Centre International d'Etudes Supérieures en Sciences Agronomiques de Montpellier - SupAgro

Contribution des informations expérimentales et expertes à l'amélioration des modèles linéaires d'étalonnage multivarié en spectrométrie

THÈSE

présentée et soutenue publiquement le 13 Décembre 2010

pour l'obtention du grade de Docteur en

Sciences des Procédés - Sciences des Aliments

(Spécialité: Biochimie, Chimie, Technologie des Aliments)

par

Jean-Claude Boulet

Composition du jury:

Président :	M. El Mostofa Qannari	Professeur, Oniris, France
Rapporteurs :	M. Achim Kohler	Docteur-Ingénieur, Nofima, Norvège
	M. Douglas N. Rutledge	Professeur, AgroParisTech, France
Examinateur:	M. Nicolas Molinari	Maître de conférences, UM1, France
Directeur de Thèse:	M. Jean-Michel Roger	Professeur consultant, Cemagref, France

UMR1201 Ingénierie et Technologie pour Agro-Procédés **UMR1083 Sciences pour l'Oenologie**

Mis en page avec la classe thloria.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué directement ou indirectement à ce travail.

Claude Flanzy et Michel Moutounet ont été pour moi des exemples à suivre dans l'engagement et la rigueur scientifique.

Guy Albagnac est à l'origine de ma réorientation professionelle en 2005 vers une activité de recherche et a toujours soutenu mon projet de thèse.

Christine Lacroix m'a apporté une méthode plus qu'un enseignement pour progresser en Anglais.

Thierry Doco et Pascale Williams m'ont témoigné leur confiance dès le début. Le travail sur les polysaccharides réalisé ensemble a été le fil d'Ariane conduisant aux principales idées développées dans ce mémoire.

Les membres de mon comité de thèse, Dominique Bertrand, Gérard Mazerolles et Robert Sabatier ont également été un soutien précieux. Bien que nous nous connaissions déjà un peu, j'ai pu découvrir leurs qualités humaines comme leurs qualités scientifiques.

Et j'ai eu la chance d'avoir un formidable directeur de thèse en la personne de Jean-Michel Roger. Il a bien voulu accepter de diriger mon travail en plus de ses autres engagements. Son optimisme sans faille, son ouverture d'esprit et sa patience sont déjà légendaires au Cemagref. Ses choix se se sont toujours révélés excellents. Ce mémoire lui doit beaucoup, gagnant en clarté d'expression et de raisonnement.

Je suis également reconnaissant envers Achim Kohler, Nicolas Molinari, El Mostopha Qannari, Douglas Rutledge, pour avoir bien voulu se pencher avec attention sur ce travail, et avoir fait pour certains un long voyage malgré de nombreuses autres sollicitations.

Enfin une pensée pour les nombreuses personnes de la recherche à l'UMR-SPO, l'UE-Pech Rouge, au CEMAGREF, ou dans la filière Bag-In-Box, avec qui j'ai collaboré avec plaisir sur des sujets pas toujours en lien direct avec le thème de ce mémoire.

A mes parents Guy et Odile.

A mes filles Laurence et Lucie.

A Catherine.

Ce mémoire s'appuie sur les communications suivantes.

Publications :
- J.C.Boulet, T.Doco, J.M.Roger, Improvement of calibration models using two successive orthogonal projection methods. Application to quantification of wine mannoproteins, Chemometrics and Intelligent Laboratory Systems 87 (2007) 295-302
- J.C.Boulet, J.M.Roger, Improvement of Direct Calibration in spectroscopy, Analytica Chimica Acta 668 (2010) 130-136

Présentations orales :
- J.C.Boulet, T.Doco, J.M.Roger, Improvement of calibration models using two successive orthogonal projection methods. Application to quantification of wine mannoproteins, Macromolecules in wine (Macrowine) 2008 Montpellier
- J.C.Boulet, C.Barron, N.Gorretta, J.M.Roger, IDC-Improved Direct Calibration : a new direct calibration method applied to hyperspectral image analysis, IEEE GRSS Workshop on Hyperspectral Image and Signal Processing (WHISPERS) 2009 Grenoble
- J.C.Boulet, J.M.Roger, A new direct calibration method : IDC-Improved Direct Calibration, Chimiometrie 2009 Paris

Posters :
- J.C.Boulet, T.Doco, J.M.Roger, Improvement of calibration models using two successive orthogonal projection methods. Application to quantification of wine mannoproteins, Chemometrics in Analytical Chemistry (CAC) 2008 Montpellier
- J.C.Boulet, J.M.Roger, IDC-Improved Direct Calibration. Application to ethanol quantification in musts and wines, Near Infrared Spectroscopy (NIR) 2009 Bangkok
- J.C.Boulet, D.Bertrand, G.Mazerolles, R.Sabatier, J.M.Roger, VODKA-PLSR, a new family of PLS models based on the NIPALS algorithm, Chemometrics in Analytical Chemistry (CAC) 2010 Anvers

Table des matières

Introduction

1 L'importance croissante de l'analyse en ligne dans le contrôle de procédés 1
2 Bases de la spectroscopie, le modéle linéaire général de mélange 2
3 Plan du mémoire 3

Chapitre 1
Place des informations expérimentales et expertes dans les modèles d'étalonnage

1.1 Définitions, champ d'application 6
1.1.1 Informations expérimentales, informations expertes 6
1.1.2 Méthodes d'étalonnage, méthodes de prétraitement 6
1.1.3 Champ d'application 7
1.1.4 Notations 7
1.2 Les étalonnages 8
1.2.1 Les étalonnages directs, utilisant une information experte 8
1.2.2 Les étalonnages inverses, utilisant une information expérimentale 12
1.3 Les prétraitements spectraux 15
1.3.1 Prétraitements utilisant de l'information expérimentale différente du jeu d'étalonnage 15
1.3.2 Prétraitements utilisant le jeu d'étalonnage comme information expérimentale 17
1.3.3 Prétraitements utilisant de l'information experte 18
1.3.4 Prétraitements utilisant conjointement des informations expérimentales et expertes 20
1.3.5 Conclusion sur les prétraitements 22
1.4 Discussion 22
1.4.1 Elimination d'information spectrale 23
1.4.2 Informations spectrales utiles, nuisibles, neutres 23
1.5 Conclusion 25

Chapitre 2
Un modèle linéaire général d'étalonnage et de prétraitement

2.1 Théorie du modèle général . . . 28
2.2 Calcul des étalonnages et prétraitements . . . 28
2.2.1 Calcul des étalonnages . . . 28
2.2.2 Calcul des prétraitements . . . 30
2.3 Validation de l'insertion des étalonnages et prétraitements dans le modèle général 30
2.3.1 Modèle général et étalonnages. . . . 30
2.3.2 Modèle général et prétraitements. . . . 31
2.3.3 Le cas de la PLSR . . . 32
2.3.4 Le cas de l'OSC . . . 35
2.4 Discussion et conclusion . . . 37

Chapitre 3
Première implémentation : IDC, une nouvelle méthode d'étalonnage direct

3.1 Théorie de l'IDC . . . 42
3.2 Premier exemple d'application de l'IDC : quantification de l'éthanol en fermentation . . . 44
3.2.1 Matériels et méthodes . . . 44
3.2.2 Résultats . . . 45
3.2.3 Conclusion sur la première application de l'IDC . . . 50
3.3 Deuxième exemple d'application de l'IDC : analyse des parois de la couche à aleurones du grain de blé . . . 53
3.3.1 Matériels et méthodes . . . 53
3.3.2 Résultats . . . 54
3.3.3 Conclusion sur la deuxième application de l'IDC . . . 55
3.4 Discussion . . . 55
3.4.1 Les fondements spectroscopiques de l'étalonnage direct . . . 55
3.5 Conclusion sur l'IDC . . . 61

Chapitre 4
Deuxième implémentation : VODKA-PLSR, une famille de modèles de régression

4.1 NIPALS-P une nouvelle version de NIPALS . . . 64
4.2 Le modèle VODKA-PLSR . . . 65
4.3 Application : quantification de l'éthanol dans des moûts de raisin en fermentation 67
4.3.1 Les données . . . 67

4.3.2 Paramétrage et validation des modèles de régression 67
4.3.3 Résultats 68
4.4 Discussion 70
4.4.1 Informations expérimentales et informations expertes dans le modèle PLSR 70
4.4.2 Choix de l'algorithme NIPALS 71
4.4.3 Présence d'une incohérence dans NIPALS ? 71

Chapitre 5
Discussion et conclusion

5.1 Place centrale des informations utiles et nuisibles 73
5.1.1 L'information utile, pour les étalonnages 75
5.1.2 L'information nuisible, pour les prétraitements 77
5.2 La notion de métrique introduite par $\mathbf{\Sigma}$ 78
5.2.1 Construction de $\mathbf{\Sigma}$ 78
5.2.2 Utilité fonctionnelle de $\mathbf{\Sigma}$ pour les étalonnages directs 79
5.2.3 Utilité fonctionnelle de $\mathbf{\Sigma}$ pour les étalonnages inverses 80
5.2.4 Perspectives d'une métrique $\mathbf{S}$ dans $\mathbb{R}^N$ 80
5.3 Combinaison de modèles d'étalonnage et de prétraitement 81
5.4 Le NAS, concentré d'information experte pour l'IDC et VODKA-PLSR 82
5.5 Gestion par les projections orthogonales de plusieurs informations nuisibles . . . 83
5.6 Conclusion générale 83

Annexe A Script Matlab et Scilab de la fonction VODKA-PLSR **85**

Glossaire **87**

Index **89**

Bibliographie **91**

Introduction

1 L'importance croissante de l'analyse en ligne dans le contrôle de procédés

Un procédé est une opération unitaire au cours de laquelle des éléments peuvent apparaitre, disparaitre ou être modifiés. La cuisson du pain est un exemple de procédé. Tout produit est le résultat de l'action d'un ou plusieurs procédés. Cette notion s'applique à de nombreuses industries, par exemple en agriculture, agro-alimentaires, chimie, pharmacie. La qualité du contrôle du procédé est l'élément décisif qui conduit à l'obtention d'un produit final conforme ou non conforme à un cahier des charges. Dans un concept artisanal, ce contrôle est dépendant de l'expertise de l'opérateur : boulanger estimant le moment où le pain est cuit, vigneron dégustant le vin pour décider de la durée d'une macération. La personne responsable de l'opération peut être qualifiée d'expert dans son domaine, son choix s'appuie sur l'expérience et sur les sens : vue, toucher, perception sensorielle. Cette approche très ancienne et relativement performante dans des cas simples n'est pas du tout transposable dans le concept industriel. En premier lieu, les produits sont beaucoup plus complexes et ne peuvent pas être évalués simplement : une méthode d'analyse est nécessaire, par exemple pour quantifier la matière active dans un médicament. En second lieu, la production industrielle recquiert une régularité de production toujours difficile quand l'homme décide seul. Les dires d'experts doivent être accompagnés d'éléments analytiques objectifs. En troisième lieu, d'autres exigences règlementaires sont apparues. Les règlements ISO 31000-2009, ISO 9001-2008 et ISO 14001 sont destinés respectivement à prévenir les risques, à garantir une qualité de produit, à protéger l'environnement. A ces normes générales se superposent des règlements spécifiques par filière. Le secteur alimentaire doit respecter le Paquet Hygiène, règlement CE178/2002. Le secteur pharmaceutique est incité à appliquer le Process Analytical Technology, une norme issue de la Food and Drug Administration (Etats-Unis) demandant un contrôle analy-

tique en ligne lors de la fabrication de médicaments. Ces exemples illustrent pourquoi la demande analytique est en constante augmentation, et pourquoi elle évolue de méthodes discontinues avec prélèvement d'échantillon vers des méthodes en ligne non destructives, sans prélèvement. Les spectrométres dans l'ultra-violet, le visible, le proche et le moyen infra-rouge, ont des spécificités techniques qui répondent très bien à ces contraintes : peu encombrants, robustes, peu coûteux, spectres très répétables.

2 Bases de la spectroscopie, le modéle linéaire général de mélange

La spectrométrie d'absorption est basée sur le principe d'absorption du rayonnement lumineux par les molécules. Sous l'effet d'un apport d'énergie précis, les liaisons covalentes peuvent être déformées : étirement de la liaison, rotation des atomes par exemple. A chaque liaison et type de déformation correspond un ou plusieurs niveaux d'énergie, donc une ou plusieurs longueurs d'onde pouvant être absorbées. Ainsi le spectre d'une molécule pure est le résultat de la contribution des différentes liaisons qui la constituent. Par exemple le spectre du méthanol est théoriquement la somme des absorbances des liaisons C-H, C-O et O-H de la molécule, plus les interactions entre liaisons. Or la plupart des molécules organiques sont formées avec ces 3 liaisons. En conséquence toutes les molécules d'une même famille présentent des absorbances dans les mêmes plages spectrales, la différence entre molécules réside dans la forme du spectre. Chaque composé chimique a un spectre qui lui est propre, c'est son empreinte digitale. La spectroscopie quantitative analyse des échantillons, donc des milieux complexes formés de très nombreuses molécules ou composés chimiques et prédit la concentration de l'un des composés. Deux notions importantes apparaissent : (1) le signal d'un composé est rarement explicite dans le spectre de l'échantillon, il est géneralement environné du bruit produit par les autres composés ; (2) d'après la théorie, le spectre final est l'addition des absorbances apportées par les différents composés. La première notion explique pourquoi la quantification d'un composé d'intérêt n'est jamais obtenue de manière simple et directe comme cela est possible en analyse chimique ou chromatographique, elle est toujours issue d'un minimum de calculs, d'un étalonnage. Et la deuxième notion énonce le modèle linéaire général de mélange, c'est à dire le modéle linéaire de la loi des mélanges issue de Beer-Lambert (Linear Mixture Model, [1]). Elle justifie le choix d'un modèle linéaire pour l'étalonnage ou le prétraitement des spectres. Aucun modèle non linéaire, comme les réseaux de neurones ou les Support Vector Machine (SVM) ne sera considéré dans la suite de ce tra-

vail focalisé uniquement sur les méthodes multivariées prédisant une seule grandeur d'intérêt quantitative.

3 Plan du mémoire

Les modèles linéaires sont très populaires en chimiométrie. Les modèles d'étalonnage permettent une prédiction d'une grandeur d'intérêt, alors que les modèles de prétraitement préparent les données pour obtenir ensuite un meilleur étalonnage. De nombreuses méthodes ont été proposées. La première partie de la thèse est une étude bibliographique sur les modèles linéaires, orientée en fonction de la nature de l'information experte ou expérimentale mise en oeuvre dans chaque modèle. La partie 2 présente la proposition scientifique : les principales méthodes d'étalonnage et de régression peuvent s'écrire sous forme d'un modèle général utilisant une ou plusieurs entrées : connaissances expérimentales tels des spectres acquis sur des échantillons ; connaissances expertes telles des spectres purs. Deux paramètres $\mathbf{P}$ et $\mathbf{\Sigma}$ sont déduits des entrées. Les parties 3 et 4 sont des implémentations du modèle général. La troisième partie est une nouvelle méthode d'étalonnage direct consistant en l'utilisation de deux informations expertes. La quatrième partie présente VODKA-PLSR, une familles de modèles de type PLSR directement issue d'une présentation différente de l'algorithme NIPALS de la PLSR.

Chapitre 1

Place des informations expérimentales et expertes dans les modèles d'étalonnage

Sommaire

1.1 Définitions, champ d'application 6

1.1.1 Informations expérimentales, informations expertes 6

1.1.2 Méthodes d'étalonnage, méthodes de prétraitement 6

1.1.3 Champ d'application 7

1.1.4 Notations 7

1.2 Les étalonnages 8

1.2.1 Les étalonnages directs, utilisant une information experte 8

1.2.2 Les étalonnages inverses, utilisant une information expérimentale 12

1.3 Les prétraitements spectraux 15

1.3.1 Prétraitements utilisant de l'information expérimentale différente du jeu d'étalonnage 15

1.3.2 Prétraitements utilisant le jeu d'étalonnage comme information expérimentale 17

1.3.3 Prétraitements utilisant de l'information experte 18

1.3.4 Prétraitements utilisant conjointement des informations expérimentales et expertes 20

1.3.5 Conclusion sur les prétraitements 22

1.4 Discussion 22
1.4.1 Elimination d'information spectrale 23
1.4.2 Informations spectrales utiles, nuisibles, neutres 23
1.5 Conclusion 25

1.1 Définitions, champ d'application

1.1.1 Informations expérimentales, informations expertes

De manière très générale, chaque loi, mesure, valeur, est porteuse d'une information. Le choix d'identifier deux types d'informations : expérimentales et expertes, est guidé par le niveau de généralisation de l'information considérée.

- Une information experte est une information universelle, elle n'est pas rattaché à un échantillon en particulier. Des exemples d'informations expertes sont des rapports stoechiométriques, des spectres purs, des masses molaires, des lois s'appuyant sur une théorie. Des informations expertes utilisées dans ce travail sont par exemple des spectres purs.
- Une information expérimentale est une information rattachée à un échantillon. Elle n'est pas du tout utilisable pour un autre échantillon, et elle est dépendante de l'expérimentation, seule manière d'en obtenir une estimation. Ainsi par exemple un jeu d'étalonnage $(\mathbf{X}, \mathbf{y})$ contient deux informations expérimentales : les spectres et les valeurs de référence de la grandeur d'intérêt. D'autres informations expérimentales sont représentées par des spectres acquis après un plan d'expérience.

1.1.2 Méthodes d'étalonnage, méthodes de prétraitement

Deux familles complémentaires d'outils sont utilisables pour construire un étalonnage. La première famille d'outils est constituée des étalonnages proprement dits donnant une prédiction $\widehat{\mathbf{y}}$ à partir des spectres $\mathbf{X}$. Chaque étalonnage est représenté par un vecteur $\mathbf{b}$ de dimension P, la prédiction $\widehat{\mathbf{y}}$ est donnée par la relation :

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} + y_0\mathbf{1}_N$$

Le terme y_0 est l'ordonnée à l'origine, une valeur constante pour chaque échantillon. La

qualité prédictive de chaque modèle est apprécié par l'écart entre $\mathbf{y}$ et $\widehat{\mathbf{y}}$, entre valeurs réelles et valeurs estimées.

La deuxième famille d'outils est constituée des méthodes linéaires de prétraitement. Ces méthodes n'ont pas de capacité de prédiction, elle permettent de corriger les spectres d'un matrice $\mathbf{X}$ pour obtenir une matrice $\mathbf{X}^*$ de même dimension. La différence $\mathbf{X} - \mathbf{X}^*$ correspond généralement à une information indésirable de $\mathbf{X}$ que l'on ne retrouve plus dans $\mathbf{X}^*$. Des modèles d'étalonnage peuvent être calculés sur $\mathbf{X}^*$ avec souvent des meilleures performances que s'ils étaient appliqués sur $\mathbf{X}$. La transformation de $\mathbf{X}$ en $\mathbf{X}^*$ est toujours positionnée *avant* l'étalonnage, d'où le nom de prétraitement.

1.1.3 Champ d'application

Les variables spectrales sont des absorbances mesurées à différentes longueurs d'onde. L'écart entre deux longueurs d'onde consécutives est le pas de résolution du spectrophotomètre, valeur choisie la plus petite possible compte-tenu des contraintes matérielles : performance du spectrophotométre, temps d'acquisition. Les variables spectrales forment un continuum sur une plage spectrale, deux variables spectrales proches sont très fortement corrélées, ainsi la représentation graphique d'un spectre est une courbe continue.

Une autre propriété des spectres est que les perturbations spectrales, ou grandeurs d'influence, sont structurées. Une perturbation sur une variable spectrale n'est pas indépendante de la perturbation sur une autre variable spectrale. Par exemple, après ajout d'un composé chimique dans un échantillon, la déformation spectrale induite par cet apport aura la forme du spectre pur du composé chimique. Autre exemple, il a été montré [2] que le trouble entraîne des déformations spectrales de forme apparentée à un polynome. Dans tous les cas une grandeur d'influence est théoriquement modélisable par une information expérimentale ou experte.

Ce mémoire est dédié à des applications en spectroscopie. Toutefois les méthodes qui y sont décrites sont applicables de manière plus générale à tout type d'information dès lors qu'elle vérifie les conditions d'excellente répétabilité des acquisitions, de continuité entre variables et de structure des perturbations ou grandeurs d'influence.

1.1.4 Notations

Sauf indication contraire, les vecteurs sont notés en caractères minuscule gras et les matrices en caractères majuscule gras. Les scalaires sont en caractères normaux, majuscule pour les

paramètres, minuscule pour les indices.

Les spectres expérimentaux sont disposés en ligne, par exemple dans les matrices $\mathbf{X}$ ou $\mathbf{X}_G$. Par contre les spectres purs ou les vecteurs-propres sont disposés en colonne.

Les mêmes identifiants tels $\mathbf{P}$, $\mathbf{\Sigma}$ ou $\mathbf{X}_G$ se retrouvent dans des méthodes différentes et correspondent à des matrices différentes. La raison est que la signification de la matrice est la même. Ainsi la similarité de notation facilite la mise en évidence des similarités de fonctionnement entre méthodes. Les principales notations sont regroupées dans le tableau 1.1.

1.2 Les étalonnages

Etalonner selon un modèle linéaire consiste à déterminer le vecteur des b-coefficients $\mathbf{b}$ et l'ordonnée à l'origine y_0 tels que $\widehat{y}$ défini par :

$$\widehat{y} = \mathbf{x}'\mathbf{b} + y_0$$

soit la meilleure estimation de y, c'est à dire minimise $|y - \widehat{y}|$ sous certaines contraintes.

Les méthodes d'étalonnage peuvent être classées selon la présence ou absence d'informations expertes [1]. Les étalonnages directs utilisent une information experte connue *a priori*, au moins le spectre pur de la grandeur d'intérêt. L'information utile est clairement identifiée et *directement* introduite dans le modèle. Les étalonnages indirects n'utilisent pas d'information experte, mais une information expérimentale sous forme d'un jeu d'étalonnage. Les informations nécessaires à la construction des modèles sont extraites du jeu d'étalonnage, d'où le qualificatif *indirect*.

