et 13 variables latentes pour les spectres SNVDETDER2. Ce niveau de complexité peut être expliqué en considérant d'une part, la variabilité présente dans les échantillons naturels du lot de données (Chapitre 2, paragraphe I.1) et, d'autre part, le fait que l'analyse proche infrarouge est effectuée directement sur des échantillons textiles intacts qui sont par nature des produits finis. Certains auteurs comme par exemple, Sohn *et al*[3] ont montré qu'en travaillant avec des échantillons broyés, on peut réduire de façon significative le nombre de variables latentes puisqu'on améliore l'homogénéité des échantillons analysés. Afin d'illustrer la relation entre les données chimiques et les prédictions du modèle quantitatif, nous présentons sur la Figure 22, les valeurs prédites en validation croisée en fonction des valeurs de la mesure de référence des 186 échantillons du lot d'entraînement et les valeurs prédites en prédiction en fonction des valeurs de la mesure de référence des 109 échantillons du lot de prédiction, pour les deux prétraitements.

Nous observons que pour les deux prétraitements, les valeurs de prédiction en fonction des valeurs de la mesure de référence sont satisfaisantes. Cependant, la prédiction des

échantillons purs et plus particulièrement des 100% PES (échantillon noté 1, par exemple) reste difficile aussi bien en validation croisée qu'en prédiction.

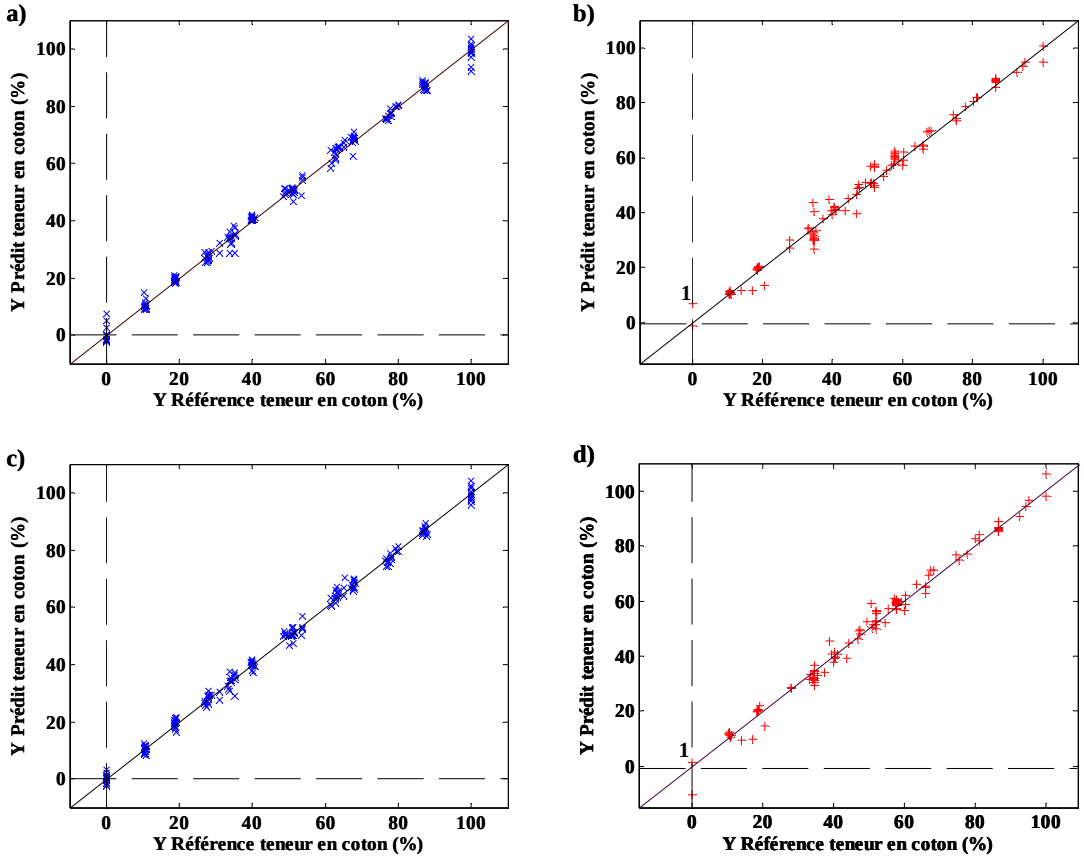


Figure 22 : Validation des modèles PLS sur les spectres complets. a) et b) respectivement, échantillons du lot d'entraînement en validation croisée et échantillons du lot de prédiction pour les spectres prétraités SNV, c) et d), respectivement, échantillons du lot d'entraînement en validation croisée et échantillons du lot de prédiction pour les spectres prétraités SNVDETDER2.

Nous remarquons que les résultats obtenus avec le prétraitement SNV ne sont pas significativement différents de ceux obtenus avec le prétraitement SNVDETDER2. Pour la sélection de variables qui va suivre, nous avons pris le parti de ne retenir que le prétraitement SNV, puisqu'il s'agit d'un prétraitement transférable aux données même, après la sélection des variables pertinentes. Au contraire, la SNVDETDER2 n'est applicable que sur les

spectres complets et ne s'appliquerait pas au cas de mesures spectroscopiques discrètes, du type des mesures fournies par un capteur, ce qui est un des objectifs important de ce travail.

II.2. Résultats de la sélection de variables

- ***Procédure des algorithmes génétiques***

Les algorithmes génétiques sont appliqués sur les échantillons du lot d'entraînement. Les paramètres des algorithmes génétiques sont précisés ci-dessous. La population initiale est composée de 256 individus. La largeur du gène (***m***) correspond à une fenêtre de 10 variables (~77 cm^{-1}). Ce choix prend en compte l'allure des spectres mais aussi les temps de calcul. Le nombre de variables présentes (la valeur du gène est égale à 1) initialement dans chaque individu correspond à 30% du nombre total de variables. Le nombre maximal de générations est limité à 100 et le pourcentage de convergence vaut 50%. Les paramètres génétiques sont le *point de croisement double*, C_d, et le *taux de mutation*, *Mu*, égal à 0,005. Cinq itérations sont réalisées pour lesquelles les erreurs de prédiction (RMSEP) seront moyennées. Nous noterons que les variables finalement conservées pour la construction des modèles PLS sont celles qui correspondent aux fenêtres prises en compte dans au moins 80% des modèles calculés.

La Figure 23 représente les 22 régions spectrales sélectionnées par la méthode AG-PLS[(89)]. Afin de faciliter l'interprétation, les spectres d'un échantillon pur coton et d'un échantillon pur polyester sont ajoutés sur le graphique. Nous constatons que les 22 régions spectrales sont distribuées sur tout le domaine spectral de 3800 à 10 000 cm^{-1}. Nous pouvons observer également que la majorité des intervalles sélectionnés correspond aux bandes d'absorption du PES, notamment celle située à ~ 6000 cm^{-1} attribuée précédemment. Nous constatons enfin que l'information contenue dans la région comprise entre 6000 et 7000 cm^{-1}, qui est associée à la teneur en coton, n'est pas sélectionnée. Les échantillons de coton purs sont très affectés par les variabilités physico-chimiques mises en évidence dans le chapitre 2, paragraphe I.1. De plus, nous rappellons qu'il s'agit d'un mélange binaire de constituants ayant des caractéristiques chimiques très différentes. Cela tend à expliquer le fait que la sélection de bandes d'absorption du coton n'est pas nécessaire. Néanmoins, vue la complexité des modèles construits, la nature de l'échantillon reste certainement la difficulté principale.

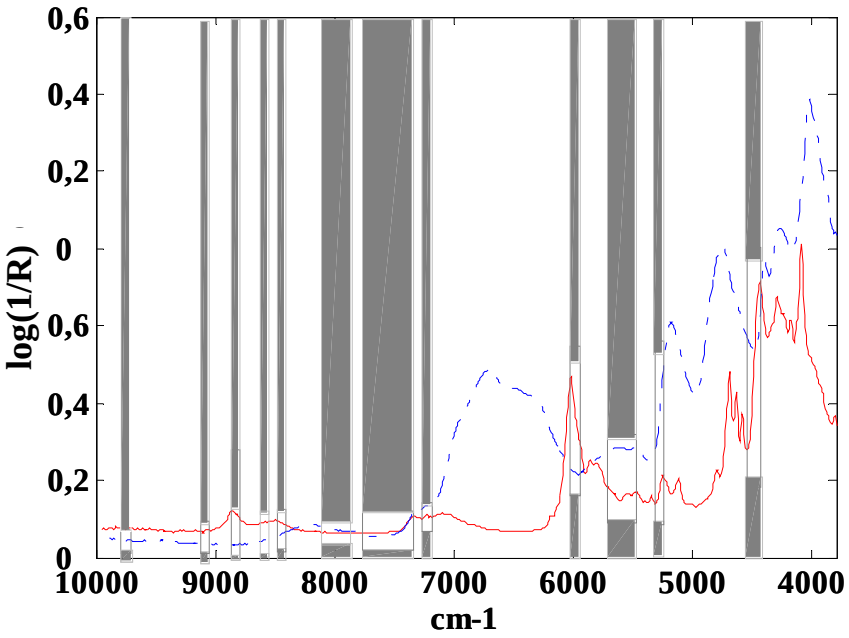


Figure 23 : Sélection de variables par la méthode des AG-PLS. Les régions spectrales sélectionnées sont représentées par des rectangles gris. Les spectres d'un échantillon pur coton et d'un échantillon pur polyester sont reportés en pointillé et en trait plein, respectivement.

- ***Procédure de l'information mutuelle***

Comme nous l'avons décrit dans le chapitre 1 au paragraphe II.2, la sélection de variables par le calcul de l'information mutuelle (IM) permet de déterminer la dépendance statistique entre deux variables. Elle est réalisée en estimant cette dernière sur la base des *k* plus proches voisins. Dans notre étude, nous avons pris *k*=6 pour les deux lots de données. La procédure de sélection par IM est réalisée sur les 186 échantillons du lot d'entraînement. Afin de pouvoir comparer les zones sélectionnées par AG-PLS et par IM, les spectres considérés ici sont prétraités SNV. Dans un premier temps, la procédure B conduit à la sélection d'un ensemble de quatre variables. Cet ensemble restreint contient les variables qui correspondent aux nombres d'onde suivants : 5994, 5147, 4839, 4454 cm^{-1}. Une seconde étape est ensuite réalisée afin de compléter éventuellement cet ensemble. Les variables ajoutées correspondent aux variables ayant les valeurs d'information mutuelle les plus grandes. Ainsi, onze variables

supplémentaires sont finalement considérées de façon à construire un ensemble de 15 variables (M=15). Une recherche exhaustive est ensuite réalisée pour toutes les combinaisons possibles de ces 15 variables et, finalement, la combinaison de huit variables est optimale. Les huit variables sélectionnées par la méthode de l'information mutuelle sont représentées sur la Figure 24. Nous remarquons que sur les quatre variables sélectionnées par la procédure directe-indirecte, seules deux variables sont conservées à 5994 cm^{-1} et à 4839 cm^{-1}. Les variables situées à 5994 cm^{-1}, 6071 cm^{-1} et 8843 cm^{-1} correspondent aux bandes du PES. A la différence de la sélection par la procédure AG-PLS, des variables sont choisies entre 6000 et 7500 cm^{-1}, c'est-à-dire dans la large bande d'absorption du coton.

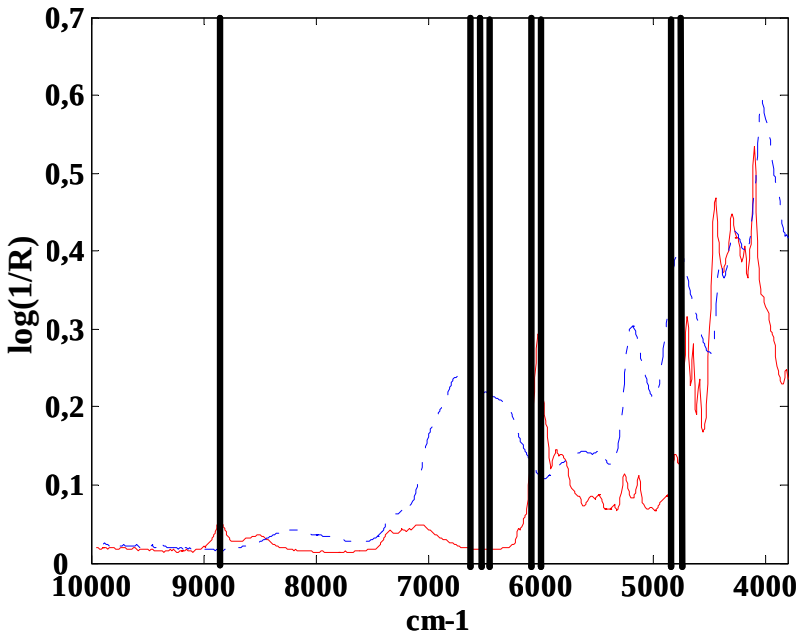


Figure 24 : Variables sélectionnées par IM (4762 cm^{-1}, 4839 cm^{-1}, 5994 cm^{-1}, 6071 cm^{-1}, 6456 cm^{-1}, 6533 cm^{-1}, 6610 cm^{-1}, 8843 cm^{-1}).

II.3. Modèles prédictifs sur les spectres réduits

Les variables retenues par la procédure AG-PLS étant sélectionnées en fonction de l'erreur de validation croisée, seule une régression PLS est construite sur les 22 fenêtres de 10 variables identifiées par AG-PLS. A l'opposé, la mesure de la pertinence des variables par l'information mutuelle est une méthode de sélection indépendante qui n'est pas fondée sur le résultat d'un modèle particulier. La construction d'un modèle PLS ou d'un modèle ANN est donc envisageable sur les huit variables sélectionnées par IM. Les résultats obtenus pour les modèles AG-PLS, IM-PLS et IM-ANN sont présentés dans le Tableau 3.

	AG-PLS			IM-PLS			IM-ANN		
	C	VC	P	C	VC	P	T	V	P
Nombre d'échantillons	186		109	186		109	124	62	109
Variables latentes (PLS)	11			4					
Architecture (ANN)							8*3*1		
Nombre de variables	22x10			8			8		
R^2	0,997		0,995	0,977		0,973	0,997		0,995
RMSE(%)	1,75	2,07	2,30	4,21	4,45	4,60	2,25	2,39	2,53

Tableau 3 : Résultats obtenus sur les spectres réduits. C : lot d'entraînement, VC : lot de validation croisée, P : lot de prédiction, T : lot d'apprentissage, V : lot de validation.

Sur la Figure 25, les valeurs de *Y* prédit en fonction de *Y* référence sont tracées pour les 186 échantillons du lot d'entraînement et pour les 109 échantillons du lot de prédiction.
Par comparaison aux prédictions obtenues pour les spectres complets, nous constatons que les échantillons purs, et particulièrement les 100% PES, sont mieux prédits après la sélection de variables (échantillon 1).

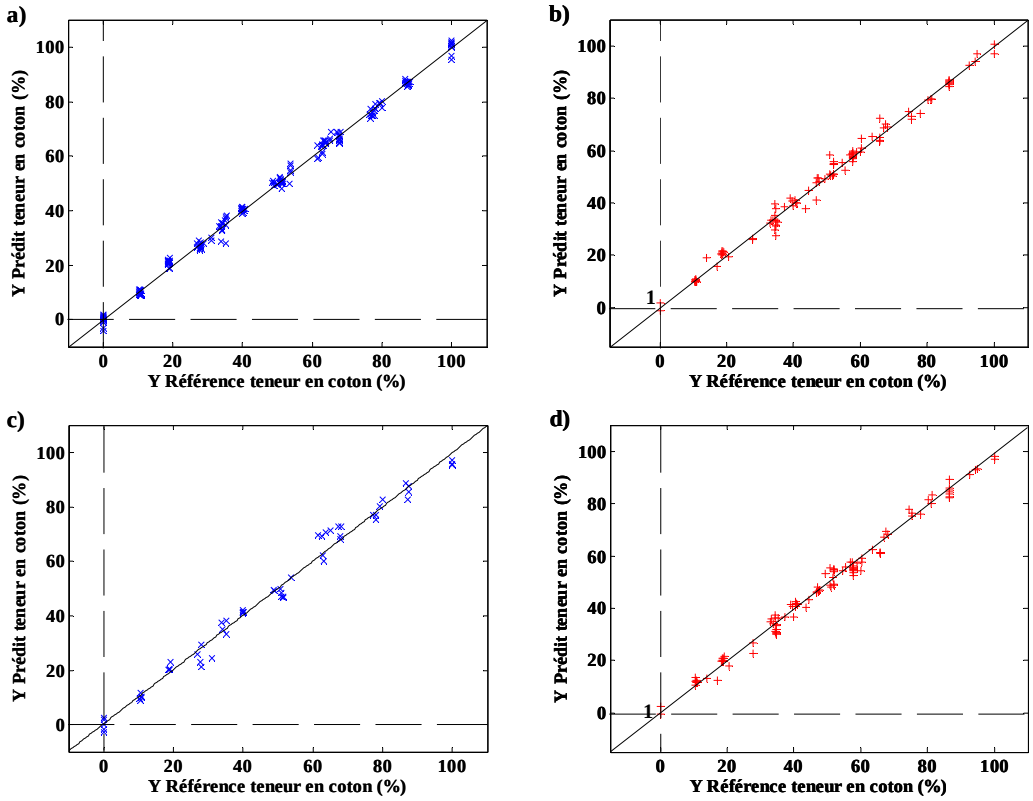


Figure 25 : a) Validation du modèle AG-PLS sur les 22 fenêtres de 10 variables en validation croisée sur les 186 échantillons du lot d'entraînement, b) en prédiction, c) validation du modèle IM-ANN sur les 8 variables pour les 62 échantillons du lot de validation, d) en prédiction.

Les erreurs de validation croisée et de prédiction obtenues sur les 22 fenêtres de 10 variables sont de 2,07% et 2,30%, respectivement, alors qu'elles étaient de 2,20% et 2,93% pour le modèle sur spectres complets (chapitre 2, paragraphe II.1). La dimension du modèle est réduite à 11 variables latentes au lieu de 15 variables sur les spectres complets. Cette amélioration est appréciable par rapport au modèle sur les spectres complets. Ceci peut être discuté en terme de biais-variance [67,89]. La complexité du modèle (et donc la variance) diminuant, la robustesse du modèle d'étalonnage doit augmenter, ce que semblent traduire les erreurs obtenues en validation croisée.

En ce qui concerne le modèle IM-PLS (Tableau 3), les résultats sont proposés à titre de comparaison. La complexité du modèle est réduite par rapport au modèle sur les spectres complets. Par contre, les erreurs de validation croisée et de prédiction sont multipliées par

deux. Par conséquent, comme nous pouvions le prévoir (Chapitre 1, paragraphe II.2), le modèle PLS n'est pas le modèle le plus adapté après une sélection de variables par IM.

Par ailleurs, nous avons réalisé un réseau de neurones sur les huit variables sélectionnées par IM. Dans cette étude, l'algorithme d'entraînement est une propagation inverse élastique (*resilient back-propagation*, Rprop) qui est appliquée avec une procédure d'*early stopping*. Les paramètres d'apprentissage sont Δ_0, η^-, η^+, $\Delta_{max}^{(90)}$. Δ_0 est égal à 0,005, η^- et η^+ valent, respectivement, 0,5 et 1,2 et Δ_{max} est égal à 50. On notera dans le Tableau 3 que la procédure d'*early stopping* requiert la séparation du lot d'entraînement en un lot d'apprentissage et un lot de validation contenant, respectivement, 124 et 62 échantillons. Les capacités prédictives du modèle ANN sont estimées sur le lot de prédiction qui lui reste inchangé (109 échantillons). L'apprentissage des réseaux de neurones est répété dix fois en initialisant avec des distributions aléatoires de poids. L'architecture du réseau de neurones est la suivante, 8 neurones sur la couche d'entrée, 3 neurones sur la couche cachée et 1 neurone sur la couche de sortie. Les erreurs de prédiction obtenues sont un peu plus élevées que celles du modèle AG-PLS. En effet, l'erreur de validation croisée est de 2,39% et l'erreur de prédiction est de 2,53%. Ces valeurs restent satisfaisantes compte tenu de l'objectif de réduction du nombre de variables qui lui est pleinement rempli. En effet, le modèle n'est construit que sur les huit variables sélectionnées par la procédure IM.

- ***Discussion***

Les sélections de variables par AG-PLS et par IM sont deux procédures fondées sur des concepts différents. Par conséquent, les variables sélectionnées ne sont pas nécessairement les mêmes. Le principal avantage de l'information mutuelle est que la mesure de la pertinence peut être appliquée comme un prétraitement, en plus d'un prétraitement classique tel que la SNV, et surtout indépendamment du modèle d'étalonnage. De plus, aucune supposition sur la relation entre les variables X et les valeurs de la mesure de référence Y n'est nécessaire. D'un point de vue de la sélection de variables, la procédure d'information mutuelle peut être préférable à la procédure AG-PLS car elle ne représente au final que 8 variables contre 22 fenêtres de 10 variables dans le premier cas. Cependant, d'un point de vue prédictif, la procédure AG-PLS permet d'améliorer les résultats. Pour ce mélange de fibres binaires

coton /polyester, nous noterons que les différentes méthodes satisfont les exigences de la directive CE 96/74 de la commission européenne qui requiert une tolérance à ± 3%.

III. Analyse quantitative de mélanges textiles coton/viscose

Nous avons montré dans le paragraphe précédent, que pour le lot de données coton/polyester, la sélection de variables permet de maintenir de bonnes capacités prédictives. Ce travail est répété pour le second lot de données ($\boldsymbol{X_2}$) qui contient des mélanges de fibres de coton et de fibres de viscose. La tâche est plus difficile pour ce lot du fait de la similitude des bandes d'absorption des fibres de coton et de viscose, toutes deux d'origine cellulosique.

III.1. Régressions PLS sur spectres complets

Nous reprenons pour le lot $\boldsymbol{X_2}$, la démarche détaillée précédemment sur le lot $\boldsymbol{X_1}$. Un modèle de régression PLS est réalisé sur les spectres complets. La région spectrale au-delà de 7500 cm^{-1} est très bruitée et ne présente pas d'information spectrale spécifique. Par conséquent, nous travaillons sur le domaine 3800-7500 cm^{-1}. Les spectres comportent ainsi 480 variables. Le modèle PLS construit requiert 11 variables latentes et les erreurs obtenues en entraînement, en validation croisée et en prédiction sont, respectivement, 2,12%, 2,84% et 3,74% pour des échantillons avec des teneurs en coton variant de 0 à 100%. Ces résultats sont très encourageants compte tenu de la similitude entre les fibres de coton et les fibres de viscose. Sur la Figure 26, nous présentons, pour les 146 échantillons du lot d'entraînement et pour les 53 échantillons du lot de prédiction, les valeurs des teneurs en coton en fonction des valeurs cibles. Cependant, nous pointerons deux échantillons qui doivent être considérés avec précaution puisque leur teneur en coton n'est pas représentée dans le lot d'entraînement du fait de la procédure de sélection aléatoire des échantillons. Il s'agit de l'échantillon noté 18 qui contient 52% de coton, mais qui est prédit à 64%. L'autre échantillon noté 26 est un mélange de 95% coton mais est prédit à 83%.

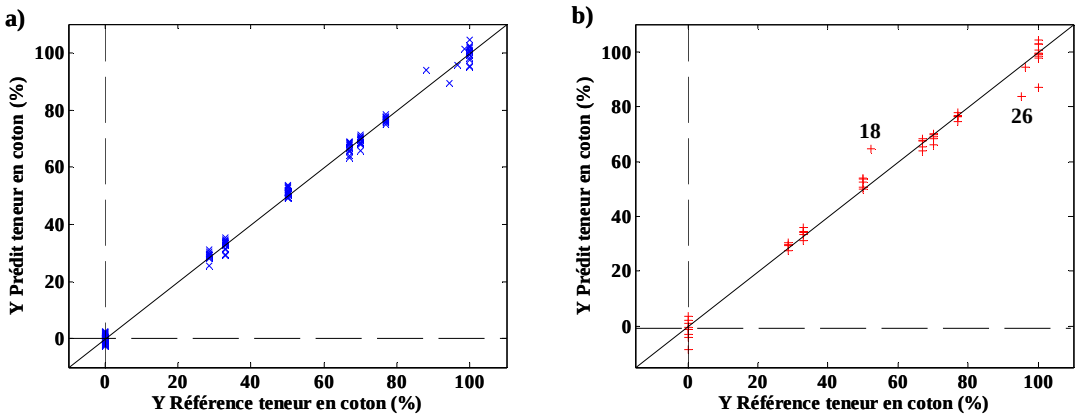


Figure 26 : a) Validation du modèle PLS sur les spectres complets pour les mélanges coton/viscose en validation croisée, b) en prédiction.

La remarque faite précédemment sur les échantillons purs reste valable. Les valeurs prédites à 0 et à 100% s'étendent sur un large intervalle autour des valeurs cibles.

III.2. Résultats de la sélection de variables

- ***Procédure des algorithmes génétiques (AG-PLS)***

La procédure AG-PLS est appliquée au lot $\boldsymbol{X_2}$ (coton/viscose), le paramétrage est identique à celui proposé précédemment (lot $\boldsymbol{X_1}$). Les zones sélectionnées sont représentées sur la Figure 27, les variables retenues sont codées par des rectangles gris. Les AG-PLS ont sélectionné 12 régions de 10 variables chacune[91]. Ce qui représente environ 25% des variables de départ. Ces 12 régions couvrent tout le domaine spectral. La similitude entre les spectres d'un échantillon pur coton et d'un échantillon pur viscose, rend difficile l'interprétation des zones sélectionnées car il s'agit de deux fibres d'origines cellulosiques. Cependant, les spectres présentent essentiellement des différences dans les bandes d'absorption situées entre 6000 et 7500 cm^{-1}. Nous noterons qu'à la différence du lot $\boldsymbol{X_1}$, la large bande du coton (6000-7500 cm^{-1}) est sélectionnée. En effet, cette gamme spectrale correspond à la région des premières harmoniques des liaisons C-H, ce qui est en accord avec

la formule semi-développée de la viscose où la différence entre le coton et la viscose est le groupement $-C_3H_6OH$ (chapitre 2, paragraphe I).