Le centrage des données est préconisé pour plusieurs méthodes, toutefois les calculs restent possibles sans centrage. Lorsque cela n'est pas précisé explicitement, les données peuvent être centrées, ou pas.

1.2.1 Les étalonnages directs, utilisant une information experte

Deux étalonnages directs ont été décrits. Tous deux utilisent une information experte, le spectre pur $\mathbf{k}$ de la grandeur d'intérêt. Ils diffèrent entre eux selon la manière dont l'effet des grandeurs d'influence est identifié puis caractérisé.

La Direct Calibration

La Direct Calibration (DC) reprise par [1] part du principe qu'un spectre est le résultat de la seule influence des composés chimiques présents dans l'échantillon. Le calcul de la DC est

$\mathbf{X}$	matrice $N \times P$ de N individus et P variables explicatives
$\mathbf{y}$	vecteur $N \times 1$ contenant les valeurs de la grandeur d'intérêt
$\mathbf{X}_i$	projeté de $\mathbf{X}$ orthogonalement à $\{ \mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_i\}$
$\mathbf{y}_i$	projeté de $\mathbf{y}$ orthogonalement à $\{ \mathbf{t}_1, \mathbf{t}_2, ...\mathbf{t}_i\}$
$\mathbf{T}$	matrice $N \times A$ de composantes (scores) pour $\mathbf{X}$
$\mathbf{P}$	matrice $P \times A$ d'axes (loadings) pour $\mathbf{X}$
$\mathbf{W}$	matrice $P \times A$ de poids pour $\mathbf{X}$
$\mathbf{c}$	vecteur $A \times 1$ des poids $c_1, c_2...c_A$ pour $\mathbf{y}$
$\mathbf{\Sigma}$	matrice $P \times P$ symétrique
$\mathbf{I}_N$, $\mathbf{I}_P$	matrices identité $N \times N$, $P \times P$
$\mathbf{O}_N$	matrice nulle $N \times N$
$\mathcal{P}_i$	projecteur $N \times N$ sur $\mathbf{t}_i$
$\mathcal{P}_i^{\perp}$	projecteur $N \times N$ orthogonalement à $\mathbf{t}_i$
$\mathcal{P}_{1:i}^{\perp}$	projecteur $N \times N$ orthogonalement à $\{ \mathbf{t}_1...\mathbf{t}_i\}$
$\mathcal{Q}_i^{\perp}$	projecteur $P \times P$, $\mathbf{\Sigma}$-orthogonal à $\mathbf{p}_i$
$\mathcal{Q}_{1:i}^{\perp}$	projecteur $P \times P$, $\mathbf{\Sigma}$-orthogonal à $\{ \mathbf{p}_1...\mathbf{p}_i\}$
$\mathbf{t}_i$	i^{eme} vecteur colonne de $\mathbf{T}$
$\mathbf{p}_i$	i^{eme} vecteur-colonne de $\mathbf{P}$
$\mathbf{w}_i$	i^{eme} vecteur-colonne de $\mathbf{W}$
$\mathbf{1}_N$, $\mathbf{1}_P$	vecteurs composés respectivement de N et P termes 1
$\mathcal{E}^U$, $\mathcal{E}^N$, $\mathcal{E}^R$, $\mathcal{E}^I$	sous-espaces vectoriels utiles, nuisibles, résiduels, inutiles
$\mathcal{E}_X$	sous-espace vectoriel de $\mathbb{R}^N$ décrit par les vecteurs-colonne de $\mathbf{X}$
$\mathcal{F}_X$	sous-espace vectoriel de $\mathbb{R}^P$décrit par les vecteurs-ligne de $\mathbf{X}$

TAB. 1.1 – Principales notations

basé sur le modèle linéaire de la loi des mélanges ou Linear Mixture Model (LMM). Soient $\mathbf{x}$ un spectre acquis sur un échantillon et y la valeur de la grandeur d'intérêt associée à cet échantillon ; $\mathbf{k}$ le vecteur $(P, 1)$ du spectre pur de la grandeur d'intérêt ; $\mathbf{K}$ la matrice (P, Q) des spectres purs des Q constituants du produit analysé autres que la grandeur d'intérêt et $\mathbf{t}_\chi$ de dimensions $(Q, 1)$ leurs concentrations. La contribution de la grandeur d'intérêt à $\mathbf{x}$ est $y\mathbf{k}$. Les grandeurs d'influence chimiques sont le résultat spectral des concentrations de tous les composés chimiques autres que le composé d'intérêt présents dans l'échantillon analysé. Dans l'hypothèse où tous les composés chimiques suivent la loi de Beer-Lambert, chacun apporte au spectre final un profil égal à son spectre pur pondéré par sa concentration. Leur contribution au spectre est donnée par : $\mathbf{K}\mathbf{t}_\chi$. Le LMM s'écrit finalement :

$$\mathbf{x}' = y\mathbf{k}' + \mathbf{t}'_\chi\mathbf{K}' + \varepsilon' \tag{1.1}$$

oú ε est un vecteur de bruit dont les P variables sont indépendantes les unes des autres et présentent des amplitudes faibles suivant la même distribution. La Direct Calibration (DC) propose de projeter $\mathbf{x}$ sur $\mathbf{k}$, orthogonalement à l'information experte représentée par $\mathbf{K}$. Soit $\mathbf{\Sigma}_{DC}$ de dimension (P, P) le projecteur orthogonal à $\mathbf{K}$:

$$\mathbf{\Sigma}_{DC} = (\mathbf{I} - \mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}')$$

En transposant puis en multipliant à droite les membres de l'équation (1.1) par $\mathbf{\Sigma}_{DC}$:

$$\mathbf{x}'\mathbf{\Sigma}_{DC} = y\mathbf{k}'\mathbf{\Sigma}_{DC} + \mathbf{t}'_\chi\mathbf{K}'\mathbf{\Sigma}_{DC} + \varepsilon'\Sigma_{DC}$$

Par construction, le terme $\mathbf{K}'\mathbf{\Sigma}_{DC}$ est nul, ainsi :

$$\mathbf{x}'\mathbf{\Sigma}_{DC} = y\mathbf{k}'\mathbf{\Sigma}_{DC} + \varepsilon'\Sigma_{DC}$$

Nous supposons que ε est suffisamment petit pour que $\varepsilon'\Sigma_{DC}$ soit négligeable. La multiplication à droite par $\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{DC}\mathbf{k})^{-1}$ permet de retrouver la formule de la DC dans le cas de la prédiction d'une seule grandeur d'intérêt ([3]) :

$$\widehat{y} = \mathbf{x}'\mathbf{\Sigma}_{DC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{DC}\mathbf{k})^{-1}$$

d'oú :

$$\mathbf{b}_{DC} = \mathbf{\Sigma}_{DC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{DC}\mathbf{k})^{-1}$$

Pour appliquer la DC, deux conditions sont supposées remplies : (1) les spectres purs de toutes les grandeurs chimiques sont connus et linéairement indépendants de manière à ce que $(\mathbf{K}'\mathbf{K})$ soit inversible ; (2) l'effet sur les spectres des grandeurs physiques ϕ_j est supposé négligeable. Ces deux hypothèses sont très contraignantes. Ainsi l'hypothèse (1) est rarement remplie car très souvent les spectres purs d'un ou plusieurs composés présents dans les échantillons sont inconnus. Et l'hypothèse (2) sur l'absence d'effet des grandeurs d'influence physiques est rarement vérifiée en dehors d'un laboratoire où l'environnement est contrôlé. Ces raisons expliquent pourquoi la DC est un modèle pour lequel les conditions d'application sont rarement remplies.

En conclusion, la DC est une méthode d'étalonnage direct n'utilisant que de l'information experte constituée de spectres purs. Le terme $\mathbf{\Sigma}_{DC}\mathbf{k}$ correspond à la définition du NAS-Net Analyte signal ([4]), le spectre pur de la grandeur d'intérêt projeté orthogonalement aux grandeurs d'influence dans le cas où celles-ci sont uniquement de nature chimique. L'information spectrale relative à la grandeur d'intérêt se trouve dans un sous-espace vectoriel de dimension 1 dont $\mathbf{\Sigma}_{DC}\mathbf{k}$ est une base.

La Science-Based Calibration

La Science-Based Calibration (SBC) [3] part du principe qu'un spectre $\mathbf{x}$ est la somme de deux contributions : (1) la contribution $y\mathbf{k}$ de la grandeur d'intérêt ; (2) une erreur $\mathbf{x}^N$ dûe aux grandeurs d'influence physiques et chimiques. La SBC détermine $\mathbf{\Sigma}_{SBC}$ au moyen d'un plan d'expérience. Soient N spectres acquis sur N échantillons pour lesquels la grandeur d'intérêt est constante. Le centrage de ces spectres donne une matrice $\mathbf{X}_G$ de dimensions $(N \times P)$ ne contenant que de l'information liée au bruit. La matrice $\mathbf{\Sigma}$ est déduite de la formule suivante :

$$\mathbf{\Sigma}_{SBC} = [\mathbf{X}'_G\mathbf{X}_G(N-1)^{-1}]^{-1}$$

Considérons maintenant l'ensemble de la population sur lesquels les spectres peuvent être acquis, et appelons σ l'écart-type mesurant la variabilité de la grandeur d'intérêt dans cette population. Les b-coefficients de la SBC sont donnés par [3] :

$$\mathbf{b}_{SBC} = \sigma^2\mathbf{\Sigma}_{SBC}\mathbf{k}(1 + \sigma^2\mathbf{k}'\mathbf{\Sigma}_{SBC}\mathbf{k})^{-1}$$

Si l'échelle de variation de la grandeur d'intérêt est importante, alors la constante 1 est négligeable et l'équation précédente se simplifie :

$$\mathbf{b}_{SBC} = \mathbf{\Sigma}_{SBC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{SBC}\mathbf{k})^{-1} \tag{1.2}$$

Nous remarquons que la division par $N-1$ dans le calcul de $\mathbf{\Sigma}_{SBC}$ est inutile pour le calcul des b-coefficients, et par la suite :

$$\mathbf{\Sigma}_{SBC} = (\mathbf{X'}_G\mathbf{X}_G)^{-1}$$

Ainsi la SBC est une méthode d'étalonnage direct utilisant conjointement de l'information expérimentale et de l'information experte. L'information experte est représentée par le spectre pur $\mathbf{k}$, l'information expérimentale par la matrice $\mathbf{\Sigma}_{SBC}$ obtenue grâce à $\mathbf{X}_G$. Bien que $\mathbf{X}_G$ contienne de l'information expérimentale, l'utilisation de la SBC apporte deux avantages sur les méthodes inverses : 1) il n'est pas nécessaire de connaitre les valeurs de référence pour les spectres de $\mathbf{X}_G$; 2) le nombre de spectres de $\mathbf{X}_G$ peut être petit si ces spectres sont bien choisis. La SBC ne s'appuie pas sur la notion de NAS, par contre l'information spectrale sur la grandeur d'intérêt est dans un espace de dimension 1 dont $\mathbf{\Sigma}_{SBC}\mathbf{k}$ est une base.

Conclusion sur les étalonnages directs

Les étalonnages directs utilisent toujours de l'information experte, et parfois de l'information expérimentale. Ces informations sont portées par un vecteur et une matrice : (1) le spectre pur $\mathbf{k}$ qui fixe à 1 la dimension de l'espace contenant l'information sur la grandeur d'intérêt ; (2) une matrice $\mathbf{\Sigma}$ symétrique, de dimension (P, P), qui donne avec $\mathbf{k}$ une base de l'espace contenant l'information sur la grandeur d'intérêt.

1.2.2 Les étalonnages inverses, utilisant une information expérimentale

Les étalonnages inverses partent du principe que l'information experte représentée par $\mathbf{k}$ ou $\mathbf{K}$ n'est pas connue ou pas utilisée. Par contre, une information expérimentale $(\mathbf{X}, \mathbf{y})$ est disponible, représentée respectivement par les spectres et les valeurs de référence de la grandeur d'intérêt acquis sur N échantillons. Plusieurs méthodes de régression ont été proposées, toutes utilisent simultanément $\mathbf{X}$ et $\mathbf{y}$ pour déterminer leurs paramètres propres.

Méthode des moindres carrés : l'Ordinary Least Square Regression

Soit $\widehat{\mathbf{y}}$ une estimation de $\mathbf{y}$. Nous supposons que $\widehat{\mathbf{y}}$ est obtenu à partir de $\mathbf{X}$ par une combinaison linéaire des colonnes de $\mathbf{X}$. Si B est le rang de $\mathbf{X}$, $\widehat{\mathbf{y}}$ appartient au sous-espace vectoriel de $\mathbb{R}^N$ de dimension B défini par les colonnes de $\mathbf{X}$. La méthode des moindres carrés ou Ordinary

Least Square Regression (OLSR) reprise par [1] minimise $\| \mathbf{y} - \mathbf{Xb} \|$ dans $\mathbb{R}^N$, ce qui équivaut à ce que $\widehat{\mathbf{y}}$ soit la projection orthogonale de $\mathbf{y}$ sur $\mathbf{X}$:

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.3}$$

L'inversion de $(\mathbf{X}'\mathbf{X})$ de dimensions (P, P) est une contrainte forte de l'OLSR. Souvent en spectroscopie cette opération n'est pas possible. C'est systématiquement le cas lorsque $N < P$, lorsque le nombre d'individus est inférieur au nombre de variables. C'est parfois le cas alors que $N > P$, lorsque les spectres sont fortement colinéaires ; on parle alors de mauvais conditionnement. Dès lors l'OLSR n'est pas applicable.

L'OLS est une méthode qui utilise de l'information expérimentale, le jeu d'étalonnage. Elle ne nécessite aucun paramétrage, aucune hypothèse sur l'espace où se trouve l'information spectrale relative à la grandeur d'intérêt. Plus exactement, cet espace dans $\mathbb{R}^N$ est supposé être le même que celui décrit par les vecteurs-colonne de $\mathbf{X}$.

Méthode des moindres carrés pondérés (WLSR) et généralisés (GLSR)

L'équation 1.3 est ré-écrite en introduisant une métrique $\mathbf{S}$ dans $\mathbb{R}^N$ et devient :

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{S}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}\mathbf{y} \tag{1.4}$$

Si $\mathbf{S}$ est une matrice diagonale dont les élements de la diagonale sont les poids associés aux individus correspondants, alors l'équation 1.4 est celle de la WLSR-Weighted Least Square Regression ([1]). Si $\mathbf{S}$ est l'inverse de la matrice de covariance de l'erreur $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$, alors l'équation 1.4 est celle de la GLSR-Generalised Least Square Regression ([1]).

Pour la WLSR, l'information contenue dans $\mathbf{S}$ peut être experte ou expérimentale selon son origine. Pour la GLSR, l'information contenue dans $\mathbf{S}$ est expérimentale puisqu'elle dépend des échantillons.

Il est important de souligner l'utilisation d'une matrice $\mathbf{S}$ à vocation de métrique dans $\mathbb{R}^N$ pour déformer la projection orthogonale de manière à favoriser l'information concernant la grandeur d'intérêt. L'impact est potentiellement plus important pour la WLSR, puisqu'en attribuant des poids nuls dans $\mathbf{S}$, des individus peuvent être supprimés. *A contrario*, la GLSR peut uniquement pondérer faiblement des individus, pas les éliminer.

Projection sur un sous-espace vectoriel : la Principal Component Regression

La Principal Component Regression (PCR) propose de projeter $\mathbf{X}$ sur un sous-espace vectoriel de $\mathbb{R}^P$ de dimension A, puis de réaliser l'OLSR dans ce sous-espace vectoriel. Soit $\mathbf{P}$ une matrice $P \times A$ contenant les A premiers vecteurs-propres de l'ACP sur $\mathbf{X}$. Les coordonnées des individus de $\mathbf{X}$ dans le sous-espace vectoriel défini par les colonnes de $\mathbf{P}$ sont :

$$\mathbf{T} = \mathbf{XP}$$

L'OLSR appliquée à $\mathbf{T}$ donne :

$$\widehat{\mathbf{y}} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$$

En substituant $\mathbf{T}$ par sa valeur $\mathbf{XP}$:

$$\widehat{\mathbf{y}} = \mathbf{XP}(\mathbf{P}'\mathbf{X}'\mathbf{XP})^{-1}\mathbf{P}'\mathbf{X}'\mathbf{y}$$

L'espace de la PCR est celui des A premiers vecteurs propres de l'ACP sur $\mathbf{X}$. Selon la forme classique de la PCR, chaque variable est prise en compte selon un critère de variabilité, pas en fonction de son explication de $\mathbf{y}$. Finalement, l'espace défini par les colonnes de $\mathbf{P}$ représente la variabilité spectrale, mais pas nécessairement l'information qui permet d'expliquer $\mathbf{y}$. Ce problème est reconnu, une solution proposée est la CPCR Correlation Principal Component Regression [5], une forme modifiée de la PCR où le choix d'inclure une nouvelle composante dans le modèle d'étalonnage dépend de la capacité de cette composante à prédire $\mathbf{y}$. Toutefois malgré ce choix *a posteriori* basé sur $\mathbf{y}$, la construction des vecteurs propres reste basée sur l'ACP, donc sur la notion de variabilité dans $\mathbf{X}$.

Projection sur un sous-espace vectoriel de structures latentes : la Partial Least Square Regression

Deux groupes de modèles PLSR ont été proposés. Les modèles PLS1 ne prédisent qu'une grandeur d'intérêt à la fois, alors que les modèles PLS2 prédisent simultanément plusieurs grandeurs d'intérêt. Nous n'aborderons pas les PLS2, et par la suite le terme PLSR désigne systématiquement un modèle PLS1.

Tout comme la PCR, la PLSR [6] propose de projeter $\mathbf{X}$ dans $\mathbb{R}^P$ sur un sous-espace vectoriel défini par les colonnes d'une matrice $\mathbf{P}$ $(P \times A)$. Les coordonnées des individus dans ce

sous-espace vectoriel constituent $\mathbf{T}$ $(N \times A)$. Mais contrairement à la PCR qui identifie un sous-espace vectoriel selon la variabilité des spectres de $\mathbf{X}$, la PLSR identifie un sous-espace vectoriel qui apporte de l'information sur $\mathbf{y}$. Le calcul des paramètres de la PLSR est détaillé plus loin. L'objectif de l'algorithme est de maximiser simultanément la covariance et la corrélation entre valeurs de référence $\mathbf{y}$ et vecteurs $\mathbf{t}_i$ [7]. Toutefois comme l'augmentation du nombre A de variables latentes (la dimension du modèle) augmente la corrélation mais diminue la covariance [7], le modèle optimum est un compromis. Plusieurs matrices intermédiaires sont construites : deux matrices de scores $\mathbf{T}$ $(N \times A)$ et $\mathbf{c}$ $(N \times 1)$, une matrice de vecteurs $\mathbf{P}$ $(P \times A)$ et une matrice de poids $\mathbf{W}$ $(P \times A)$. La formule des b-coefficients est donnée par [7] :

$$\mathbf{b}_{PLSR} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{c}'$$

et les scores $\mathbf{T}$:

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}$$

Conclusion sur les étalonnages inverses

L'OLSR reste dans l'espace vectoriel de dimension $\mathbb{R}^P$. Par contre les méthodes PCR et PLSR sont calculées dans des espaces de dimension A bien inférieure à P. Ces sous-espaces vectoriels contiennent l'information spectrale relative à la grandeur d'intérêt. Ils ont pour base les vecteurs d'une matrice $\mathbf{P}$.

1.3 Les prétraitements spectraux

Tous les prétraitements enlèvent de l'information spectrale indésirable dans $\mathbf{X}$ pour donner $\mathbf{X}^*$. Ils utilisent de l'information expérimentale ou experte.

1.3.1 Prétraitements utilisant de l'information expérimentale différente du jeu d'étalonnage

Supposons que des spectres ont été acquis dans des conditions telles que seules une ou plusieurs grandeurs d'influence s'expriment. La contribution spectrale de la grandeur d'intérêt est rendue nulle grâce au plan d'expérience et/ou des opérations de centrage. Les spectres ainsi obtenus sont réunis dans une matrice $\mathbf{X}_G$. Alors une SVD de $\mathbf{X}_G$ donne une matrice de vecteurs-propres $\mathbf{P}$ de dimensions $(P \times A)$ dont les vecteurs colonne constituent une base de ce sous-espace

vectoriel nuisible. Pour toute matrice $\mathbf{X}$ de spectres, il est possible d'enlever l'information nuisible en projetant $\mathbf{X}$ orthogonalement à $\mathbf{P}$, donnant une matrice $\mathbf{X}^*$ de spectres corrigés selon la formule :

$$\mathbf{X}^* = \mathbf{I} - \mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$$

et puique les vecteurs de $\mathbf{P}$ forment une base orthonormée :

$$\mathbf{X}^* = \mathbf{I} - \mathbf{P}\mathbf{P}' \tag{1.5}$$

Plusieurs approches ont été proposées pour déterminer $\mathbf{X}_G$.

L'Independant Interference Reduction

L'Independant Interference Reduction (IIR) [8] utilise un ensemble de spectres pour lesquels la grandeur d'intérêt prend une valeur nulle. Ainsi ces spectres ne peuvent exprimer que de la variabilité liée aux grandeurs d'influence physiques ou chimiques, ils sont regroupés dans $\mathbf{X}_G$.

L'External Parameter Orthogonalisation et Transfer Orthogonal Projection

L'External Parameter Orthogonalisation (EPO) [9] et Transfer Orthogonal Projection (TOP) [10] sont deux méthodes identiques. Soit un même groupe de M échantillons pour lesquels les spectres ont été acquis à R niveaux d'une seule grandeur d'influence. Un échantillon est donc représenté par R spectres pour lesquels, selon le modèle linéaire, la contribution des autres grandeurs d'influence est constante. Le centrage de ces R spectres ne garde que l'effet de la grandeur d'influence étudiée. La matrice $\mathbf{X}_G$ de dimensions $(MR \times P)$ est obtenue en compilant les R spectres centrés de chaque échantillon, pour les M échantillons.

Dynamic Orthogonal Projection

Dynamic Orthogonal Projection (DOP) [11] est une méthode issue de EPO destinée à corriger en ligne une grandeur d'influence apparaissant de manière imprévue. Elle suppose de connaitre le spectre et la valeur de référence d'au moins un échantillon. Soit $(\mathbf{X}, \mathbf{y})$ un jeu d'étalonnage acquis avant l'apparition de la grandeur d'influence. Soit $\mathbf{x}_1$ un spectre acquis après l'apparition de la grandeur d'influence, et y_1 la valeur de la grandeur d'intérêt associée à $\mathbf{x}_1$. Le spectre $\widehat{\mathbf{x}}_1$ qui aurait été obtenu à la place de $\mathbf{x}_1$ en l'absence de la grandeur d'influence est estimé par une moyenne pondérée de spectres de $\mathbf{X}$ choisis pour leur proximité avec y_1. Le spectre de différence

$\widehat{\mathbf{x}}_1$-$\mathbf{x}_1$ caractérise uniquement la grandeur d'influence. La même opération est répétée avec $\mathbf{x}_2$, $\mathbf{x}_3$,...les spectres de différence sont regroupés dans $\mathbf{X}_G$.

Error Removal by Orthogonal Subtraction

L'Error Removal by Orthogonal Subtraction [12] est une méthode issue de TOP. Elle permet de prendre en compte les répétitions. Pour un même échantillon, les spectres des différentes répétitions sont centrés. L'ensemble des spectres centrés de tous les échantillons est regroupé dans une matrice $\mathbf{X}_G$ de moyenne nulle par construction.

1.3.2 Prétraitements utilisant le jeu d'étalonnage comme information expérimentale

Deux méthodes de prétraitement utilisent l'information expérimentale fournie par un jeu d'étalonnage $(\mathbf{X}, \mathbf{y})$: l'Orthogonal Signal Correction (OSC) [13] et le Net Analyte Preprocessing (NAP) [14].

L'Orthogonal Signal Correction

Plusieurs méthodes de calcul de l'OSC ont été proposées, mais diffèrent peu entre elles. Une approche directe [15] a été choisie comme support pour sa simplicité algorithmique et aussi parce qu'elle est la base de l'Orthogonal Projection to Latent Structures (OPLS) [7]. Contrairement à la PLSR dont elle est inspirée, l'OSC identifie puis élimine de $\mathbf{X}$ une information expliquant le maximum de variabilité dans $\mathbf{X}$ tout en étant orthogonale à $\mathbf{y}$. L'algorithme calcule les matrices $\mathbf{P}$ et $\mathbf{W}$ de vecteurs et de poids, puis la correction est obtenue ainsi :

$$\mathbf{X}^* = \mathbf{X}(\mathbf{I} - \mathbf{W}\mathbf{P}')$$

Le Net Analyte Preprocessing

Le Net Analyte Preprocessing [14] est également une méthode qui enlève une information de $\mathbf{X}$ orthogonale à $\mathbf{y}$ avec une approche plus directe que l'OSC. L'information à enlever $\mathbf{X}^N$ est définie par la projection de $\mathbf{X}$ orthogonalement à $\mathbf{y}$:

$$\mathbf{X}^N = (\mathbf{I}_N - \mathbf{y}(\mathbf{y}'\mathbf{y})^{-1}\mathbf{y}')\mathbf{X}$$

Les A premiers vecteurs-propres d'une ACP sur $\mathbf{X}^N$ donnent la matrice $\mathbf{P}$. La correction de $\mathbf{X}$ en $\mathbf{X}^*$ est obtenue par projection de $\mathbf{X}$ orthogonalement à $\mathbf{P}$ selon la formule 1.5.

1.3.3 Prétraitements utilisant de l'information experte

Deux algorithmes basés sur de l'information experte sont dédiés aux déformations spectrales de la ligne de base.

Dérivée et lissage par Savitsky-Golay

L'algorithme de Savitsky-Golay (SG) [16] permet de lisser une courbe, c'est à dire d'enlever du bruit non structuré. Il permet aussi de calculer les dérivées première et seconde, soit les corrections respectives des décalages à l'origine et de la pente de la ligne de base.

Dans sa version initiale [16], l'algorithme de SG est proposé sous forme de fonctions de convolution de type polynomial. Soit un intervalle $[-M : +M]$ utilisé comme abscisse pour un segment de spectre $\mathbf{z}_i$ de même dimension. L'algorithme détermine des vecteurs $\mathbf{c}_0$, $\mathbf{c}_1$,... $\mathbf{c}_R$ regroupés dans une matrice $\mathbf{C}$ de coefficients de convolution telle que, après normalisation, le produit $\mathbf{z'}_i\mathbf{C}$ donne pour le point central de $\mathbf{z}_i$ les coefficients $\mathbf{t}_i$ du polynome de degré R qui s'ajuste le mieux à $\mathbf{z}_i$. Savitsky et Golay indiquent très clairement que leur méthode est basée sur l'approximation de $\mathbf{z}_i$ selon un polynome d'ordre R en utilisant la méthode des moindres carrés. Mais ils n'expliquent pas leur méthode d'un point de vue géométrique, alors que les problèmes de régression par les moindres carrés peuvent être résolus par des projections orthogonales. Bien que non décrite à notre connaissance dans la littérature, la présentation de SG qui suit est sous-entendue dans la description originale et à ce titre nous paraît logiquement positionnée dans la partie bibliographique.

Sur un intervalle $[-M : +M]$ (M petit nombre entier), un spectre peut être approximé à un polynome. Un polynome de degré R est calculé, ses paramètres sont obtenus pour le point central de l'intervalle. L'intervalle est ensuite décalé d'une variable, le calcul repris pour estimer le point d'à coté. De cette manière tout le spectre est balayé par l'intervalle, sauf aux bordures.

Soit M un nombre entier, et λ le vecteur de dimension $(2M + 1)$ constitué des valeurs de l'intervalle $[-M : +M]$. Une matrice $\mathbf{\Lambda}_R$ de dimensions $((2M + 1) \times (R + 1))$ est construite en prenant pour i^{eme} colonne les valeurs de λ portées à la puissance $i - 1$ ([17]). Ainsi les colonnes de $\mathbf{\Lambda}_R$ décrivent le sous-espace vectoriel de $\mathbb{R}^P$ contenant les polynomes de degré R.

Soit $\mathbf{z}_i$ de dimension $(2M + 1)$ la partie du spectre $\mathbf{x}$ de rayon M centrée sur x_i. Chaque valeur $z_i(\lambda)$ de $\mathbf{z}_i$ est modélisée selon SG par un polynome basé sur λ, c'est à dire qu'il existe

des coefficients $t_{i0}, t_{i1}, ...t_{iR}$ tels que pour toute valeur de λ :

$$\widehat{z}_i(\lambda) = t_{i0} + t_{i1}\lambda + t_{i2}\lambda^2 + ... + t_{iR}\lambda^R \tag{1.6}$$

Regroupons les valeurs $\{t_{i0}t_{i1}...t_{iR}\}$ dans le vecteur $\mathbf{t}_i$ de dimension $(R+1)$, et les $\widehat{z}_i(\lambda)$ dans $\widehat{\mathbf{z}}_i$. Ces matrices sont liées par la relation suivante :

$$\widehat{\mathbf{z}'}_i = \mathbf{t'}_i\mathbf{\Lambda'}_R$$

Par ailleurs $\widehat{\mathbf{z}}_i$, la meilleure approximation de $\mathbf{z}_i$ au sens des moindres carrés dans l'espace engendré par les colonnes de $\mathbf{\Lambda}_R$, est la projection de $\mathbf{z}_i$ sur $\mathbf{\Lambda}_R$:

$$\widehat{\mathbf{z}'}_i = \mathbf{z'}_i\mathbf{\Lambda}_R(\mathbf{\Lambda'}_R\mathbf{\Lambda}_R)^{-1}\mathbf{\Lambda'}_R$$

Des deux équations précedentes il découle :

$$\mathbf{t'}_i = \mathbf{z'}_i\mathbf{\Lambda}_R(\mathbf{\Lambda'}_R\mathbf{\Lambda}_R)^{-1}$$

Le terme $\mathbf{\Lambda}_R(\mathbf{\Lambda'}_R\mathbf{\Lambda}_R)^{-1}$ est égal à la fonction de convolution $\mathbf{C}$ décrite par [18].

Les dérivées successives de $\widehat{z}_i(\lambda)$ sont facilement calculables avec l'équation 1.6. Par exemple, les dérivées première et seconde sont :

$$d(\widehat{z}_i(\lambda))/d(\lambda) = t_{i1} + 2t_{i2}\lambda + 3t_{i3}\lambda^2 + ... + Rt_{iR}\lambda^{R-1}$$
$$d^2(\widehat{z}_i(\lambda))/d(\lambda)^2 = 2t_{i2} + 3t_{i3}\lambda + ... + R(R-1)t_{iR}\lambda^{R-2}$$

Pour le point central de l'intervalle, $\lambda = 0$ et les calculs des dérivées sont évidents :

$$\widehat{z}_i(0) = t_{i0}$$
$$d(\widehat{z}_i(0))/d(\lambda) = t_{i1}$$
$$d^2(\widehat{z}_i(0))/d(\lambda)^2 = 2t_{i2}$$
$$...$$
$$d^r(\widehat{z}_i(0))/d(\lambda)^r = r!t_{ir}$$

Ainsi l'algorithme de Savitsky-Golay utilise uniquement de l'information experte représentée par $\mathbf{\Lambda}_R$. Cette information experte est mise en oeuvre en utilisant une projection orthogonale.

Detrend

La méthode Detrend a été proposée par [19] pour corriger deux types de déformations de ligne de base : un décalage de l'origine, et l'apparition d'une pente non nulle. Soient les vecteurs λ_0 un vecteur de dimension $(P \times 1)$ ne contenant que la valeur 1, et λ_1 un vecteur de même dimension contenant les P premiers entiers dans l'ordre. Soit $\mathbf{\Lambda}_1$ la matrice de dimension $(P \times 2)$ construite avec λ_1 et λ_2. La correction par Detrend consiste en une projection orthogonale à $\mathbf{\Lambda}_1$. La matrice corrigée $\mathbf{X}^*$ est obtenue ainsi :

$$\mathbf{X}^* = \mathbf{X}^*(\mathbf{I}_P - \mathbf{\Lambda}_1(\mathbf{\Lambda}_1'\mathbf{\Lambda}_1)^{-1}\mathbf{\Lambda}_1')$$

Detrend utilise l'information experte de $\mathbf{\Lambda}_1$ pour réaliser une projection orthogonale.

1.3.4 Prétraitements utilisant conjointement des informations expérimentales et expertes

La SNV

La standardisation des spectres (Standard Normal Variate scaling) aussi proposée par [19] est une transformation donnant à chaque spectre une moyenne de 0 et un écart-type de 1. Elle est conçue pour corriger le décalage à l'origine de la ligne de base associée à des variations globales d'intensité des spectres.

La SNV est composée de deux opérations successives : (1) un centrage par ligne ; (2) une normalisation.

- **Le centrage par ligne** Le centrage est une projection orthogonale [20]. Soit $\mathbf{x}$ un vecteur de dimensions $(P \times 1)$. Soit $\mathbf{1}_P$ le vecteur $(P \times 1)$ dont chaque élément a la valeur 1. La moyenne $\bar{x}$ des élements de $\mathbf{x}$ est :

  $$\bar{x} = \mathbf{x}'\mathbf{1}_P P^{-1}$$

  Le vecteur centré $\mathbf{x}_c$ est obtenu en retirant $\bar{x}$ à chaque élement de $\mathbf{x}$:

  $$\mathbf{x}'_c = \mathbf{x}' - \bar{x}\mathbf{1}'_P$$

  En compilant ces deux équations et en remarquant que $P^{-1} = (\mathbf{1}'_P\mathbf{1}_P)^{-1}$:

  $$\mathbf{x}'_c = \mathbf{x}'(\mathbf{I}_P - \mathbf{1}_P(\mathbf{1}'_P\mathbf{1}_P)^{-1}\mathbf{1}'_P)$$

– **La normalisation** Le vecteur normalisé $\mathbf{x}_{cn}$ est obtenu en divisant $\mathbf{x}_c$ par sa norme :

$$\mathbf{x}_{cn} = (\mathbf{x'}_c\mathbf{x}_c)^{-1/2}\mathbf{x}_c$$

Par cette transformation linéaire dépendante de chaque vecteur, $\mathbf{x}_c$ et $\mathbf{x}_{cn}$ sont colinéaires dans le même sous-espace vectoriel.

Ainsi la SNV est un prétraitement utilisant conjointement de l'information expérimentale et de l'information experte. L'information experte est représentée par le vecteur $\mathbf{1}_P$, sa mise en oeuvre est une projection orthogonale. L'information expérimentale est représentée par la norme de chaque vecteur $\mathbf{x}$.

L'Extended Multiplicative Signal Correction

L'Extended Multiplicative Signal Correction (EMSC) est une amélioration de la Multiplicative Signal Correction (MSC). Elle a été proposée [2] pour corriger des déformations de la ligne de base dûes à la diffusion de la lumière (sous l'effet du trouble ou de la granulométrie par exemple). Soient v_1 à v_P les P variables spectrales. La matrice $\mathbf{\Lambda}$ de dimensions $(P, 3)$ est créée à partir des 3 vecteurs suivants : (1) un vecteur λ_0 composé uniquement de valeurs 1 ; (2) un vecteur λ_1 composé des valeurs nominales des P variables spectrales, de v_1 à v_P, et (3) un vecteur λ_2 composé des valeurs nominales des P variables spectrales élevées au carré, de v_1^2 à v_P^2. Les échantillons analysés contiennent R composés dont les R spectres purs sont supposés connus et forment la matrice $\mathbf{K}$ de dimensions (P, R).

Soit un échantillon i, $\mathbf{x}_{ichem}$ de dimensions $(P \times 1)$ le spectre théorique de cet échantillon, obtenu en conditions idéales, et $\mathbf{x}_i$ de dimensions $(P \times 1)$ le spectre obtenu en conditions réelles. Le modèle EMSC postule qu'il existe α_i, β_i, γ_i et δ_i tels que :

$$\mathbf{x'}_i = \alpha_i\mathbf{x'}_{ichem} + \beta_i\lambda'_0 + \gamma_i\lambda'_1 + \delta_i\lambda'_2 + \varepsilon' \tag{1.7}$$

La correction par EMSC consiste à estimer les quatre coefficients α_i, β_i, γ_i et δ_i, afin de les introduire dans l'equation 1.7 pour en déduire $\mathbf{x}_{ichem}$ le spectre corrigé. Une idée importante de l'EMSC est d'identifier simultanément les effets additifs et multiplicatifs des déformations spectrales observées. L'effet multiplicatif est donné par α_i. Afin de rendre α_i indépendant de $\mathbf{x'}_{ichem}$, l'équation 1.7 est re-arrangée et deux termes nouveaux apparaissent : (1) le vecteur $\mathbf{m}$ représentant le spectre moyen de la population d'échantillons ; (2) la matrice $\mathbf{K}^*_m$ obtenue en gardant $(R-1)$ colonnes de $\mathbf{K}$ puis en retranchant $\mathbf{m}$ à chacune des $(R-1)$ colonnes.

L'association de $\mathbf{m}$ et de $\mathbf{K}_m^*$ donne une matrice dont les colonnes décrivent le même sous-espace vectoriel que $\mathbf{K}$, la perte d'un vecteur évite le problème de colinéarité entre $\mathbf{m}$ et $\mathbf{K}$ [21]. De nouveaux coefficients $\alpha_{i,1}...\alpha_{i,R-1}$ conduisent à une nouvelle expression de l'équation 1.7 :

$$\mathbf{x'}_i = \alpha_i \mathbf{m'} + \alpha_{i,1}(\mathbf{k'}_1 - \mathbf{m}) + ... + \alpha_{i,R-1}(\mathbf{k'}_{R-1} - \mathbf{m}) + \beta_i \lambda'_0 + \gamma_i \lambda'_1 + \delta_i \lambda'_2 + \varepsilon'$$

La concaténation des matrices $\mathbf{K}_m^*$, $\mathbf{m}$ et $\mathbf{\Lambda}$ donne $\mathbf{M}$, de dimensions $(P, R+3)$. Le vecteur $\mathbf{x}_i$ appartient en théorie au sous-espace vectoriel défini par les colonnes de $\mathbf{M}$. La meilleure estimation $\widehat{\mathbf{x}}_i$ de $\mathbf{x}_i$ dans ce sous-espace vectoriel est la projection de $\mathbf{x}_i$ sur $\mathbf{M}$:

$$\widehat{\mathbf{x}'}_i = \mathbf{x'}_i \mathbf{M}(\mathbf{M'M})^{-1}\mathbf{M'} \tag{1.8}$$

Les $R+3$ scores de $\widehat{\mathbf{x}}_i$ dans ce sous-espace vectoriel, soit α_i, $\alpha_{i,1}$,...$\alpha_{i,R-1}$, β_i, γ_i et δ_i, sont regroupés dans un vecteur $\mathbf{t}_i$ vérifiant :

$$\widehat{\mathbf{x}'}_i = \mathbf{t'}_i \mathbf{M'} \tag{1.9}$$

Il est immédiat que :

$$\mathbf{t'}_i = \mathbf{x'}_i \mathbf{M}(\mathbf{M'M})^{-1} \tag{1.10}$$

En conclusion l'EMSC utilise de l'information experte : la matrice $\mathbf{\Lambda}$ qui modélise des déformations polynomiales de ligne de base ; le spectre moyen $\mathbf{m}$; la matrice $\mathbf{K}$ des spectres purs des composés chimiques présents dans l'échantillon. L'EMSC utilise aussi de l'information expérimentale représentée par les coefficients α_i, β_i, γ_i et δ_i déduits de $\mathbf{x}_i$ par une transformation qui dépend de chaque échantillon $\mathbf{x}_i$.

1.3.5 Conclusion sur les prétraitements

Les prétraitements identifient une matrice ou un vecteur nommés $\mathbf{P}$, $\mathbf{\Lambda}$ ou $\mathbf{1}_P$ selon les cas. Le cas général (à l'exception de Savitsky-Golay) est que cette matrice représente l'information à éliminer. L'information spectrale corrigée est le résidu d'une projection orthogonale sur cette matrice.

1.4 Discussion

Vue sous l'angle de la gestion globale de l'information, cette revue bibliographique fait apparaitre des convergences entre les différentes méthodes.

1.4.1 Elimination d'information spectrale

Etalonnages comme prétraitements enlèvent de l'information à la matrice $\mathbf{X}$, donnant $\mathbf{X}^*$. Cette élimination peut être dure (Hard) ou douce (Soft) selon [22].

- Hard Correction : des dimensions sont enlevées à $\mathbb{R}^P$. Toute l'information contenue dans les dimensions enlevées est réduite à néant et ne se retrouve plus dans $\mathbf{X}^*$, d'où le terme Hard. Les prétraitements ont une approche directe puisqu'ils identifient et enlèvent les dimensions non souhaitées. Ainsi l'EPO et l'OSC réalisent une projection de $\mathbf{X}$ orthogonalement à une matrice $\mathbf{P}$ dont les colonnes forment une base de l'espace à éliminer. Detrend projette $\mathbf{X}$ orthogonalement à une matrice modélisant les polynomes. L'EMSC identifie l'information spectrale à éliminer puis la soustrait des données initiales $\mathbf{X}$. Les étalonnages ont une approche indirecte : ils identifient une base de l'espace à conserver. Ainsi toute l'information extérieure à cet espace est éliminée. Cette base est représentée par la matrice $\mathbf{P}$ pour des étalonnages inverses tels la PCR ou la PLSR. Elle est représentée par le vecteur $\mathbf{\Sigma k}$ pour un étalonnage direct, la DC. Il est à noter que $\mathbf{\Sigma}$ est une projection orthogonale, d'où une correction de $\mathbf{k}$ en $\mathbf{\Sigma k}$ qualifiée de Hard.
- Soft Correction : des dimensions de $\mathbb{R}^P$ sont pondérées par des coefficients non nuls. Ces dimensions conservent dans $\mathbf{X}^*$ au moins une partie de l'information qu'elles contenaient dans $\mathbf{X}$, d'où l'adjectif de Soft. Un premier exemple est donné par la GLS, la pondération Soft est dûe à la matrice $\mathbf{S}$ de poids. Un deuxième exemple est donné par la SBC, la pondération Soft est dûe à l'inverse d'une matrice de variance-covariance.

En conclusion, nous observons que les étalonnages et les prétraitements éliminent de manière plus ou moins radicale une partie de l'information mesurée, soit en projetant sur un sous-espace vectoriel, soit en déformant l'espace d'origine par une métrique adaptée.