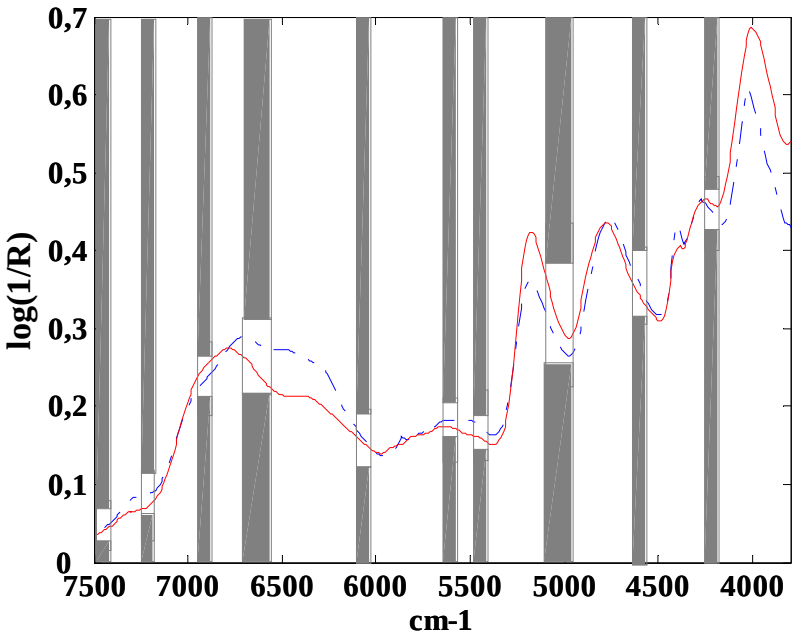


Figure 27 : Régions spectrales sélectionnées par la méthode AG-PLS.

- ***Procédure de l'information mutuelle***

La sélection de variables par IM est appliquée sur les 146 spectres des échantillons du lot d'entraînement. La procédure B (Chapitre 1, paragraphe II.2) sélectionne successivement un sous-ensemble de cinq variables qui sont les suivantes : 5301, 5532, 3992, 4070 et 5147 cm^{-1}. Une seconde étape est ensuite réalisée afin de compléter cet ensemble. Les variables qui sont ajoutées correspondent aux variables ayant les valeurs d'information mutuelle les plus élevées. Dix nouvelles variables sont incluses dans cet ensemble de façon à avoir M=15 et après la recherche exhaustive 12 variables sont finalement retenues[(91)] (Figure 28). Nous pouvons observer que seule la variable située à 5147 cm^{-1} qui correspond à la dernière variable sélectionnée par la procédure directe-indirecte n'est pas prise en compte dans les variables finales.

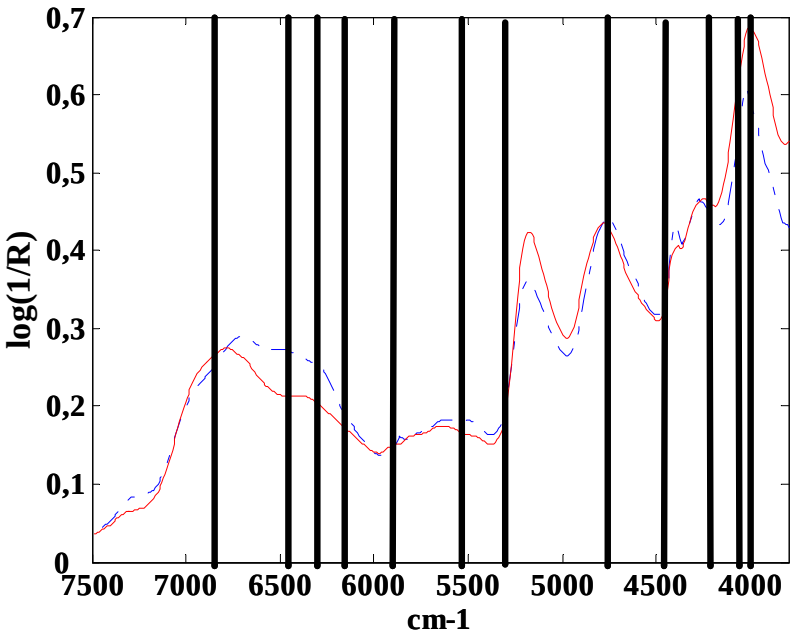


Figure 28 : Variables sélectionnées par IM (3992 cm^{-1}, 4070 cm^{-1}, 4223 cm^{-1}, 4454 cm^{-1}, 4750 cm^{-1}, 5301 cm^{-1}, 5532 cm^{-1}, 5917 cm^{-1}, 6100 cm^{-1}, 6302 cm^{-1}, 6450 cm^{-1}, 6841 cm^{-1}).

Pour l'interprétation, nous mettrons l'accent sur l'observation des pseudo-absorbances en fonction de la teneur en coton à ces longueurs d'onde sélectionnées. Pour certaines variables, en particulier celles situées à 4070, 4750, 6100 et 6450 cm^{-1}, la corrélation entre les valeurs d'absorbance et la teneur en coton n'est pas idéalement linéaire (Figure 29). Seule l'IM sélectionne ces variables. En effet, elles ne sont pas considérées par la méthode AG-PLS qui sélectionne essentiellement les variables pour lesquelles les valeurs des absorbances sont linéairement corrélées à la teneur en coton. De plus, nous constatons une dispersion relativement importante des valeurs des absorbances pour les échantillons pur coton et pur PES, même après l'application de prétraitements.

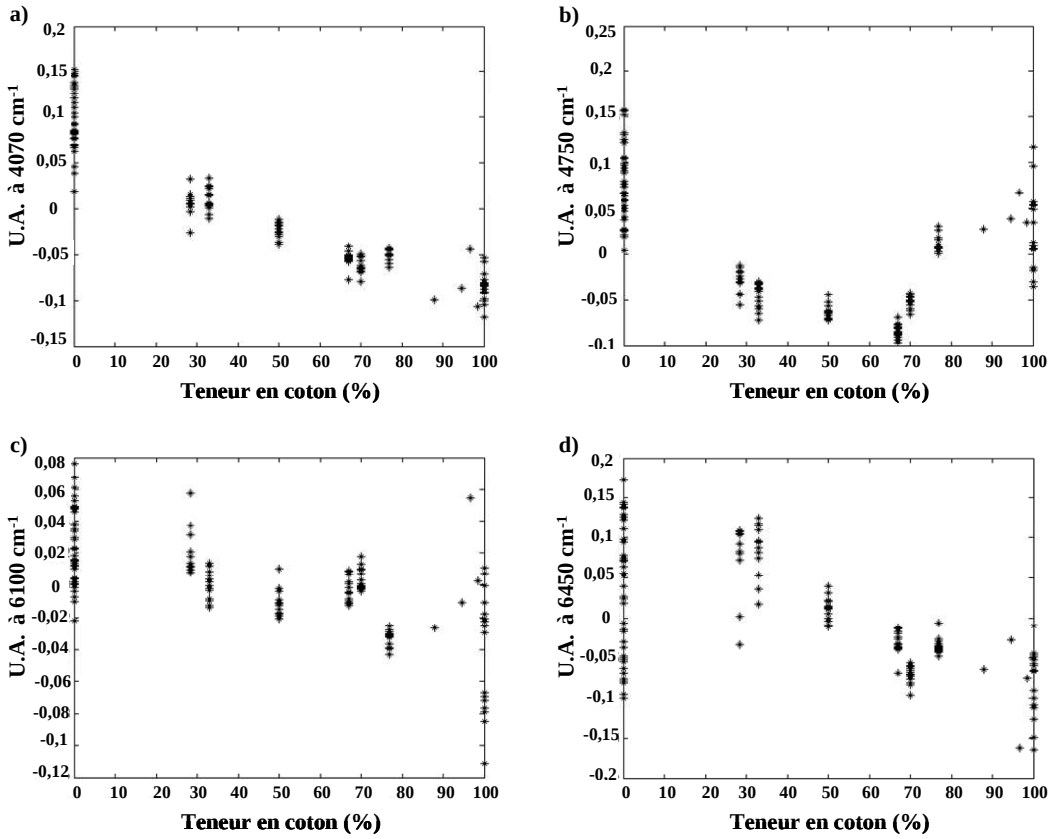


Figure 29 : Valeurs des pseudo-absorbances représentées en fonction de la teneur en coton (%) aux nombres d'onde a) 4070 cm^{-1}, b) 4750 cm^{-1}, c) 6100 cm^{-1}, d) 6450 cm^{-1}[91].

III.3. Modèles prédictifs sur les spectres réduits

Le Tableau 4 regroupe les résultats obtenus avec les modèles PLS et ANN construits sur les variables sélectionnées par AG-PLS et par IM. Afin de faciliter la comparaison des résultats, les valeurs prédites obtenues pour les modèles AG-PLS et IM-ANN sont présentées en fonction des valeurs de la mesure de référence sur la Figure 30.

	AG-PLS			IM-PLS			IM-ANN		
	C	VC	P	C	VC	P	T	V	P
Nombre d'échantillons	146		53	146		53	98	48	53
Variables latentes (PLS)	9			4					
Architecture (ANN)							12*4*1		
Nombre de variables	12x10			12			12		
R^2	0,993		0,990	0,979		0,967	0,996		0,991
RMSE(%)	3,02	3,43	3,44	5,04	5,56	6,33	2,30	2,61	3,43

Tableau 4 : Résultats obtenus sur les spectres réduits. C : lot d'entraînement, VC : lot de validation croisée, P : lot de prédiction, T : lot d'apprentissage, V : lot de validation.

Un modèle PLS est construit sur les 12 fenêtres de 10 variables (120 variables) sélectionnées par les AG-PLS. Ces 120 variables, identifiées dans le paragraphe précédent, représentent environ 25% du spectre complet. Les erreurs d'entraînement, de validation croisée et de prédiction sont respectivement égales à 3,02%, 3,43% et 3,44%[(91)]. Nous constatons une légère amélioration de l'erreur de prédiction par rapport à celle obtenue avec le modèle sur les spectres complets (RMSEP=3,74%).

Si nous comparons les résultats AG-PLS (Figure 30a et b) avec ceux obtenus pour les spectres complets, nous remarquons que la prédiction de l'échantillon 18 est maintenant acceptable. Au contraire, la teneur en coton de l'échantillon 26 est encore sous-estimée. Comme précédemment, nous conservons une dispersion des valeurs prédites pour les échantillons purs et spécialement pour les 100% coton. En ce qui concerne la Figure 30d, les deux échantillons pointés précédemment sont maintenant prédits plus correctement. De plus, nous pouvons remarquer que le modèle IM-ANN est capable de prendre en compte la variabilité contenue dans les échantillons purs, en particulier pour les échantillons 100% viscose. Au contraire, le modèle IM-ANN conduit à l'addition d'une source de variance inexpliquée pour les échantillons avec une teneur en coton de 70-80%.

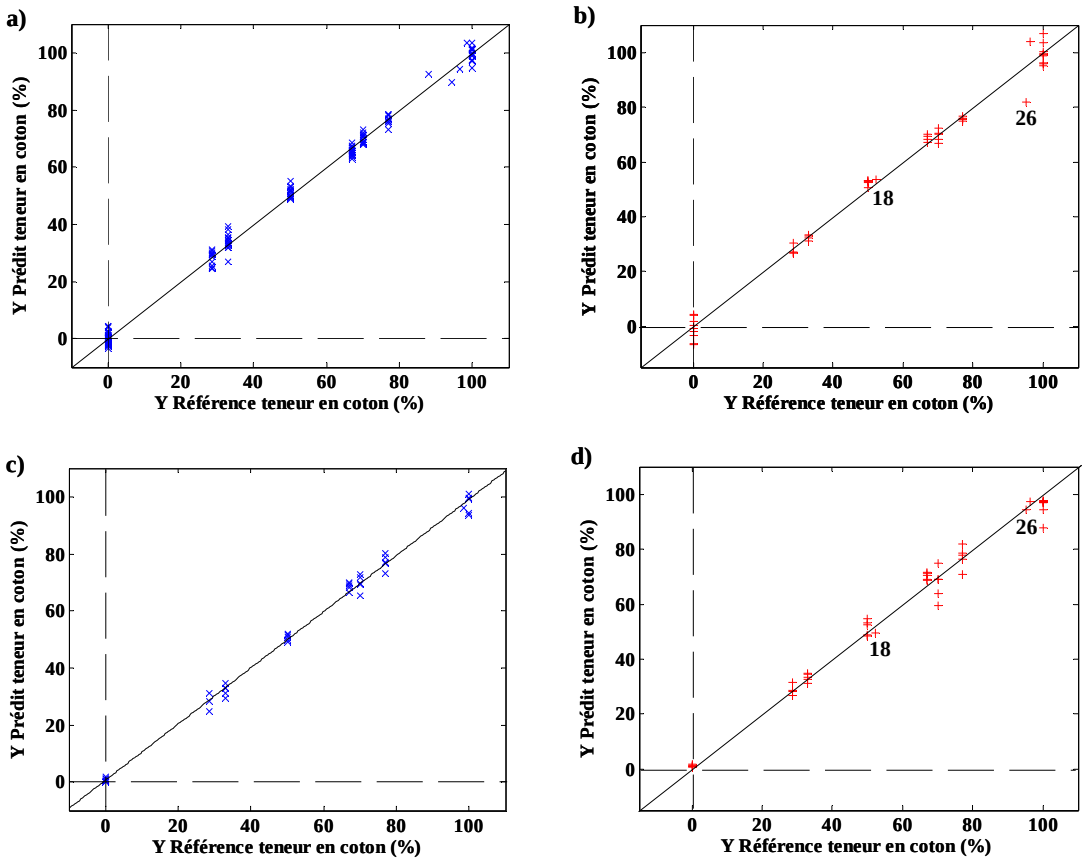


Figure 30 : a) validation du modèle AG-PLS en validation croisée sur le lot d'entraînement, b) en prédiction, c) validation du modèle IM-ANN sur le lot de validation, d) en prédiction.

- ***Discussion***

La sélection de variables par information mutuelle a retenu 12 variables. Une régression PLS et un modèle ANN sont construits sur les 12 variables sélectionnées. Néanmoins, l'erreur de prédiction obtenue avec le modèle IM-PLS vaut 6,33%. Ce très mauvais résultat était attendu du fait de la structure des données (Figure 29).

Nous avons donc construit un modèle ANN sur les 12 variables sélectionnées par IM. L'architecture optimale du réseau de neurones est 12*4*1. Les résultats obtenus sont présentés dans le Tableau 4. Les erreurs obtenues pour le lot d'apprentissage (2,30%) et pour le lot de validation croisée (2,61%) sont légèrement améliorées par rapport aux résultats du modèle AG-PLS. Les résultats obtenus par les modèles AG-PLS et IM-ANN sont du même

ordre de grandeur. Cependant, l'amélioration majeure apportée par le modèle IM-ANN est une réduction du nombre de variables à considérer (12 variables seulement). Il y a aussi un gain en termes de complexité et d'interprétation du modèle.

- ***Conclusion***

Nous avons déterminé la teneur en coton dans deux mélanges binaires de matières textiles. Les résultats présentés dans ce chapitre montrent que le recours à la spectroscopie PIR et à l'étalonnage multivarié est une alternative intéressante (par rapport aux mesures chimiques de référence actuelles, longues et polluantes) en vue d'une meilleure connaissance, plus rapide et plus directe, permettant éventuellement une analyse en vue d'un contrôle de routine. Les capacités prédictives obtenues sur les spectres complets pour les mélanges coton/PES satisfont la directive de la commission européenne. La prédiction de la teneur en coton reste plus délicate pour le lot de données coton/viscose car il s'agit d'un mélange de fibres cellulosiques. Nous avons cependant obtenu des résultats proches des ± 3 % en masse. Ces deux résultats obtenus à l'échelle du laboratoire ont fait l'objet de deux publications [89,91].

En vue d'une application pour un contrôle de routine ou d'une instrumentation spectroscopique simplifiée, deux procédures de sélection de variables ont été réalisées afin de sélectionner les variables pertinentes. Ces deux procédures sont fondées sur des concepts différents. L'information mutuelle peut être appliquée comme un prétraitement et ce indépendamment d'un modèle d'étalonnage. De fait, l'ensemble des variables sélectionnées ne sont pas forcément les mêmes, car le choix de l'estimation de la pertinence et le choix de l'algorithme pour réaliser l'optimisation sont basés sur des critères différents. De plus, les capacités prédictives obtenues pour les modèles construits sur les variables sélectionnées par AG-PLS ou par IM sont légèrement améliorées par rapport à celles obtenues sur les spectres complets.

En conclusion, la sélection des variables pertinentes permet de réduire considérablement le nombre de variables jusqu'à 8 variables sélectionnées par IM pour le lot coton/PES et 12 variables pour le lot coton/viscose.

Chapitre 3

Analyse qualitative de tissus par spectroscopie proche infrarouge

---

Ce chapitre illustre la classification de tissus textiles sur la base de leurs spectres PIR. La variable *Y* correspond à une propriété physique d'intérêt. Afin de respecter la confidentialité de ce travail, la nature exacte de cette propriété n'est pas communiquée. Outre les objectifs en terme de prédiction, le projet s'oriente vers l'utilisation d'une instrumentation simplifiée de type capteurs qui nécessitera la réduction du nombre de longueurs d'onde analysées, tout en conservant un rapport signal sur bruit acceptable pour l'analyse embarquée.

Ce chapitre présente les résultats de l'étude de faisabilité, au sens où ces modèles de classification sont construits sur les spectres proche infrarouge acquis sur un spectrophotomètre de laboratoire. Comme nous allons le montrer, les variabilités chimiques et physiques des échantillons ont dirigé notre étude vers le choix de méthodes de discrimination non linéaires. Dans un second temps, prenant en compte le cahier des charges défini par le partenaire industriel, l'effet de la sélection de variables a été simulé sur les spectres de laboratoire.

I. Présentation des échantillons

Les tissus textiles composant la base de données utilisée pour l'étude ont été recensés et sélectionnés par le partenaire industriel. Ils proviennent essentiellement du secteur de l'habillement. Les matières textiles sont constituées de fibres issues de trois grandes familles : les fibres naturelles, les fibres chimiques et les fibres minérales. Ces dernières, notamment l'amiante, ne sont pas utilisées dans le domaine de l'habillement. A l'intérieur de ces grandes familles, les fibres ont des origines variées (Figure 31). Historiquement, les premières fibres utilisées pour la confection de vêtements ont été les fibres naturelles. Elles peuvent provenir du règne végétal ou du règne animal. Les fibres végétales sont essentiellement cellulosiques (le coton, le lin, le chanvre, la ramie). Les fibres du règne animal sont protéiniques (la laine et la soie). La seconde famille de fibres correspond aux fibres chimiques. Certaines telles que la viscose et l'acétate peuvent être obtenues à partir de matières premières naturelles et dans ce cas elles sont dites artificielles. Si elles sont, au contraire, fabriquées à partir de réactions chimiques, on parle alors de fibres synthétiques. Parmi ces fibres synthétiques, on retrouve principalement le polyamide (PA), le polyester (PES), l'acrylique et l'élasthanne.

Le lot de données spectrales contient 227 échantillons qui peuvent être des fibres pures ou des mélanges de deux, trois ou quatre fibres différentes. Dans le domaine de l'habillement, huit fibres sont couramment utilisées. La répartition des échantillons en fonction de leur nature est présentée dans le Tableau 5. Le lot de données est constitué de 95 échantillons purs et 132 échantillons contenant un mélange de fibres. Cette distribution des échantillons en fonction de leur nature tient compte autant que possible de la représentativité des matières textiles dans les vêtements du prêt-à-porter.

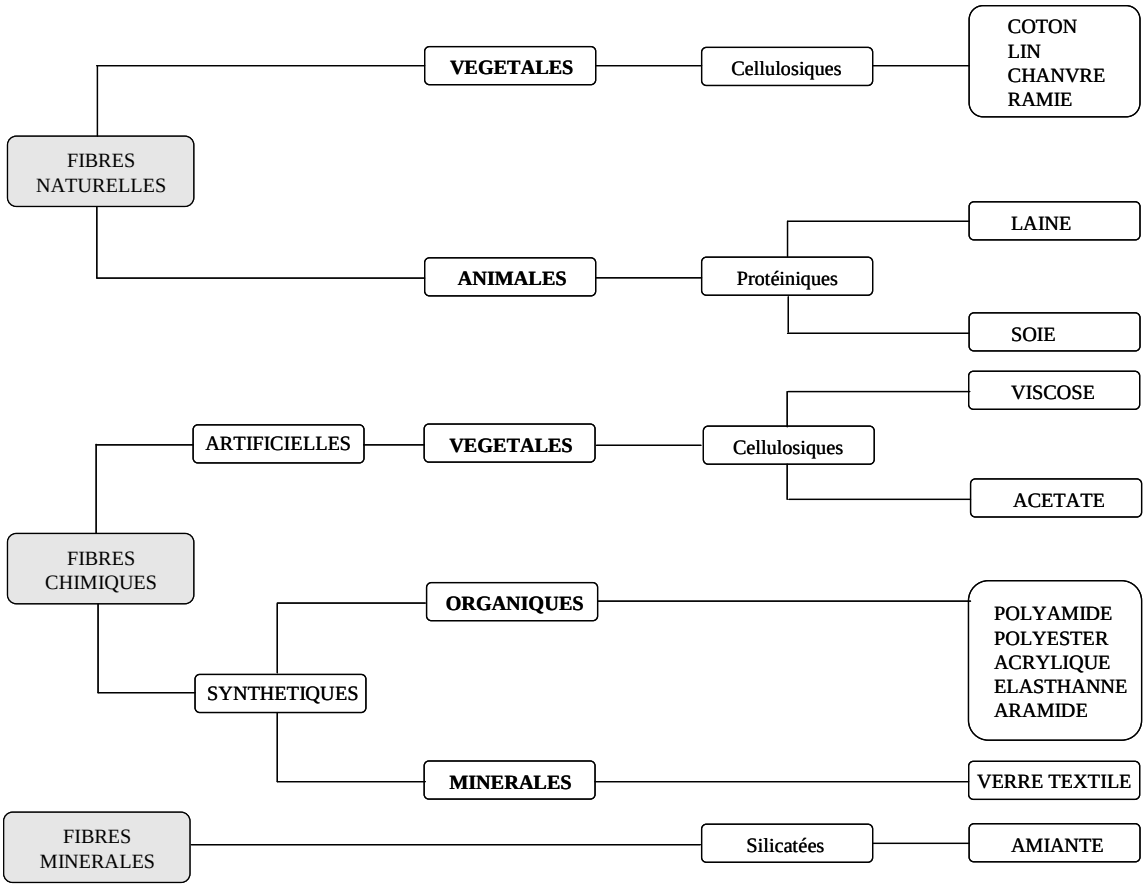


Figure 31 : Origine des matières textiles (d'après I. Brossard [(92)]).

	Composition	Nombre d'échantillons
ECHANTILLONS : PURS	Acrylique	6
	Coton	32
	Laine	15
	Lin	3
	Polyamide (PA)	6
	Polyester (PES)	23
	Soie	7
	Viscose	3
	Total	**95**
ECHANTILLONS : MELANGES	**Mélanges binaires** dont	**100**
	Coton/PES	22
	Coton/PA	7
	Mélanges ternaires dont	**24**
	Coton/PES/viscose	2
	Mélanges quaternaires dont	**8**
	Coton/PES/acrylique/viscose	1
	Total	**132**

Tableau 5 : Répartition des échantillons en fonction de leur composition.

Le mélange de fibres le plus répandu dans le domaine textile est le mélange coton/PES. En effet, les fibres synthétiques de type PES possèdent une bonne élasticité et une forte résistance qui permettent d'obtenir, en mélange avec les fibres naturelles, des fils plus faciles à travailler en filature, plus réguliers pour le tissage et donc des tissus plus solides. Cependant, certains mélanges de fibres ne sont pas présents sur le marché du textile car ils n'apportent pas de propriétés supplémentaires au tissu. C'est le cas par exemple des mélanges soie/viscose, soie/acrylique, soie/laine et coton/laine. Par conséquent, nous noterons que le lot de données ne cherche pas à couvrir tous les mélanges existants mais à représenter correctement ceux qui sont significativement présents sur le marché du textile.

La matrice de données *Y* code une propriété physique que l'on souhaite à terme prédire. Les échantillons sont répartis en trois classes, notée P1, P2 ou P3, en fonction de la caractéristique prise par cette propriété. La répartition des échantillons en fonction des classes est présentée dans le Tableau 6. L'attribution d'un échantillon à une classe n'est pas

uniquement dépendante de la nature de l'échantillon que d'ailleurs, on ne connait pas forcément. De plus, la présence d'un traitement chimique peut, dans certains cas, perturber le classement. Par exemple, l'ennoblissement des tissus par l'ajout d'apprêts peut modifier la classe de l'échantillon. Pour certains échantillons, nous reviendrons néanmoins à la nature exacte du tissu après l'avoir fait doser par un organisme acrédité (IFTH). Cependant, en aucun cas les objectifs définis ici ne peuvent être transcrits en terme d'analyse quantitative.

Propriété *Y*	**Nature**	**Nombre d'échantillons**
P1	Acrylique, PA, mélanges divers…	64
P2	PES, laine, soie, mélanges divers…	117
P3	Coton, lin, mélanges divers…	46

Tableau 6 : Répartition des échantillons du lot de données en fonction de la propriété Y.

Nous remarquons que le nombre d'échantillons présents dans la classe P2 est plus important par rapport à celui des deux autres classes. Ceci vient du fait que les variabilités présentes dans cette classe sont plus importantes. Sans rentrer dans les détails, cette classe contient une diversité en terme de nature des tissus (fibres d'origines naturelles, artificielles ou synthétiques pures ou en mélanges binaires et ternaires), alors que la classe P3 à l'opposé ne contient que des tissus composés de fibres naturelles.

II. Échantillonnage et acquisition des données spectrales

Les spectres proche infrarouge ont été acquis sur un spectrophotomètre commercial. L'appareil utilisé est le FOSS-XDS *near infrared* avec le module *Rapid Content Analyzer*. Il s'agit d'un instrument dispersif à réseau holographique. Ce spectrophotomètre est constitué d'une source tungstène et de huit détecteurs, quatre détecteurs en silicium (Si) pour le domaine du visible et le début du proche infrarouge (400-1100 nm) et quatre détecteurs en sulfure de plomb (PbS) pour le domaine du proche infrarouge (1100-2500 nm). Le domaine spectral s'étend de 400 à 2500 nm avec une résolution spectrale de 0,5 nm. L'acquisition des spectres est réalisée avec le logiciel VisionMC®. Le spectre de la référence se fait de façon

automatique. Il s'agit de la référence interne de l'appareil (céramique possédant un taux de réflectance ~ 80%).

En fonction des objectifs de l'analyse et de la nature physique de l'échantillon, il convient de choisir une technique d'échantillonnage[93]. Cette étape est très importante car le mode de présentation conditionne la faisabilité d'une analyse qualitative par la répétabilité, la sensibilité et la spécificité des informations spectrales. Les échantillons textiles analysés étant très diffusants, nous considérerons que même sur des chemins optiques faibles, il n'y a pas ou peu de transmission derrière la cellule. L'étude est réalisée en réflexion diffuse (Figure 32). Le mode d'échantillonnage utilisé est une géométrie 0-45°. L'énergie lumineuse est envoyée sous une incidence normale par rapport à la surface de l'échantillon. La lumière diffusée dans tout le demi-espace est collectée par 4 détecteurs placés à 45° par rapport à la normale. Les spectres obtenus en réflexion diffuse sont calculés en pseudo-absorbance (log(1/R)).

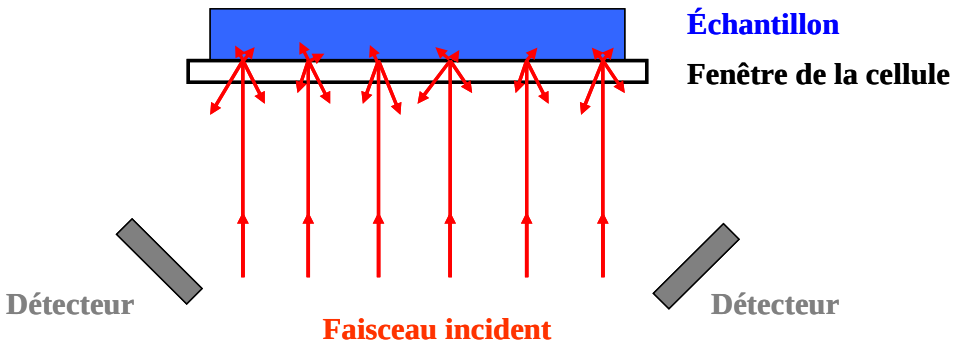


Figure 32 : Réflexion diffuse.

Compte-tenu des objectifs de l'application, les échantillons sont mesurés directement, sans préparation ni destruction de l'échantillon. La majorité des échantillons textiles analysés sont des mélanges de fibres de compositions différentes. Les mélanges de fils non intimes peuvent faire apparaître dans certains cas des inhomogénéités sur l'ensemble du tissu. Des différences spectrales peuvent être visibles par exemple entre l'endroit et l'envers du tissu (Figure 33). L'échantillon dont le spectre est présenté sur la Figure 33, est un velours de couleur verte constitué d'un mélange de fibres coton/PES. Le spectre du tissu côté velours présente les bandes d'absorption caractéristiques d'un tissu 100% coton. Le spectre de l'envers du tissu met en évidence la présence de polyester avec en particulier la bande

d'absorption située à 1660 nm. Une expertise de l'échantillon par des méthodes de référence décrites dans le chapitre 2, paragraphe I.2, révèle que la surface extérieure correspond à des fils de coton qui sont maintenus par des fils de polyester et de coton tandis que la composition en masse de cet échantillon est de 84% coton, 16% PES.

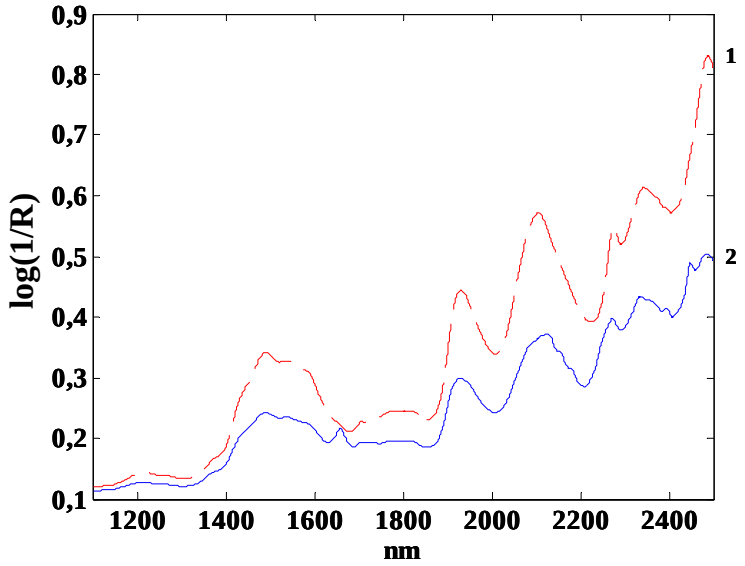


*Figure 33 : Spectres bruts d'un mélange coton/PES, (**1**) de l'endroit du tissu et (**2**) de l'envers du tissu.*

Pour s'affranchir de ces sources de variabilité, pour chaque échantillon, trois mesures sont réalisées, deux mesures sur l'endroit de l'échantillon avec une rotation de l'échantillon d'un angle de 90° entre les deux mesures et une mesure sur l'envers du tissu. Les absorbances de ces trois spectres sont ensuite moyennées afin de mieux prendre en compte les variabilités dans l'homogénéité ou l'aspect de surface de l'échantillon.

III. Interprétation physico-chimique des spectres

La demande du marché du textile est telle que ce domaine est sans cesse en évolution. En particulier, la recherche de textiles dits *intelligents* [92] est en nette expansion. A l'échelle grand public, cela se traduit en particulier par l'arrivée de traitements chimiques donnant aux

tissus des propriétés spécifiques (antitache, antibactérien...). L'influence de l'ennoblissement des tissus n'est pas l'objet de notre étude car la plupart du temps, la nature des apprêts n'est pas connue. Mais c'est bien sûr un paramètre influent qui peut induire des différences spectrales locales, ou plus globalement des dérives de la ligne de base. A titre d'exemple, la Figure 34 propose un échantillon avec un apprêt chimique spécifique. Nous constatons un décalage en longueur d'onde pour la bande d'absorption située à 1477 nm pour l'échantillon 100% coton avec un apprêt par rapport à celle de l'échantillon pur coton non traité qui pointe à 1491 nm. De plus, des similarités avec la bande d'absorption d'un échantillon pur viscose sont visibles, ce qui peut générer des erreurs pour la classification. Nous noterons donc que l'ajout d'un apprêt chimique peut modifier, dans certains cas, l'allure et la position des bandes d'absorption.

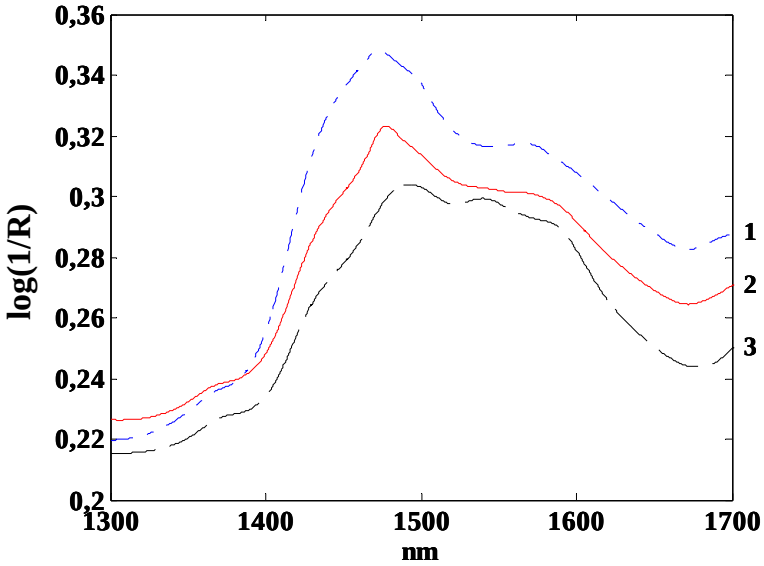


*Figure 34 : Spectres bruts sur le domaine 1300-1700 nm, (**1**) un échantillon 100% viscose sans apprêt, (**2**) un échantillon 100% coton avec un apprêt chimique et (**3**) un échantillon 100% coton sans apprêt.*

D'autre part, l'ennoblissement des tissus par des teintures permet de donner à une fibre une coloration uniforme différente de sa teinte naturelle. Nous constatons sur la Figure 35, une dérive de la ligne de base relativement importante jusqu'à environ 1300 nm pour le tissu de couleur noire par rapport à un tissu blanc de composition identique. Cette dérive de la ligne de base n'est pas liée à la nature chimique du tissu car d'une part, de nombreux échantillons

de compositions chimiques différentes sont affectés, et d'autre part, nous n'observons pas de différences dans la région spectrale des bandes de combinaisons. Cette dérive est liée à l'absorption des longueurs d'onde du visible.

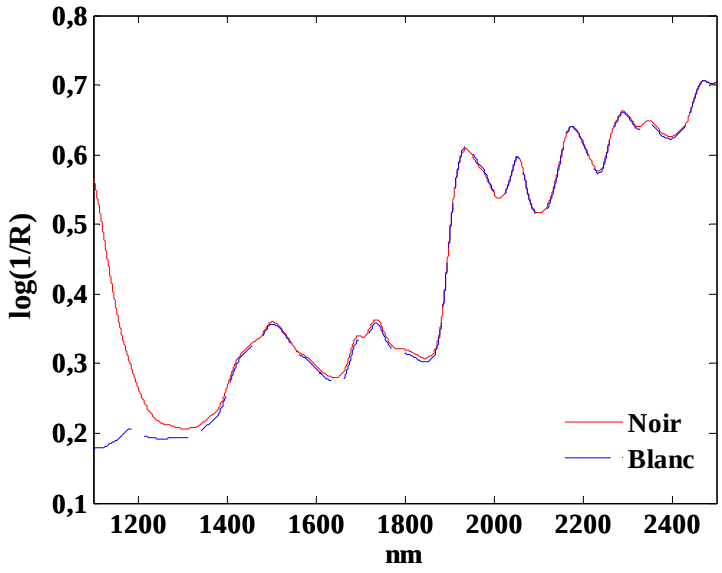


Figure 35 : Spectres bruts d'échantillons pur laine de couleur noire ou blanche.

Certaines des variabilités physico-chimiques mises en évidence dans ce paragraphe pourront être réduites par l'application de traitements chimiométriques que nous allons maintenant aborder.

IV. Traitements chimiométriques

IV.1. Exploration des données spectrales

Les spectres bruts des 227 échantillons du lot de données sont acquis sur le spectrophotomètre de laboratoire et sont présentés sur la Figure 36. L'analyse exploratoire du lot complet des données est réalisée avec l'analyse PCA qui est rappelée en annexe 1. La Figure 37 montre la projection des 227 échantillons du lot de données dans le plan des deux

premières composantes principales (CP1-CP2) et dans le plan formé par les deuxième et troisième composantes principales (CP2-CP3).

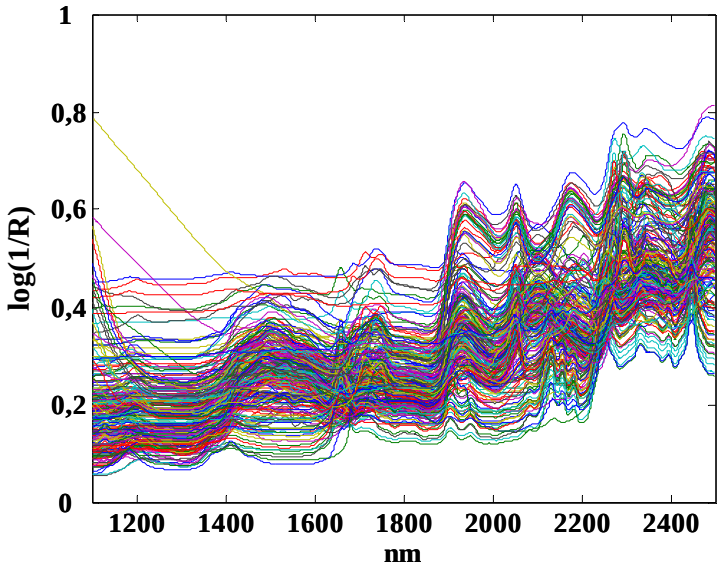


Figure 36 : Spectres bruts du lot complet de données.

La projection des échantillons dans les plans (CP1-CP2) et (CP2-CP3), respectivement, n'est pas homogène. Nous noterons la présence d'échantillons aberrants, ou tout du moins atypiques qui déforment le nuage suivant les axes CP1 et CP2. Ces échantillons sont identifiés en réalisant des tests statistiques, le test T^2 de Hotelling [85] et le test statistique Q[86]. Pour rappel, ces deux tests sont redéfinis en annexe 2. Nous constatons sur la Figure 37, que certains échantillons sont pointés par les deux tests et seront considérés comme échantillons aberrants, c'est le cas par exemple de l'échantillon noté 90. Cet échantillon est un tissu de couleur bleu foncé et de composition 63% coton/37% PES. Il présente une forte dérive de la ligne de base jusqu'à environ 1880 nm difficilement explicable comme nous l'avons observé précédemment. Nous remarquons également des différences dans les bandes d'absorption au-dessous de 1600 nm entre cet échantillon et un échantillon ayant la même composition (Figure 38).

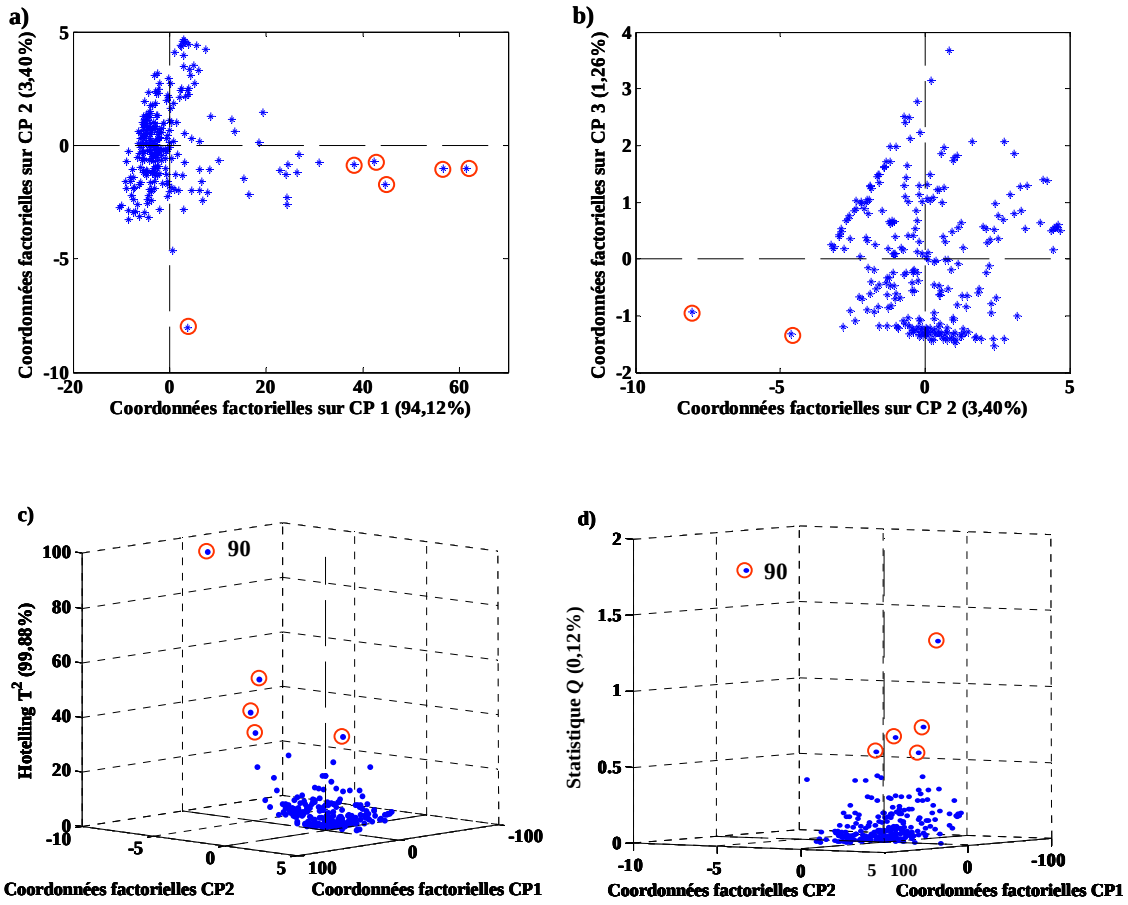


Figure 37 : a) PCA dans le plan (CP1-CP2), b) PCA dans le plan (CP2- CP3), c) test T^2 de Hotelling, d) test Q sur les 227 échantillons du lot de données. Les échantillons aberrants sont entourés.

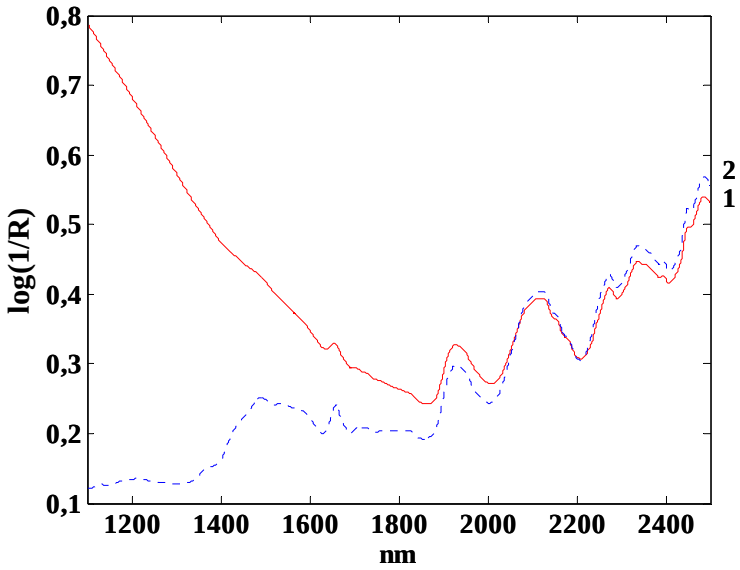


*Figure 38 : Spectres bruts (**1**) de l'échantillon n°90 et (**2**) d'un échantillon avec la même composition.*

Une fois les échantillons aberrants éliminés en accord également avec l'expertise du partenaire industriel, une PCA est réalisée sur les 221 échantillons composant le lot de données. Finalement, les trois premières composantes principales décrivent 98,78% de la variance totale. Les résultats de la PCA dans le plan (CP2-CP3) sont présentés sur la Figure 39. Les vecteurs propres présentant les longueurs d'onde les plus influentes sont repris sur la Figure 40. Le vecteur propre de la première composante principale (CP1) présente la même allure que le spectre moyen des échantillons du lot de données (inversé). Nous remarquons par exemple, la présence de l'information reliée à la bande d'absorption caractéristique du PES située à ~ 1660 nm. Le vecteur propre de la CP1 présente une dérive de la ligne de base importante due au fait que les spectres ne sont pas prétraités. Le vecteur propre de la CP2 s'explique par l'opposition de l'information reliée à une des bandes d'absorption du PES (contribution négative) avec celles de la laine (partie positive du vecteur propre). Le vecteur propre de la CP3 s'interprète par l'opposition des informations du coton et de l'acrylique. Nous constatons que la partie négative du vecteur propre présente essentiellement les informations liées aux bandes d'absorption du coton entre 1386 et 1657 nm, à 1791 nm et à 2486 nm. La partie positive du vecteur propre contient quant à elle principalement les

informations liées aux bandes d'absorption de l'acrylique situées entre 1658 et 1790 nm et à 2272 nm.

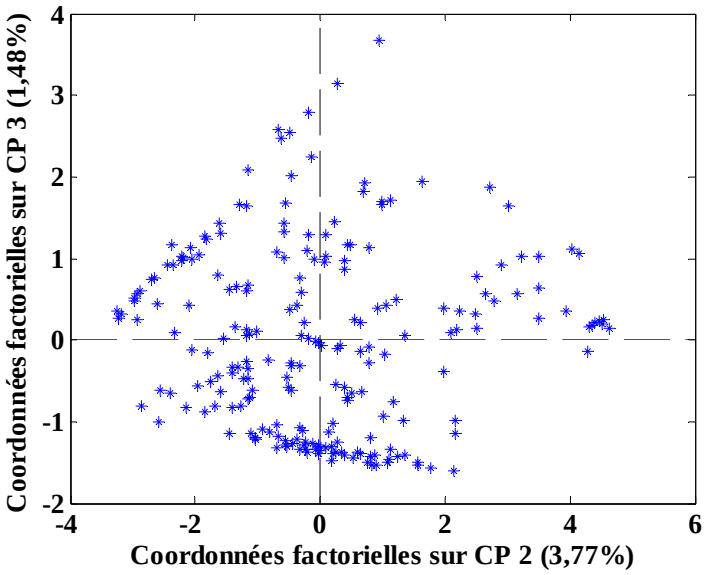


Figure 39 : PCA du lot de données dans le plan (CP2-CP3).

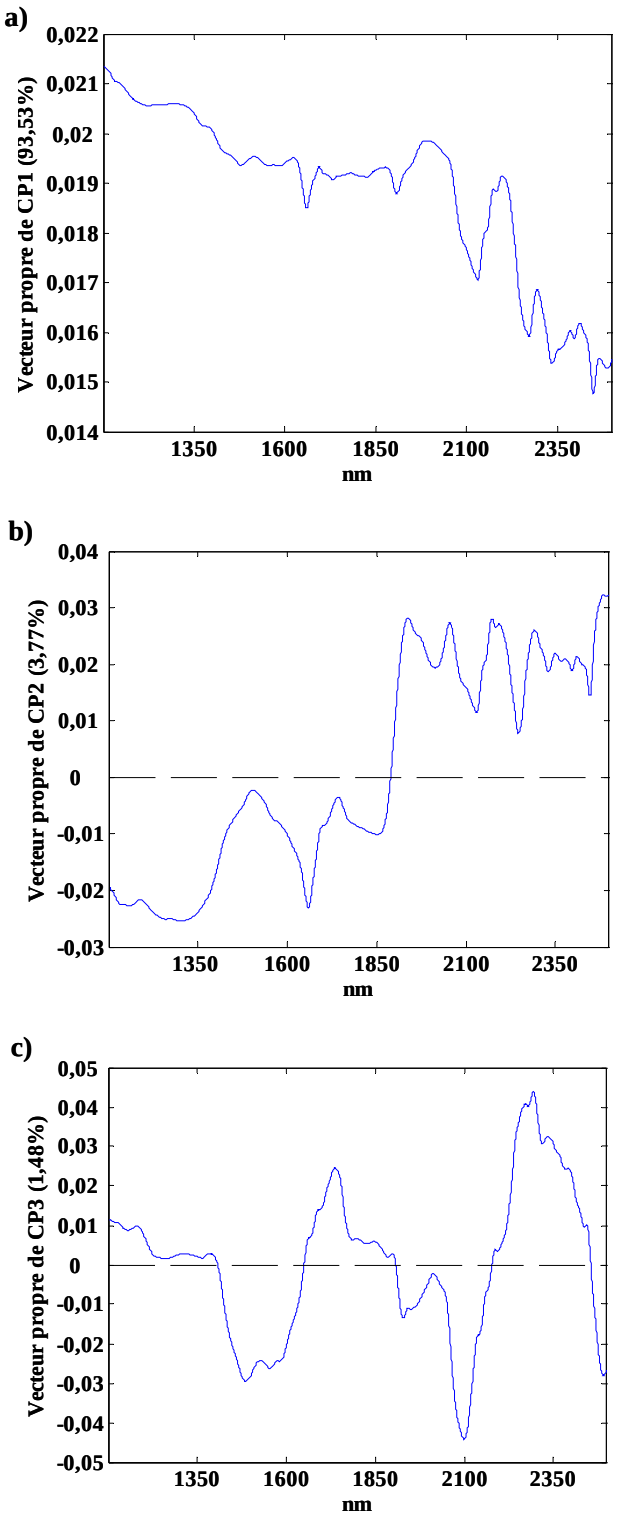


Figure 40 : Vecteurs propres a) de CP1, b) de CP2 et c) de CP3.

IV.2. Distribution et répartition des échantillons

Les échantillons du lot de données sont généralement répartis en deux lots, un lot d'entraînement et un lot de prédiction. Le terme prédiction sous-entend que les échantillons concernés n'ont jamais été utilisés lors de la construction du modèle. L'optimisation du modèle est effectuée sur le lot d'entraînement dont les échantillons doivent couvrir toutes les variabilités rencontrées dans le lot de données. Par conséquent, le choix des échantillons du lot d'entraînement est crucial pour la généralisation du modèle d'étalonnage aux données futures, notamment à celles du lot de prédiction. Diverses méthodes de sélection d'échantillons existent. Nous avons utilisé dans notre étude en particulier une procédure aléatoire de sélection d'échantillons ainsi qu'une sélection avec l'algorithme de Kennard et Stone (KS)[(94)]. Ces deux méthodes de sélection sont détaillées en annexe 4. Il faut noter l'existence d'autres approches telles que l'algorithme Optisim[(95)] ou encore Duplex [(96)].

La répartition des échantillons du lot de données en un lot d'entraînement et en un lot de prédiction est présentée dans le cas d'une distribution aléatoire (Figure 41a) et dans le cas d'une distribution avec l'algorithme de Kennard et Stone (Figure 41b). Le lot d'entraînement contient 130 échantillons et le lot de prédiction contient 91 échantillons.

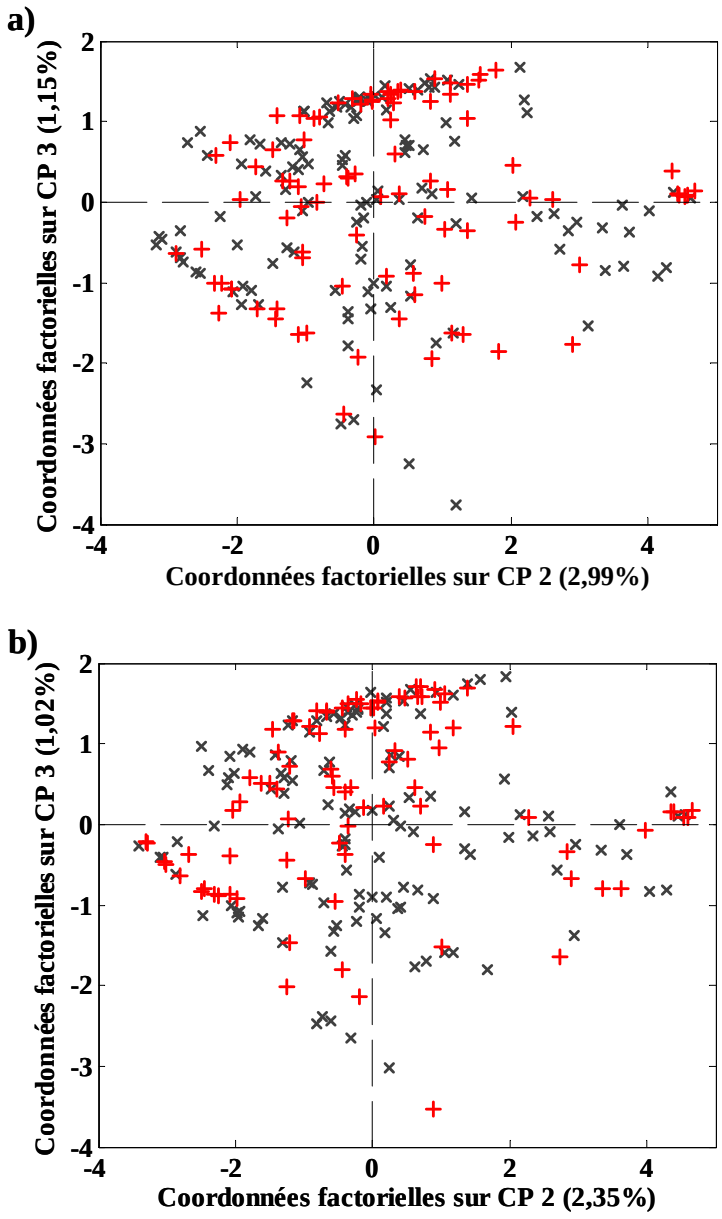


Figure 41 : Répartition des échantillons pour un lot d'entraînement (x) *et pour un lot de prédiction* (+)*, a) sélection aléatoire et b) sélection avec l'algorithme de Kennard et Stone.*

Nous observons que pour les deux méthodes les lots d'entraînement et de prédiction couvrent uniformément l'espace. La répartition des échantillons en fonction de leur classe est présentée dans le Tableau 7.