1.4.2 Informations spectrales utiles, nuisibles, neutres

Quatre natures d'informations peuvent être identifiées dans un spectre. L'information spectrale utile contenue dans le sous-espace vectoriel $\mathcal{E}^U$ est l'information utile à la construction d'un modèle d'étalonnage. L'information spectrale nuisible contenue dans le sous-espace vectoriel $\mathcal{E}^N$ est l'information nuisible à la construction d'un modèle d'étalonnage. L'information spectrale résiduelle contenue dans le sous-espace vectoriel $\mathcal{E}^R$ est l'information de $\mathbb{R}^P$ qui n'est ni utile, ni nuisible. Ces sous-espaces vectoriels sont représentés par la figure 1.1. L'information résiduelle est

orthogonale aux informations utiles et nuisibles. Enfin l'information spectrale inutile contenue dans le sous-espace vectoriel $\mathcal{E}^I$ est l'information inutile à la construction d'un modèle d'étalonnage, c'est à dire que $\mathcal{E}^U \oplus \mathcal{E}^I = \mathbb{R}^P$. L'information inutile est orthogonale à l'information utile. Bien évidemment une même information peut appartenir à plusieurs sous-espaces vectoriels. Il

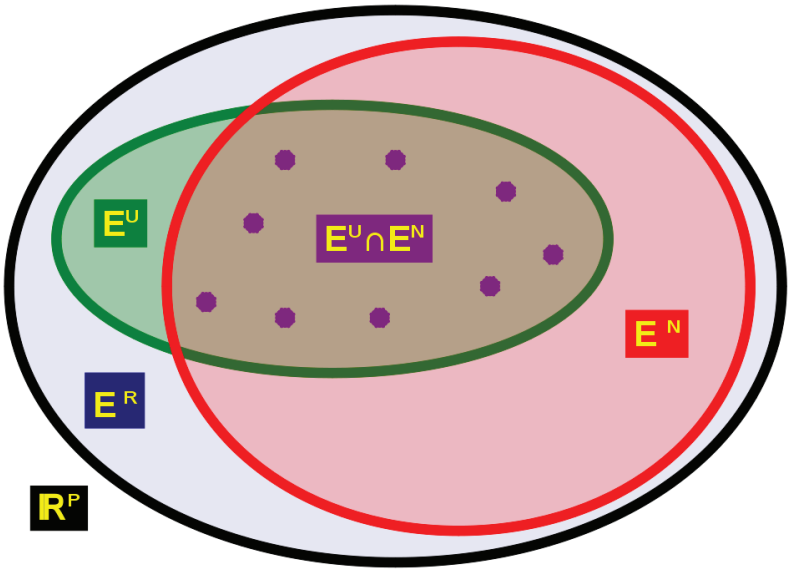


FIG. 1.1 – Représentation dans $\mathbb{R}^P$ (noir) des espaces utiles (vert), nuisibles (rouge), résiduels (bleu) et l'intersection des espaces utiles et nuisibles (violet)

est important de noter que l'intersection entre $\mathcal{E}^U$ et $\mathcal{E}^N$, notée $\mathcal{E}^U \cap \mathcal{E}^N$, n'est pas nulle, alors que l'intersection entre $\mathcal{E}^U$ et $\mathcal{E}^I$ est nulle.

Les différentes méthodes ont chacune leur approche dans la gestion de ces sous-espaces vectoriels. Les étalonnages identifient $\mathcal{E}^U - (\mathcal{E}^U \cap \mathcal{E}^N)$ au moyen d'une base, les vecteurs de $\mathbf{P}$ pour les étalonnages inverses, le vecteur $\mathbf{\Sigma k}$ pour les étalonnages directs. Certains prétraitements identifient $\mathcal{E}^I$, qu'ils enlèvent de $\mathbb{R}^P$. Deux exemples sont donnés par l'OSC et le NAP. D'autres prétraitements identifient $\mathcal{E}^N$ et l'enlèvent. Trois exemples sont donnés par l'EPO, l'EMSC et Detrend. Les performances des prétraitements peuvent dès lors être expliquées par la nature de l'information enlevée. Ainsi il a été montré que le NAP est équivalent à l'OSC [14], et que l'OSC n'améliore pas les performances de la PLSR ([23],[24]). L'OSC conduit au même sous-espace vectoriel que la PLSR puisqu'elle enlève de l'information inutile, ce qui explique pourquoi les

combinaisons OSC-PLSR ou OPLSR par exemple ont les mêmes performances que la PLSR seule. Au contraire, les pétraitements de type EPO, EMSC, Detrend sont performants puisqu'ils enlèvent de l'information nuisible, sans se limiter à de l'information inutile.

1.5 Conclusion

Cette revue bibliographique a permis d'analyser comment les méthodes d'étalonnage et de prétraitement les plus courantes gèrent les informations expérimentales et expertes. Une première analyse fait ressortir qu'un point commun aux différentes méthodes consiste à réduire l'information globale, en s'appuyant sur les notions d'espaces utiles et nuisibles. Une deuxième analyse montre que les méthodes identifient les espaces utiles et nuisibles soit avec l'information expérimentale, soit avec l'information experte, très peu de modèles utilisent les deux. Une troisième analyse fait ressortir de nombreux points communs entre les différentes méthodes, comme l'utilisation de sous-espaces vectoriels, de projections orthogonales, de métriques.

Nous allons donc étudier dans un cadre unifié l'utilisation des informations utiles et nuisibles. Une meilleure compréhension du fonctionnement commun aux étalonnages et prétraitements permettra dès lors de concevoir de nouvelles stratégies aptes à caractériser au mieux les espaces utiles et nuisibles à partir de la complémentarité entre informations expérimentales et expertes.

Chapitre 2

Un modèle linéaire général d'étalonnage et de prétraitement

Sommaire

2.1 Théorie du modèle général 28

2.2 Calcul des étalonnages et prétraitements 28

2.2.1 Calcul des étalonnages 28

2.2.2 Calcul des prétraitements 30

2.3 Validation de l'insertion des étalonnages et prétraitements dans le modèle général 30

2.3.1 Modèle général et étalonnages. 30

2.3.2 Modèle général et prétraitements. 31

2.3.3 Le cas de la PLSR 32

2.3.4 Le cas de l'OSC 35

2.4 Discussion et conclusion 37

Un modèle général regroupe la plupart des modèles d'étalonnage et de prétraitement. Les informations expérimentales et expertes sont utilisées pour deux objectifs. En premier lieu, elles servent à définir le sous-espace vectoriel contenant l'information utile ou nuisible, dont une base est constituée par les colonnes d'une matrice $\mathbf{P}$. En second lieu, elles peuvent contribuer à définir une notion de distance qui avantage l'information utile au détriment de l'information nuisible. Cette distance est associée à une métrique ou pseudo-métrique $\mathbf{\Sigma}$. Un formalisme commun aux étalonnages directs, inverses et aux prétraitements est ainsi proposé. L'information spectrale utile

ou nuisible est obtenue par projection des spectres sur l'information utile ou nuisible ($\mathbf{P}$) selon la métrique $\mathbf{\Sigma}$ définie dans $\mathbb{R}^P$.

2.1 Théorie du modèle général

Soit $\mathbf{P}$ une matrice $(P \times A)$ dont les A colonnes définissent une base d'un sous-espace vectoriel de l'espace vectoriel $\mathbb{R}^P$ muni d'une métrique ou pseudo-métrique $\mathbf{\Sigma}$. Soit $\mathbf{X}$ une matrice de spectres. L'information de $\mathbf{X}$ relative à $\mathbf{P}$ est notée $\mathbf{X}^{U/N}$ selon la nature utile / nuisible de $\mathbf{P}$. Elle est obtenue par projection orthogonale de $\mathbf{X}$ sur $\mathbf{P}$ avec la métrique $\mathbf{\Sigma}$ soit :

$$\mathbf{X}^{U/N} = \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'}$$

L'objectif des étalonnages et prétraitements est de transformer $\mathbf{X}$ en $\mathbf{X}^*$ telle que l'accès à l'information utile soit plus facile dans $\mathbf{X}^*$ que dans $\mathbf{X}$. Deux cas se présentent selon la nature utile ou nuisible de l'information contenue dans $\mathbf{P}$:

- si $\mathbf{P}$ contient de l'information utile, l'information extraite par la projection de $\mathbf{X}$ sur $\mathbf{P}$ est l'information utile de $\mathbf{X}$, elle est conservée :

$$\mathbf{X}^* = \mathbf{X}^U = \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'} \tag{2.1}$$

  $\mathbf{X}^*$ est la projection de $\mathbf{X}$ sur $\mathbf{P}$ selon $\mathbf{\Sigma}$;
- si $\mathbf{P}$ contient de l'information nuisible, l'information extraite par la projection de $\mathbf{X}$ sur $\mathbf{P}$ est l'information nuisible de $\mathbf{X}$, elle est enlevée de $\mathbf{X}$:

$$\mathbf{X}^* = \mathbf{X} - \mathbf{X}^N = \mathbf{X}(\mathbf{I}_P - \mathbf{\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'}) \tag{2.2}$$

  $\mathbf{X}^*$ est la projection de $\mathbf{X}$ orthogonalement à $\mathbf{P}$ selon $\mathbf{\Sigma}$.

2.2 Calcul des étalonnages et prétraitements

Etalonnages et prétraitements se différencient principalement selon la nature des informations contenues dans les paramètres $\mathbf{P}$ et $\mathbf{\Sigma}$.

2.2.1 Calcul des étalonnages

Pour les étalonnages, les vecteurs-colonne de $\mathbf{P}$ décrivent un sous-espace vectoriel. Chaque méthode a sa manière propre de calculer $\mathbf{P}$ et $\mathbf{\Sigma}$ à partir des entrées disponibles. Trois formes

de matrices $\mathbf{\Sigma}$ ont été identifiées : (1) la matrice identité ; (2) une projection orthogonale ; (3) une pondération de type Mahalanobis. Une fois $\mathbf{P}$ et $\mathbf{\Sigma}$ connues, les scores ou coordonnées $\mathbf{T}$ des individus dans la base formée par les colonnes de $\mathbf{P}$ sont déduits de la formule 2.1. Ainsi :

$$\mathbf{T} = \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1} \tag{2.3}$$

Une régression aux moindres carrés, ou projection de $\mathbf{y}$ sur $\mathbf{T}$ donne une estimation $\widehat{\mathbf{y}}$ de $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{T}(\mathbf{T'T})^{-1}\mathbf{T'y} \tag{2.4}$$

d'où le modèle d'étalonnage représenté par le vecteur des b-coefficients :

$$\mathbf{b} = \mathbf{\Sigma P}(\mathbf{P'\Sigma P})^{-1}[(\mathbf{P'\Sigma P})^{-1}\mathbf{P'\Sigma X'X\Sigma P}(\mathbf{P'\Sigma P})^{-1}]^{-1}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'\Sigma X'y} \tag{2.5}$$

Cette formule est utilisée indifféremment par les étalonnages directs ou inverses.

- Les étalonnages inverses aussi appelés régressions utilisent comme entrées une information expérimentale sous forme d'un jeu d'étalonnage $(\mathbf{X}, \mathbf{y})$, éventuellement complété d'un vecteur $\mathbf{r}$. Chaque méthode d'étalonnage inverse utilise ces entrées à sa façon de manière à déterminer les paramètres $\mathbf{\Sigma}$ et $\mathbf{P}$. Le modèle est obtenu en appliquant les équations 2.3 et 2.4.
- Les étalonnages directs utilisent comme entrées une information experte, le spectre pur $\mathbf{k}$ de la grandeur d'intérêt, associée à la valeur 1 représentant la fraction massique, volumique ou molaire de la grandeur d'intérêt dans le spectre pur (ce point est discuté page 55). Les matrices $\mathbf{X}$ et $\mathbf{y}$ sont remplacées respectivement par $\mathbf{k'}$ et 1. D'autres entrées sont constituées d'informations expertes (spectres purs de grandeurs d'influence) ou expérimentales (spectres suivant un plan d'expérience), elles permettent à chaque étalonnage direct de déterminer le paramètre $\mathbf{\Sigma}$. Le paramètre $\mathbf{P}$ est déterminé par l'opérateur, c'est toujours un vecteur $\alpha\mathbf{k}$ avec α un coefficient dépendant de l'unité. Lorsque l'unité est une fraction (molaire, volumique ou massique), la prédiction de la grandeur d'intérêt dans le spectre pur doit donner 1, ce qui conduit à prendre : $\mathbf{P} = \mathbf{k}(\mathbf{k'\Sigma k})^{-1}$. Ainsi l'équation 2.5 se simplifie et conduit à la formule commune à toutes les méthodes d'étalonnage direct :

$$\mathbf{b} = \mathbf{\Sigma k}(\mathbf{k'\Sigma k})^{-1} \tag{2.6}$$

2.2.2 Calcul des prétraitements

Les prétraitements utilisent $\mathbf{P}$ pour identifier puis enlever l'information nuisible selon l'équation 2.2

$$\mathbf{X}^* = \mathbf{X}(\mathbf{I}_P - \mathbf{\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'}) \tag{2.7}$$

2.3 Validation de l'insertion des étalonnages et prétraitements dans le modèle général

La plupart des méthodes d'étalonnage ou prétraitement décrites au chapitre 1 sont expliquées par le modèle linéaire général. Les cas de la PLSR et de l'OSC sont traités à part du fait d'une démonstration plus complexe.

2.3.1 Modèle général et étalonnages.

Les principales méthodes d'étalonnage direct ou indirect sont revues sous l'angle du modèle linéaire général.

Application aux methodes d'étalonnage indirect

Toutes les méthodes indirectes ou de régression utilisent un jeu d'étalonnage $(\mathbf{X}, \mathbf{y})$. Les différences portent sur la manière de déterminer l'information utile $\mathbf{P}$ et d'utiliser l'information nuisible pour calculer $\mathbf{\Sigma}$.

L'OLSR A partir de $\mathbf{X}$, aucune hypothese ni calcul n'est fait pour déterminer l'information utile ou l'information nuisible, donc par defaut : $\mathbf{P} = \mathbf{I}$ et $\mathbf{\Sigma} = \mathbf{I}$. Ainsi $\mathbf{T} = \mathbf{X}$ et la formule 2.5 des b-coefficients correspond à celle de l'OLS :

$$\mathbf{b}_{OLS} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

L'OLS est expliquée par le modéle général.

La Principal Component Regression Une décomposition en valeurs singulières de $\mathbf{X}$ donne $\mathbf{P}_{PCA}$ la matrice (P, A) des A premiers vecteurs propres. Par defaut $\mathbf{\Sigma} = \mathbf{I}$. Ainsi la formule 2.3 donne $\mathbf{T} = \mathbf{XP}_{PCA}$ et la formule 2.5 des b-coefficients est celle de la PCR :

$$\mathbf{b}_{PCR} = \mathbf{P}_{PCA}[\mathbf{P'}_{PCA}\mathbf{X'XP}_{PCA}]^{-1}\mathbf{P'}_{PCA}\mathbf{X'y}$$

La PCR est bien expliquée par le modéle général.

Application aux methodes d'étalonnage direct

Nous avons vu que l'équation 2.6 est commune aux méthodes d'étalonnage direct. Pour mémoire :

$$\mathbf{b} = \mathbf{\Sigma k}(\mathbf{k}'\mathbf{\Sigma k})^{-1}$$

Cette formule est exactement celle identifiée pour les méthodes d'étalonnage direct, chapitre 1. Les différences entre méthodes portent sur la manière dont les grandeurs d'influence sont utilisées dans la construction de $\mathbf{\Sigma}$.

Ainsi, la DC et la SBC sont bien expliquées par le modèle général.

2.3.2 Modèle général et prétraitements.

Trois groupes de prétraitements ont été identifiés selon la nature expérimentale ou experte des informations.

Application aux prétraitements utilisant une information expérimentale issue d'un plan d'expérience

Les méthodes de projection orthogonale, soient IIR, EPO, TOP, DOP, EROS ont la même formule :

$$\mathbf{X}^U = \mathbf{X}(\mathbf{I} - \mathbf{P}\mathbf{P}')$$

Nous avons vu au chapitre 1 que $\mathbf{P}$ contient des vecteurs orthonormés, donc il est immédiat que la formule précédente correspond à l'équation 2.7 avec $\mathbf{\Sigma} = \mathbf{I}_P$. Ainsi les méthodes IIR, EPO, DOP, TOP, EROS appartiennnent bien au modèle général.

Application aux prétraitements utilisant comme information expérimentale un jeu d'étalonnage

Un jeu d'étalonnage sert à calculer la matrice $\mathbf{P}$ pour corriger les grandeurs d'influence.

Le Net Analyte Preprocessing La formule de correction par le Net Analyte Preprocessing [14] noté NAP :

$$\mathbf{X}^{NAP} = \mathbf{X}(\mathbf{I}_P - \mathbf{P}\mathbf{P}')$$

montre clairement son appartenance au modèle général. La différence avec les projections orthogonales réside dans un calcul différent pour la matrice $\mathbf{P}$, ici extraite du jeu d'étalonnage.

Application aux prétraitements utilisant conjointement des informations expérimentales et expertes

La SNV La première partie de la SNV, le centrage par ligne, se rattache au modèle général puisqu'il s'agit d'une projection orthogonale à l'information experte $\mathbf{1}_P$, en accord avec l'équation 2.7 du modèle général. Par contre la deuxième partie, la normalisation, ne s'y rattache apparemment pas.

L'EMSC Le calcul des coefficients de l'EMSC repose sur la projection d'un spectre $\mathbf{x}_i$ sur une matrice $\mathbf{M}$, voir équation 1.8. L'EMSC utilise le principe du modèle général pour le calcul des coefficients $\mathbf{t}_i$, voir équations 1.8, 1.9 et 1.10 page 22, mais s'en éloigne lors de l'utilisation de ces coefficients pour réaliser les corrections spectrales selon l'équation 1.7 page 21.

2.3.3 Le cas de la PLSR

L'appartenance de la PLSR au modèle général est beaucoup plus complexe que pour les autres méthodes, c'est pourquoi elle fait l'objet d'une partie indépendante. L'algorithme NIPALS a été choisi comme support de la démonstration car cet algorithme est souvent pris comme référence lors de comparaisons de différents algorithmes de PLSR. Une nouvelle propriété de NIPALS est proposée : le calcul de $\mathbf{T}$ à partir des matrices $\mathbf{X}$, $\mathbf{P}$ et $\mathbf{\Sigma} = (\mathbf{X}'\mathbf{X})^+$ (pseudo-inverse au sens de Moore-Penrose).

L'algorithme NIPALS comprend d'abord une phase d'initialisation : $\mathbf{X}_0 = \mathbf{X}$ et $\mathbf{y}_0 = \mathbf{y}$, ensuite une boucle calcule les paramètres de la PLSR à chaque itération. Pour $i = 1, 2, 3, ..A$:

$$\mathbf{w}_i = \mathbf{X}'_{i-1}\mathbf{y}_{i-1}({\mathbf{y}'}_{i-1}\mathbf{y}_{i-1})^{-1} \tag{2.8}$$

$$\| \mathbf{w}_i \| = 1 \tag{2.9}$$

$$\mathbf{t}_i = \mathbf{X}_{i-1}\mathbf{w}_i \tag{2.10}$$

$$c_i = \mathbf{y}'_{i-1}\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1} \tag{2.11}$$

$$\mathbf{p}_i = \mathbf{X}'_{i-1}\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1} \tag{2.12}$$

$$\mathbf{X}_i = \mathbf{X}_{i-1} - \mathbf{t}_i\mathbf{p}'_i = (\mathbf{I}_N - \mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i)\mathbf{X}_{i-1} \tag{2.13}$$

$$\mathbf{y}_i = \mathbf{y}_{i-1} - \mathbf{t}_i\mathbf{c}'_i = (\mathbf{I}_N - \mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i)\mathbf{y}_{i-1} \tag{2.14}$$

puis retour à l'équation (2.8) en incrémentant i de 1.

Expression de $\mathbf{t}_i$ fonction de $\mathbf{X}$, $\mathbf{p}_i$ et $\mathbf{\Sigma}$

Avec les notations du tableau 1.1 page 9, l'équation 2.12 s'écrit :

$$\mathbf{p}_i = \mathbf{X}'\mathcal{P}^{\perp}_{1:i-1}\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}$$

Et comme $\mathbf{t}_i$ est déjà orthogonal à $\{\ \mathbf{t}_1 ... \mathbf{t}_{i-1}\}$, le projecteur $\mathcal{P}^{\perp}_{1:i-1}$ est inutile ce qui permet de simplifier, comme proposé également par [25] :

$$\mathbf{p}_i = \mathbf{X}'\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1} \tag{2.15}$$

Par ailleurs, calculons le produit $\mathbf{p}'_i\mathbf{\Sigma}\mathbf{p}_i$:

$$\mathbf{p}'_i\mathbf{\Sigma}\mathbf{p}_i = (\mathbf{t}'_i\mathbf{t}_i)^{-1}\mathbf{t}'_i\mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{t}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1} \tag{2.16}$$

Le terme $\mathbf{X}(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'$ est égal à $\mathbf{X}\mathbf{X}^{+}$ le projecteur orthogonal sur $\mathbf{X}$ ([26]). Soit $\mathcal{E}_X$ le sous-espace vectoriel de $\mathbb{R}^N$ décrit par les colonnes de $\mathbf{X}$. Montrons par récurrence que pour tout i, $\mathbf{t}_i$ et les vecteurs-colonne de $\mathbf{X}_{i-1}$ appartiennent à $\mathcal{E}_X$. La relation est vraie au rang 1 puisque $\mathbf{X}_0 = \mathbf{X}$ et que $\mathbf{t}_1$ est une combinaison linéaire des colonnes de $\mathbf{X}$. Supposons la relation vraie au rang i, pour $\mathbf{X}_{i-1}$ et $\mathbf{t}_i$. Alors $\mathbf{t}_i\mathbf{p}'_i$ est une matrice dont tous les vecteurs-colonne sont proportionnels à $\mathbf{t}_i$, donc ils appartiennent tous à $\mathcal{E}_X$. Ainsi, selon l'équation 2.13, les vecteurs-colonne de $\mathbf{X}_i$ appartiennent à $\mathcal{E}_X$. L'équation 2.10 montre alors que $\mathbf{t}_{i+1}$ appartient à $\mathcal{E}_X$. La récurrence est vérifiée au rang $i+1$.