Propriété *Y*	Aléatoire	Kennard et Stone
P1	33	37
P2	69	68
P3	28	25

Tableau 7 : Distribution des échantillons du lot d'entraînement dans les classes Y.

La répartition des échantillons pour une classe donnée est semblable pour les deux méthodes. Nous remarquons cependant que le nombre d'échantillons contenus dans la classe P2 est plus important que celui des deux autres classes, ceci tient compte de la répartition des échantillons dans la base de données (chapitre 1, paragraphe I, Tableau 6).

IV.3. Choix des prétraitements

Les données spectrales brutes, telles qu'elles sont acquises sur un spectrophotomètre proche infrarouge, ne possèdent pas obligatoirement la forme la mieux adaptée aux traitements chimiométriques. Sur les spectres complets, il est nécessaire d'appliquer un prétraitement aux données spectrales avant leur exploitation. Dans cette étude, les prétraitements retenus pour la présentation des résultats sont le centrage par le spectre moyen (*Mean cente*r, MC), la déviation normale standardisée (SNV)[97] et la dérivée première de Savitzky-Golay (DER1)[84]. Ces prétraitements sont rappelés en annexe 3 et illustrés ci-dessous.

- ***Déviation normale standardisée***

La déviation normale standardisée (SNV)[97,98] permet de réduire les variations d'intensité générale des spectres. L'avantage de ce prétraitement est que chaque échantillon est transformé indépendamment des autres échantillons présents dans le lot de données. La Figure 42 permet d'apprécier l'effet de la transformation SNV sur les spectres proche infrarouge d'échantillons purs PES. Les variations de signal dues aux phénomènes physiques, en particulier, aux aspects liés à la diffusion[99] (voile, tissu polaire), sont amoindries ce qui exalte les informations chimiques.

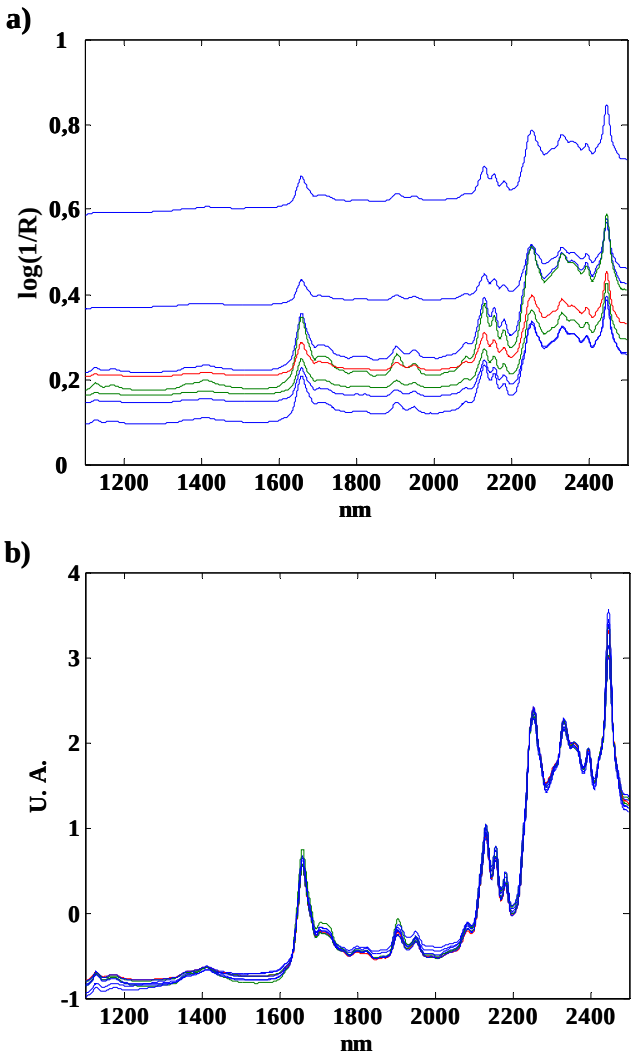


Figure 42 : a) Spectres bruts d'échantillons purs PES, b) spectres prétraités SNV des échantillons purs PES présentés en a).

- ***Dérivation***

La dérivation réduit les variations de la ligne de base [100,101] dues par exemple, à la composition chimique des tissus (cellulose, chapitre 2, paragraphe I), l'ennoblissement (chapitre 3, paragraphe III) et permet une meilleure séparation des bandes d'absorption[102]. On exalte ainsi des bandes présentes dans le recouvrement observé sur les spectres proche infrarouge[103]. Les spectres du lot de données sont prétraités avec une dérivée de Savitzky-

Golay[84]. Les paramètres de l'algorithme sont les suivants, dérivée d'ordre 1, polynôme du second degré et une fenêtre de 15 longueurs d'onde ce qui correspond à environ 7 nm.

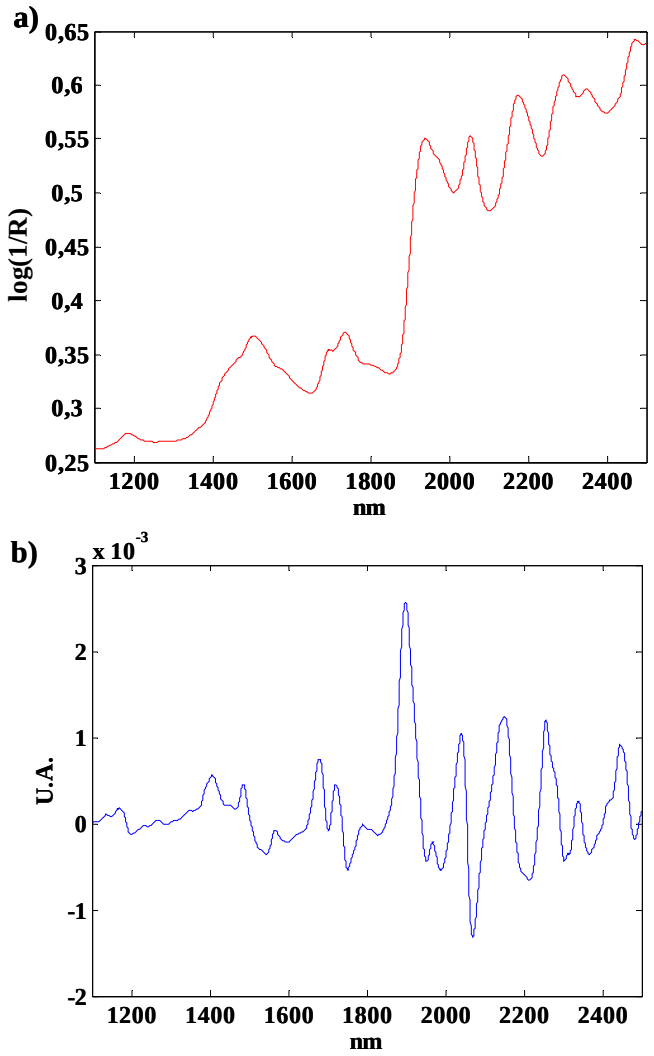


Figure 43 : Spectre d'un échantillon 100% laine, a) spectre brut, b) spectre prétraité DER1.

La dérivation d'ordre 1 des spectres permet d'exalter les bandes d'absorption et en particulier celle dans la région comprise entre 1350 et 1650 nm (Figure 43). Cette bande d'absorption correspond aux vibrations d'élongation et de déformation des liaisons C-H des groupements CH_2[104].

V. Résultats obtenus en classification sur spectres complets et discussion

L'analyse qualitative du lot de données est dans un premier temps réalisée sur le domaine spectral 1100-2500 nm. Seuls les résultats obtenus par les méthodes de classification, LDA et SVM, sont présentés dans le manuscrit. Les capacités prédictives obtenues pour les différents modèles sont toujours accompagnées d'un intervalle de confiance établi à partir de répétitions sur dix lots d'entraînement et de prédiction correspondant à des répartitions différentes des échantillons[105] et calculé à partir d'une distribution de *Student*[106].

V.1. Analyse Discriminante Linéaire

La méthode LDA est appliquée sur les scores de la PCA puisque la matrice ***X*** contient un nombre de variables supérieur au nombre d'échantillons. On utilise ensuite les paramètres du modèle pour la prédiction. En parallèle, diverses combinaisons de prétraitements sont testées. Les résultats présentés dans le Tableau 8 correspondent aux pourcentages moyens des échantillons bien classés obtenus en prédiction pour les dix modèles construits. Ils estiment donc la quantité d'échantillons bien prédits. Les pourcentages moyens ont été calculés à partir de l'Équation 40.

Équation 40 $$Pourcentage(P^*, P\ ) = \left(\frac{1}{N} \sum_{i=1}^{N} \delta(i) \right) \times 100$$

où $\delta(i) = \begin{cases} 1\ si\ P^*(i) = P\ (i) \\ 0\ si\ P^*(i) \neq P\ (i) \end{cases}$, N le nombre d'échantillons du lot d'échantillons considéré, $P(i)$ et $P^*(i)$ sont, respectivement, la valeur de référence et la valeur prédite de la propriété P pour l'échantillon i.

Prétraitements	P en %
MC	71,4 (± 3,2)
SNV	77,7 (± 2,5)
DER1	82,5 (± 1,9)
SNVDER1	78,7 (± 1,9)

Tableau 8 : Pourcentages moyens des échantillons bien classés en prédiction (P en %) et entre parenthèses, intervalles de confiance à 95%.

Les résultats obtenus sont présentés sous la forme d'un pourcentage moyen d'échantillons bien classés associés à un intervalle de confiance à 95% sur cette valeur moyenne. Nous constatons qu'un simple centrage des spectres permet d'obtenir un taux d'échantillons bien classés en prédiction de 71,4% (± 3,2%). Cette valeur est significative pour un problème de classification à trois classes compte tenu du nombre d'échantillons[105]. Les capacités prédictives obtenues avec les prétraitements SNV, DER1 et SNVDER1 sont améliorées par rapport à celles obtenues avec le prétraitement MC. Afin de déterminer si ces résultats sont significativement différents, un test de comparaison des moyennes expérimentales basé sur la statistique *t* du test de *Student* a été réalisé après avoir vérifié l'homogénéité des variances de tous les échantillons[106]. Nous remarquons que nous avons une différence significative entre les résultats pour les spectres prétraités MC et les résultats pour les spectres prétraités SNV, DER1 et SNVDER1. Il semble donc qu'un simple centrage des données ne soit pas le prétraitement optimal. Les meilleures performances prédictives sont obtenues avec les prétraitements DER1 et SNVDER1, nous notons toutefois que les résultats ne sont pas significativement différents. Néanmoins, nous retiendrons par la suite que sur les spectres complets prétraités DER1, la méthode LDA permet d'obtenir des résultats en prédiction de 82,5% (± 1,9%).

- ***Interprétation du modèle***

L'interprétation du modèle est réalisée pour les spectres prétraités DER1, pour un lot de prédiction donné, noté W_P. Afin d'identifier les échantillons mal prédits par le modèle, la

Figure 44 présente les échantillons du lot de prédiction projetés dans le plan (CP2-CP3) de la PCA.

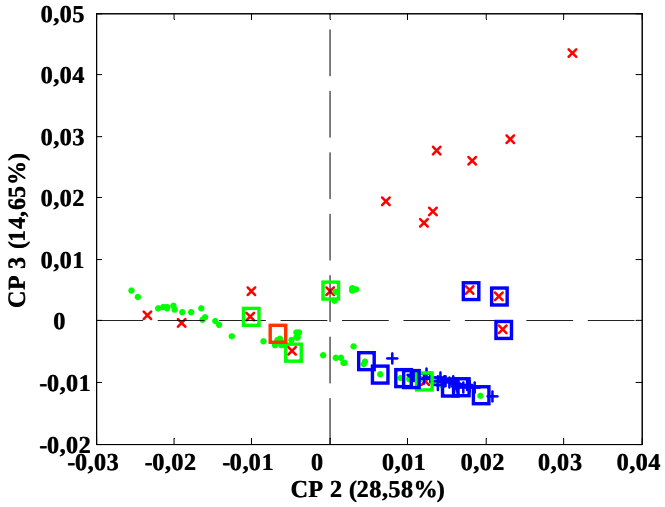


Figure 44 : Echantillons d'un lot de prédiction W_P projetés dans le plan (CP2-CP3) de la PCA. Echantillons de la classe P1(x), P2 (●), P3 (+). Les échantillons entourés d'un carré correspondent aux échantillons mal prédits par la méthode LDA, la couleur du carré indique la classe prédite.

Nous remarquons que 15 échantillons ne sont pas prédits correctement par le modèle. Parmi ces échantillons, 14 échantillons sont constitués de mélanges de fibres. Cela correspond bien à une tendance observée tout au long de l'étude, les échantillons purs sont plus faciles à modéliser, par conséquent, leur propriété *Y* est plus facile à prédire. Ces échantillons sont situés en bordure des classes P1-P2 et P2-P3 dans le plan (CP2-CP3) de la PCA. Seul un échantillon pur est mal prédit, il s'agit de l'échantillon comportant un apprêt chimique présenté dans le chapitre 3, paragraphe III. La distribution et la complexité des échantillons sont telles que les méthodes linéaires comme LDA ne peuvent établir des frontières permettant une discrimination correcte entre les classes.

V.2. Méthode des Support Vector Machine

Le principe de la méthode des SVM a été expliqué dans le premier chapitre de ce manuscrit. Nous avons insisté sur le potentiel de l'algorithme des SVM pour les données non linéairement séparables. L'espace des données d'entrée est transformé en un nouvel espace de plus grande dimension dans lequel des méthodes linéaires peuvent être appliquées. Cette transformation est réalisée par une fonction non linéaire ϕ. L'avantage est que seule la fonction noyau *K* doit être explicite. La fonction noyau retenue dans notre étude, et en général pour les applications des SVM en spectroscopie PIR[19,24,107], est une fonction de base radiale (RBF) de la forme suivante :

Équation 41 $$K(x,y) = \exp\left(-G\|x-y\|^2\right) \text{ avec } G=1/(2\sigma^2)$$

- ***Optimisation de C et de G***

Deux paramètres, le terme de pénalité *C* lié à la marge et le terme *G* lié à la largeur σ de la gaussienne, doivent être optimisés (chapitre 1, paragraphe III.2.2). Leurs valeurs sont dépendantes du problème donné et une grille d'optimisation est utilisée. Un exemple d'optimisation est présenté sur la Figure 45. Sur ce graphique est reporté le pourcentage d'échantillons bien classés en validation croisée en fonction des valeurs de *C* et *G*. La grille est constituée de 100 points dont les valeurs minimale et maximale de *C* sont, respectivement, égales à 0,01 et 100 000 et celles de *G* valent 0,00001 et 100 000. Les valeurs de *C* et de *G* testées varient de façon logarithmique.

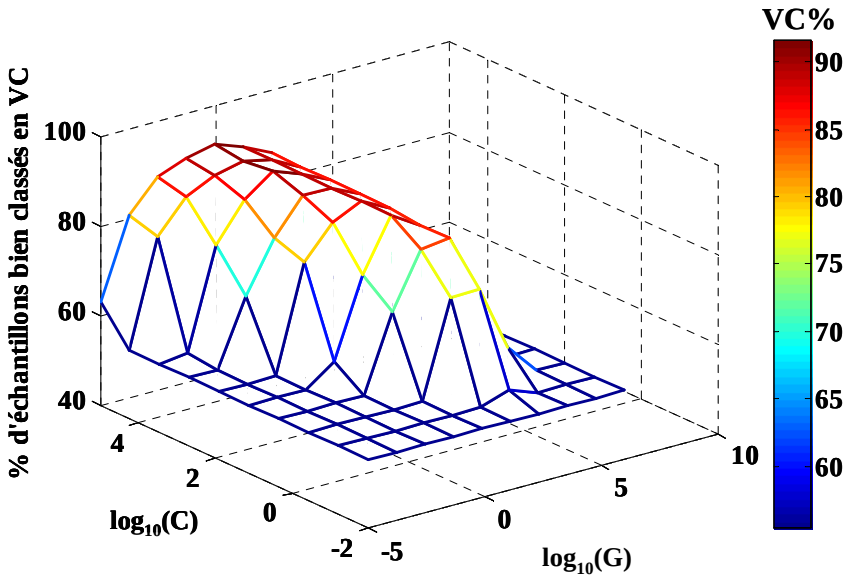


Figure 45 : Grille d'optimisation. Le code couleur correspond aux pourcentages d'échantillons bien classés en validation croisée (VC).

La valeur de *G* qui correspond à l'inverse de la largeur de la gaussienne au carré, dépend de la distribution des données. Afin de trouver le *C* et le *G* optimaux, la grille d'optimisation initiale présentée sur la Figure 45 est affinée. En effet, une nouvelle grille est calculée pour des valeurs de *C* et de *G* minimales et maximales qui valent respectivement 77 et 16681 pour *C* et 0,0027 et 215 pour *G*. Ceci permet d'obtenir un nombre de valeurs de *C* et de *G* plus important uniquement dans la zone optimale tout en limitant le temps de calcul.

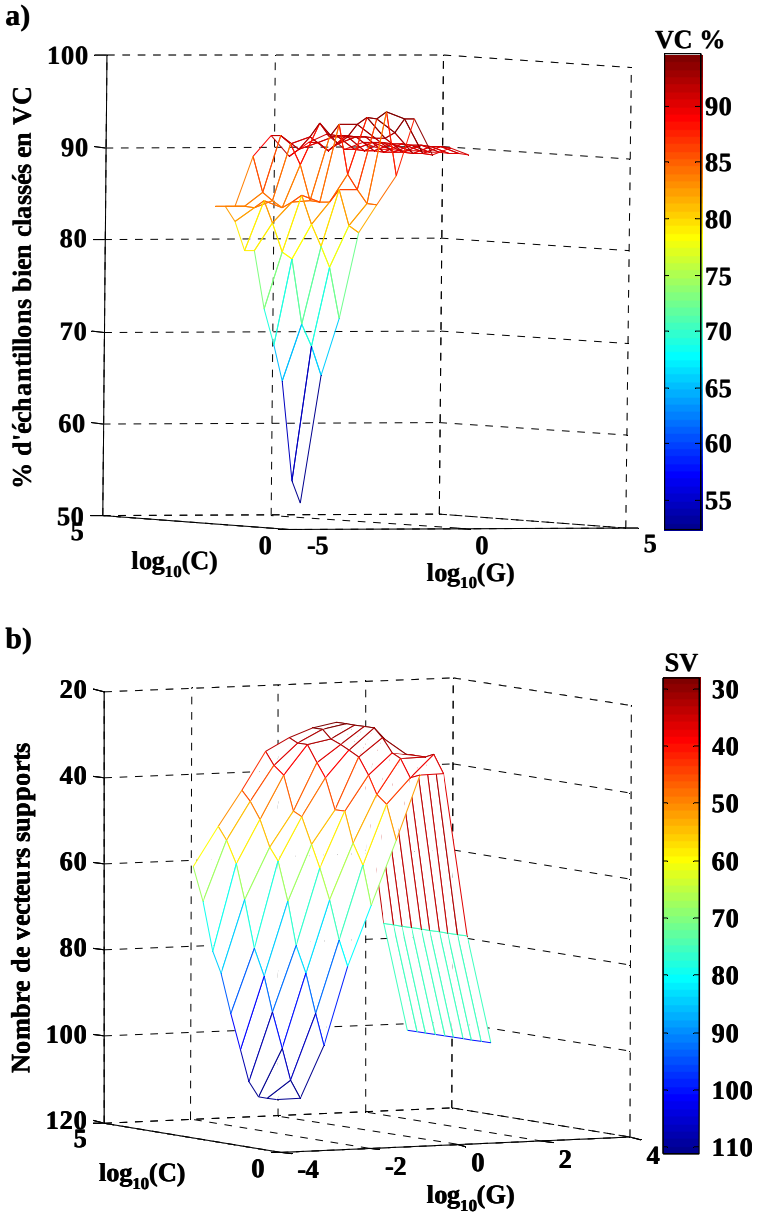


Figure 46 : Grille d'optimisation, a) pourcentages d'échantillons bien classés en fonction de $log_{10}(C)$ et $log_{10}(G)$ et b) nombre de vecteurs supports en fonction de $log_{10}(C)$ et $log_{10}(G)$.

Nous constatons sur la Figure 46a que la valeur du pourcentage maximal d'échantillons bien classés en validation croisée peut correspondre à plusieurs valeurs de *C* et de *G*. Dans ce cas, nous sommes face à une indétermination et il peut donc être intéressant de représenter le nombre de vecteurs supports en fonction des valeurs des logarithmes décimaux de *C* et *G* (Figure 46b). Les échantillons du lot d'entraînement identifiés par le modèle

comme étant des vecteurs supports suffisent à construire le modèle. Par conséquent, pour la plupart des applications le nombre de vecteurs supports est un paramètre à prendre en compte. Dans le cadre de notre étude, nous cherchons parmi les valeurs de *C* et de *G* optimales le nombre de vecteurs supports le plus faible. Dans l'exemple présenté sur la Figure 46 les valeurs de *C* et de *G* optimales tenant compte du nombre de vecteurs supports sont *C*=464 et *G*=12,91. Le nombre de vecteurs supports est égal à 36. Il s'agit donc d'un compromis entre les valeurs de *C* et de *G* correpondant d'un côté aux meilleurs résultats en validation croisée et de l'autre côté au nombre minimum de vecteurs supports. Lors de la construction des modèles SVM en fonction des différents prétraitements et de la répartition des échantillons dans les lots d'entraînement, nous avons remarqué, que la valeur du paramètre *G*, qui est liée à la largeur de la gaussienne, est identique pour les dix lots d'entraînement mais dépend du prétraitement utilisé c'est-à-dire du niveau des valeurs numériques associées aux échantillons. Les valeurs du paramètre *C* qui est lié à la marge, dépend plutôt de la distribution des échantillons et reste indépendante du prétraitement utilisé.

- ***Résultats***

Les résultats obtenus pour les différents prétraitements sont présentés dans le Tableau 9. Il s'agit des pourcentages moyens d'échantillons bien classés en validation croisée et en prédiction, obtenus comme précédemment sur les dix répartitions des lots d'échantillons et présentés accompagnés de leur intervalle de confiance au seuil α=5%. Comme précédemment, un simple centrage des spectres permet d'obtenir des résultats acceptables (89,1% (±1,8%) en validation croisée et 87,3% (± 2,4%) en prédiction) mais pas optimaux. Un test de comparaison des moyennes n'a pas mis en avant des différences significatives entre les résultats obtenus en validation croisée et en prédiction pour les prétraitements MC et SNV. Nous notons toutefois, une différence significative entre les résultats obtenus avec le prétraitement MC et avec le prétraitement SNVDER1. Nous retiendrons par la suite que sur les spectres complets prétraités SNVDER1, la méthode des SVM permet d'obtenir un taux de prédiction de 93,2% (±2,2%).

Prétraitements	VC en %	P en %
MC	89,1 (± 1,8)	87,3 (± 2,4)
SNV	90,5 (± 1,9)	88,8 (±2,5)
DER1	91,9 (± 1,1)	91,1 (± 2)
SNVDER1	93,9 (± 1,3)	93,2 (± 2,2)

Tableau 9 : Pourcentages moyens des échantillons bien classés obtenus sur les 10 lots d'entraînement et de prédiction en fonction du prétraitement utilisé avec un intervalle de confiance égal à 95%. VC : en validation croisée et P : en prédiction.

- ***Interprétation du modèle***

Le modèle interprété ci-dessous correspond au modèle construit avec les spectres prétraités SNVDER1. Le lot de prédiction (W_P) choisi pour l'interprétation est celui qui a servi à l'interprétation du modèle LDA.

– En terme de vecteurs supports

Afin de visualiser les vecteurs supports, les échantillons du lot d'entraînement sont représentés dans le plan (CP2-CP3) de la PCA (Figure 47). Les échantillons correspondant aux vecteurs supports sont entourés. Le nombre de vecteurs supports est de 36 échantillons répartis de la façon suivante 12 échantillons de la classe P1, 18 échantillons de la classe P2 et 6 échantillons de la classe P3. Le nombre de vecteurs supports de la classe P3 est relativement faible par rapport à celui des classes P1 et P2. Ceci peut s'expliquer par le fait que la variabilité chimique contenue dans cette classe est plus faible que dans les autres. En effet, les vecteurs supports de la classe P3 sont des échantillons 100% coton ou 100% lin, très représentatifs de cette classe. Les vecteurs supports de la classe P2 sont constitués de quelques échantillons purs mais surtout d'échantillons de composition plus complexe. En ce qui concerne la classe P1, le constat est identique. Même si les vecteurs supports sont déterminés dans un espace de plus grande dimension, nous remarquons que dans le plan (CP2-CP3) les vecteurs supports correspondent essentiellement aux échantillons situés en bordure de classes.

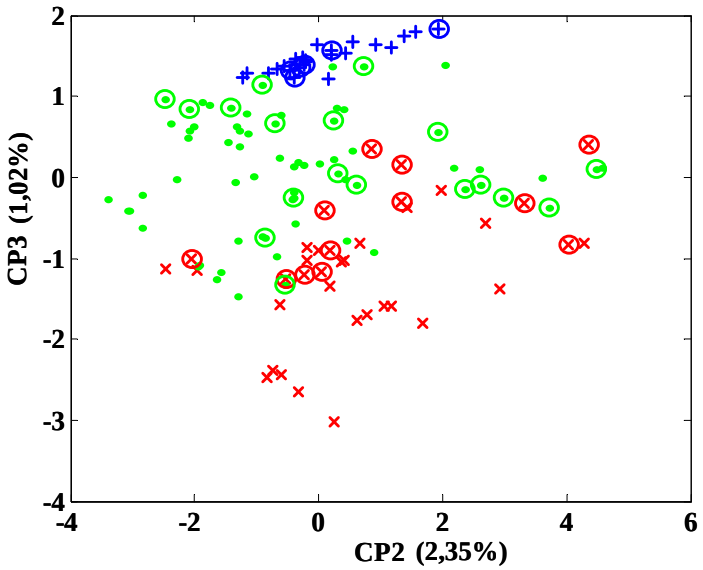


Figure 47 : Echantillons du lot d'entraînement projetés dans le plan (CP2-CP3) de la PCA. Echantillons de la classe P1(x), P2 (●), P3 (+). Les échantillons entourés correspondent aux vecteurs supports.

– En terme d'échantillons mal prédits

Afin d'identifier les échantillons mal classés par le modèle, les échantillons du lot de prédiction sont projetés dans le plan (CP2-CP3) de la PCA et le résultat est proposé sur la Figure 48. Cinq échantillons sont mal classés par le modèle. Ils sont situés en bordure des classes P1-P2 et P2-P3 dans le plan (CP2-CP3) de la PCA. Il faut noter que ces cinq échantillons étaient déjà mal classés par la méthode LDA. Cependant, tous les échantillons de la classe P2 sont bien classés. Un seul échantillon de la classe P3 est mal classé. Il s'agit de l'échantillon, noté 70, qui contient un apprêt chimique et qui a été interprété dans le chapitre 3, paragraphe III. Les quatre autres échantillons mal classés appartiennent à la classe P1. Parmi ceux-ci, deux échantillons, notés 18 et 48, sont à considérer avec précaution car leur composition exacte n'est pas représentée dans le lot d'entraînement.