En conclusion pour tout i le vecteur $\mathbf{t}_i$ appartient à $\mathcal{E}_X$, le sous-espace vectoriel défini par les colonnes de $\mathbf{X}$. Par conséquent la projection de $\mathbf{t}_i$ sur $\mathbf{X}$ donne $\mathbf{t}_i$. L'équation 2.16 se simplifie et donne après réarrangement des termes :

$$(\mathbf{t}'_i\mathbf{t}_i)^{-1} = \mathbf{p}'_i\mathbf{\Sigma}\mathbf{p}_i \tag{2.17}$$

Reprenons l'équation 2.15 et multiplions de chaque coté à gauche par $\mathbf{X\Sigma}$ soit $\mathbf{X}(\mathbf{X'X})^+$:

$$\mathbf{X\Sigma p}_i = \mathbf{X}(\mathbf{X'X})^+\mathbf{X't}_i(\mathbf{t}'_i\mathbf{t}_i)^{-1}$$

Comme vu précédemment, cette équation se simplifie et se réarrange :

$$\mathbf{t}_i = \mathbf{X\Sigma p}_i(\mathbf{t}'_i\mathbf{t}_i)$$

La relation 2.17 permet d'exprimer $\mathbf{t}_i$ en fonction de $\mathbf{p}_i$, $\mathbf{X}$ et $\mathbf{\Sigma}$:

$$\mathbf{t}_i = \mathbf{X\Sigma p}_i(\mathbf{p'}_i\mathbf{\Sigma p}_i)^{-1} \tag{2.18}$$

Montrons également que les $\mathbf{p}_i$ sont strictement orthogonaux entre eux au sens de $\mathbf{\Sigma}$. Si $i \neq j$, $\mathbf{t}_i$ et $\mathbf{t}_j$ sont orthogonaux au sens Euclidien, leur produit scalaire est nul, donc en utilisant l'équation 2.18 :

$$\mathbf{p'}_i\mathbf{\Sigma X'X\Sigma p}_j = 0$$

Comme $\mathbf{\Sigma} = (\mathbf{X'X})^+$ et d'après une des quatre propriétés d'une pseudo-inverse au sens de Moore-Penrose : $\mathbf{\Sigma X'X\Sigma} = \mathbf{\Sigma}$, l'équation précédente se simplifie ainsi :

$$\mathbf{p'}_i\mathbf{\Sigma p}_j = 0$$

En conclusion, si $i \neq j$, les vecteurs $\mathbf{p}_i$ et $\mathbf{p}_j$ sont strictement orthogonaux entre eux au sens de la métrique $\mathbf{\Sigma}$.

Expression de T en fonction de P et Σ

Puisque les $\mathbf{p}_i$ sont orthogonaux entre eux au sens de $\mathbf{\Sigma}$, la matrice $\mathbf{P'\Sigma P}$ est une matrice diagonale dont le terme de la i^{eme} ligne et de la i^{eme} colonne est $\mathbf{p'}_i\mathbf{\Sigma p}_i$. En conséquence l'équation 2.18 conduit à :

$$\mathbf{T} = \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1} \tag{2.19}$$

Cette expression est exactement celle du modèle général décrit au chapitre 2, équation 2.3.

Calcul des b-coefficients

Une fois la matrice $\mathbf{T}$ connue, le modèle est calculé par une régression aux moindres carrés de $\mathbf{y}$ sur $\mathbf{T}$, soit :

$$\widehat{\mathbf{y}} = \mathbf{T}(\mathbf{T'T})^{-1}\mathbf{T'y} \tag{2.20}$$

L'information contenue dans $\mathbf{X}$ et utilisée par la PLSR est celle contenue dans $\mathbf{TP'}$ ([7], que nous appellerons $\mathbf{X}^U$ définie ainsi :

$$\begin{aligned}\mathbf{X}^U &= \mathbf{TP'} \\ \mathbf{X}^U &= \mathbf{X\Sigma P}(\mathbf{P'\Sigma P})^{-1}\mathbf{P'}\end{aligned}$$

soit la projection de $\mathbf{X}$ sur $\mathbf{P}$ au sens de la métrique $\mathbf{\Sigma}$, toujours en accord avec le modèle général.

2.3.4 Le cas de l'OSC

Plusieurs algorithmes de calcul de l'OSC ont été proposés depuis [13]. L'approche directe [15] a été choisie pour son lien direct avec l'algorithme NIPALS de la PLSR. L'OSC identifie l'information nuisible $\mathbf{X}^N$ dans l'espace des variables par projection de $\mathbf{X}$ orthogonalement à $\mathbf{X'y}$:

$$\mathbf{X}^N = \mathbf{X}(\mathbf{I}_P - \mathbf{X'y}(\mathbf{y'XX'y})^{-1}\mathbf{y'X})$$

Les vecteurs $\mathbf{w}_i$ sont les A premiers vecteurs propres d'une SVD sur $\mathbf{X}^N$. Les vecteurs $\mathbf{p}_i$ sont déduits de $\mathbf{w}_i$ [15] :

$$\mathbf{p}_i = \mathbf{X'Xw}_i(\mathbf{w'}_i\mathbf{X'Xw}_i)^{-1} \tag{2.21}$$

La correction par OSC [15] est obtenue par la formule :

$$\mathbf{X}^{OSC} = \mathbf{X} - \sum_{i=1}^{A} \mathbf{t}_i\mathbf{p'}_i$$

et sachant que $\mathbf{t}_i = \mathbf{Xw}_i$ [15] :

$$\mathbf{X}^{OSC} = \mathbf{X} - \mathbf{X}\sum_{i=1}^{A} \mathbf{w}_i\mathbf{p'}_i$$

Il a été montré par [15] que les vecteurs $\mathbf{w}_i$ sont orthogonaux entre eux et avec les $\mathbf{p}_j$, $i \neq j$, d'où :

$$\mathbf{X}^{OSC} = \mathbf{X}(\mathbf{I}_P - \mathbf{WP'})$$

Posons $\mathbf{\Sigma} = (\mathbf{X'X})^+$. Grâce à l'équation 2.21, il devient possible d'exprimer $\mathbf{w}_i$ en fonction de $\mathbf{p}_i$ et $\mathbf{\Sigma}$. Calculons d'abord le produit $\mathbf{p'}_i\mathbf{\Sigma p}_i$:

$$\mathbf{p'}_i\mathbf{\Sigma p}_i = (\mathbf{w'}_i\mathbf{X'Xw}_i)^{-1}\mathbf{w'}_i\mathbf{X'X\Sigma X'Xw}_i(\mathbf{w'}_i\mathbf{X'Xw}_i)^{-1}$$

$$\begin{aligned}
\mathbf{p'}_i\boldsymbol{\Sigma}\mathbf{p}_i &= (\mathbf{t'}_i\mathbf{t}_i)^{-1}\mathbf{w'}_i\mathbf{X'}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X'}\mathbf{X}\mathbf{w}_i(\mathbf{t'}_i\mathbf{t}_i)^{-1} \\
&= (\mathbf{t'}_i\mathbf{t}_i)^{-1}\mathbf{w'}_i\mathbf{X'}\mathbf{X}\mathbf{w}_i(\mathbf{t'}_i\mathbf{t}_i)^{-1} \\
&= (\mathbf{t'}_i\mathbf{t}_i)^{-1}\mathbf{t'}_i\mathbf{t}_i(\mathbf{t'}_i\mathbf{t}_i)^{-1} \\
&= (\mathbf{t'}_i\mathbf{t}_i)^{-1}
\end{aligned}$$

Ainsi d'après l'équation 2.21 :

$$\mathbf{X'}\mathbf{X}\mathbf{w}_i = \mathbf{p}_i(\mathbf{p'}_i\boldsymbol{\Sigma}\mathbf{p}_i)^{-1}$$

Après multiplication à gauche par $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}\mathbf{X'}\mathbf{X}\mathbf{w}_i = \boldsymbol{\Sigma}\mathbf{p}_i(\mathbf{p'}_i\boldsymbol{\Sigma}\mathbf{p}_i)^{-1}$$

Supposons maintenant que la SVD de $\mathbf{X}$ s'écrit avec les matrices $\mathbf{U}$, $\mathbf{D}$ et $\mathbf{V}$ avec $\mathbf{U'U} = \mathbf{V'V} = \mathbf{I}$ et $\mathbf{D}$ diagonale sans termes nuls. Par convention, $\mathbf{D}^2$ est obtenue en portant tous les termes de la diagonale de $\mathbf{D}$ au carré, $\mathbf{D}^{-2}$ est obtenue en inversant les termes de la diagonale de $\mathbf{D}^2$. Les relations suivantes sont successivement déduites :

$$\begin{aligned}
\mathbf{X} &= \mathbf{UDV'} \\
\mathbf{X'X} &= \mathbf{VD}^2\mathbf{V'} \\
(\mathbf{X'X})^+ &= \mathbf{VD}^{-2}\mathbf{V'} \\
(\mathbf{X'X})^+\mathbf{X'X} &= \mathbf{VV'}
\end{aligned}$$

Le terme $\boldsymbol{\Sigma}\mathbf{X'}\mathbf{X}\mathbf{w}_i$ est donc la projection de $\mathbf{w}_i$ sur les vecteurs-propres de $\mathbf{X}$. Par ailleurs soit $\mathcal{F}_X$ le sous-espace vectoriel de $\mathbb{R}^P$ décrit par les vecteurs-ligne de $\mathbf{X}$. Le vecteur $\mathbf{X'y}$ est une combinaison linéaire des lignes de $\mathbf{X}$, donc il appartient à $\mathcal{F}_X$. Comme $\mathbf{X}^N$ est obtenue par projection de $\mathbf{X}$ orthogonalement à un vecteur de $\mathcal{F}_X$, les lignes de $\mathbf{X}^N$ décrivent un sous-espace vectoriel inclus dans $\mathcal{F}_X$. Les vecteurs $\mathbf{w}_i$ étant vecteurs propres de $\mathbf{X}^N$, ils appartiennent à $\mathcal{F}_X$, dont une base est constituée des vecteurs de $\mathbf{V}$. En conclusion : $\boldsymbol{\Sigma}\mathbf{X'}\mathbf{X}\mathbf{w}_i = \mathbf{w}_i$ et conduit aux relations recherchées :

$$\begin{aligned}
\mathbf{w}_i &= \boldsymbol{\Sigma}\mathbf{p}_i(\mathbf{p'}_i\boldsymbol{\Sigma}\mathbf{p}_i)^{-1} \\
\mathbf{W} &= \boldsymbol{\Sigma}\mathbf{P}(\mathbf{P'}\boldsymbol{\Sigma}\mathbf{P})^{-1}
\end{aligned}$$

Il est à noter que ces relations, vraies pour l'OSC, ne sont absolument pas vérifiées par la PLSR. La matrice $\mathbf{X}^{OSC}$ corrigée par OSC s'exprime en fonction de $\mathbf{X}$, $\boldsymbol{\Sigma}$ et $\mathbf{P}$:

$$\mathbf{X}^{OSC} = \mathbf{X}(\mathbf{I}_P - \boldsymbol{\Sigma}\mathbf{P}(\mathbf{P'}\boldsymbol{\Sigma}\mathbf{P})^{-1}\mathbf{P'})$$

Nous retrouvons exactement la formule 2.7. En conséquence l'OSC se rattache bien au modèle général.

2.4 Discussion et conclusion

La plupart des méthodes décrites au chapitre 1 se rattachent au modèle général, voir tableau 2.1.

Etalonnages directs	DC, SBC
Etalonnages inverses	OLSR, PCR, PLSR
Prétraitements	IIR, EPO, TOP, DOP, EROS OSC, NAP, Detrend

TAB. 2.1 – Méthodes d'étalonnage et prétraitements rattachés au modèle général

Les méthodes d'étalonnage et prétraitement consacrées par l'expérience comme les plus performantes sont entièrement expliqués par le modèle général. Ce modèle linéaire a donc une signification forte et incontournable dans la construction des étalonnages et des prétraitements.

Au delà d'une classification des méthodes existantes, ce modèle général ouvre deux axes d'innovation.

- **Indépendance entre $\mathbf{\Sigma}$ et $\mathbf{P}$**

  Le modèle général ne contient pas de contrainte explicite reliant $\mathbf{\Sigma}$ et $\mathbf{P}$. Pourtant pour les deux principales méthodes de régression, la PCR comme la PLSR, les vecteurs $\mathbf{p}_i$ sont orthogonaux entre eux au sens de $\mathbf{\Sigma}$. Cette propriété n'apparait pas indispensable, toutefois elle pourrait avoir l'avantage de simplifier les calculs.

  La figure 2.1 donne un aperçu des couples $(\mathbf{\Sigma}, \mathbf{P})$ formés par les modèles d'étalonnage et de prétraitement précedemment décrits. Si nous partons sur le postulat qu'il n'y a pas de contrainte, alors il devient possible de créer de nouveaux modèles à partir des équations 2.3 et 2.4 page 29 en utilisant $\mathbf{P}$ indépendemment de $\mathbf{\Sigma}$. Par exemple, un premier modèle utiliserait $\mathbf{P}$ de la PLSR et $\mathbf{\Sigma}$ de la PCR soit l'identité. Un deuxième modéle utiliserait $\mathbf{P}$ de la PCR et $\mathbf{\Sigma}$ de la PLSR soit $(\mathbf{X}'\mathbf{X})^{+}$. De nouveaux modèles de régression pourraient être obtenus par des choix indépendants de $\mathbf{P}$ et $\mathbf{\Sigma}$.
- **Gestion des informations expérimentales et expertes** Le modèle général montre bien l'importance de décrire les sous-espaces vectoriels utiles et nuisibles pour les étalonnages

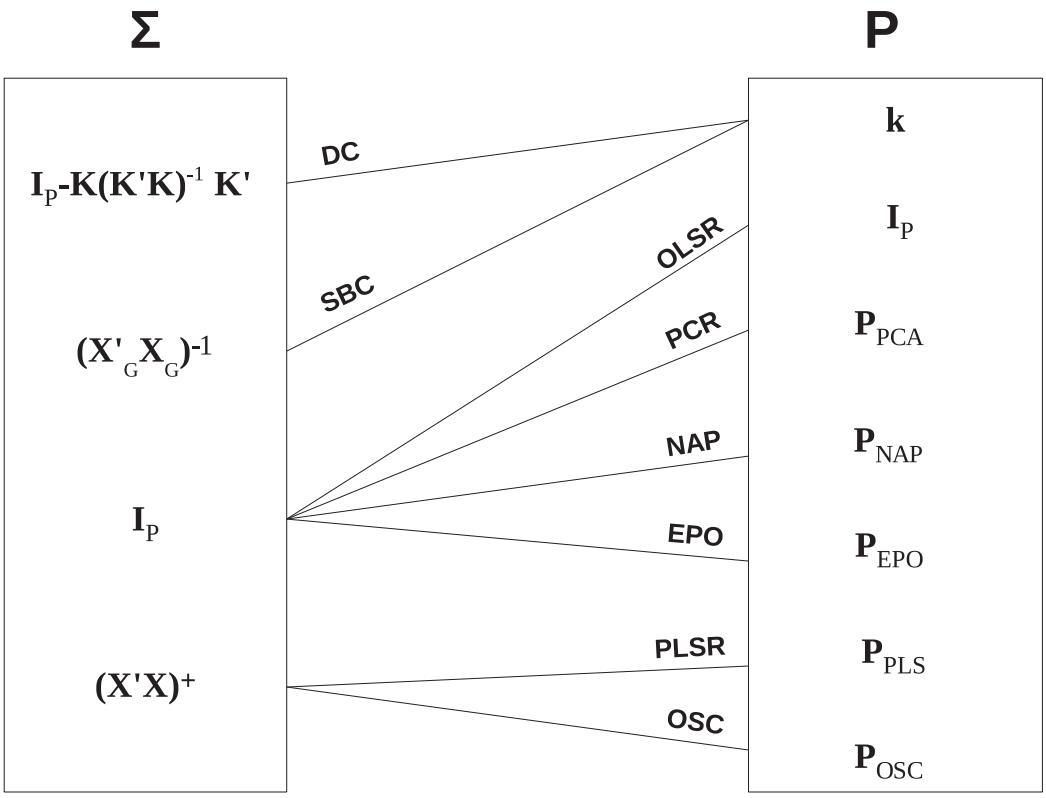


FIG. 2.1 – Etalonnages et prétraitements : couples $(\mathbf{P}, \mathbf{\Sigma})$ décrits dans la littérature

et prétraitements respectivement. Il introduit aussi l'importance de la métrique utilisée pour les projections orthogonales. Il n'y a plus de frontière théorique entre informations expérimentales et expertes : les unes comme les autres ont la même fonction et la même capacité pour décrire une base d'un sous-espace vectoriel et une métrique.

Nous n'avons pas étudié le premier axe, la gestion indépendante des matrices $\mathbf{\Sigma}$ et $\mathbf{P}$. Nous avons choisi de privilégier le second axe, un point faible mis en évidence par l'état de l'art : la sous-représentation des informations expertes. Ainsi deux nouvelles méthodes favorisant l'utilisation d'informations expertes sont proposées, elles viennent implémenter le modèle général.

Chapitre 3

Première implémentation : IDC, une nouvelle méthode d'étalonnage direct

Sommaire

3.1 Théorie de l'IDC 42

3.2 Premier exemple d'application de l'IDC : quantification de l'éthanol en fermentation 44

3.2.1 Matériels et méthodes 44

3.2.2 Résultats 45

3.2.3 Conclusion sur la première application de l'IDC 50

3.3 Deuxième exemple d'application de l'IDC : analyse des parois de la couche à aleurones du grain de blé 53

3.3.1 Matériels et méthodes 53

3.3.2 Résultats 54

3.3.3 Conclusion sur la deuxième application de l'IDC 55

3.4 Discussion 55

3.4.1 Les fondements spectroscopiques de l'étalonnage direct 55

3.5 Conclusion sur l'IDC 61

Les deux méthodes d'étalonnage direct proposées par la littérature, la DC et la SBC, utilisent toutes deux une information experte, le spectre pur $\mathbf{k}$ de la grandeur d'intérêt à partir duquel la matrice $\mathbf{P}$ est construite. Les différences entre DC et SBC portent sur $\mathbf{\Sigma}$. La DC construit $\mathbf{\Sigma}$ avec de l'information experte, les spectres purs des grandeurs d'influence chimique. Cette approche

permet d'identifier le sous-espace vectoriel nuisible dû aux grandeurs d'influence chimiques. La SBC construit $\mathbf{\Sigma}$ avec de l'information expérimentale issue d'un plan d'expérience. Cette approche permet d'identifier le sous-espace vectoriel dû aux grandeurs d'influence physiques, du moins selon la présentation faite par [3].

L'IDC-Improved Direct Calibration est une nouvelle méthode qui cumule les avantages respectifs de la DC et de la SBC. Le principe est d'utiliser conjointement des informations expertes (spectres purs) et des informations expérimentales (spectres issus d'un plan d'expérience) afin de déterminer au mieux une base du sous-espace vectoriel nuisible dont les vecteurs formeront une matrice $\mathbf{R}$. Le calcul de $\mathbf{\Sigma}$ est le même que pour la DC, une projection orthogonale que l'on peut qualifier de Hard Correction. L'IDC est une DC comportant simplement plus d'informations, d'où le nom de la méthode.

3.1 Théorie de l'IDC

Rappelons le modèle linéaire de mélange vu au chapitre 1 page 8 au sujet de la DC :

$$\mathbf{x}' = y\mathbf{k}' + \mathbf{t}'_{\chi}\mathbf{K}' + \varepsilon' \tag{3.1}$$

oú ε est un vecteur de bruit. Ce modèle ne prend en compte que les grandeurs d'influence chimiques. Supposons maintenant que des grandeurs d'influence physiques induisent des perturbations spectrales structurées évoluant dans un sous-espace vectoriel dont une base est constitué par les colonnes d'une matrice $\mathbf{Q}$. L'équation 3.1 devient :

$$\mathbf{x}' = y\mathbf{k}' + \mathbf{t}'_{\chi}\mathbf{K}' + \mathbf{t}'_{\phi}\mathbf{Q}' + \varepsilon' \tag{3.2}$$

avec $\mathbf{t}_{\phi}$ le vecteur des coordonnées relatives aux grandeurs d'influence physiques.

La résolution de cette équation implique d'annuler l'effet des grandeurs d'influence chimiques et physiques, en pratique $\mathbf{K}$ et $\mathbf{Q}$. Une information experte permet d'obtenir $\mathbf{K}$, les spectres purs des grandeurs chimiques. Nous proposons d'utiliser une information expérimentale pour déterminer $\mathbf{Q}$. Les matrices $\mathbf{K}$ et $\mathbf{Q}$ sont alors jointes ensemble pour donner une matrice $\mathbf{R}$. Soit $\mathbf{\Sigma}_{IDC}$ le projecteur orthogonal à $\mathbf{R}$:

$$\mathbf{\Sigma}_{IDC} = (\mathbf{I} - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}')$$

Le modèle général est appliqué avec $\mathbf{k}$ et $\mathbf{\Sigma}_{IDC}$:

$$\widehat{y} = \mathbf{x}'\mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1} \tag{3.3}$$

ce qui est équivalent à multiplier l'équation 3.2 à droite par $\mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1}$, et à supposer que le bruit ε est suffisamment faible pour que $\varepsilon\mathbf{\Sigma}_{IDC}\mathbf{k}(\mathbf{k}'\mathbf{\Sigma}_{IDC}\mathbf{k})^{-1}$ puisse être négligé. Le modèle IDC peut être construit à partir du moment où $\mathbf{k}$, $\mathbf{K}$ et $\mathbf{Q}$ sont connues. Les matrices $\mathbf{k}$ et $\mathbf{K}$ sont données par la connaissance experte. Le moyen le plus simple d'identifier $\mathbf{Q}$ est de construire une matrice $\mathbf{X}_G$ ne contenant que de l'information expérimentale sur les grandeurs d'influence physiques, de la même manière que pour l'IIR ou l'EPO, chapitre 1. Une décomposition en valeurs singulières sur $\mathbf{X}_G$ donne $\mathbf{Q}$ de dimensions $(P \times A)$.

Le choix de la dimension A est une étape importante de la méthode. En théorie, comme il n'y a pas de variation de la grandeur d'intérêt dans $\mathbf{X}_G$, A devrait pouvoir prendre une valeur très grande de manière à capturer toute l'information nuisible. En pratique, ce serait dangereux car une information résiduelle sur la grandeur d'intérêt peut être présente dans $\mathbf{X}_G$ et être capturée par $\mathbf{P}$ si A est trop grand. D'un autre coté, si A est trop petit, toutes les grandeurs d'influence ne seront pas corrigées. Pour choisir A, il est possible d'utiliser un jeu de prédiction et d'examiner l'erreur de prédiction en fonction de A. Cependant, cette approche fait perdre l'avantage majeur de l'étalonnage direct qui est justement de ne pas nécessiter de base d'étalonnage. Nous proposons donc une autre approche qui consiste à appliquer le modèle IDC sur $\mathbf{X}_G$ (pour lequel la grandeur d'intérêt est constante) pour différentes valeurs de A, et d'examiner l'évolution de l'erreur de prédiction.

Si en théorie la matrice $\mathbf{K}$ doit contenir les spectres purs de tous les composés présents dans l'échantillon, en pratique cette approche se heurte à un certain nombre de difficultés. Ainsi dans des échantillons très complexes, certains produits purs ne peuvent pas être extraits et stabilisés en quantités suffisantes pour être mesurés. L'essentiel est que toute l'information relative aux grandeurs d'influence soit présente soit dans $\mathbf{X}_G$, soit dans $\mathbf{K}$. Une règle simple consiste à considérer qu'il faut mettre dans $\mathbf{K}$ les spectres purs des produits dont la concentration ne varie pas dans $\mathbf{X}_G$; ou bien, qu'il faut faire varier dans $\mathbf{X}_G$ les grandeurs d'influence non contenues dans $\mathbf{K}$.

3.2 Premier exemple d'application de l'IDC : quantification de l'éthanol en fermentation

3.2.1 Matériels et méthodes

L'étude portait sur le suivi de la fermentation alcoolique lors de la vinification. La spectroscopie proche infra-rouge a été utilisée pour quantifier la production d'éthanol. La base de données expérimentales était constituée par les spectres et valeurs de référence en éthanol de 1480 échantillons de moûts et vins plus ou moins fermentés. Les spectres avaient été acquis sur un spectrophotomètre Jasco (trajet optique $1mm$, plage $500-2500nm$, pas d'acquisition de $2nm$, référence eau) aux établissements Skalli-Fortant de France (Sète, France). Le choix de la plage spectrale permet d'englober toute la région proche infra-rouge. Le seul prétraitement appliqué a consisté à décaler les lignes de base de manières à ce que tous les spectres expérimentaux passent par un point commun donné par une absorbance nulle à $1170nm$, soit la 336eme variable. Les valeurs de référence en éthanol de ces mêmes échantillons étaient mesurées par spectrométrie moyen infra-rouge (Foss). L'ensemble des spectres formait une matrice de spectres $\mathbf{X}$ de dimensions $(N = 1480, P = 1001)$, et un vecteur $\mathbf{y}$ de valeurs de référence en éthanol de dimensions $(N = 1480, 1)$. Par ailleurs les spectres purs de l'éthanol ($\mathbf{k}$), du glycérol, de l'acide lactique et de l'eau ont été acquis sur le même spectrophotométre Jasco avec les mêmes paramètres sauf la référence qui était l'air. Aucun prétraitement n'a été réalisé sur ces spectres purs.

Traitement des données

Les données ont été traitées avec le logiciel Scilab. Les données expérimentales $(\mathbf{X}, \mathbf{y})$ ont été réparties en trois jeux :

- $\mathbf{X}_G$, contenait les 165 spectres d'échantillons pour lesquels la concentration en éthanol est nulle ;
- $(\mathbf{X}_{etal}, \mathbf{y}_{etal})$, contenait, dans l'ordre chronologique d'acquisition, les 315 premiers échantillons dont les teneurs en éthanol sont non nulles ;
- $(\mathbf{X}_{test}, \mathbf{y}_{test})$, contenait, dans l'ordre chronologique d'acquisition, les 1000 derniers échantillons dont les teneurs en éthanol sont non nulles.

Ce découpage chronologique a été choisi pour assurer la plus grande indépendance entre les jeux de données d'étalonnage et de validation. Il a été vérifié que les histogrammes de $\mathbf{y}_{etal}$ et $\mathbf{y}_{test}$ étaient comparables.

Le spectre pur de l'éthanol anhydre a été divisé par 100 pour exprimer les résultats en p.cent volumique. La matrice $\mathbf{P}$ était le résultat d'une SVD sur $\mathbf{X}_G$, et $\mathbf{K}$ a été construite en fonction des grandeurs d'influence non représentées dans $\mathbf{X}_G$.

Sept modèles d'étalonnage ont été calculés puis testés sur $(\mathbf{X}_{test}, \mathbf{y}_{test})$. Les trois premiers modèles étaient destinés à l'explication du fonctionnement de l'IDC. Le premier modèle $m1$ était une simple projection sur $\mathbf{k}$. Le deuxième modèle $m2$ utilisait l'IDC avec uniquement $\mathbf{k}$ et $\mathbf{K}$, ce qui correspond à une DC avec peu de spectres purs. Le troisième modèle $m3$ utilisait l'IDC avec $\mathbf{k}$ et $\mathbf{X}_G$. Le quatrième modèle $m4$ utilisait l'IDC avec $\mathbf{k}$, $\mathbf{K}$ et $\mathbf{X}_G$. Le cinquième modèle $m5$ utilisait la PLSR, calculée sur $(\mathbf{X}_{etal}, \mathbf{y}_{etal})$ par validation croisée de l'algorithme NIPALS. Le nombre de variables latentes a été choisi de manière à minimiser le RMSECV. Le sixième modèle $m6$ était une IDC avec $\mathbf{k}$, $\mathbf{K}$ et $\mathbf{X}_G$, et une dimension A nettement plus élevée que la valeur optimale choisie dans les modèles (m3) et (m4). Enfin le septième modèle $m7$ reprenait le modèle $m4$ après que le spectre de l'eau, référence air, ait été éliminé de $\mathbf{K}$.

Comparaison des modèles

Les modèles ont d'abord été évalués visuellement selon leur aptitude générale de prédiction, c'est à dire par l'alignement des valeurs prédites par rapport aux valeurs de référence le long de la droite $(\widehat{y} = y)$. Pour chaque modèle, le RMSEP et sa décomposition entre biais et RMSEPc corrigé du biais ont été calculés. Une interprétation des pics des b-coefficients a été réalisée par comparaison avec le spectre pur de l'éthanol.

3.2.2 Résultats

Construction de la matrice K

Il est inhabituel en spectrocopie d'utiliser des spectres n'ayant pas été acquis dans les mêmes conditions. C'est pourtant le cas ici, puisque les données expérimentales représentées par $\mathbf{X}$ et $\mathbf{X}_G$ et les données expertes représentées par $\mathbf{k}$ et $\mathbf{K}$ ont respectivement l'eau et l'air comme référence. L'explication tient au fait que les spectres expérimentaux sont généralement acquis avec la référence eau pour des raisons pratiques, alors que les données expertes sont acquises avec la référence air pour se rapprocher des spectres purs. Pour un spectre quelconque, notons $\mathbf{x}^w$ sa valeur avec la référence eau, et $\mathbf{x}^a$ sa valeur avec la référence air. Si $\mathbf{k}^a_{water}$ est le spectre

de l'eau avec la référence air, alors :

$$\mathbf{x}^w = \mathbf{x}^a - \mathbf{k}^a_{water}$$

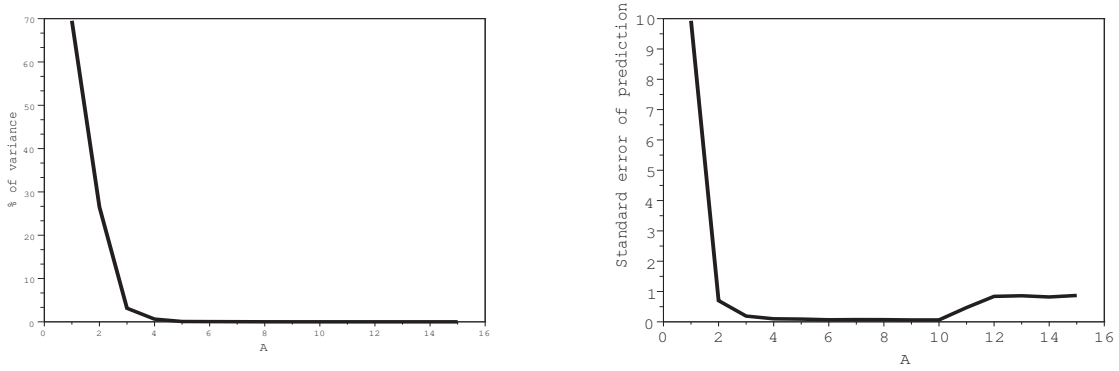
Ainsi, la différence entre les différents spectres $\mathbf{x}$ mesurés avec les références eau et air est tout simplement le spectre de l'eau par rapport à l'air. Dès lors que ce même spectre est introduit dans $\mathbf{K}$, du fait que $\mathbf{\Sigma}_{IDC}$ est une projection orthogonale à $\mathbf{K}$, le produit $\mathbf{\Sigma}_{IDC}\mathbf{k}^a_{water}$ est nul. En conclusion, l'incorporation dans $\mathbf{K}$ du spectre de l'eau $\mathbf{k}^a_{water}$ avec la référence air permet d'utiliser directement les données expérimentales acquises avec la référence eau, à la place de données expérimentales avec la référence air qu'il aurait pu paraitre plus logique d'utiliser.

Les principaux composés naturels des moûts et des vins, hormis l'éthanol, sont : l'eau, le glucose, le fructose, le glycérol, les acides tartrique, malique et lactique. Or les moûts ayant servi à constituer $\mathbf{X}_G$ contiennent en quantités variables du glucose et du fructose, ainsi que des acides tartrique et malique. Les spectres purs de ces composés n'ont donc pas été mis dans $\mathbf{K}$. Par contre, le glycérol et l'acide lactique sont absents des moûts donc ils ne sont pas représentés dans $\mathbf{X}_G$. C'est pourquoi leurs spectres purs ont été mis dans $\mathbf{K}$. Au final $\mathbf{K}$ contient les spectres de l'eau, du glycérol et de l'acide lactique, mesurés par rapport à l'air.

Détermination des paramètres

La figure 3.1 permet de choisir la dimension des modèles $m3$, $m4$, $m6$ et $m7$. La figure 3.1a représente l'évolution du pourcentage d'inertie de $\mathbf{X}_G$ capturée par les vecteurs de $\mathbf{P}$. La figure 3.1b représente l'évolution de l'erreur standard du modèle $m4$ appliqué sur $\mathbf{X}_G$ en fonction de A. La valeur $A = 4$ a été retenue car elle permet de capturer pratiquement toute l'information de $\mathbf{X}_G$ tout en présentant une erreur de prédiction minimale. Cette valeur optimale de $A = 4$ a été aussi appliquée à $m2$ et $m7$. La remontée de l'erreur de prédiction à partir de $A = 10$, figure 3.1b, confirme bien le risque d'incorporer de l'information liée à la grandeur d'intérêt, évoqué dans la partie théorie. Pour le vérifier, le modèle $m6$ avec $A = 12$ a été également construit.

La figure 3.2 permet de régler la dimension du modèle PLS. Elle représente l'erreur standard de prédiction de la PLSR en validation croisée sur le jeu d'étalonnage. Le RMSECV est stabilisé à partir de 5 variables latentes, donc le modèle PLSR est construit avec 5 variables latentes.



(a) (b)

FIG. 3.1 – (a) évolution du pourcentage d'inertie de $\mathbf{X_G}$ capturée par les $A = 1$ à 15 premiers vecteurs de $\mathbf{P}$; (b) erreur standard de prédiction du modèle IDC appliqué sur $\mathbf{X_G}$, pour $A = 1$ à 15.

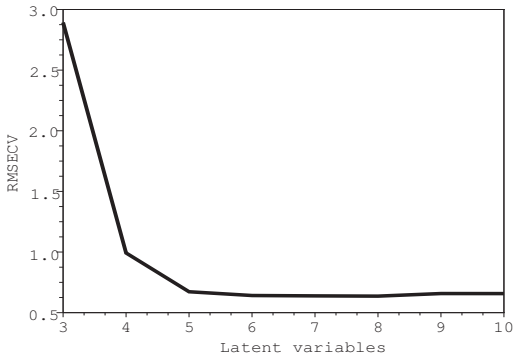


FIG. 3.2 – RMSECV de la PLSR pour les variables latentes 3 à 10.

Analyse des modèles $m1$ à $m7$

La figure 3.3-m2 présente la prédiction obtenue par le modèle $m2$, une DC avec peu de spectres purs. La prédiction est certes trop bruitée pour être utilisable en pratique, mais elle est sensible à $\mathbf{y}$: le coefficient de détermination entre valeurs prédites et valeurs de référence est de 0.87. Le tableau 3.1 représente les performances obtenues par différents modèles dérivés de $m2$ par élimination de 0 à 2 spectres de la matrice $\mathbf{K}$. Il montre clairement que cette corrélation élevée est apportée par la présence du spectre pur de l'eau, et d'au moins un des deux autres spectres purs : glycérol ou acide lactique. Dans tous les autres cas, le R^2 est inférieur ou égal à 0.20 et peut être considéré comme nul. La présence du spectre de l'eau est discutée dans le paragraphe suivant consacré au biais. Quant aux spectres du glycérol et de l'acide lactique, ils ont des cosinus respectifs avec le spectre de l'éthanol de 0.93 et 0.92. Cette forte colinéarité a pour conséquence d'enlever du spectre de l'éthanol une grande partie d'information sans intérêt puisque non spécifique à l'éthanol. L'information spécifique à l'éthanol se retrouve dans le spectre après correction, donc elle a un plus grand poids relatif ce qui explique l'apparition de la sensibilité de la prédiction à l'éthanol.

La figure 3.3-m3 présente la prédiction obtenue par le modèle $m3$, une IDC sans spectres purs. Ce modèle ne contient aucun des trois spectres évoqués précédemment : eau, acide lactique, glycérol. Il est donc normal que la corrélation entre valeurs prédites et valeurs de référence soit voisine de 0. Toutefois l'introduction de $\mathbf{X}_G$ conduit à supprimer toute la variabilité de prédiction, donc à supprimer du bruit. Au final, ce modèle prédit une valeur proche de 0 pour tous les échantillons, comparativement à $m1$.

La figure 3.3-m1 présente la prédiction obtenue sans aucune correction. L'absence de $\mathbf{K}$ conduit à une prédiction insensible à l'éthanol. Et l'absence de $\mathbf{P}$ conduit à une prédiction fortement bruitée. Logiquement la prédiction de $m1$ est fortement bruitée de part et d'autre de 0.

Le modèle IDC complet $m4$ donne des prédictions très satisfaisantes, tout à fait comparables avec celles de la PLSR, voir figure 3.3-$m4$ et $m5$, tables 3.2 et 3.3. Le RMSEP de l'IDC est meilleur que celui de la PLSR pour les teneurs en éthanol inférieures à 10% vol ; respectivement 0.87 et 0.90 % vol. La situation est inversée pour les teneurs en éthanol supérieures à 10% vol., les RMSEP respectifs de l'IDC et de la PLSR sont alors de 1.01 et 0.92% vol. Plus généralement, les modèles $m2$ à $m4$ présentent une erreur plus forte dans la zone des hautes teneurs en éthanol,

correspondant à des vins en fin de fermentation ou des vins finis. Le problème est certainement dû à une évolution des vins en fin de fermentation ou à l'effet de la stabilisation physique et chimique des vins finis ; cela se traduit par des composés non pris en compte dans $\mathbf{X}_G$ ou $\mathbf{K}$. Le jeu d'étalonnage de la PLSR contenait des vins finis, d'où une plus grande robustesse pour la PLSR dans cette situation. Pour améliorer le modèle IDC, il faudrait compléter $\mathbf{X}_G$ par d'autres spectres acquis sur vins finis. Le problème est alors de disposer d'échantillons ayant tous la même teneur en éthanol, de manière à ce qu'un simple centrage élimine l'effet de l'éthanol sur ces spectres. Cela n'a pas été possible, trop peu d'échantillons avaient la même valeur nominale d'éthanol, et l'imprécision de la mesure de référence ajoutait une incertitude sur la teneur réelle en éthanol.

Le modèle $m6$, IDC avec une valeur de A volontairement élevée, figure 3.3-$m6$ a une prédiction mauvaise, bien moins bonne que celle du modèle $m4$. La différence entre ces modèles est que la valeur optimale $A = 4$ a été choisie pour $m4$, alors qu'une valeur excessive, $A = 12$, a été choisie pour $m6$. Ceci confirme que, tout comme la PLSR, le réglage de l'IDC doit être fait précisément.

Comparé à $m4$, le modèle $m7$, IDC sans le spectre de l'eau, présente un léger biais et une pente différente de 1 (voir tableau 3.2), ainsi qu'une variabilité de prédiction nettement plus élevée. La différence entre ces deux modèles est uniquement la présence du spectre de l'eau dans $\mathbf{K}$ pour $m4$ et son absence pour $m7$. Cela atteste de l'importance du spectre de l'eau avec référence air dans $\mathbf{K}$.

La figure 3.4 (a) représente les b-coeffificents de l'IDC et le spectre pur de l'éthanol. Les 4 principaux pics du spectre de l'éthanol (1580, 1710, 2085 et 2295 nm) se retrouvent dans les b-coefficients de l'IDC, figure 3.4. Le pic de l'éthanol à 2085 nm se retrouve atténué dans les b-coefficients. Une explication est que les sucres ont également un fort pic d'absorbance à cette longueur d'onde [27]. En plus des pics de l'éthanol, deux autres pics sont visibles dans les b-coefficients. Le pic négatif à 1450 nm peut être relié à l'absorbance de l'eau dans cette zone : une forte absorbance traduit une forte teneur en eau, donc une moindre teneur en éthanol. Le pic positif à 1940 nm est beaucoup plus compliqué à interpréter puisque, outre un fort pic de l'eau, plusieurs composés absorbent dans cette zone, par exemple le glycérol et l'acide lactique. C'est une région spectrale très complexe, comme en attestent les premières composantes de $\mathbf{P}$, non présentées.

La comparaison des b-coefficients de la PLSR et de l'IDC, figure 3.4 (b) montre qu'ils sont nettement différents. Leur cosinus est de 0.47, donc l'angle entre ces deux vecteurs est voisin de

60 °. Les différences portent sur la plage visible comme sur la plage infra-rouge. Dans le visible, les b-coefficients de l'IDC sont pratiquement nuls, en accord avec la non-absorbance de l'éthanol. Par contre les b-coefficients de la PLSR ne sont pas nuls. Ils vont donc réagir à la présence de couleur rouge dans les vins. Cela comporte un risque d'erreur. Dans l'infra-rouge, le modéle IDC différe de celui de la PLSR par un pic plus important vers $1580nm$, et surtout par l'écart entre 2230 et $2300nm$.

Cet exemple illustre la non-unicité des modèles : des prédictions équivalentes sont obtenues par des modèles très différents.

Spectres dans **K**	R^2
Eau (W)	0.20
Acide lactique (L)	0.00
Glycerol (G)	0.06
L + G	0.03
W + L	0.74
W + G	0.85
W + L + G	0.87

TAB. 3.1 – Coefficients de corrélation R^2 entre valeurs prédites et valeurs de référence, pour des modèles obtenus à partir de $m2$ en enlevant 0, 1 ou 2 spectres

3.2.3 Conclusion sur la première application de l'IDC

Cette première application est un exemple simple pour lequel toutes les informations nécessaires à la construction des modèles IDC et PLSR étaient disponibles. Il permet une comparaison didactique entre ces deux méthodes d'étalonnage.

La PLSR utilise une information expérimentale, un jeu d'étalonnage. L'IDC utilise simultanément une information experte, des spectres purs, et une information expérimentale, un ensemble de spectres caractérisant les grandeurs d'influence non prises en compte parmi les spectres purs. Ces deux approches sont donc très différentes, et pourtant elles conduisent à des modèles équivalents. Cela démontre donc tout le potentiel de l'IDC, et cela confirme aussi l'importance d'utiliser la complémentarité entre informations expérimentales et expertes afin de caractériser complètement l'espace nuisible.