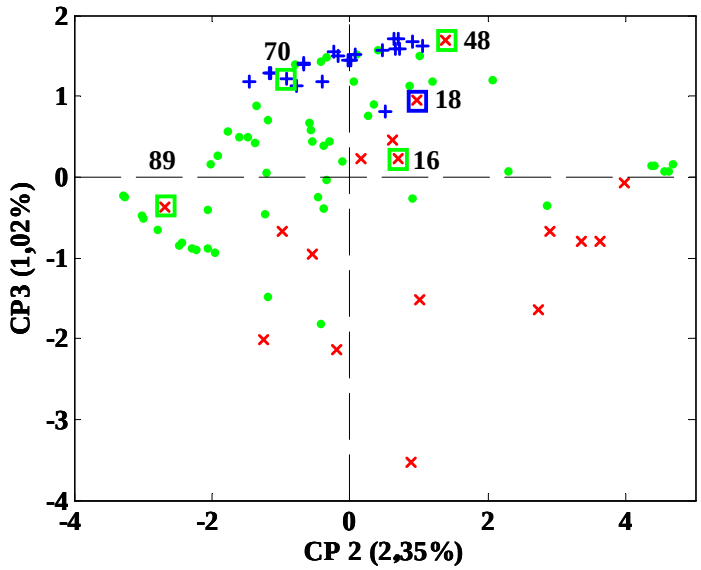


Figure 48 : Echantillons du lot de prédiction W_P projetés dans le plan (CP2- CP3) de la PCA. Echantillons de la classe P1(x), P2 (●), P3 (+). Les échantillons entourés d'un carré correspondent aux échantillons mal prédits, la couleur du carré indique la classe prédite et les numéros correspondent aux numéros des échantillons du lot de prédiction.

En résumé, les meilleures capacités prédictives obtenues sont de 93,2% (±2,2%) en moyenne pour le prétraitement SNVDER1 sur les spectres complets. Nous avons constaté précédemment sur le lot de prédiction W_P qu'un modèle construit sur 36 vecteurs supports (échantillons) classe de façon incorrecte 5 échantillons sur les 91 échantillons du lot de prédiction. Les résultats obtenus sur les spectres complets sont très encourageants en vue d'un passage à une instrumentation simplifiée.

VI. Résultats obtenus en classification sur les spectres réduits et discussion

Comme nous le verrons au chapitre 4, le cahier des charges défini par les partenaires industriels a orienté les choix instrumentaux vers un prototype à filtres interférentiels. Afin de montrer une faisabilité de la réduction du nombre de variables spectroscopiques des spectres

du spectrophotomètre de laboratoire, nous avons volontairement dégradé la qualité des spectres en simulant arbitrairement à pas régulier la position de 27 filtres théoriques (λ_1, ...,λ_{27}) sur le domaine spectral 1100-2500 nm. Les filtres interférentiels commerciaux couramment utilisés possèdent une largeur à mi-hauteur d'environ 10 nm. Afin d'obtenir la valeur simulée pour chacun des 27 filtres, nous avons moyenné pour chaque filtre les valeurs des 21 valeurs d'absorbance autour de la longueur d'onde cible (la résolution spectrale du spectrophotomètre étant de 0,5 nm). La matrice des données spectrales $\boldsymbol{X}$ est réduite de 2800 variables (spectres complets) à 27 variables. Comme nous l'avons déjà remarqué, la réduction du nombre de variables ne permet plus d'utiliser des prétraitements basés sur des moyennes mobiles tels que la dérivée. Les spectres sont donc ici prétraités uniquement SNV et les résultats sont repris uniquement pour les SVM.

Nous rappelons dans le Tableau 10, les capacités prédictives moyennes obtenues sur les spectres complets prétraités SNV. Nous présentons aussi le nombre de vecteurs supports utilisés pour la construction du modèle sur le lot d'entraînement W_C et le nombre d'échantillons mal prédits pour le lot de prédiction W_P.

Prétraitement	VC en %	P en %	Nbre de vecteurs supports pour W_C	Nbre d'échantillons mal prédits pour W_P
SNV	90,5 (± 1,9)	88,8 (±2,5)	29	9

Tableau 10 : Résultats obtenus sur les spectres complets prétaités SNV. W_C : lot d'entraînement, W_P : lot de prédiction.

- ***Résultats***

Les pourcentages d'échantillons bien classés en validation croisée et en prédiction sont, respectivement, de 91,3% (± 1,6%) et 90,1% (± 1,7%) pour les spectres réduits. Nous noterons qu'un test de comparaison des moyennes ne permet pas de mettre en évidence une différence significative entre ces résultats, pour un intervalle de confiance au seuil α=5%. La réduction des spectres n'affecte pas le niveau de prédiction moyen obtenu sur le spectrophotomètre de laboratoire.

- ***Interprétation du modèle***

L'interprétation des modèles est réalisée en comparant deux modèles construits pour le même lot d'entraînement (W_C), l'un sur les spectres complets (2800 variables) et l'autre sur les spectres réduits (27 variables).

Malgré la réduction importante du nombre de variables, le nombre de vecteurs supports nécessaires à la construction du modèle reste du même ordre de grandeur pour les deux modèles. En effet, en ce qui concerne les spectres complets, le nombre de vecteurs supports est égal à 29 et, pour les spectres réduits, il vaut 23. Afin de visualiser les vecteurs supports, les échantillons du lot d'entraînement sont représentés dans le plan (CP2-CP3) de la PCA (Figure 49a) pour les spectres complets et sont représentés sur la Figure 49b dans le plan (CP1-CP2) de la PCA pour les spectres réduits. Nous noterons que les plans des composantes principales sont choisis de façon à optimiser la représentation des vecteurs supports. Les échantillons correspondant aux vecteurs supports sont entourés.

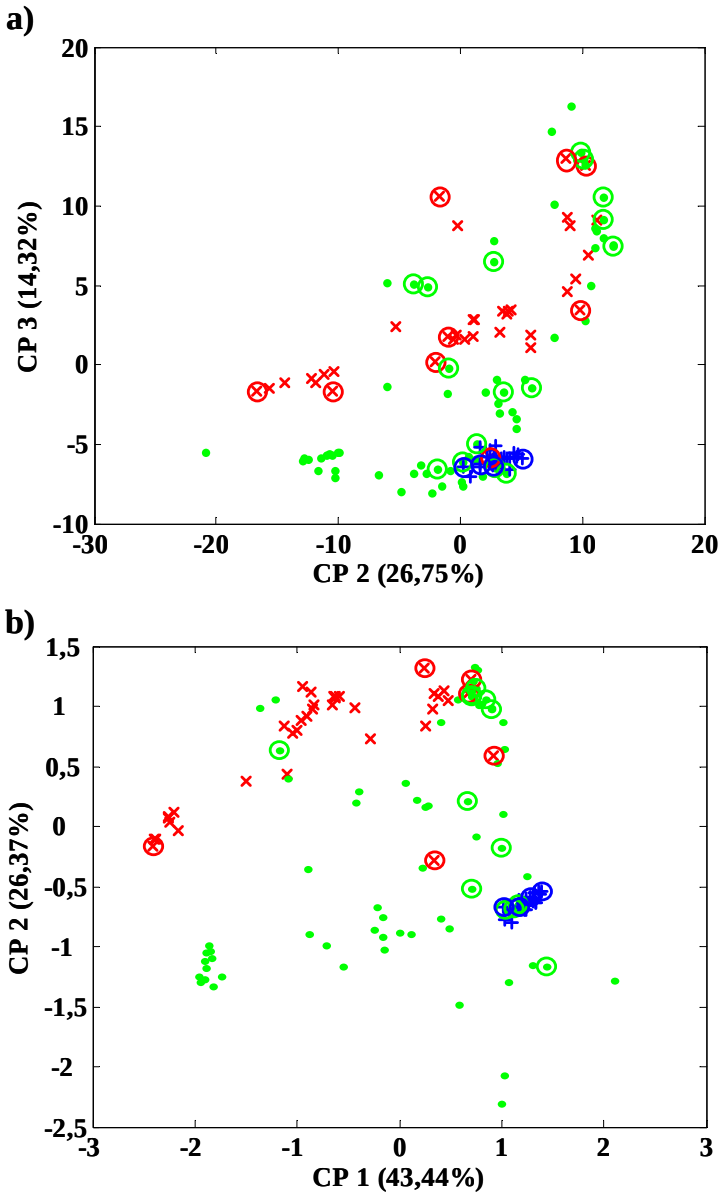


Figure 49 : a) PCA sur les spectres complets, échantillons du lot d'entraînement W_C projetés dans le plan (CP2-CP3), b) PCA sur les spectres réduits, échantillons du lot d'entraînement W_C projetés dans le plan (CP1-CP2). Echantillons de la classe P1(x), P2 (●), P3 (+). Les échantillons entourés correspondent aux vecteurs supports.

Afin d'identifier les échantillons mal classés par le modèle, les échantillons du lot de prédiction pour les spectres complets et pour les spectres réduits sont projetés dans le plan (CP2-CP3) et (CP1-CP2) de la PCA, respectivement, sur la Figure 50a et c.

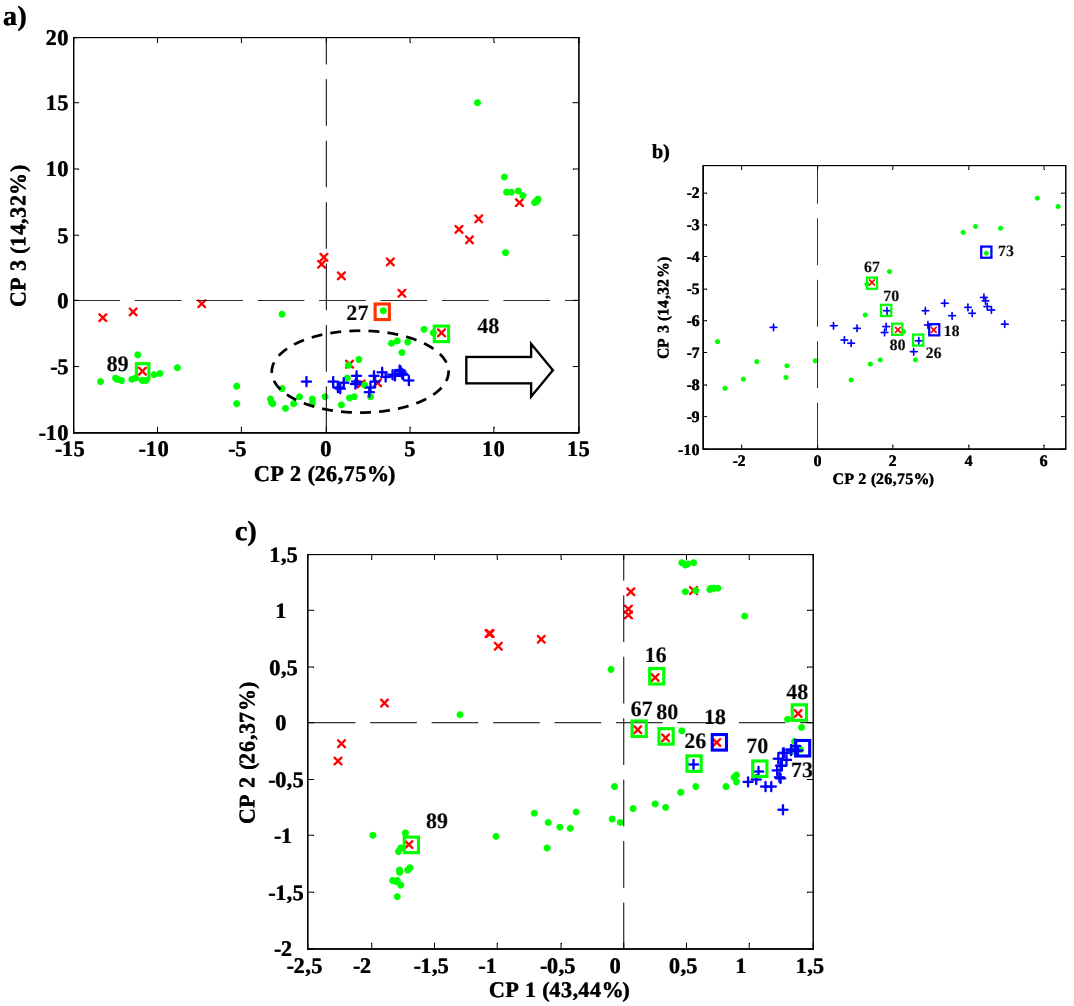


Figure 50 : a) PCA sur les spectres complets, échantillons du lot de prédiction W_P projetés dans le plan (CP2-CP3), b) zoom correspondant à l'ellipse, c) PCA sur les spectres réduits, échantillons du lot de prédiction W_P projetés dans le plan (CP1-CP2). Echantillons de la classe classe P1(x), P2 (●), P3 (+). Les échantillons entourés d'un carré correspondent aux échantillons mal prédits et les numéros correspondent aux numéros des échantillons du lot de prédiction.

Le nombre d'échantillons mal prédits est identique pour les deux modèles, il est égal à 9. Par conséquent, les capacités prédictives obtenues sur les spectres réduits restent convenables. Ces échantillons sont situés, par principe de la méthode, en bordure de classes. A l'exception d'un échantillon, les échantillons qui sont mal attribués sont les mêmes c'est-à-

dire que les échantillons qui sont mal classés par le modèle sur spectres complets ne sont pas mieux prédits avec les spectres réduits.

- ***Conclusion***

Pour développer une analyse qualitative satisfaisante, le choix des échantillons du lot d'entraînement, la mesure spectrale, la mesure de référence, le choix et la construction du modèle de classification sont déterminants dans sa mise en place. Les échantillons utilisés lors de l'étalonnage doivent être d'une part, en nombre suffisant afin de couvrir les variabilités physico-chimiques, et d'autre part, être représentatifs de la base de données à analyser. C'est-à-dire pour les tissus textiles, il faut à la fois avoir une diversité des fibres, en terme de natures chimiques, d'origines, filage, tissage...

La technique est dépendante des objectifs de l'analyse et de la nature physique des échantillons. Les mesures spectrales ont été réalisées en réflexion diffuse. Les valeurs de la mesure de référence doivent être déterminées dans les mêmes conditions pour tous les tissus du lot de données car les modèles d'étalonnage sont établis à partir des mesures de référence.

Deux méthodes de discrimination ont été utilisées dans ce chapitre, la méthode LDA et la méthode des SVM. Les meilleurs résultats obtenus en prédiction sur les spectres complets avec la méthode LDA sont de 82,5% (±1,9%) alors que ceux obtenus avec la méthode des SVM sont de 93,2% (±2,2%). Nous avons constaté que la distribution et la complexité des tissus textiles sont telles que les méthodes établissant des frontières linéaires entre les classes (LDA) ne permettent pas une discrimination optimale entre les classes.

Nous avons montré une faisabilité avec les méthodes SVM sur les spectres complets. Dans l'objectif de l'utilisation d'une instrumentation simplifiée, les SVM ont, en plus, permis la réduction par un échantillonage en longueurs d'onde du domaine spectral. Il n'a pas pu être mis en évidence de dégradation des capacités prédictives statistiquement significatives pour un intervalle de confiance au seuil α=5%.

Chapitre 4

Analyse qualitative de tissus sur spectrophotomètre prototype

---

L'objectif des partenaires industriels est de réaliser à terme une instrumentation simplifiée, de type capteur, à moindre coût de production. Il reste que pour l'instant, le prix de la miniaturisation en spectroscopie PIR est encore relativement élevé, notamment par rapport au système de détection. A l'heure actuelle, un prototype intermédiaire a été mis au point par les partenaires industriels sur la gamme spectrale qui s'étend de 1100 à 1800 nm. Dans cette phase de développement, le choix d'une technologie basée sur l'utilisation de filtres interférentiels permet de conserver une certaine modularité dans le choix des longueurs d'onde et d'avoir ainsi un bon outil de « paillasse ».

Nous verrons dans un premier temps, une description sommaire du prototype. Mon travail a concerné la caractérisation métrologique de ce dernier lors de la phase de développement, puis la construction de modèles qualitatifs pour la prédiction.

I. Description du prototype

Les choix, notamment celui de la gamme spectrale 1100 à 1800 nm, ont été dépendant du système de détection pour lequel la marge de manœuvre est faible (pour des raisons de coût de revient). Au cours de la phase de développement, plusieurs campagnes de mesures ont été réalisées afin d'améliorer la conception optique, l'électronique ou la présentation de l'échantillon. Un schéma de principe du montage optique du prototype est présenté sur la Figure 51. Les principaux éléments optiques sont détaillés ci-dessous.

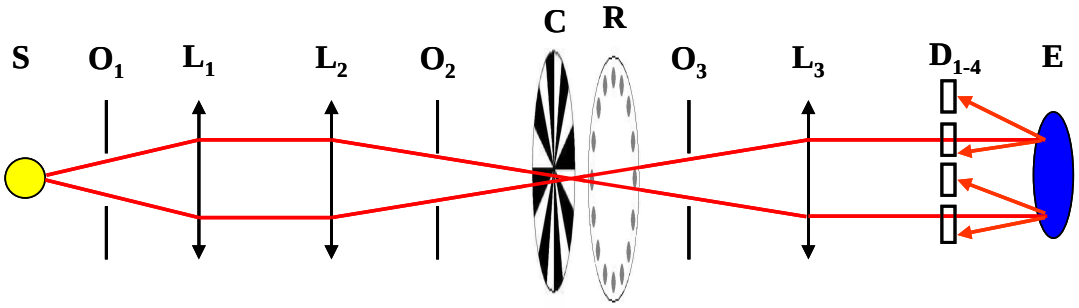


*Figure 51 : **S** : source, **O**$_{1-3}$: obturateurs, **L**$_{1-3}$: lentilles, **C** : chopper, **R** : roue à filtres, **D**$_{1-4}$: détecteurs, **E** : échantillon.*

La source **S** du prototype est une ampoule halogène quartz. Un obturateur $\mathbf{O_1}$ limite l'énergie lumineuse incidente injectée dans le système. Le faisceau lumineux passe ensuite à travers deux lentilles plano-convexes en verre borosilicaté (BK7) qui possèdent une transmission de 85 à 90% sur la gamme spectrale. Un second obturateur $\mathbf{O_2}$ est placé après la seconde lentille. La lumière est segmentée dans le temps à l'aide d'un modulateur (*chopper*) qui est situé dans le plan focal image de la lentille $\mathbf{L_2}$. Le *chopper* améliore la détection et l'amplification du signal lumineux [108]. Les longueurs d'onde de la lumière sont sélectionnées par des filtres optiques interférentiels qui sont portés par un barillet (roue à filtres) placé perpendiculairement au rayon lumineux au plus près de ce plan. Un système de codage en relation avec un moteur pas à pas permet de placer successivement chacun des filtres devant les rayons lumineux. Les rayons lumineux sont perpendiculaires à la surface du filtre et l'angle d'incidence reste constant pour toutes les mesures. Dans la configuration du spectrophotomètre présentée, la roue reçoit 13 filtres interférentiels dont les longueurs d'onde,

numérotées λ_1, λ_2, …, λ_{13}, sont contenues dans la gamme spectrale 1100-1800 nm et ne sont pas explicitées afin de respecter la confidentialité. Ces choix tiennent compte aussi du cahier des charges défini par les partenaires industriels. Les largeurs à mi-hauteur des bandes passantes de ces filtres ont été caractérisées et sont typiquement de l'ordre de 10 nm. Un troisième obturateur **O_3** est situé après la roue à filtres. Les rayons lumineux parallèles illuminent l'échantillon. L'énergie lumineuse provenant de la réflexion diffuse de l'échantillon est mesurée par quatre détecteurs inclinés avec un angle de 45° par rapport au faisceau incident. La position des détecteurs est schématisée sur la Figure 52. Pour plus de clarté, le schéma de la Figure 52 représente seulement deux détecteurs, les autres sont situés à 90° de part et d'autre de ces deux détecteurs. Afin de s'affranchir du milieu extérieur, le faisceau lumineux avant le *chopper* est confiné dans un tube.

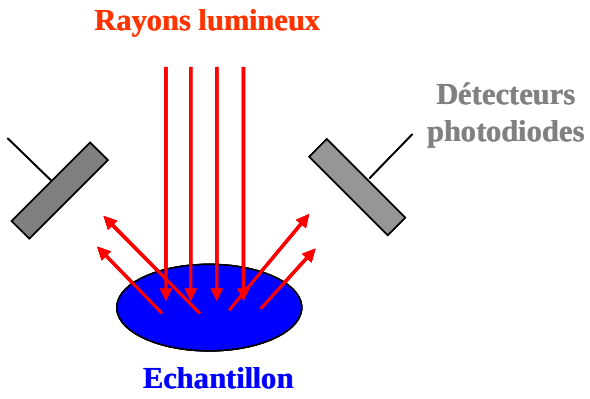


Figure 52 : Positionnement des détecteurs.

La nature du détecteur dépend de la gamme spectrale d'utilisation. Les détecteurs utilisés dans cette application sont des photodiodes en arséniure de gallium et d'indium (InGaAs étendu) qui sont sensibles dans la région spectrale comprise entre 1100 et 1800 nm. La surface de chaque détecteur est de 1 mm^2. Les valeurs en volts obtenues pour chacun des détecteurs sont moyennées. Une carte d'acquisition de chez *National Instrument* permet de récupérer ces valeurs et de les mettre en forme par l'intermédiaire du logiciel *Labview*. Un échantillon est caractérisé par un spectre discret constitué par les 13 mesures obtenues.

Le spectre de la référence est réalisé sur un disque circulaire de Spectralon® de caractéristiques proches de la référence interne des spectrophotomètres de laboratoire. Il s'agit d'une résine thermoplastique dont la réflectance est supérieure à 95% sur le domaine spectral

d'intérêt. Le spectre de la référence est acquis avant chaque mesure d'un échantillon. On peut alors convertir les mesures en unités usuelles d'absorption lumineuse de la forme $log(I_0/I)$, I_0 étant l'intensité de la référence et I l'intensité de l'échantillon à une longueur d'onde donnée. La méthode d'échantillonnage est identique à celle employée sur le spectrophotomètre de laboratoire.

II. Caractérisation métrologique du prototype

II.1. Fidélité d'une méthode de mesure

La fidélité[109] représente l'écart entre les résultats obtenus en appliquant à plusieurs reprises la même méthode sur un même échantillon dans des conditions déterminées. Lorsque les conditions de mesures sont les plus identiques possible, par exemple un même opérateur, des mêmes conditions d'analyses, un intervalle de temps réduit, on parle de répétabilité. Par contre, lorsque les conditions entre les différentes répétitions de mesures sont changées, par exemple des analyses effectuées dans des laboratoires différents, positionnements différents de l'échantillon, on parle alors de reproductibilité. Ces deux grandeurs sont mathématiquement définies, d'une part par les écarts-types de répétabilité et de reproductibilité et, d'autre part, par le coefficient de variation (CV) (ou coefficient relatif de déviation standard, RSD) qui correspond à l'écart-type. Dans notre cas, l'écart-type et le CV seront calculés à chaque longueur d'onde et déterminés selon les Équations 42 et 43[109] :

Équation 42 $$s = \sqrt{\frac{\sum_i \left(a_i - \bar{a}\right)^2}{n-1}}$$ *écart-type*

avec n le nombre de répétitions de la mesure, c'est-à-dire le nombre de spectres réalisés pour la reproductibilité ou la répétabilité, a_i l'absorbance du spectre i, $\bar{a}$ l'absorbance moyenne des n spectres.

Équation 43 $$CV = 100\frac{s}{\bar{a}}$$ coefficient de variation

II.2. Reproductibilité

La reproductibilité est réalisée sur sept échantillons (E1-E7) contenant un mélange de fibres coton/PES avec une teneur en coton variant de 0 à 100%. Les valeurs de la mesure de référence de ces échantillons sont connues en terme de classe (analyse qualitative) et en terme de teneur en coton (analyse quantitative). Pour chacun des échantillons dix mesures sont acquises, dans un temps court, en repositionnant l'échantillon avant chaque mesure. Le Tableau 11 présente les coefficients de variation des 7 échantillons à deux longueurs d'onde, une caractéristique de la bande d'absorption du coton (λ_8) et une caractéristique de la bande du polyester (λ_{11}).

Échantillons	E1	E2	E3	E4	E5	E6	E7
Composition (%) coton/PES	100/0	70/30	65/35	55/45	50/50	35/65	0/100
CV (λ_8)	0,65	0,56	0,47	2,21	1,77	1,29	2,25
CV (λ_{11})	0,70	0,59	0,56	2,11	1,68	1,32	2,08

Tableau 11 : Coefficient de variation (CV) en reproductibilité.

A titre de comparaison, pour l'échantillon E5, la valeur du CV du prototype est égale à 1,77% à λ_8 et à 1,68% à λ_{11} alors que la valeur du CV du spectrophotomètre de laboratoire est égale à 1,47% à λ_8 et à 1,57% à λ_{11}. Le premier constat, d'ordre général est le suivant, le CV calculé pour ces deux longueurs d'onde reste acceptable par rapport à celui calculé sur le spectrophotomètre de laboratoire. Toutefois, il est difficile d'interpréter la reproductibilité en considérant seulement deux longueurs d'onde, ou même plusieurs considérées individuellement.

Afin de visualiser la reproductibilité obtenue pour chaque échantillon, une PCA est réalisée sur la matrice contenant les valeurs des absorbances des spectres discrets (13 points de mesure). Les échantillons sont projetés dans un espace à trois dimensions formé par les trois premières composantes principales (Figure 53).

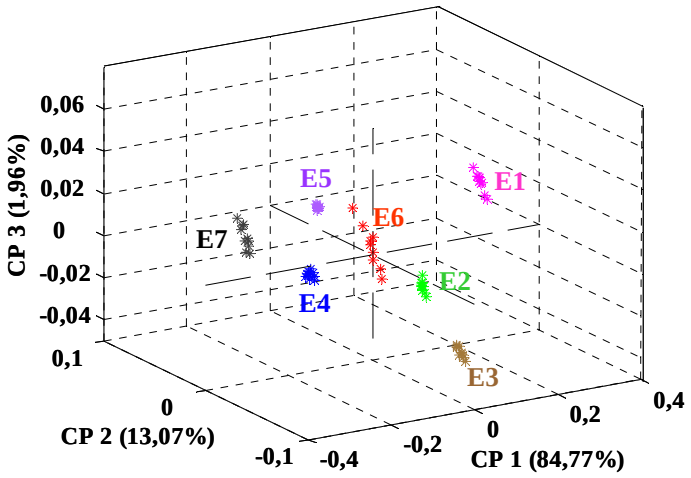


Figure 53 : PCA concernant la reproductibilité sur les échantillons E1 à E7.

Nous remarquons une direction de variation dans le plan (CP2-CP3) due à la nature des échantillons analysés et au repositionnement de l'échantillon avant chaque mesure.

- ***Interprétation des vecteurs propres***

Les vecteurs propres ont la dimension des spectres discrets, c'est-à-dire qu'ils comportent 13 points. Néanmoins, afin de conserver la confidentialité concernant le positionnement précis en longueurs d'onde des 13 filtres, ils sont représentés, après un lissage, sur la Figure 54.

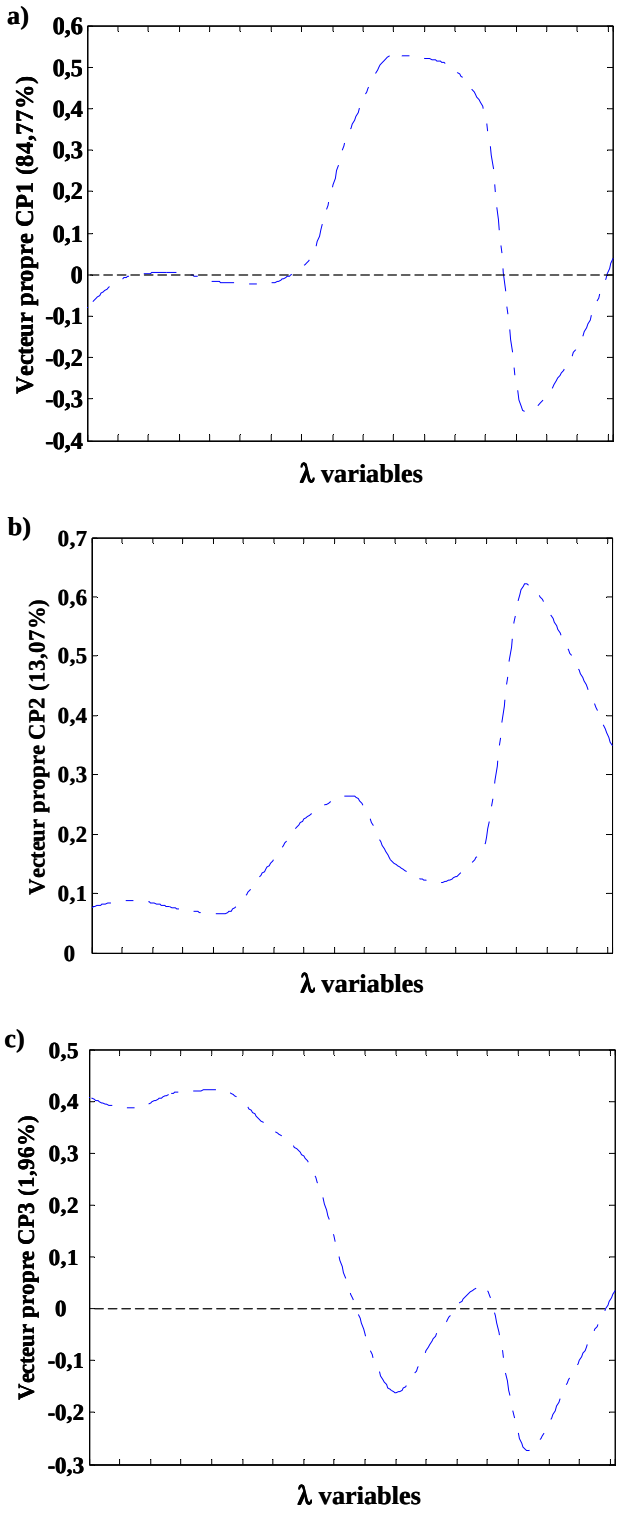


Figure 54 : Vecteurs propres a) CP1, b) CP2 et c) CP3.

Le vecteur propre de la CP1 oppose l'information spectrale contenue dans un échantillon pur coton à celle contenue dans un échantillon pur polyester (chapitre 2). L'information contenue dans le vecteur propre de la CP2 est essentiellement liée au polyester. La partie positive du vecteur propre de la CP3 exprime essentiellement les variabilités physico-chimiques contenues dans le lot de données, alors que la partie négative exprime l'information contenue dans un échantillon PES.

II.3. Répétabilité

La répétabilité est réalisée sur les sept échantillons utilisés pour la reproductibilité (E1 à E7). Une fois l'échantillon positionné, dix mesures successives sont répétées dans un temps relativement court. Le coefficient de variation (CV) est calculé pour les sept échantillons en fonction des 13 longueurs d'onde. Les valeurs des CV obtenues sont reportées dans le tableau pour deux longueurs d'onde, λ_8 et λ_{11}.

Échantillons	E1	E2	E3	E4	E5	E6	E7
Composition (%) coton/PES	100/0	70/30	65/35	55/45	50/50	35/65	0/100
CV (λ_8)	-	0,2	0,2	-	0,23	-	-
CV (λ_{11})	0,16	0,11	0,07	0,09	0,07	-	-

Tableau 12 : Coefficient de variation (CV) en répétabilité, (-) représente les valeurs inférieures à 0,001.

Nous constatons pour les longueurs d'onde λ_8 à λ_{11}, que les valeurs de CV sont très faibles voire même quasi-nulles pour certains échantillons (-). Or les valeurs de répétabilité obtenues avec le spectrophotomètre de laboratoire sont quasiment nulles. En spectroscopie proche infrarouge, le recouvrement spectral important oblige à avoir une bonne répétabilité et c'est le cas ici[110-112].

- ***Visualisation des données***

Une PCA est réalisée sur les données de répétabilité. La Figure 55 représente la projection des dix mesures pour les sept échantillons dans l'espace à trois dimensions (CP1, CP2, CP3).

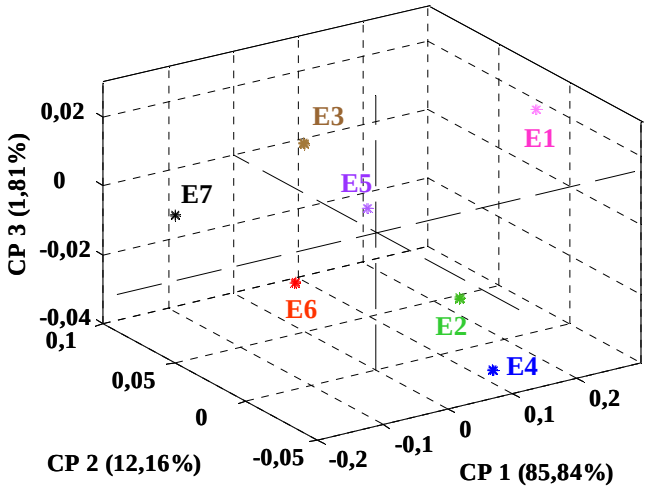


Figure 55 : PCA concernant la répétabilité sur les échantillons E1 à E7.

Nous constatons que les mesures de répétabilité pour un échantillon donné se projettent dans un voisinage très proche. Globalement, la répétabilité est très acceptable. Contrairement au graphique concernant la reproductibilité, nous ne constatons pas de direction privilégiée de la variance. En effet, ce qui change ici c'est qu'il n'y a pas de repositionnement de l'échantillon avant chaque mesure, il n'y a donc pas de différences dues aux aspects physiques de l'échantillon. Par conséquent, l'optimisation qui serait à apporter au prototype serait une meilleure prise en compte de la nature physique de l'échantillon analysé.

Les trois premiers vecteurs propres de la PCA ne sont pas représentés ici car leur interprétation chimique est très similaire à ceux de la PCA réalisée sur les mesures de reproductibilité. En effet, il s'agit des mêmes échantillons. Par conséquent, les variabilités chimiques sont très proche.

II.4. Dynamique du prototype

Pour évaluer la dynamique du prototype dans les conditions réelles d'acquisition des spectres, nous avons réalisé trois mesures sur chacun des 7 échantillons coton/PES. Ces mesures ont été ensuite moyennées. Afin d'observer cette dynamique, les absorbances des échantillons (E1 à E7) mesurées à une longueur d'onde caractéristique du coton (λ_8) sont représentées en fonction de la teneur en coton sur la Figure 56a. De même, les absorbances de ces mêmes échantillons sont mesurées à une longueur d'onde caractéristique du PES (λ_{11}) et sont représentées en fonction de la teneur en coton sur la Figure 56b.

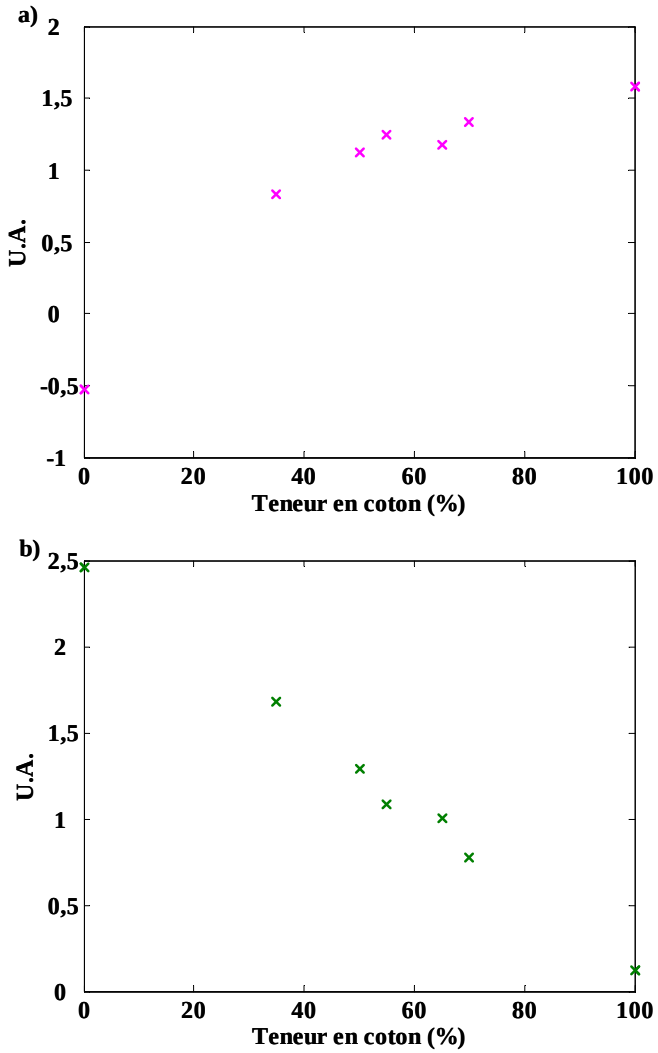


Figure 56 : Absorbances en fonction de la composition.

Nous constatons que pour ces deux longueurs d'onde la dynamique du prototype est acceptable ici sur des spectres prétraités SNV. Néanmoins, nous remarquons que la corrélation entre les valeurs des absorbances et la teneur en coton à la longueur d'onde λ_8 n'est pas parfaitement linéaire alors qu'à λ_{11}, elle est quasiment linéaire. Ce résultat est en accord avec ceux obtenus sur le lot industriel coton/PES (chapitre 2, paragraphe II.2).

II.5. Etude préliminaire de matières textiles coton/PES

Avant d'acquérir les spectres du lot complet de données, nous avons réalisé dans un premier temps une analyse préliminaire sur 47 échantillons constitués uniquement d'échantillons purs coton, polyester et de mélanges coton/PES et pour lesquels sont associées différentes valeurs de la propriété physique *Y*. Nous enregistrons les spectres dans les mêmes conditions à la fois sur le spectrophotomètre de laboratoire et sur le prototype. Les spectres acquis sur le spectrophotomètre de laboratoire sont réduits sur la gamme spectrale 1100-1800 nm et sont regroupés dans la matrice notée $\boldsymbol{X_{L1}}$ (laboratoire). Les spectres discrets des 47 échantillons acquis sur le prototype constituent la matrice notée $\boldsymbol{X_{P1}}$ (prototype). L'analyse PCA est réalisée indépendamment sur les deux matrices. Nous comparons les vecteurs propres de la CP1, CP2 et CP3 des matrices $\boldsymbol{X_{P1}}$ et $\boldsymbol{X_{L1}}$ (Figure 57). Les vecteurs propres de la matrice $\boldsymbol{X_{P1}}$ sont représentés après un lissage. Les scores des deux matrices ne sont pas représentés ici, l'idée étant, après la caractérisation métrologique présentée précédemment, de s'intéresser à l'information physico-chimique.

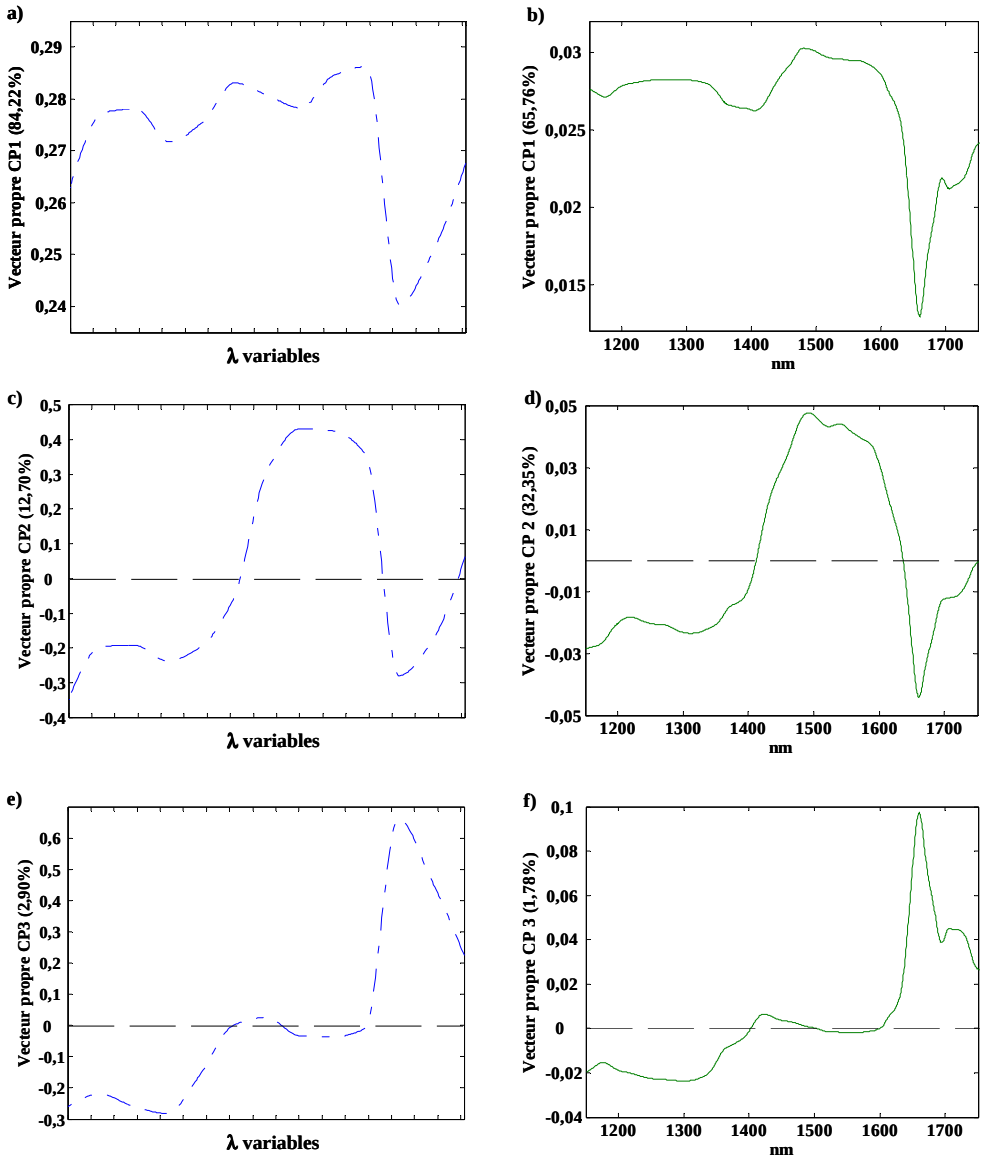


Figure 57 : Vecteurs propres des composantes principales, respectivement, pour la matrice $\boldsymbol{X_{P1}}$*, a) CP1, c) CP2, e) CP3 et pour la matrice* $\boldsymbol{X_{L1}}$*, b) CP1, d) CP2, f) CP3.*

L'interprétation des vecteurs propres est similaire à celle faite dans le chapitre 4, paragraphe II.2. Nous constatons que l'allure des vecteurs propres est identique pour les deux matrices $\boldsymbol{X_{P1}}$ et $\boldsymbol{X_{L1}}$. En d'autres termes, l'information accessible sur cette matrice est la même pour les deux instruments, ce qui révèle un bon accord entre le prototype et le spectrophotomètre de laboratoire sur cette matrice de données.

Les résultats obtenus pour la caractérisation métrologique du prototype semblent valider ce dernier que ce soit par des mesures numériques d'échantillons standard (reproductibilité, répétabilité, CV) ou avec un lot de données coton/PES comportant des variabilités physico-chimiques. Ces résultats laissent entrevoir un bon potentiel pour l'analyse qualitative du lot complet des données.

III. Résultats obtenus en classification sur les spectres du prototype et discussion

Les 221 échantillons du lot de données sont acquis sur le prototype dans les mêmes conditions que sur le spectrophotomètre de laboratoire. On rappelle ici que pour chaque échantillon trois mesures sont réalisées pour différentes positions de l'échantillon et ce sans préparation préalable de l'échantillon. Les absorbances de ces trois spectres sont ensuite moyennées pour mieux prendre en compte les différences d'homogénéité ou d'aspects de la surface, par exemple. Les spectres des 221 échantillons sont représentés sur la Figure 58. Ils sont regroupés dans la matrice des données spectrales, notée $\boldsymbol{X_P}$.

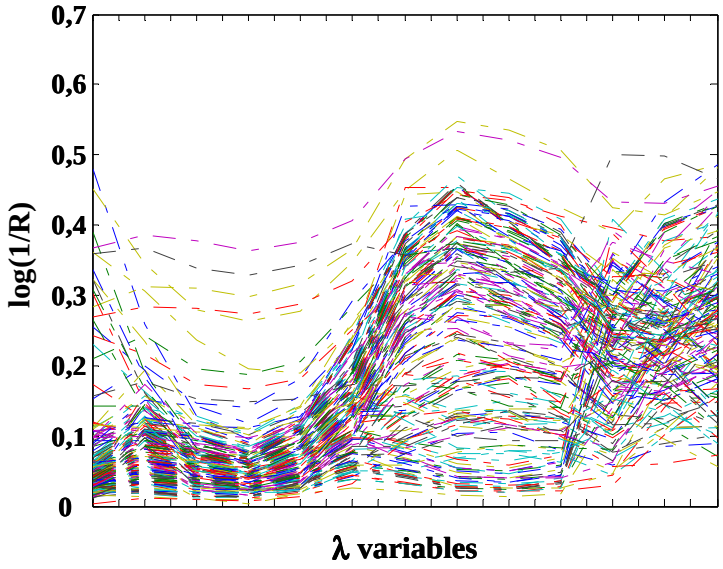


Figure 58 : Spectres discrets des échantillons du lot de données sur le prototype.

- ***Visualisation des données***

Une PCA est réalisée sur la matrice $\boldsymbol{X_P}$. La Figure 59 présente les échantillons de la matrice $\boldsymbol{X_P}$ projetés dans le plan (CP2-CP3) de la PCA qui représente 5,28% de la variance totale. Nous remarquons que les échantillons sont répartis de façon homogène dans l'espace (CP2-CP3) qui représente 5,28% de la variance totale.

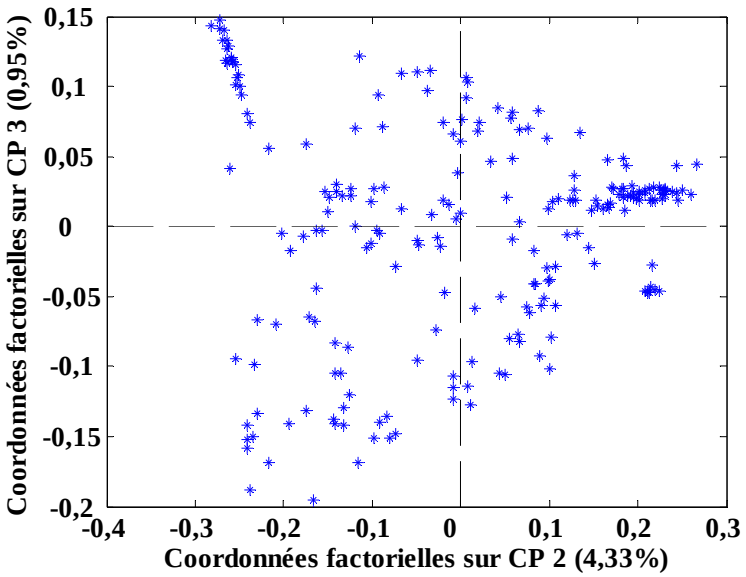


Figure 59 : PCA de la matrice $\boldsymbol{X_P}$ dans le plan (CP2-CP3).

Pour l'interprétation, nous avons comparé les vecteurs propres de la matrice ***Xp*** (221x13) avec ceux de la matrice ***X*** (pour les spectres acquis sur le spectrophotomètre, matrice définie au chapitre 3) sur le domaine 1100-1800 nm (221x1401). Les vecteurs propres des trois premières composantes principales sont représentés sur la Figure 60.

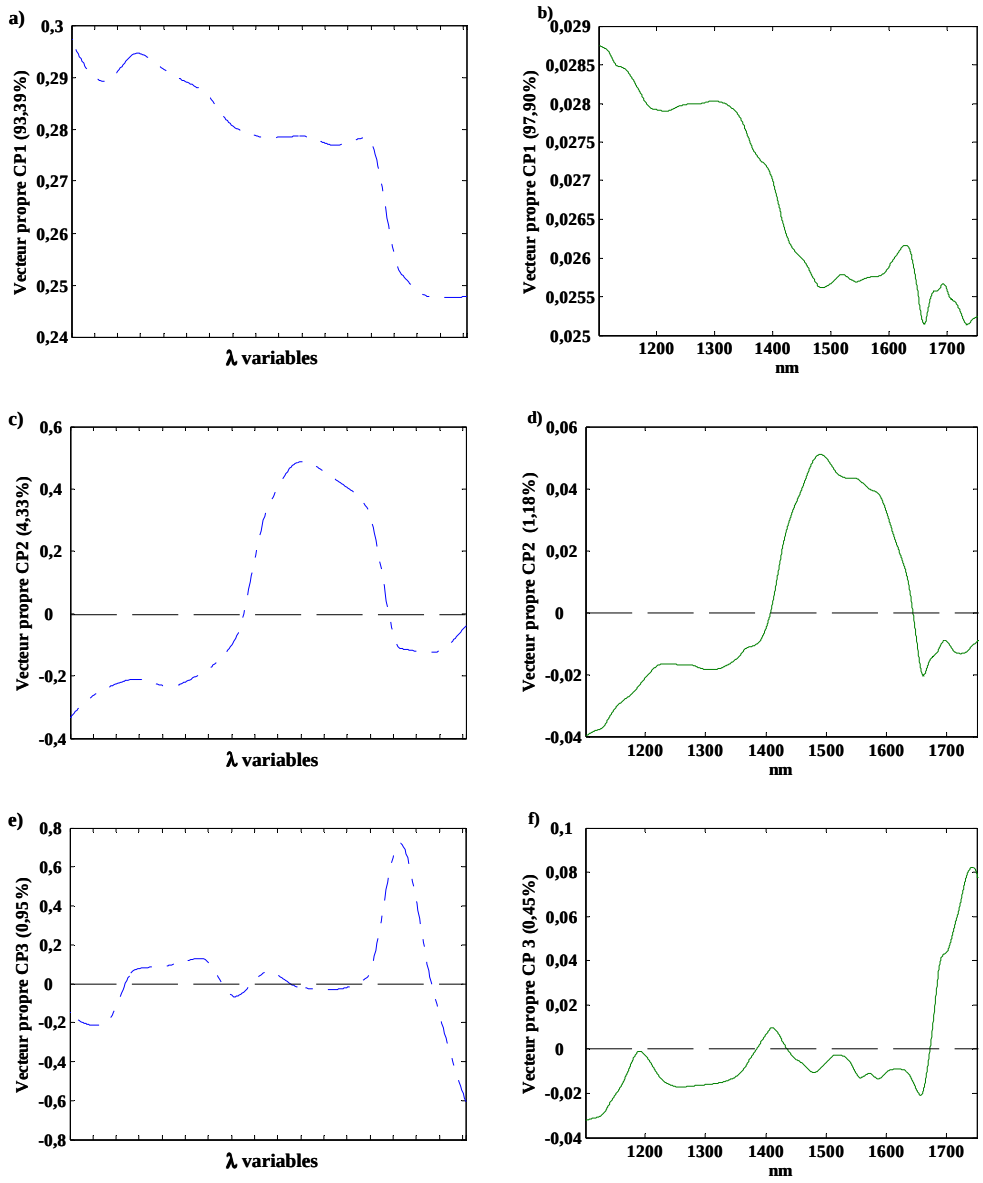


*Figure 60 : Vecteurs propres de la matrice **Xp** après un lissage, CP1 en a), CP2 en c) et CP3 en e) et vecteurs propres de la matrice **X**, CP1 en b), CP2 en d) et CP3 en f).*

Nous constatons une bonne adéquation entre les résultats obtenus sur les deux spectrophotomètres. Le vecteur propre de la CP1 exprime les variabilités contenues dans les spectres. Nous remarquons une dérive de la ligne de base importante due au fait que les spectres ne sont pas encore prétraités. L'interprétation des vecteurs propres de CP2 et CP3 a déjà été détaillée dans le chapitre 3, paragraphe IV.

- ***Résultats***

Un modèle SVM a été développé sur les spectres de la matrice $\boldsymbol{X_P}$. Les répartitions des échantillons dans les lots d'entraînement et de prédiction sont conservées, pour permettre la comparaison des résultats. Seuls les résultats obtenus avec le prétraitement SNV sont retenus. Les pourcentages moyens d'échantillons bien classés en validation croisée et en prédiction sont, respectivement, de 89% (± 2,2%) et de 88,1% (± 3,1%). Compte tenu de la réduction de la dimension du problème et du cahier des charges défini par les partenaires industriels, les capacités prédictives obtenues avec le protype sont acceptables.

- ***Interprétation du modèle***

L'interprétation du modèle est réalisée pour le lot d'entraînement W_C et W_P. Les échantillons du lot d'entraînement sont projetés dans le plan (CP2-CP3) de la PCA sur la Figure 61. Les échantillons correspondant aux vecteurs supports sont entourés. Le nombre de vecteurs supports nécessaires à la construction du modèle est égal à 37 échantillons alors qu'il était égal à 29 pour les spectres complets. La répartition des vecteurs supports en fonction des trois classes est la suivante : 17 vecteurs supports pour la classe P1, 15 vecteurs supports pour la classe P2 et 5 vecteurs supports pour la classe P3.