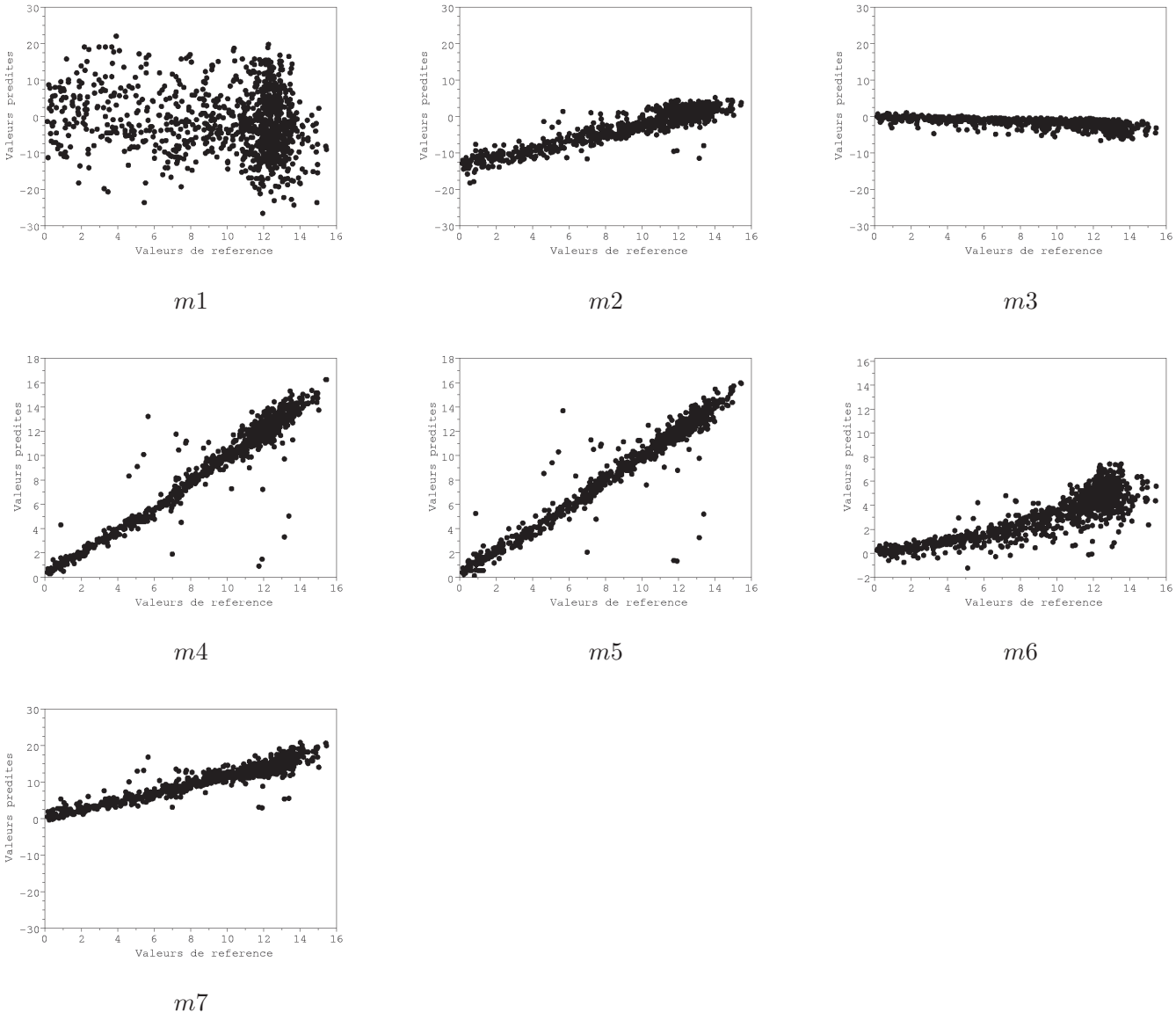


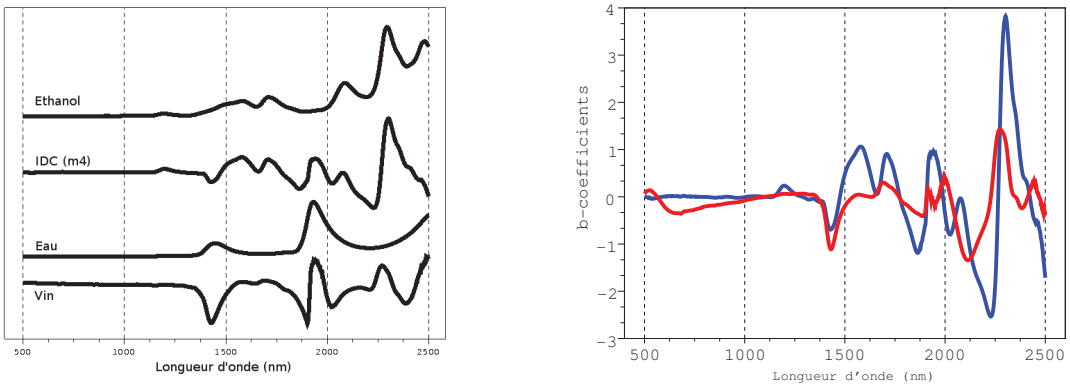
FIG. 3.3 – Test des modèles : $m1$ projection simple sans correction ; $m2$ IDC n'utilisant que $\mathbf{K}$; $m3$ IDC n'utilisant que $\mathbf{X_G}$; $m4$ IDC complète avec $A = 4$; $m5$ PLSR avec 5 variables latentes ; $m6$ IDC complète avec $A = 12$; $m7$ modèle $m4$ après retrait du spectre de l'eau. La deuxième ligne a une échelle différente des première et troisième lignes.

Modèle	Pente	Ordonnée à l'origine	$RMSEP_c$ % vol.	$RMSEP$ % vol.	R^2
$m1$ $\mathbf{\Sigma} = \mathbf{I}$	-0.36	10.89	10.00	14.79	0.025
$m2$ $\mathbf{\Sigma} = \mathbf{K}$	1.16	11.88	1.86	12.08	0.867
$m3$ $\mathbf{\Sigma} = \mathbf{P}, A = 4$	-0.153	9.55	3.75	10.26	0.048
$m4$-IDC $\mathbf{\Sigma} = [\mathbf{KP}], A = 4$	0.990	0.105	0.96	0.96	0.938
$m5$-PLSR PLSR, 5VL	0.974	0.012	0.92	0.92	0.943
$m6$ $\mathbf{\Sigma} = [\mathbf{KP}], A = 12$	0.436	6.15	2.34	6.59	0.785
$m7$	1.11	1.539	1.465	2.12	0.902

TAB. 3.2 – Indices de caractérisation des modèles

Model	Ethanol < 10 %	Ethanol ≥ 10 %
IDC (m4)	0.87	1.01
PLS (m5)	0.90	0.92

TAB. 3.3 – RMSEP des modèles IDC (m4) et PLSR (m5) détaillé selon la teneur en éthanol de l'échantillon



(a) (b)

FIG. 3.4 – (a) Spectres de l'éthanol, de l'eau, d'un vin, b-coefficients de l'IDC, modèle $m4$; (b) b-coefficients de l'IDC (bleu) et de la PLSR (rouge), modèles respectifs $m4$ et $m5$

L'application suivante montre les performances de l'IDC en imagerie hyperspectrale, dans des conditions où la PLSR n'st pas applicable.

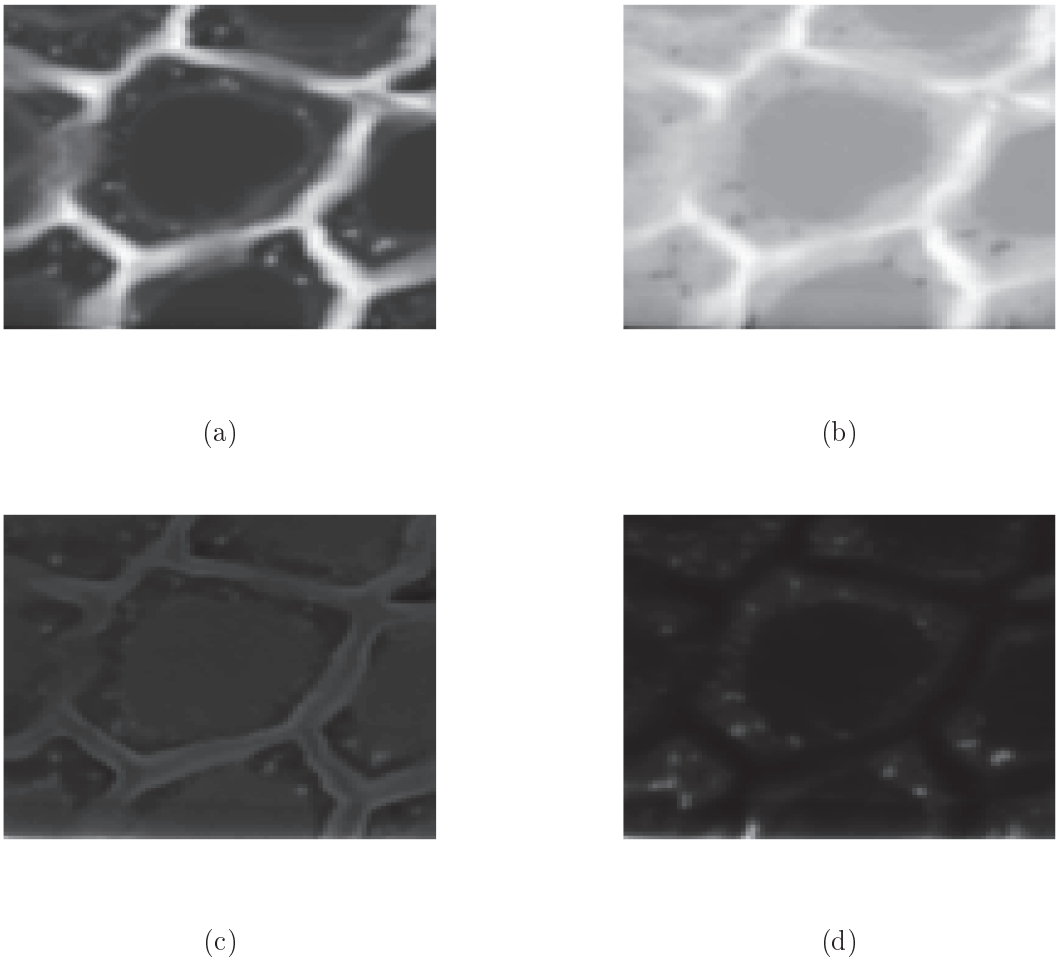
3.3 Deuxième exemple d'application de l'IDC : analyse des parois de la couche à aleurones du grain de blé

3.3.1 Matériels et méthodes

Une image hyperspectrale de cellules à aleurone de blé a été réalisée sur une coupe mince obtenue manuellement avec un microscope Raman confocal (Almega, ThermoElectron) ayant la configuration suivante : excitation du laser He-Ne $\lambda = 633nm$, réseau 1800 fentes/mm, fente $25\mu m$, objectif x100. Les données obtenues représentaient une aire de 70x70μm, pas de $0.8\mu m$ soient 89x89 pixels. Chaque pixel correspondait à un spectre de 461 nombres d'onde entre 862.7 et 1749.8 cm^{-1}. La matrice $\mathbf{X}_G$ a été choisie dans l'image. Lors de l'acquisition, la mise au point a été faite sur les parois, objets de l'analyse. La partie centrale d'une cellule correspond au parois vues de l'extérieur. Les parois sont normalement recouvertes de la membrane cytoplasmique, qui ne contient pas les molécules spécifiques aux parois. C'est pourquoi $\mathbf{X}_G$ est représentée par les spectres d'un carré de 11x11 pixels au centre de la cellule. Par ailleurs, les spectres purs ont été acquis à partir de fractions pures des 3 principaux composants de la paroi cellulaire :

arabino-xylans (Ax), β -glucans (Bg) and arabinose esterifieé à l'acide ferulique (Ara-fe). Trois modèles IDC ont été calculés pour quantifier respectivement Ax, Ara-fe and Bg en prenant successivement chacun des 3 spectres comme grandeur d'intérêt, les deux autres donnant la matrice **K**. Un quatrième modèle a été calculé pour quantifier Bg avec la méthode DC, ce qui revient à n'utiliser que **K** contenant les spectres de Ax et Ara-fe.

3.3.2 Résultats



(a) (b)

(c) (d)

FIG. 3.5 – Tests des modèles : Prédictions de Ara-Fe (a), Ax(b) et Bg (c) par IDC ; prédiction de Bg(d) par DC

La matrice **P** a été construite avec les 3 premiers vecteurs-propres d'une ACP sur $\mathbf{X}_G$. Ensuite les modèles IDC ont été calculés comme décrit précédemment. Les figures 3.5a, 3.5b et

3.5c représentent les estimations des concentrations respectives par IDC de Ara-fe, Ax and Bg. La figure 3.5d donne la prédiction par DC de Bg, soit une IDC sans $\mathbf{P}$. Les couleurs blanches représentent les plus hautes concentrations de chaque composé. Ara-fe et Ax sont localisés au centre, alors que les Bg sont localisés principalement sur les bords de la paroi cellulaire. Aucune conclusion ne peut être extraite de la figure 3.5d : l'absence de l'information contenue dans $\mathbf{X}_G$ enlève au modèle DC toute sa capacité de prédiction.

La localisation majoritaire de Ara-fe, Ax et Bg donnée par les modèles IDC est en accord avec de précédents travaux ([28], [29]) qui ont abouti à la même conclusion avec d'autres méthodes analytiques basées sur l'immunochimie.

3.3.3 Conclusion sur la deuxième application de l'IDC

En imagerie hyperspectrale, les étalonnages inverses ou régressions ne sont pas applicables par manque de jeu d'étalonnage. Par contre les étalonnages directs peuvent être utilisés. L'IDC a des performances nettement supérieures à celles de la DC pour l'application présentée ici, ce qui confirme l'importance de l'identification et de la correction de l'information nuisible. La possibilité d'utiliser des informations expérimentales et expertes d'origines très diverses donne une grande souplesse et une excellente adaptabilité de l'IDC vis à vis des différentes situations expérimentales. L'IDC est particulièrement adaptée à l'imagerie hyperspectrale.

3.4 Discussion

Quatre propriétés importantes de l'IDC sont discutées : les fondements spectroscopiques de l'étalonnage direct ; la gestion de l'ordonnée à l'origine et de la pente ; la qualité des spectres de référence ; les liens de l'IDC avec le Net Analyte Signal.

3.4.1 Les fondements spectroscopiques de l'étalonnage direct

Nous avons vu que les étalonnages directs sont une application du modèle général avec les entrées $\mathbf{X} = \mathbf{k}'$ et $\mathbf{y} = 1$. La valeur $\mathbf{y} = 1$ est justifiée par la concentration égale à 1 ou 100 p.cent dans l'échantillon dont le spectre est $\mathbf{k}$. Cependant, la notion de concentration est floue, elle peut recouvrir différentes définitions : concentration massique, concentration molaire, concentration volumique par exemple. Si la valeur est toujours 1 dans un composé pur, elle peut prendre des valeurs différentes selon la définition considérée lorsque le composé n'est plus pur. Cela peut être

illustré par un exemple.

Soient les solutions (Eau), (EtOH) et (mélange) contenant respectivement : de l'eau pure ; de l'éthanol pur ; un mélange Eau-Ethanol pour moitié en volume. Les densités de l'eau et de l'éthanol sont respectivement de 1 et 0.8, leurs masses molaires respectives 18 et 46 g, la rétractation du mélange est négligée.

	Fraction volumique	Fraction massique	Fraction molaire
Eau	0.50	0.556	0.762
Ethanol	0.50	0.444	0.238

TAB. 3.4 – Fractions molaires, massiques et volumiques en eau et éthanol dans une solution moitié eau moitié éthanol en volume

Les fractions volumiques, massiques et molaires de l'éthanol sont égales entre elles en solutions pures. Elles ont pour valeur 1 dans la solution (EtOH) et 0 dans la solution (Eau). Mais elles diffèrent sensiblement dans la solution (mélange), voir tableau 3.4. La question posée est de savoir quelle est l'unité sur laquelle est basé l'étalonnage direct : fraction volumique ? massique ? molaire ? autre ? La loi de Beer-Lambert n'est applicable que dans le cas de solutions fortement diluées, pour lesquelles un doublement ou triplement de la concentration en grandeur d'intérêt ne modifie pas significativement le volume, la masse ou le nombre total de moles de la solution. Dans ces conditions les fractions volumiques, massiques et molaires de la grandeur d'intérêt restent proportionelles entre elles en fonction de la concentration. Le fait que Beer-Lambert soit généralement exprimée en masse n'est qu'une convention, ce n'est pas en soi une réponse à la question posée.

Ce questionnement a été abordé par [30]. Les auteurs s'appuient sur un exemple, un mélange de toluène, dichlométhane et n-heptane en proportions massiques connues. Un exemple de calcul est donné dans le tableau 3.5 pour l'échantillon 9. Les prédictions par DC sont confrontées aux pourcentages massiques et molaires, ainsi qu'à H% le pourcentage d'atomes H apportés par chaque composé avec deux calculs : brut ou corrigé. Les pourcentages volumiques ne figurent pas dans [30]. Ils ont été rajoutés. La formule i et la suivante sont données par [30].

Une lecture rapide de ce tableau montre que les compositions exprimées en pourcentage volumique sont les valeurs les plus proches des prédictions par DC. Le calcul a donc été refait

Produits purs :		Toluène	Dichlorométhane	n-Heptane	Total
Masse molaire (g)	a	92.13	84.94	100.20	
Masse volumique (g/mL)	b	0.8669	1.336	0.6837	
Moles/100mL	c=100*b/a	0.941	1.573	0.682	
Nombre de H	d	8	2	16	
Echantillon :					
Pourcentage massique (g/100g)	e	23.86	26.45	49.69	100.00
Nombre de moles/100g	f=e/a	0.259	0.311	0.496	1.066
Pourcentage molaire		24.3	29.2	46.5	100.0
Volume en mL/100g	g=e/b	28.0	19.8	72.7	120.5
Pourcentage volumique		23.2	16.4	60.4	100.0
Nombre de moles de H/100g	h=d*f	2.07	0.62	7.94	10.63
Pourcentage molaire de H		19.5	5.8	74.7	100
Nombre de moles de H/100mL	i=c*d*e 100	1.79	0.83	5.42	8.04
Pourcentage de H corrigé		22.3	10.3	67.4	100
Prédictions par DC		23	17	56	

TAB. 3.5 – Concentrations en toluène, dichlorométhane et n-heptane exprimées selon différentes unités, pour l'échantillon 9

pour 15 échantillons issus de mélanges en proportions variables des mêmes composés. Les données, tableau 3.6, montrent une bonne corrélation entre les pourcentages volumiques et les prédictions par DC. Ce résultat est en accord avec l'exemple donné pour l'application de l'IDC, chapitre 3, où les modèles DC et IDC donnaient bien une prédiction de l'éthanol en pourcentage volumique.

	Toluène			Dichlorométhane			n-Heptane		
Echantillon	masse p.cent	volume p.cent	DC x100	masse p.cent	volume p.cent	DC x100	masse p.cent	volume p.cent	DC x100
1	100	100	100	0	0	0	0	0	0
2	76.4	83.4	83.1	23.6	16.6	14.4	0	0	2.0
3	74.1	69.3	70.4	0	0	0.7	25.9	30.7	29.2
4	50.3	61.0	61.0	49.7	39.0	35.4	0	0	3.0
5	48.9	49.8	50.8	25.1	16.6	15.8	26.0	33.6	33.8
6	49.9	44.0	45.6	0	0	0.8	50.1	56.0	54.3
7	25.3	33.8	34.8	74.8	66.2	62.4	0	0	2.7
8	25.3	28.4	28.2	49.7	36.0	35.2	25.0	35.6	36.5
9	23.9	23.2	23.7	26.4	16.4	15.5	49.7	60.4	61.3
10	25.2	21.0	22.0	0	0	0.5	74.8	79.0	77.9
11	0	0	0	100	100	100	0	0	0
12	0	0	−0.5	75.0	60.6	61.2	25.0	39.4	40.9
13	0	0	0.5	49.5	33.4	34.8	50.4	66.6	67.2
14	0	0	0.3	24.3	14.1	14.7	75.7	85.9	86.5
15	0	0	0	0	0	0	100	100	100

TAB. 3.6 – Pourcentages massiques et volumiques comparées aux prédictions par DC pour le Toluène, le Dichlorométhane et le n-Heptane en proportions variables dans différentes solutions

Ainsi, ces résultats montrent que les prédictions par étalonnage direct donnent une valeur correspondant au pourcentage volumique de la grandeur d'intérêt. Toutefois cela n'est pas généralisable. En effet les deux exemples étudiés concernaient des mélanges de produits miscibles entre eux, c'est à dire qu'un volume v_A de masse m_A d'un produit A ajouté à un volume v_B de masse m_B d'un produit B donne un volume $v_A + v_B$ en première approximation, de masse $m_A + m_B$, où les produits A et B sont en mélange. Nous n'avons pas d'exemple de produits

solubles, c'est à dire pour lesquels une masse m_A d'un produit A solide ajouté à un volume v_B de masse m_B d'un produit B liquide donne en première approximation un volume v_B de masse $m_A + m_B$ où les produits A et B sont en mélange. Un exemple très simple est celui du glucose dissous dans l'eau. Il est impossible de définir le pourcentage volumique du glucose en solution dans l'eau, et donc d'appliquer l'étalonnage direct selon la règle définie précédemment. Le questionnement reste posé.

Gestion de l'ordonnée à l'origine et de la pente

Dans la théorie du modèle IDC, la reprèsentation graphique des valeurs prédites $\widehat{\mathbf{y}}$ contre les valeurs de référence $\mathbf{y}$ doit donner une droite de pente 1 passant par l'origine. En pratique, des pentes et ordonnées à l'origine différentes respectivement de 1 et 0 peuvent s'expliquer par une mauvaise prise en compte des grandeurs d'influence. Prenons un exemple simple.

Soit un spectre $\mathbf{x}_i$ acquis sur un échantillon i dans d'excellentes conditions expérimentales, c'est à dire que les seules grandeurs d'influence sont chimiques et représentées par $\mathbf{K}$. La métrique $\mathbf{\Sigma}_0$ est la projection orthogonale à $\mathbf{K}$. Supposons qu'il apparait une grandeur d'influence qui rajoute systématiquement une contribution constante $\mathbf{e}$ au spectre initial. Le spectre observé au final est $\mathbf{x}_i^*$ tel que :

$$\mathbf{x}_i^* = \mathbf{x}_i + \mathbf{e}$$

Supposons aussi que les spectres purs ont été mesurés sur le même spectrophotométre. La matrice des spectres purs mesurés est :

$$\mathbf{K}^* = \mathbf{K} + \mathbf{e}\mathbf{1}'$$

Deux situations sont possibles selon que cette grandeur d'influence ait été identifiée ou pas.