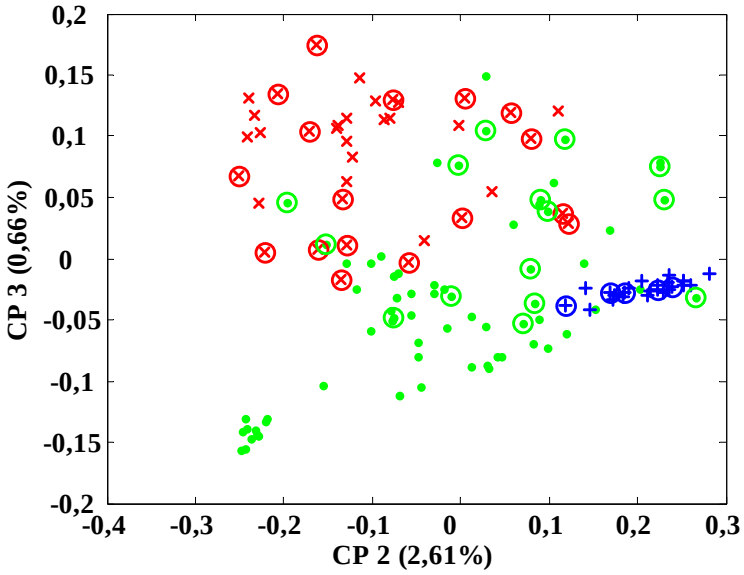


Figure 61 : Echantillons du lot d'entraînement W_C projetés dans le plan (CP2-CP3) de la PCA. Echantillons de la classe P1(x), P2 (●), P3 (+). Les échantillons entourés correspondent aux vecteurs supports.

Le modèle construit avec les 37 vecteurs supports classe de façon incorrecte 9 échantillons. Ces échantillons sont identifiés par des rectangles sur la Figure 62. Nous constatons qu'en dépit de la réduction du nombre de variables spectroscopiques, le nombre d'échantillons mal prédits reste identique pour la même distribution des échantillons du lot de prédiction. Parmi ces 9 échantillons, 7 échantillons étaient mal classés par le modèle sur les spectres complets. C'est le cas des échantillons n°18, 26, 48, 70, 73, 80 et 89. De plus, malgré l'augmentation du nombre de vecteurs supports appartenant à la classe P1, le nombre d'échantillons P1 mal prédits reste similaire.

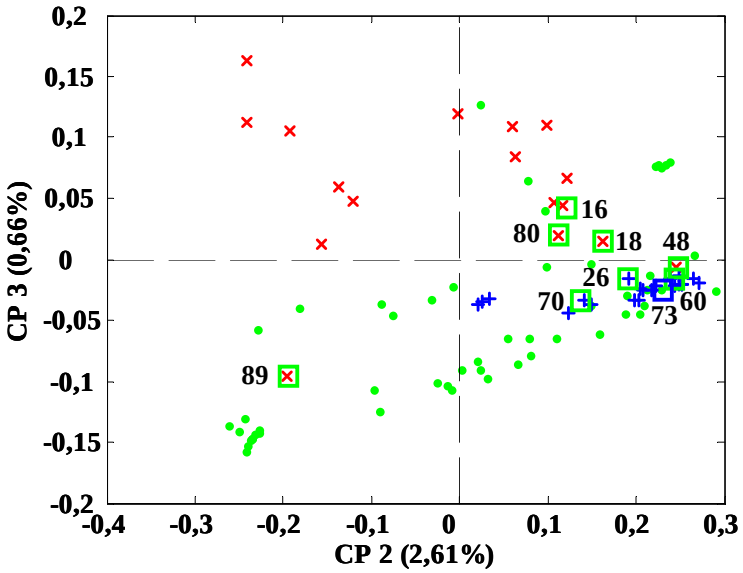


Figure 62 : Echantillons du lot de prédiction W_P projetés dans le plan (CP2-CP3) de la PCA. Echantillons de la classe P1(x), P2 (●), P3 (+). Les échantillons entourés d'un carré correspondent aux échantillons mal prédits, la couleur du carré indique la classe prédite. Les numéros correspondent aux numéros des échantillons du lot de prédiction.

- ***Conclusion***

Ce chapitre a montré le potentiel d'une instrumentation spectroscopique simplifiée pour la classification d'échantillons textiles dans trois catégories associées à une propriété physique d'intérêt. Le prototype mis au point en parallèle présente des caractéristiques métrologiques acceptables en terme de répétabilité, reproductibilité, nous laissant envisager une analyse qualitative satisfaisante.

Les capacités prédictives des modèles de classification obtenues à partir des spectres acquis sur le prototype, pour un prétraitement donné, sont acceptables compte tenu du nombre restreint de variables spectroscopiques.

Conclusion

Les thématiques abordées dans ce manuscrit sont variées : chimiométrie, statistiques, chimie analytique, spectroscopie et instrumentation. Concernant les aspects de nos recherches en chimiométrie, nous nous sommes focalisés, d'une part, sur l'identification et la sélection de l'information pertinente, la sélection de variables spectroscopiques, et, d'autre part, sur la construction de modèles de classification performants de type *support vector machine* (SVM), gardant à l'esprit que les deux aspects sont corrélés. En étalonnage multivarié, la première démarche est d'identifier les variables significatives du lot de données et de ne conserver que celles-ci. En général, seules certaines combinaisons de variables sont nécessaires pour résoudre le problème. La sélection de variables permet dans un premier temps de réduire les redondances des variables d'origine, puis d'améliorer les performances des capacités prédictives des modèles construits, et enfin d'interpréter les variables sélectionnées d'un point de vue de l'information chimique contenue dans les données spectrales. Cependant, trouver la combinaison de variables spectroscopiques optimales est une tâche difficile. Tout d'abord compte tenu de la dimension des données spectroscopiques, il n'est pas envisageable de tester de manière exhaustive toutes les combinaisons de variables. Nous avons donc eu recourt à des algorithmes stochastiques du type des algorithmes génétiques qui réalisent une recherche aléatoire et globale dans un espace de grande dimension. Il se peut également que compte tenu de la précision du résultat, plusieurs combinaisons de variables conduisent à des performances en prédiction qui ne sont pas significativement différentes d'un point de vue statistique. Une autre approche, basée sur la mesure de la pertinence des variables spectroscopiques par information mutuelle, a été mise en place. L'avantage de cette méthode est que l'estimation est réalisée indépendamment du choix d'un modèle de régression. D'autre part, nous avons montré que d'un point vue du nombre de variables sélectionnées, l'information mutuelle est la procédure la plus adaptée. Il faut noter que les méthodes présentées et les résultats obtenus peuvent être généralisés à d'autres domaines de la chimie où le nombre de variables ou d'échantillons peut être encore bien plus important[(113)]. Ces méthodes de sélection de variables ont été mises en oeuvre pour la détermination de la teneur en coton dans les mélanges textiles coton/PES et coton/viscose. Dans le premier cas,

l'information mutuelle a sélectionné 8 variables sur les 806 variables spectroscopiques et pour le lot de données coton/viscose, 12 variables sur les 480 variables spectroscopiques sont conservées. Les capacités prédictives obtenues pour les modèles construits sur ces variables sélectionnées ne sont pas dégradées par rapport à celles obtenues sur les spectres complets. Des résultats significatifs ont été obtenus pour le lot de données coton/PES puisque l'erreur de prédiction de la teneur en coton a été réduite de 2,93% sur les spectres complets à 2,53% pour le modèle construit sur les 8 variables.

D'autre part, concernant le problème de classification des données textiles en fonction d'une propriété physico-chimique d'intérêt, la méthode des *support vector machine* s'est avérée la plus performante. En effet, elle a permis de prendre en compte les multiples challenges liés au problème de classification de ces données en terme de nombre de classes, de distribution, de complexité et de variabilité des échantillons analysés. Les principes de cette méthode ont été détaillés dans ce manuscrit. Lorsque les données ne sont pas séparables linéairement, le calcul de la frontière nécessite la transformation de l'espace des données de départ en un nouvel espace de plus grande dimension dans lequel les séparateurs linéaires pourront être déterminés. La construction des modèles SVM requiert alors l'optimisation de deux paramètres, un paramètre de régularisation de la fonction de coût et le paramètre lié à la fonction noyau utilisée. Les modèles SVM reposent sur la sélection de l'information significative mais l'approche concerne cette fois-ci l'identification des échantillons les plus représentatifs. Ces échantillons, appelés vecteurs supports, déterminent la position du séparateur. Les capacités prédictives obtenues sur les spectres complets de laboratoire sont de 90,5% (± 2,2%) en tenant compte du fait que la réduction du nombre de variables spectroscopiques ne permet plus d'utiliser certains prétraitements. Ayant montré une faisabilité satisfaisante pour la classification de données textiles sur les spectres complets et sur un domaine spectral réduit à 27 variables par un échantillonage en longueurs d'onde, la décision a pu être prise de tendre vers l'utilisation d'une instrumentation simplifiée, c'est-à-dire concrétement la réalisation d'un prototype. La technologie mise en place par les partenaires industriels est développée sur une gamme spectrale plus étroite échantillonnée par 13 variables. Les caractéristiques métrologiques de cette instrumentation étant acceptables, des modèles de classification ont été développés sur ces spectres proche infrarouge discrets. Les résultats montrent une bonne tenue des modèles à la réduction du nombre de variables spectroscopiques puisque les capacités prédictives obtenues sont de 88,1% (± 3,1%). Enfin, nous retiendrons que l'étude laisse entrevoir la possibilité de transférer l'analyse et ces modèles sur des capteurs pour une application rapide. La technologie du prototype pouvait

être miniaturisée mais à des coûts de revient encore trop élevés à l'heure actuelle par rapport au cahier des charges défini, ce qui est un frein pour une application grand public.

ANNEXE 1 : Analyse en composantes principales

L'analyse en composantes principales[114,115] (PCA) est une approche très utilisée pour l'étude exploratoire des données spectrales [116]. Elle calcule de nouvelles variables, appelées composantes principales (CP) qui sont des combinaisons linéaires des variables de départ correspondant aux directions de plus grande variance. Ces CP sont mutuellement orthogonales. Elle projette le nuage de points dans un espace de représentation de faible dimension. Puisque l'objectif de l'analyse est la simplification, il faut choisir un compromis entre deux objectifs contradictoires : prendre un espace de faible dimension et conserver une variance maximale. La PCA permet une condensation et une visualisation rapide des données.

Les composantes principales sont les vecteurs propres de la matrice de variance-covariance centrée XX^T. La matrice de variance-covariance est diagonalisée et les données sont projetées sur ces vecteurs propres. Pour des variables centrées, on peut approcher la matrice des données spectrales $\boldsymbol{X}$ par l'Équation 44 :

$$\text{Équation 44} \quad \boldsymbol{X} = T_A P_A^T + E$$

où $\boldsymbol{X}$ est la matrice des données spectrales (nxm), T_A est la matrice des coordonnées factorielles (nxA), P_A la matrice des vecteurs propres (mxA), E la matrice des résidus (nxm) et A le nombre de composantes principales retenues.

Pour chaque composante principale, T_A et P_A sont calculées à partir des opérations suivantes :

$$\text{Équation 45} \quad P_A^T = \left(T_A^T T_A\right)^{-1} T_A^T \boldsymbol{X}$$

$$\text{Équation 46} \quad T_A = \boldsymbol{X} P_A \left(P_A^T P_A\right)^{-1} = \boldsymbol{X} P_A$$

ANNEXE 2 : Test de Hotelling et test statistique Q

Le test de Hotelling et le test statistique Q permettent d'identifier les échantillons aberrants présents dans le lot de données.

Le test T^2 de Hotelling correspond à une mesure de la variation d'un échantillon à l'intérieur d'un modèle PCA, pour le $i^{ème}$ échantillon de la matrice $\boldsymbol{X}$, la statistique est calculée à partir de l'Équation 47.

$$\text{Équation 47} \quad T_i^2 = t_i \left(\frac{T_A^T T_A}{n-1} \right)^{-1} t_i^T$$

où t_i est la $i^{ème}$ ligne de la matrice des coordonnées factorielles T_A. La valeur élevée de T_i^2 pour le $i^{ème}$ échantillon indique sa forte influence sur la construction du modèle.

La statistique Q pour l'échantillon i est calculée à partir de la matrice des résidus E comme l'Équation 48.

$$\text{Équation 48} \quad Q_i = e_i e_i^T$$

Le paramètre Q indique le manque d'ajustement des données au modèle PCA. En effet, le test statistique Q est une mesure de la différence entre un échantillon et de sa projection dans le modèle PCA. Une valeur élevée de Q indique les échantillons extrêmes qui ne sont pas bien modélisés. La valeur limite de Q est calculée de façon à avoir intervalle de confiance au seuil α=5%[(86)].

ANNEXE 3 : Prétraitements

- ***Dérivation***

Le but de la dérivation est d'augmenter numériquement la résolution apparente des spectres. En effet, elle permet une meilleure séparation des bandes d'absorption et une correction des variations de la ligne de base[(101)]. L'algorithme de Savitzky-Golay[(84)] est très utilisé car il réalise un lissage avant le calcul de la dérivée. Il s'agit d'une méthode basée sur la moyenne mobile. En effet, une fenêtre de *j* variables est sélectionnée. Ces *j* variables sont ensuite ajustées par un polynôme de degré *d* sur au moins *d*+1 points du spectre autour du point *j*. Le point central de la fenêtre des *j* variables est remplacé par la valeur du polynôme. La dérivée en ce point est calculée. Par exemple, dans la littérature de la spectroscopie proche infrarouge, la largeur de la fenêtre varie pour une dérivée seconde entre 13 et 17 longueurs d'onde [(98,117,118)], typiquement le degré du polynôme est souvent égal à 2[(119)] et la dérivée est généralement d'ordre 1 ou 2.

- ***Déviation normale standardisée***

La déviation normale standardisée (SNV)[(97)] élimine les différences multiples dues aux effets de diffusion et à la taille des particules[(98)]. Considérons un spectre d'absorbance constitué de *m* longueurs d'onde. L'absorbance à une longueur d'onde *j* est notée a_j. La moyenne des absorbances est notée $\bar{a}$. Celle-ci est retranchée à l'absorbance pour chaque longueur d'onde *j*. C'est l'étape de centrage des données. On norme finalement par la déviation standard *s* des absorbances du spectre *J* pour obtenir l'absorbance corrigée à la longueur d'onde *j* notée x^c_j :

Équation 49 $$x_j^c = \left(a_j - \bar{a}\right) / \sqrt{\frac{\sum_{j=1}^{m}\left(a_j - \bar{a}\right)^2}{m-1}} \quad \text{pour } j=1\ldots m.$$

- ***Méthode « Det-trending »***

La méthode « De-trending »[97] est une méthode pour la correction de la ligne de base. Elle élimine la courbure de la ligne de base des spectres. La ligne de base est modélisée comme une fonction du nombre de longueurs d'onde par un polynôme de degré *d*, typiquement du second degré qui est soustrait au spectre de départ.

$$\text{Équation 50} \quad x_{ij}^{c} = x_{ij} - b_{ij} ,$$

où x^{c}_{ij} correspond au spectre corrigé, x_{ij} au spectre de départ et b_{ij} est la valeur du polynôme à la longueur d'onde *j* du spectre *i*.

ANNEXE 4 : Répartition des échantillons

- ***Sélection aléatoire***

La sélection aléatoire des échantillons est la technique la plus simple. Elle suppose qu'un groupe d'échantillons extrait aléatoirement à partir d'un lot suffisamment grand suit la distribution statistique du lot entier. Cependant, il peut y avoir un risque que certaines classes d'échantillons ne soient pas représentées dans le lot d'entraînement. Afin d'éviter ce risque, Wu *et al*[120] sélectionnent aléatoirement les échantillons en procédant classe par classe. Les trois quart des objets de chaque classe sont attribués au lot d'entraînement et les échantillons restants constituent le lot de prédiction.

- ***Algorithme de Kennard et Stone***

Une alternative à la sélection aléatoire est l'utilisation de l'algorithme de Kennard et Stone[94,121]. L'algorithme maximise la distance euclidienne minimale entre les échantillons déjà sélectionnés et les échantillons restants.

La procédure est décrite ci-dessous et est illustrée sur la Figure 63 :

– a) sélection des échantillons les plus éloignés. Il s'agit ici des échantillons n°1 et 2 qui sont entourés sur la Figure 63a ;
– b) pour chaque échantillon restant, calcul de la distance euclidienne par rapport à l'échantillon le plus proche déjà sélectionné (Figure 63b) ;
– c) sélection de l'échantillon ayant la plus grande distance avec l'échantillon déjà sélectionné. Le troisième échantillon sélectionné est l'échantillon n°4.

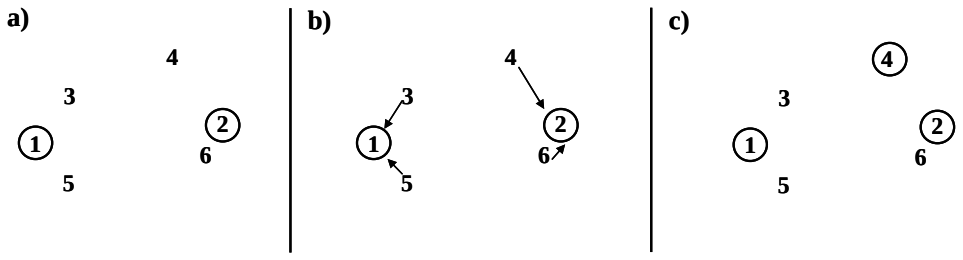


Figure 63 : Répartition des échantillons avec l'algorithme de Kennard et Stone[12]

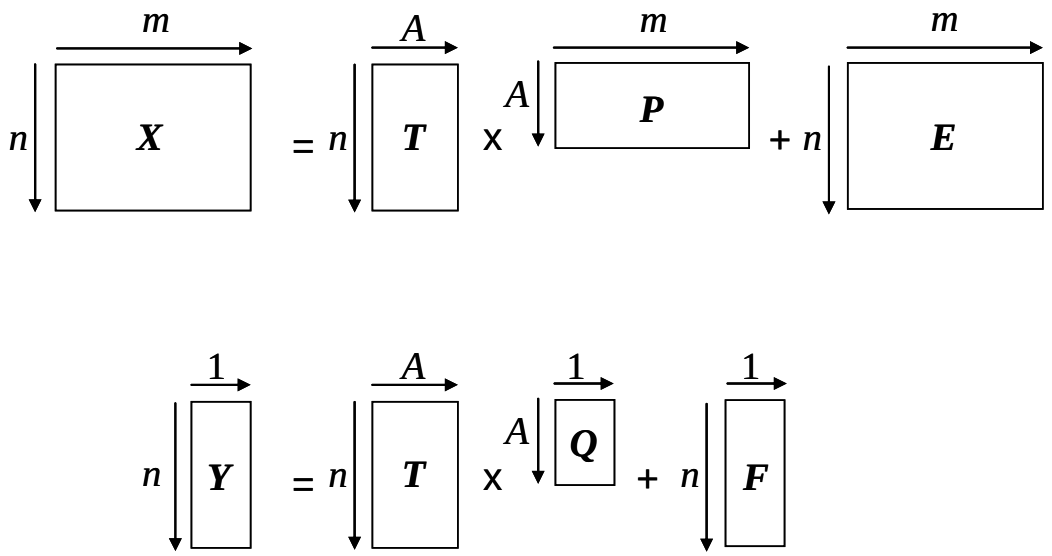
Cette procédure est répétée jusqu'à ce que le nombre d'échantillons spécifié par l'utilisateur soit atteint. Cependant, le nombre d'échantillons à inclure dans chaque lot n'est pas le problème fondamental. En effet, il est plus important de savoir qu'elles sont les variabilités présentes dans le lot de données et comment sélectionner les échantillons du lot d'entraînement afin de prendre en compte ces variabilités[122].

Dans certains cas, les résultats obtenus en classification sont meilleurs avec la distribution des spectres réalisée par l'algorithme de Kennard et Stone qu'avec une répartition aléatoire[120].

ANNEXE 5 : Méthode de régression des moindres carrés partiels (PLS)

Il existe de nombreuses versions de l'algorithme de régression PLS. Elles diffèrent au niveau des normalisations et des calculs intermédiaires, mais elles aboutissent toutes à la même régression. Parmi ces versions, on distingue deux algorithmes, l'algorithme NIPALS initialement proposé par Wold et Martens en 1984[21] et l'algorithme SIMPLS initialisé par De Jong en 1993[123]. Il existe une distinction dans l'application de la méthode de régression des moindres carrés partiels (PLS). Il faut séparer le cas où il y a une seule variable *Y* à prédire de celui où il y en a plusieurs. Dans le premier cas, on parle de régression PLS univariée (PLS1) et dans le second cas de régression PLS multivariée (PLS2)[124]. Dans le cadre de notre étude, nous avons utilisé principalement la régression PLS univariée et c'est pour cela que nous présenterons seulement cette méthode.

Dans cette procédure, on cherche à réaliser une régression d'une variable à prédire *Y* sur des variables $X_1, \ldots, X_m$, qui peuvent être hautement corrélées entre elles. En effet, il s'agit d'une régression de la variable *Y* sur les variables $t_1, t_2, \ldots t_A$, (avec $A<m$) qui sont les variables latentes. Les variables latentes sont des combinaisons linéaires de $X_1, \ldots, X_m$. A la différence de l'analyse en composantes principales que nous avons décrite en annexe 1, les variables latentes sont déterminées en tenant compte des variables prédictives *Y*. Au lieu de modéliser exclusivement les variables *X*, la matrice des spectres $\boldsymbol{X}$ et la matrice de la variable à prédire *Y* sont obtenues à partir des équations décrites dans la Figure 64.



X : la matrice des échantillons

Y : la matrice de la propriété à prédire

T : la matrice des coordonnées factorielles (« scores »)

P : la matrice des vecteurs propres (« loadings »)

Q : la matrice des poids de Y

E : la matrice des résidus de X

F : la matrice des résidus de Y

m : le nombre de variables

n : le nombre d'échantillons

A : le nombre de variables latentes

Figure 64 : Équations de la méthode PLS schématisées sous forme de matrices.

Un paramètre important pour toutes les méthodes factorielles est le nombre de variables latentes à prendre en compte dans la construction du modèle. Ce nombre dépend fortement des données. Plus le nombre de variables latentes augmente, plus le modèle va prendre en compte toute la variance des données observées. Cependant, cette variance n'étant pas forcément significative pour le problème considéré *Y*, un nombre de variables latentes trop important engendre un sur-entraînement du modèle dans la phase d'entraînement et ceci au détriment des prédictions futures. Le choix du nombre de variables latentes est déterminé

par validation croisée. Il est alors possible de tracer la variation de l'erreur de validation en fonction du nombre de variables latentes utilisées pour la construction du modèle. Le nombre optimal de variables latentes est donné par le minimum de la courbe de la validation croisée. Une fois le choix réalisé, le modèle PLS est construit puis appliqué aux échantillons du lot de prédiction.

ANNEXE 6 : Réseaux de neurones artificiels (ANN)

Un réseau de neurones artificiels[125] (ANN) est une méthode dont la conception est très schématiquement inspirée du fonctionnement de neurones biologiques (humains ou non). Les signaux d'entrée passent dans un neurone, où la valeur des poids associée aux signaux d'entrée est moyennée puis transformée par une fonction de transfert en un signal de sortie. La propagation du signal est déterminée par les connexions entre les neurones et leur poids associé. Différents types de réseaux de neurones ont été développés comme les réseaux de Kohonen, les réseaux multicouches MLF (*multilayer feed forward*)… La forme des connexions est un facteur important puisqu'il détermine la circulation de l'information à travers le réseau. Les poids associés à chaque connexion sont également essentiels puisqu'ils établissent la propagation du signal à travers le réseau. Il s'agit de nombres réels qui déterminent la *qualité* de la connexion entre deux neurones. Chaque signal qui est transporté à travers la connexion est multiplié par le poids associé à la connexion. La valeur des poids peut être positive ou négative. Les neurones sont ordonnés en couches. Il existe trois types de couches, la couche d'entrée, la couche cachée et la couche de sortie Figure 65a. Tous les neurones d'une couche sont connectés à tous les neurones de la couche suivante. Le réseau reçoit les signaux à travers la couche d'entrée. L'information est ensuite propagée aux couches cachées puis à la couche de sortie qui produit la réponse du réseau. Le nombre de neurones de la couche d'entrée est égal aux nombres de variables spectroscopiques, m, de la matrice des données spectrales $\boldsymbol{X}$ (nxm), si le nombre de variables spectroscopiques est inférieur au nombre d'échantillons. Si ce n'est pas le cas, en générale, on passe par soit une sélection de variables telle que l'information mutuelle soit par une PCA. Le nombre de neurones de la couche de sortie correspond aux nombre de variables, q, de la matrice des valeurs cibles Y (nxq). La propagation du signal à travers le réseau est similaire à celle expliquée pour un neurone. Chaque neurone de la couche cachée reçoit les signaux des neurones de la couche précédente, c'est-à-dire pour la première couche cachée, la couche d'entrée. La somme des signaux associés aux poids est calculée puis transformée par une fonction de transfert en une valeur de signal de sortie. La fonction de transfert la plus couramment utilisée est la fonction sigmoïde. Cette étape est répétée pour les neurones de la couche de sortie.

Afin de déterminer la valeur des poids optimale, une phase d'apprentissage est réalisée. L'algorithme d'apprentissage le plus utilisée est la rétro-propagation (*back-propagation*) Figure 65. L'actualisation de la valeur des poids est basée sur la différence entre la valeur de sortie actuelle et la valeur cible. La procédure est réalisée en 4 étapes :

1. Initialisation des poids.
2. Calcul de la valeur de sortie pour chaque neurone avec la valeur du poids d'origine et de l'erreur E basée sur la différence entre cette valeur et la valeur de sortie cible.
3. Application des poids d'adaptation aux neurones de sortie. Dans la stratégie de la rétro-propagation, l'adaptation des poids est réalisée afin de minimiser l'erreur.
4. Calcul de la valeur d'adaptation des poids dans les couches cachées. La valeur de sortie désirée et de l'erreur n'est pas connue directement. La valeur des poids est calculée à partir des erreurs des neurones de sortie.

Sur la Figure 65, la couche de sortie contient un seul neurone dont la réponse correspond à la valeur de $y_{prédit}$ qui peut être exprimée par l'équation suivante :

Équation 51 $$y_{predit} = f_0\left[\vartheta'' + \sum_{j=1}^{nc} w_j'' f_h\left(\sum_{i=1}^{m} w_{ij}' x_i + \vartheta'\right)\right]$$

où *nc* est le nombre de neurones cachés, *m* le nombre de variables, w_{ij}', w_j'' sont les poids et θ' et θ'' sont les biais. Les valeurs des poids et des biais sont déterminés par une procédure d'entraînement itérative. La procédure complète est répétée jusqu'à ce que le taux de convergence recherché soit atteint. L'algorithme d'apprentissage utilisé dans notre étude est l'algorithme mis au point par Riedmiller[90]. Il s'agit de l'algorithme Rprop pour *Resilient propagation* en anglais. Il existe divers avantages à utiliser cet algorithme notamment le fait qu'il ne se base pas sur les valeurs des dérivées partielles de E mais sur leurs signes.

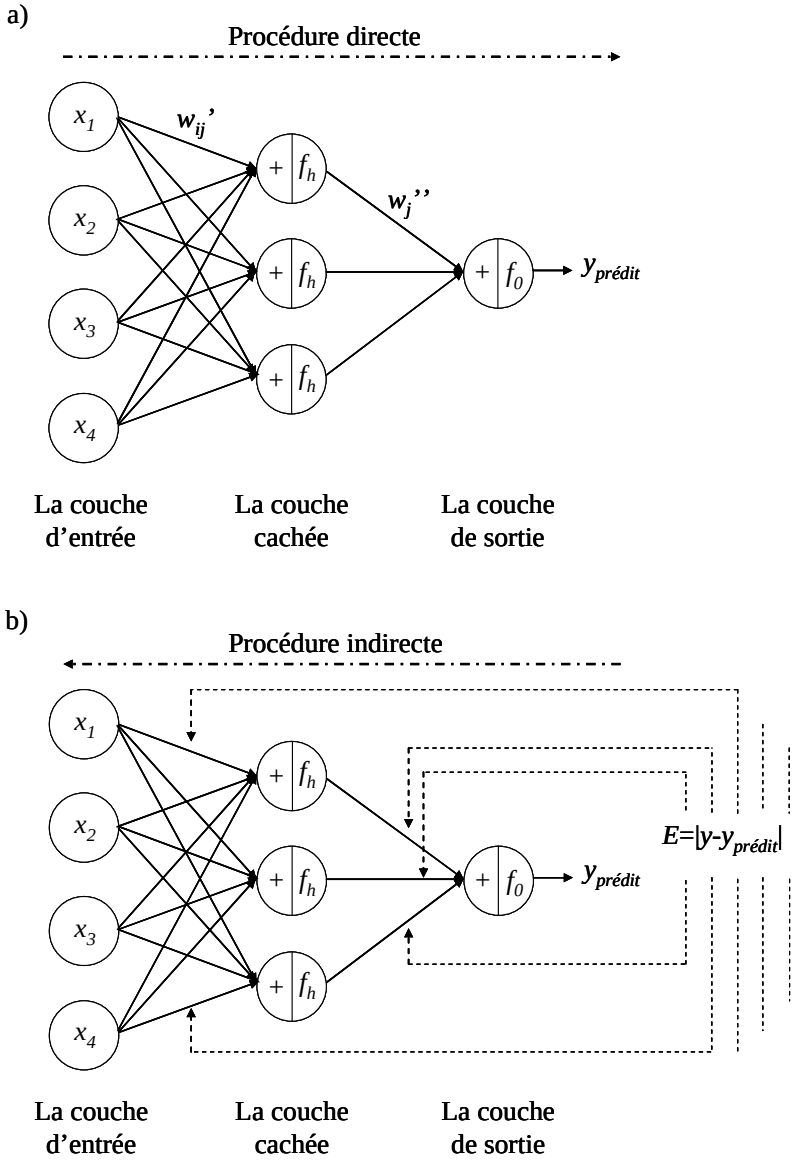


Figure 65 : Principe des réseaux de neurones à rétro-propagation de l'erreur. a) la procédure directe, b) la rétro-propagation[(126)].

ANNEXE 7 : Communications scientifiques

- ***Publications internationales à comité de lecture***

1. Quantitative analysis of cotton-polyester textile blends from near-infrared spectra.

C. Ruckebusch, F. Orhan, A.Durand, T. Boubellouta, J-P. Huvenne,

Applied Spectroscopy, **2006,** 60, 539-544.

2. Simultaneous dielectric and FT-NIR spectroscopy to monitor a polyepoxy curing process.

A. Durand, L. Hassi, G. Lachenal, I. Stevenson, G. Seytre, G. Boiteux, *Journal of Near Infrared Spectroscopy*, **2006**, 14, (3), 161-166.

3. Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton-viscose textiles.

A. Durand, O. Devos, C. Ruckebusch, J-P. Huvenne.

Analytica Chimica Acta, **2007**, 595, 72-79.

- ***Communications orales***

4. NIR quantitative analysis of textile blends : Focus on the selection of relevant variables.

A. Durand, O. Devos, C. Ruckebusch, J-P. Huvenne.

10th International Conference of Chemometrics in Analytical Chemistry (CAC 2006), Águas de Lindóia, Brésil, **2006**.

- ***Communications par affiches***

5. Influence of variable selection on multivariate models in NIR spectroscopy.

A.Durand, T. Boubellouta, C. Ruckebusch, J-P. Huvenne

32th Federation of Analytical Chemistry and Spectroscopy Societies (FACSS 2005), Montréal, Canada, **2005**.

6. Application d'algorithmes génétiques pour la sélection de variables en spectroscopie proche-infrarouge.

A. Durand, C. Ruckebusch, J-P. Huvenne.

Chimiométrie 2005, Lille, **2005**.

7. Sélection de variables par estimation de l'information mutuelle en calibration multivariée.

O. Devos, A. Durand, C. Ruckebusch, J-P. Huvenne.

Chimiométrie 2006, Paris, **2006**.

Liste des tableaux

Liste des figures

Liste des annexes

Bibliographie

1. Malley, D. F.; Hunter, K. N.; Webster, G. R. B., *Soil and Sediment Contamination* **1999**, 8, (4), 481-489.

2. Blanco, M.; Gozalez Bano, R.; Bertran, E., *Talanta* **2002**, 56, (1), 203-212.

3. Sohn, M.; Himmelsbach, D. S.; Akin, D. E.; Barton Ii, F. E., *Textile Research Journal* **2005**, 75, (8), 583-590.

4. Huck, C. W.; Guggenbichler, W.; Bonn, G. K., *Analytica Chimica Acta* **2005**, 538, (1-2), 195-203.

5. Vieira, R. A. M.; Sayer, C.; Lima, E. L.; Pinto, J. C., *Journal of Applied Polymer Science* **2002**, 84, (14), 2670-2682.

6. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier Science ed.; Amsterdam, **1997**; Vol. 20A.

7. Ilari, J. L.; Martens, H.; Isaksson, T., *Applied Spectroscopy* **1988**, 42, (7), 722-728.

8. Van Agthoven, M. A.; Fujisawa, G.; Rabbito, P.; Mullins, O. C., *Applied Spectroscopy* **2002**, 56, (5), 593-598.

9. Miller, C. E.; Svendsen, S. A.; Næs, T., *Applied Spectroscopy* **1993**, 47, (3), 346-356.

10. Bangalore, A. S.; Shaffer, R. E.; Small, G. W.; Arnold, M. A., *Analytical Chemistry* **1996**, 68, (23), 4200-4212.

11. Guyon, I.; Elisseeff, A., *Journal of Machine Learning Research* **2003**, 3, 1157-1182.

12. de Groot, P. J.; Postma, G. J.; Melssen, W. J.; Buydens, L. M. C., *Analytica Chimica Acta* **1999**, 392, (1), 67-75.

13. Esteban-Diez, I.; Gonzalez-Saiz, J. M.; Saenz-Gonzalez, C.; Pizarro, C., *Talanta* **2007**, 71, (1), 221-229.

14. Roggo, Y.; Roeseler, C.; Ulmschneider, M., *Journal of Pharmaceutical and Biomedical Analysis* **2004**, 36, (4), 777-786.

15. Van Den Broek, W. H. A. M.; Derks, E. P. P. A.; Van De Ven, E. W.; Wienke, D.; Geladi, P.; Buydens, L. M. C., *Chemometrics and Intelligent Laboratory Systems* **1996**, 35, (2), 187-197.

16. Sohn, M.; Himmelsbach, D. S.; Morrison Iii, W. H.; Akin, D. E.; Barton Ii, F. E., *Applied Spectroscopy* **2006**, 60, (4), 437-440.

17. Czekalski, J.; Patejuk-Duda, A., *Journal of Natural Fibers* **2004**, 1, (2), 25-35.

18. Fisher, G., *International Fiber Journal* **2005**, 20, (6), 14-21.

19. Langeron, Y.; Doussot, M.; Hewson, D. J.; Duchene, J., *Engineering Applications of Artificial Intelligence* **2007**, 20, (3), 415-427.

20. Neelamegam, P.; Rajendran, A., *Instrumentation Science & Technology* **2003**, 31, (4), 417-423.

21. Sjostrom, M.; Wold, S.; Lindberg, W.; Persson, J.-A.; Martens, H., *Analytica Chimica Acta* **1983**, 150, 61-70.

22. Maraboli, A.; Cattaneo, T. M. P.; Giangiacomo, R., *Journal of Near Infrared Spectroscopy* **2002**, 10, (1), 63-69.

23. Li, X. L.; He, Y.; Qiu, Z. J., *Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy and Spectral Analysis* **2007**, 27, (2), 279-282.

24. Zhao, J.; Chen, Q.; Huang, X.; Fang, C. H., *Journal of Pharmaceutical and Biomedical Analysis* **2006**, 41, (4), 1198-1204.

25. Sinnaeve, G.; Dardenne, P.; Agneessens, R.; Lateur, M.; Hallet, A., *Journal of Near Infrared Spectroscopy* **1997**, 5, (1), 1-17.

26. Woo, Y. A.; Kim, H. J.; Chung, H., *Analyst* **1999**, 124, (8), 1223-1226.

27. Pasti, L.; Jouan-Rimbaud, D.; Massart, D. L.; Noord, O. E. D., *Analytica Chimica Acta* **1998**, 364, (1-3), 253-263.

28. Wold, S.; Sjostrom, M.; Eriksson, L., *Chemometrics and Intelligent Laboratory Systems* **2001**, 58, (2), 109-130.

29. Gemperline, P. J.; Long, J. R.; Gregoriou, V. G., *Analytical Chemistry* **1991**, 63, (20), 2313-2323.

30. Poppi, R. J.; Massart, D. L., *Analytica Chimica Acta* **1998**, 375, (1-2), 187-195.

31. Balabin, R. M.; Safieva, R. Z., *Fuel* In Press, Corrected Proof.

32. Candolfi, A.; Wu, W.; Massart, D. L.; Heuerding, S., *Journal of Pharmaceutical and Biomedical Analysis* **1998**, 16, (8), 1329-1347.

33. Indahl, U. G.; Sahni, N. S.; Kirkhus, B.; Naes, T., *Chemometrics and Intelligent Laboratory Systems* **1999**, 49, (1), 19-31.

34. Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A., *Classification and Regression Trees*. Wadsworth International Group: Californie, **1984**.

35. Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J., *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier Science B.V.: Amsterdam, **1998**; Vol. 20B.

36. Kohonen, T., *Self-Organizing Maps*. second edition ed.; Springer-Verlag: Berlin, **1997**.

37. Fisher, R. A., *Annals of Eugenics* **1936**, 7, 179-188.

38. Gemperline, P. J.; Laurie, D.; Webber, F.; Cox, O., *Analytical Chemistry* **1989**, 61, (2), 138-144.

39. Burges, C. J. C., *Data Mining and Knowledge Discovery* **1998**, 2, (2), 121-167.

40. Cornuéjols, A., *Bulletin de l'AFIA* **2002**, 51.

41. Frezza-Buet, H. *Machines à Vecteurs Supports*; **2006**.

42. Webb, A., *Statistical Pattern Recognition*. John Wiley ed.; **2002**.

43. Forina, M.; Lanteri, S.; Cerrato Oliveros, M. C.; Pizarro Millan, C., *Analytical and Bioanalytical Chemistry* **2004**, 380, (3), 397-418.

44. Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B., *Applied Spectroscopy* **2000**, 54, (3), 413-419.

45. Abrahamsson, C.; Johansson, J.; Sparen, A.; Lindgren, F., *Chemometrics and Intelligent Laboratory Systems* **2003**, 69, (1-2), 3-12.

46. Despagne, F.; Massart, D.-L., *Chemometrics and Intelligent Laboratory Systems* **1998**, 40, (2), 145-163.

47. Kojadinovic, I., *Computational Statistics & Data Analysis* **2005**, 49, (4), 1205-1227.

48. Alsberg, B. K.; Woodward, A. M.; Winson, M. K.; Rowland, J. J.; Kell, D. B., *Analytica Chimica Acta* **1998**, 368, (1-2), 29-44.

49. Rossi, F.; Lendasse, A.; Francois, D.; Wertz, V.; Verleysen, M., *Chemometrics and Intelligent Laboratory Systems* **2006**, 80, (2), 215-226.

50. Lucasius, C. B.; Kateman, G., *TrAC Trends in Analytical Chemistry* **1991**, 10, (8), 254-261.

51. Jouan-Rimbaud, D.; Massart, D.-L.; Leardi, R.; De Noord, O. E., *Analytical Chemistry* **1995**, 67, (23), 4295-4301.

52. Leardi, R.; Gonzalez, A. L., *Chemometrics and Intelligent Laboratory Systems* **1998**, 41, (2), 195-207.

53. Smith, B. M.; Gemperline, P. J., *Analytica Chimica Acta* **2000**, 423, (2), 167-177.

54. Holland, J. H., *Adaptation in natural and artificial system: an introductory analysis with applications to biology, control, and artificial intelligence*. The University of Michigan Press: **1975**.

55. Lucasius, C. B.; Blommers, M. J. J.; Buydens, L. M. C.; Kateman, G., *Handbook of Genetic Algorithms*. Van Nostrand Reinhold Company: New York, **1991**.

56. Jouan-Rimbaud, D.; Massart, D. L.; de Noord, O. E., *Chemometrics and Intelligent Laboratory Systems* **1996**, 35, (2), 213-220.

57. Goicoechea, H. C.; Olivieri, A. C., *Journal of Chemometrics* **2003**, 17, (6), 338-345.

58. Lestander, T. A.; Leardi, R.; Geladi, P., *Journal of Near Infrared Spectroscopy* **2003**, 11, (6), 433-446.

59. Godon, M. *La théorie de l'évolution*. **2003**.

60. Pontes, M. J. C.; Galvao, R. K. H.; Araujo, M. C. U.; Moreira, P. N. T.; Neto, O. D. P.; Jose, G. E.; Saldanha, T. C. B., *Chemometrics and Intelligent Laboratory Systems* **2005**, 78, (1-2), 11-18.

61. Wise, B. W.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S. *PLS_Toolbox Version 3.5*, 3.5; Eigenvector Research: Manson, **2004**.

62. Benoudjit, N.; Francois, D.; Meurens, M.; Verleysen, M., *Chemometrics and Intelligent Laboratory Systems* **2004**, 74, (2), 243-251.

63. Long, J. R.; Gregoriou, V. G.; Gemperline, P. J., *Analytical Chemistry* **1990**, 62, (17), 1791-1797.

64. Marengo, E.; Bobba, M.; Robotti, E.; Lenti, M., *Analytica Chimica Acta* **2004**, 511, (2), 313-322.

65. Shannon, C. E. W. W., *The mathematical Theory of Communication*. Urbana IL, **1949**.

66. Kraskov, A.; Stögbauer, H.; Grassberger, P., *Physical Review E* **2004**, 69, 1-16.

67. Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning*. Springer: New York, **2001**.

68. Stögbauer, H.; Kraskov, A.; Astakhov, S.; Grassberger, P., *Physical Review E* **2004**, 70, 1-17.

69. Fernandez Pierna, J. A.; Baeten, V.; Renier, A. M.; Cogdill, R. P.; Dardenne, P., *Journal of Chemometrics* **2004**, 18, (7-8), 341-349.

70. Roggo, Y.; Chalus, P.; Maurer, L.; Lema-Martinez, C.; Edmond, A.; Jent, N., *Journal of Pharmaceutical and Biomedical Analysis* In Press, Corrected Proof.

71. de Groot, P. J.; Postma, G. J.; Melssen, W. J.; Buydens, L. M. C., *Analytica Chimica Acta* **2002**, 453, (1), 117-124.

72. Vapnik, V., *The Nature of Statistical Learning Theory*. **1995**.

73. Schölkopf, B.; Smola, A. J., *Learning with Kernels*. London, **2002**.

74. Kuhn, H.; Tucker, A. In *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics*, **1951**; University of California Press: 1951, 481-492.

75. Karush, W. *Minima of Functions of Several Variables with inequalities as Side Constraints*. University of Chicago, Chicago, **1939**.

76. Vert, J. M., *Introduction to support vector machines and applications to computational biology*. In Paris, **2001**.

77. Chen, Q.; Zhao, J.; Fang, C. H.; Wang, D., *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2007**, 66, (3), 568-574.

78. Hsu, C. W.; Chang, C. C.; Lin, C. J., *A practical guide to vector support classification*. In **2007**.

79. Cleve, E.; Bach, E.; Schollmeyer, E., *Analytica Chimica Acta* **2000**, 420, (2), 163-167.

80. Carrillo, F.; Colom, X.; Sunol, J. J.; Saurina, J., *European Polymer Journal* **2004**, 40, (9), 2229-2234.

81. Haensch, K.; Yates, I. *Directive 96/74/CE du Parlement européen et du Conseil du 16 décembre 1996 relative aux dénominations textiles* **1996**.

82. Flandrin-Bletty, M., *Technologie et Chimie des Textiles*. Toulouse, **1996**.

83. Ghosh, S.; Rodgers, J., *Handbook of Near-Infrared Analysis*. Marcel Dekker: New York, **1992**; Vol. 13.

84. Savitzky, A.; Golay, M., *The Perkin-Elmer Corp.* **1964**, 36, (8), 1627-1639.

85. Lillhonga, T.; Geladi, P., *Analytica Chimica Acta* **2005**, 544, (1-2), 177-183.

86. Macho, S.; Sales, F.; Callao, M. P.; Larrechi, M. S.; Rius, F. X., *Applied Spectroscopy* **2001**, 55, (11), 1532-1536.

87. AFNOR, G. *Textiles - Analyse chimique quantitative - Partie 11 : mélanges de fibres de cellulose et de polyester (méthode à l'acide sulfurique)*. **2006**.

88. AFNOR, G. *Textiles - Analyse chimique quantitive - Partie 5 : mélanges de viscose, cupro ou modal et de fibres de coton (méthode au zincate de sodium)*. **2006**.

89. Ruckebusch, C.; Orhan, F.; Durand, A.; Boubellouta, T.; Huvenne, J. P., *Applied Spectroscopy* **2006**, 60, (5), 539-544.

90. Riedmiller, M.; Braun, H., *Proceedings of the IEEE International Conference on Neural Networks* **1993**.

91. Durand, A.; Devos, O.; Ruckebusch, C.; Huvenne, J. P., *Analytica Chimica Acta* **2007**, 595, (1-2), 72-79.

92. Brossard, I., *Technologie des textiles*. Dunod ed.; Paris, **1997**.

93. Siesler, H. W.; Ozaki, Y.; Kawata, S.; Heise, H. M., *Near-Infrared Spectroscopy, Principles, Instruments, Applications*. WILEY-VCH: Weinheim, **2002**.

94. Kennard, R. W.; Stone, L. A., *Technometrics* **1969**, 11, (1), 137-148.

95. Daszykowski, M.; Walczak, B.; Massart, D. L., *Analytica Chimica Acta* **2002**, 468, (1), 91-103.

96. Snee, R. D., *Technometrics* **1977**, 19, 415-428.

97. Barnes, R. J.; Dhanoa, M. S.; Lister, S. J., *Applied Spectroscopy* **1989**, 43, (5), 772-777.

98. Luypaert, J.; Heuerding, S.; Heyden, Y. V.; Massart, D. L., *Journal of Pharmaceutical and Biomedical Analysis* **2004**, 36, (3), 495-503.

99. Davies, A. M. C.; Fearn, T., *Spectroscopy Europe* **2007**, 19, (4), 24-28.

100. Arakaki, L. S. L.; Burns, D. H., *Applied Spectroscopy* **1992**, 46, (12), 1919-1928.

101. Dhanoa, M. S.; Lister, S. J.; Sanderson, R.; Barnes, R. J., *Journal of Near Infrared Spectroscopy* **1994**, 2, (1), 43-47.

102. Ozaki, Y.; Miura, T.; Sakurai, K.; Matsunaga, T., *Applied Spectroscopy* **1992**, 46, (5), 875-878.

103. Levillain, P.; Fompeydie, D., *Analusis* **1986**, 14, (1), 1-20.

104. Osborne, B. G.; Fearn, T.; Hindle, P. H., *Practical NIR Spectroscopy with Applications in Food and Beverage Analysis*. second ed.; Longman Singapore Publishers: Singapore, **1993**.

105. de Boves Harrington, P., *TrAC Trends in Analytical Chemistry* **2006**, 25, (11), 1112-1124.

106. Ramousse, R.; Le Berre, M.; Le Guelte, L. *Introduction aux statistiques*. **1996**.

107. Belousov, A. I.; Verzakov, S. A.; von Frese, J., *Chemometrics and Intelligent Laboratory Systems* **2002**, 64, (1), 15-25.

108. Bertrand, D.; Dufour, E., *La spectroscopie infrarouge et ses applications analytiques*. TEC & DOC: Paris, **2000**.

109. Feinberg, M., *L'assurance qualité dans les laboratoires agroalimentaires et pharmaceutiques*. 2 nd ed.; Editions TEC & DOC: Paris, **2001**.

110. Dupuy, N.; Meurens, M.; Sombret, B.; Legrand, P.; Huvenne, J. P., *Applied Spectroscopy* **1992**, 46, (5), 860-863.

111. Dupuy, N.; Ruckebush, C.; Duponchel, L.; Beurdeley-Saudou, P.; Amram, B.; Huvenne, J. P.; Legrand, P., *Analytica Chimica Acta* **1996**, 335, (1-2), 79-85.

112. Szlyk, E.; Kowalczyk-Marzec, A.; Szydlowska-Czerniak, A., *Chemia Analityczna* **2007**, 52, (2), 307-325.

113. Wasim, M.; Sukri Hassan, M.; Brereton, R. G., *Analyst* **2003**, 128, (8), 1082-1090.

114. Martens, H.; Naes, T., *Multivariate calibration*. John Wiley &Sons: New York, **1989**.

115. Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J., *Journal of Chemometrics* **1987**, 1, (1), 41-56.

116. Daszykowski, M.; Stanimirova, I.; Walczak, B.; Coomans, D., *Chemometrics and Intelligent Laboratory Systems* **2005**, 78, (1-2), 19-29.

117. Pizarro, C.; Esteban-Diez, I.; Nistal, A.-J.; Gonzalez-Saiz, J.-M., *Analytica Chimica Acta* **2004**, 509, (2), 217-227.

118. Candolfi, A.; De Maesschalck, R.; Jouan-Rimbaud, D.; Hailey, P. A.; Massart, D. L., *Journal of Pharmaceutical and Biomedical Analysis* **1999**, 21, (1), 115-132.

119. Blanco, M.; Coello, J.; Montoliu, I.; Romero, M. A., *Analytica Chimica Acta* **2001**, 434, (1), 125-132.

120. Wu, W.; Massart, D. L., *Chemometrics and Intelligent Laboratory Systems* **1996**, 35, (1), 127-135.

121. Dantas Filho, H. A.; Harrop Galvao, R. K.; Ugulino Araujo, M. C.; Da Silva, E. C.; Bezerra Saldanha, T. C.; Jose, G. E.; Pasquini, C.; Raimundo, I. M.; Rodrigues Rohwedder, J. J., *Chemometrics and Intelligent Laboratory Systems* **2004**, 72, (1), 83-91.

122. Fearn, T., *NIR news* **2005**, 16, (7), 17-19.

123. de Jong, S., *Chemometrics and Intelligent Laboratory Systems* **1993**, 18, (3), 251-263.

124. Tenenhaus, M., *La régression PLS, Théorie et pratique*. Editions Technip ed.; Paris, **1998**.

125. Derks, E. P. P. A.; Pastor, M. S. S.; Buydens, L. M. C., *Chemometrics and Intelligent Laboratory Systems* **1995**, 28, (1), 49-60.

126. Despagne, F.; Massart, D. L., *Analyst* **1998**, 123, (11), 157-178.

METHODES DE SELECTION DE VARIABLES APPLIQUEES EN SPECTROSCOPIE PROCHE INFRAROUGE POUR L'ANALYSE ET LA CLASSIFICATION DE TEXTILES

Résumé Les méthodes d'analyse multivariée permettent d'extraire l'information présente dans les données spectroscopiques expérimentales pour la prédiction d'une propriété d'intérêt. La dimensionnalité des données en spectroscopie proche infrarouge est telle qu'une sélection des variables spectroscopiques et d'échantillons est nécessaire afin d'améliorer les performances, la robustesse des modèles ou de tendre vers une instrumentation simplifiée.

L'analyse rapide de la composition chimique des échantillons textiles est fondamentale dans certaines applications. Une première étude concerne la détermination de la teneur en coton dans des mélanges de fibres coton/polyester et coton/viscose par spectroscopie proche infrarouge. Afin d'améliorer les capacités prédictives obtenues sur les spectres complets, deux procédures de sélection de variables, l'information mutuelle et les algorithmes génétiques, ont été appliquées. L'erreur standard de prédiction obtenue pour le lot coton/polyester est de 2,53% sur les 8 variables sélectionnées par l'information mutuelle. Une seconde étude présente l'analyse qualitative pour la classification d'échantillons textiles dans trois classes par rapport à une propriété physico-chimique d'intérêt. La méthode des *support vector machine* présente des résultats performants avec un taux d'échantillons bien classés en prédiction de 88,8%. La réduction arbitraire du nombre de variables spectroscopiques a permis de montrer que les capacités prédictives obtenues sur les spectres complets ne sont pas dégradées. Ces résultats sont confirmés par l'utilisation d'une instrumentation simplifiée.

Mots clés : spectroscopie, chimiométrie, information mutuelle, algorithmes génétiques, support vector machine, textile.

VARIABLE SELECTION PROCEDURES IN NEAR INFRARED QUANTITATIVE ANALYSIS AND CLASSIFICATION OF TEXTILES

Abstract Multivariate analysis methods enable to extract information from spectroscopic data for the prediction of properties of interest. Due to the dimensionality of the data in near infrared spectroscopy, a selection of spectroscopic variables and samples is necessary in order to improve performances, model robustness or to use a simplified instrumentation.

Determining the composition of textile is an essential topic due to the wide range of applications. The first study relates to the determination of the cotton content in cotton/polyester and cotton/viscose blend. In order to improve the predictive capacities obtained on the full spectra, two procedures of variable selection, mutual information and the genetic algorithms were applied. The standard error of prediction obtained for the data set cotton/polyester is 2.53% on the 8 variables selected by mutual information. The second part develops the qualitative analysis for the classification of textile samples in three classes according to the physicochemical property of interest. The method of support vector machines has powerful results with a well classified samples rate of 93.2% in prediction. The arbitrary reduction of the number of spectroscopic variables demonstrates that the predictive capacities obtained on the full spectra are not degraded. These results are confirmed by the use of a simplified instrumentation.

Keywords: spectroscopy, chemometrics, mutual information, genetic algorithm, support vector machine, textile.

Auteur : Alexandra DURAND
Ecole doctorale : Sciences pour l'ingénieur
Discipline : Instrumentation et Analyses Avancées, USTL, 59 655 Villeneuve d'Ascq, France.
Laboratoire : Laboratoire de Spectrochimie Infrarouge et Raman, LASIR, CNRS UMR 8516, USTL, Bât C5, 59 655 Villeneuve d'Ascq, France.