- Premier cas : elle a été identifiée. La correction consiste naturellement à joindre $\mathbf{e}$ à $\mathbf{K}^*$ ce qui donne $\mathbf{K}^c = [\mathbf{K}\mathbf{e}]$. La métrique $\mathbf{\Sigma}_1$ est la projection orthogonale à $\mathbf{K}^c$.
- Deuxième cas : elle n'a pas été identifiée, donc pas prise en compte. La métrique $\mathbf{\Sigma}_2$ est la projection orthogonale à $\mathbf{K}^{nc}$ définie ainsi :

$$\mathbf{K}^{nc} = \mathbf{K} + \mathbf{e}\mathbf{1}'$$

$\mathbf{\Sigma}_1$ est un projecteur orthogonal à un espace contenant $\mathbf{e}$, donc il est immédiat que $\mathbf{\Sigma}_1$ enlève de $\mathbf{x}_i^*$ toute l'information apportée par $\mathbf{e}$. La correction de $\mathbf{e}$ est totale. Ce n'est pas le cas

dans la seconde situation. Comme $\mathbf{\Sigma}_2$ n'est pas un projecteur orthogonal á l'espace contenant $\mathbf{e}$, toute l'information apportée par $\mathbf{e}$ n'est pas enlevée. Cela peut donc conduire à des erreurs de prédiction de type pente non égale à 1 et ordonnée à l'origine non nulle. Un exemple est donné par la première application. Le spectre $\mathbf{e}$ est représenté par le spectre de l'eau, soustrait lors de l'acquisition. Le modéle IDC $m4$ correspond à la première situation. La correction est bonne puisque la matrice $\mathbf{R}$ contient le spectre de l'eau. Si celui-ci est enlevé, le modèle obtenu, $m7$, correspondant à la seconde situation, est nettement moins bon bien que le spectre de l'eau ait été ajouté à tous les spectres utilisés pour définir $\mathbf{X}_G$. La dispersion des points a augmenté de même que la pente.

Ce résultat confirme l'importance d'évaluer exhaustivement l'espace nuisible en s'appuyant sur toutes les sources d'informtions expérimentales et expertes disponibles. Même les déformations constantes de ligne de base sont à considérer comme des grandeurs d'influence à part entière orthogonalement auxquelles il faut projeter.

Qualité des informations expertes représentées par les spectres purs

Le résultat précédent souligne aussi toute l'importance de travailler avec des spectres purs aussi proches que possible des spectres purs de référence. Dans certaines situations il peut être tentant d'utiliser le même système spectroscopique pour faire des acquisitions sur des échantillons purs et prendre ces spectres comme des spectres purs. C'est le cas par exemple en imagerie hyperspectrale, une partie d'une image de feuille pourrrait être prise comme référence pour les feuilles. Le danger est d'inclure une grandeur d'influence sans s'en apercevoir, ensuite cette grandeur d'influence perturbera les prédictions par IDC. Pour éviter ce danger, les informations expertes de type spectres purs doivent impérativement être acquises dans des conditions les plus contrôlées possible, par exemple avec un spectrophotomètre plus performant, un échantillon plus purifié, une régulation de la température et de l'humidité... conditions *sine qua non* pour approcher une valeur de référence.

Liens avec le Net Analyte Signal (NAS)

Le terme $(\mathbf{\Sigma}_{DC}\mathbf{k})$ s'écrit $((\mathbf{I}-\mathbf{K}'(\mathbf{K}\mathbf{K}')^{-1}\mathbf{K})\mathbf{k})$, il représente la projection du spectre pur de la grandeur d'intérêt orthogonalement à la matrice des spectres purs des grandeurs d'influence chimiques. De même, $(\mathbf{\Sigma}_{IDC}\mathbf{k})$ s'écrit $((\mathbf{I}-\mathbf{R}'(\mathbf{R}\mathbf{R}')^{-1}\mathbf{R})\mathbf{k})$. Ainsi le spectre $\mathbf{k}$ de la grandeur d'intérêt est projeté orthogonalement à l'espace définissant les grandeurs d'influence chimiques

et physiques. Nous retrouvons dans ces deux cas la définition du Net Analyte Signal (NAS) ([4]) : *" the net analyte signal may be computed as the part of its spectrum orthogonal to the contribution of other coexisting constituents"*, à la différence près qu'avec l'IDC cette définition est étendue aux grandeurs d'influence physiques : l'IDC améliore la définition du NAS. Soit le scalaire $\alpha = (\mathbf{k}\mathbf{\Sigma}_{IDC}\mathbf{k}')^{-1}$. La prédiction de la grandeur d'intérêt s'écrit également :

$$\widehat{\mathbf{y}}_{IDC} = \alpha\mathbf{X}\widehat{\mathbf{NAS}}_{IDC}$$

Ainsi, avec une métrique Euclidienne et à un coefficient α près, la prédiction par IDC est le produit scalaire entre les spectres de $\mathbf{X}$ et l'estimation du NAS calculée par l'IDC. Le coefficient α a pour fonction d'ajuster l'échelle de notation de la grandeur d'intérêt qui est arbitraire : par exemple des mg/L ou g/L. Les b-coefficients de l'IDC tendent vers le NAS, ce qui n'est pas le cas des b-coefficients de la PLSR.

3.5 Conclusion sur l'IDC

La DC-Direct Calibration n'est pas applicable dès lors que certaines grandeurs d'influence chimiques ou physiques ne sont pas prises en compte (exemples 1 et 2). L'IDC permet de compléter les informations manquantes dans la DC au moyen d'informations expérimentales déterminées par un plan d'expérience. L'IDC ne nécessite pas de jeu d'étalonnage, par contre les résultats peuvent être exacts à un biais et une pente près. Il est montré également que l'IDC est une méthode de prédiction basée sur le NAS, c'est à dire que la grandeur prédite est égale au produit scalaire entre le spectre de l'échantillon et le NAS, à un facteur multiplicatif près qui tient compte de l'échelle des valeurs de la grandeur d'intérêt.

L'IDC est bien plus performante que la DC du fait que l'association entre informations expérimentales et informations expertes permet de mieux caractériser l'espace nuisible, ce que ne font ni la DC ni la SBC. Parfois, elle peut même être équivalente à la PLSR (exemple 1). Toutefois l'IDC est destinée prioritairement aux situations pour lesquelles la PLSR n'est pas ou difficilement applicable. Ainsi l'imagerie hyperspectrale offre potentiellement de nombreuses applications pour l'IDC.

Chapitre 4

Deuxième implémentation : VODKA-PLSR, une famille de modèles de régression

Sommaire

4.1 NIPALS-P une nouvelle version de NIPALS 64

4.2 Le modèle VODKA-PLSR 65

4.3 Application : quantification de l'éthanol dans des moûts de raisin en fermentation 67

4.3.1 Les données 67

4.3.2 Paramétrage et validation des modèles de régression 67

4.3.3 Résultats 68

4.4 Discussion 70

4.4.1 Informations expérimentales et informations expertes dans le modèle PLSR 70

4.4.2 Choix de l'algorithme NIPALS 71

4.4.3 Présence d'une incohérence dans NIPALS ? 71

Nous avons montré que la PLSR pouvait s'écrire selon le modèle général décrit au chapitre 2. Le paramètre $\mathbf{\Sigma}$ était calculé directement par la pseudo-inverse de $\mathbf{X'X}$ selon Moore-Penrose. Le paramètre $\mathbf{P}$ était donné par l'algorithme NIPALS.

Un nouveau mode de calcul de $\mathbf{P}$ est proposé. Il met en évidence un nouveau paramètre, un

vecteur $\mathbf{r}$ de dimension $(P \times 1)$. Des choix différents pour $\mathbf{r}$ donnent une famille d'algorithmes de régression.

4.1 NIPALS-P une nouvelle version de NIPALS

L'algorithme NIPALS est basé sur un calcul pas à pas des $\mathbf{t}_i$, $\mathbf{p}_i$ et $\mathbf{w}_i$. A la $(i+1)^{eme}$ boucle de l'algorithme, d'après les formules 2.13 et 2.14 page 33, les matrices $\mathbf{X}$ et $\mathbf{y}$ sont projetées orthogonalement aux $\{\mathbf{t}_1 ... \mathbf{t}_i\}$ déjà calculées. L'orthogonalisation se fait donc dans $\mathbb{R}^N$, dans l'espace des individus. Nous montrons que seul le calcul des $\mathbf{p}_i$ est nécessaire pour déterminer le modèle NIPALS, et qu'il peut être fait par projections $\mathbf{\Sigma}$-orthogonales dans l'espace $\mathbb{R}^P$ des variables.

Multiplions à droite chaque terme de l'équation 2.18 page 34 par $\mathbf{p'}_i$:

$$\mathbf{t}_i\mathbf{p'}_i = \mathbf{X\Sigma p}_i(\mathbf{p'}_i\mathbf{\Sigma p}_i)^{-1}\mathbf{p'}_i$$

Le terme $\mathbf{p}_i$ à gauche de l'équation est remplacé par sa valeur définie équation 2.15 page 33 :

$$\mathbf{t}_i(\mathbf{t'}_i\mathbf{t}_i)^{-1}\mathbf{t'}_i\mathbf{X} = \mathbf{X\Sigma p}_i(\mathbf{p'}_i\mathbf{\Sigma p}_i)^{-1}\mathbf{p'}_i$$

Aprés multiplication par -1 puis ajout de $\mathbf{X}$ de chaque coté :

$$(\mathbf{I}_N - \mathbf{t}_i(\mathbf{t'}_i\mathbf{t}_i)^{-1}\mathbf{t'}_i)\mathbf{X} = \mathbf{X}(\mathbf{I}_P - \mathbf{\Sigma p}_i(\mathbf{p'}_i\mathbf{\Sigma p}_i)^{-1}\mathbf{p'}_i)$$

Soient $\mathbf{P}_{1:i}$ la matrice de dimensions (P, i) contenant les vecteurs $\{\mathbf{p}_1 ... \mathbf{p}_i\}$, et $\mathbf{T}_{1:i}$ la matrice de dimensions (N, i) contenant les vecteurs $\{\mathbf{t}_1 ... \mathbf{t}_i\}$. L'équation précédente donne aussi :

$$(\mathbf{I}_N - \mathbf{T}_{1:i}(\mathbf{T'}_{1:i}\mathbf{T}_{1:i})^{-1}\mathbf{T'}_{1:i})\mathbf{X} = \mathbf{X}(\mathbf{I}_P - \mathbf{\Sigma P}_{1:i}(\mathbf{P'}_{1:i}\mathbf{\Sigma P}_{1:i})^{-1}\mathbf{P'}_{1:i}) \tag{4.1}$$

Une autre observation sur NIPALS est que les vecteurs $\mathbf{w}_i$, $\mathbf{t}_i$ et $\mathbf{p}_i$ sont calculés successivement dans cet ordre. Il y a donc un changement continuel d'espace vectoriel : $\mathbb{R}^P$ pour $\mathbf{w}_i$, puis $\mathbb{R}^N$ pour $\mathbf{t}_i$, puis finalement $\mathbb{R}^P$ pour $\mathbf{p}_i$. Or l'équation 2.18 montre qu'on peut calculer $\mathbf{t}_i$ dès lors que $\mathbf{p}_i$ est connu. Nous essayons donc de modifier NIPALS de manière à rester dans $\mathbb{R}^P$ avec pour objectif de déterminer uniquement les $\mathbf{p}_i$.

Soit $\mathbf{X}_i$ la projection de $\mathbf{X}$ orthogonalement à $\{\ \mathbf{t}_1 ... \mathbf{t}_i\}$. La combinaison des lignes 2.8, 2.10 et 2.12 de NIPALS donne :

$$\mathbf{p}_{i+1} = \alpha_{i+1}\mathbf{X'}_i\mathbf{X}_i\mathbf{X'}_i\mathbf{y} \tag{4.2}$$

avec α_{i+1} un scalaire associé à $\mathbf{p}_{i+1}$. Soit $\mathcal{Q}^{\perp}_{1:i}$ le projecteur orthogonal à $\mathbf{P}_{1:i}$ au sens de $\mathbf{\Sigma}$:

$$\mathcal{Q}^{\perp}_{1:i} = \mathbf{I}_P - \mathbf{\Sigma}\mathbf{P}_{1:i}(\mathbf{P'}_{1:i}\mathbf{\Sigma}\mathbf{P}_{1:i})^{-1}\mathbf{P'}_{1:i} \tag{4.3}$$

Reprenons la formule 4.2 en remplaçant $\mathbf{X}_i$ par $\mathbf{X}\mathcal{Q}^{\perp}_{1:i}$) :

$$\mathbf{p}_{i+1} = \alpha_{i+1}\mathcal{Q}'^{\perp}_{1:i}\mathbf{X'}\mathbf{X}\mathcal{Q}'^{\perp}_{1:i}\mathbf{X'}\mathbf{y} \tag{4.4}$$

qui peut aussi s'écrire :

$$\mathbf{p'}_{i+1} = \alpha_{i+1}(\mathbf{y'}\mathbf{X}\mathcal{Q}^{\perp}_{1:i})(\mathbf{X'}\mathbf{X}\mathcal{Q}^{\perp}_{1:i}) \tag{4.5}$$

Cette expression montre qu'il est effectivement possible de calculer $\mathbf{P}$ en restant uniquement dans $\mathbb{R}^P$. Le nouvel algorithme NIPALS-P, s'écrit donc :

- A l'étape 1 :

$$\begin{aligned} \mathbf{p}_1 &= \mathbf{X'}\mathbf{X}\mathbf{X'}\mathbf{y} \\ \mathbf{p}_1 &\leftarrow \mathbf{p}_1(\mathbf{p'}_1\mathbf{\Sigma}\mathbf{p}_1)^{-0.5} \end{aligned}$$

- A l'étape $i+1$:

$$\begin{aligned} \mathbf{p}_{i+1} &= \mathcal{Q}'^{\perp}_{1:i}\mathbf{X'}\mathbf{X}\mathcal{Q}'^{\perp}_{1:i}\mathbf{X'}\mathbf{y} \\ \mathbf{p}_{i+1} &\leftarrow \mathbf{p}_{i+1}(\mathbf{p'}_{i+1}\mathbf{\Sigma}\mathbf{p}_{i+1})^{-0.5} \end{aligned}$$

4.2 Le modèle VODKA-PLSR

Un nouveau modèle de régression est proposé. Il s'appuie sur trois entrées : une matrice $\mathbf{X}$ et une matrice $\mathbf{y}$ constituant un jeu d'étalonnage, plus un vecteur $\mathbf{r}$ de dimensions $(P \times 1)$ choisi arbitrairement. Ces entrées sont utilisées pour le calcul des deux paramètres $\mathbf{\Sigma}$ et $\mathbf{P}$. Le paramètre $\mathbf{\Sigma}$ est la pseudo-inverse de $\mathbf{X'X}$ au sens de Moore-Penrose, soit :

$$\mathbf{\Sigma} = (\mathbf{X'}\mathbf{X})^{+}$$

Le paramètre $\mathbf{P}$ est calculé par l'algorithme suivant. Soit $\mathcal{Q}^{\perp}_{1:i}$ le projecteur orthogonal au sens de $\mathbf{\Sigma}$ à la matrice $\mathbf{P}_{1:i}$ dont les colonnes sont les vecteurs $\{\mathbf{p}_1\mathbf{p}_2...\mathbf{p}_i\}$.

– A l'étape 1 :

$$\begin{aligned} \mathbf{p}_1 &= \mathbf{X}'\mathbf{X}\mathbf{r} \\ \mathbf{p}_1 &\leftarrow \mathbf{p}_1(\mathbf{p}'_1\mathbf{\Sigma}\mathbf{p}_1)^{-0.5} \end{aligned}$$

– A l'étape $i+1$:

$$\begin{aligned} \mathbf{p}_{i+1} &= \mathcal{Q}'^{\perp}_{1:i}\mathbf{X}'\mathbf{X}\mathcal{Q}'^{\perp}_{1:i}\mathbf{r} \\ \mathbf{p}_{i+1} &\leftarrow \mathbf{p}_{i+1}(\mathbf{p}'_{i+1}\mathbf{\Sigma}\mathbf{p}_{i+1})^{-0.5} \end{aligned}$$

Ainsi le vecteur $\mathbf{p}'_{i+1}$ est obtenu par le produit matriciel de $\mathbf{r}'$ avec $\mathbf{X}'\mathbf{X}$ dans l'espace orthogonal selon $\mathbf{\Sigma}$ aux vecteurs $\{\mathbf{p}_1\mathbf{p}_2...\mathbf{p}_i\}$ précédemment obtenus. Le vecteur $\mathbf{r}$ est un nouveau paramètre dont le choix permet d'intégrer des informations supplémentaires dans le modèle de régression. D'où le choix du nom : Vector Orientation Decided through Knowledge Assessment-Partial Least Square Regression (VODKA-PLSR).

Le calcul des b-coefficients est obtenu en substituant $\mathbf{T}$ dans l'équation 2.4. On obtient l'équation 2.5. Celle-ci se simplifie, d'après une propriété des pseudo-inverses de Moore-Penrose, et donne finalement :

$$\mathbf{b} = \mathbf{\Sigma}\mathbf{P}(\mathbf{P}'\mathbf{\Sigma}\mathbf{P})^{-1}\mathbf{P}'\mathbf{\Sigma}\mathbf{X}'\mathbf{y} \tag{4.6}$$

L'originalité et l'intérêt de cette méthode sont que le choix de $\mathbf{r}$ permet d'introduire soit de l'information expérimentale soit de l'information experte dans le modèle. Un premier exemple d'information expérimentale est $\mathbf{r} = \mathbf{X}'\mathbf{y}$, qui donne la PLSR selon l'algorithme NIPALS-P décrit précédemment. Un deuxième exemple d'information expérimentale est $\mathbf{r} = \mathbf{X}'\mathbf{1}_N$, un vecteur colinéaire au spectre moyen de $\mathbf{X}$. D'autres exemples d'informations expertes sont : (1) le vecteur $\mathbf{r} = \mathbf{1}_P$; (2) le vecteur $\mathbf{r} = \mathbf{k}$ où $\mathbf{k}$ est le spectre pur de la grandeur d'intérêt ; (3) le vecteur $\mathbf{r} = NAS$, le NAS étant la partie de $\mathbf{k}$ orthogonale aux grandeurs d'influence. Ces cinq exemples ne sont pas exhaustifs, en théorie les choix de $\mathbf{r}$ sont infinis. Il est donc créé une famille de modèles de régression.

Un script de VODKA-PLSR sous Scilab et Matlab est disponible en annexe.

4.3 Application : quantification de l'éthanol dans des moûts de raisin en fermentation

4.3.1 Les données

Les données sont celles décrites au chapitre 3 pour illustrer le modèle IDC. La plage spectrale a été réduite à $500 - 1898nm$. Les spectres ont ensuite été répartis dans trois matrices, pour mémoire :

- $\mathbf{X}_G$: 165 échantillons de moûts, ne contenant pas d'éthanol ;
- $\mathbf{X}$: 315 premiers échantillons de moûts en fermentation ou vins, pour construction d'un modèle d'étalonnage ;
- $\mathbf{X}_V$: 1000 derniers échantillons pour validation du modèle d'étalonnage.

Les valeurs de référence en éthanol forment les vecteurs $\mathbf{y}$ et $\mathbf{y}_V$ de dimensions respectives (315×1) et (1000×1) respectivement.

4.3.2 Paramétrage et validation des modèles de régression

Les modèles d'étalonnage ont été construits sous environnement Scilab avec l'algorithme VODKA-PLSR, avec le jeu de données $(\mathbf{X}, \mathbf{y})$. Ils ont ensuite été validés avec les données $(\mathbf{X}_V, \mathbf{y}_V)$. La qualité de prédiction est donnée par le $RMSEP$ pour les 20 premières variables latentes de chaque modèle.

Six modèles d'étalonnage ont été comparés. Les cinq premiers modèles ont été calculés à partir du jeu d'étalonnage brut $(\mathbf{X}, \mathbf{y})$, avec différents choix pour $\mathbf{r}$, voir tableau 4.1. Le NAS est calculé de la même manière que les b-coefficients du modèle IDC décrit au chapitre 3. La matrice $\mathbf{P}_{XG}$ contient les 4 premiers vecteurs propres d'une ACP sur $\mathbf{X}_G$. Une matrice $\mathbf{R}$ est obtenue en concaténant $\mathbf{P}_{XG}$ avec les spectres purs de l'eau, du glycérol et de l'acide lactique acquis avec une référence air. Le NAS est le spectre pur de l'éthanol, référence air, après projection orthogonale à $\mathbf{R}$. Le sixième modèle est une PLSR appliquée sur données centrées $(\mathbf{X}_c, \mathbf{y}_c)$. L'objectif de cette application n'est pas d'étudier l'influence des prétraitements. Toutefois cette option nous a paru nécessaire puisque le centrage de la PLSR est une pratique courante en chimiométrie.

Modèle	Valeur de $\mathbf{r}$	Notes sur $\mathbf{r}$
m_1	$\mathbf{r} = \mathbf{1}_P$	Aucune information sur $\mathbf{X}$ ou $\mathbf{y}$
m_2	$\mathbf{r} = \mathbf{X}'\mathbf{1}_N$	Moyenne algébrique des spectres de $\mathbf{X}$
m_3	$\mathbf{r} = \mathbf{X}'\mathbf{y}$	PLSR-NIPALS non centrée
m_4	$\mathbf{r} = \mathbf{k}$	Spectre pur de la grandeur d'intérêt
m_5	$\mathbf{r}$= NAS	Utilisation du Net Analyte Signal
m_6	$\mathbf{r} = \mathbf{X}'_c\mathbf{y}_c$	PLSR-NIPALS centrée

TAB. 4.1 – Choix de $\mathbf{r}$ et modèles VODKA-PLSR correspondants

4.3.3 Résultats

Comparaison des 6 modèles

Les résultats des $RMSEP$ sont présentés dans le tableau 4.2. Toutes les valeurs inférieures ou égales au plus petit $RMSEP$ du plus mauvais modèle, soit 1.02, sont représentés en gras.

Le meilleur modèle est incontestablement m_5 construit avec le NAS. Pour quatre choix de A variables latentes, il donne une erreur de prédiction meilleure que m_6, la PLSR centrée. Après m_5, les deux meilleurs modèles sont m_1 et m_2. Ils ont en commun avec m_5 de présenter une large de plage de nombre de variables latentes où le $RMSEP$ est proche de son minimum. Pour ces trois modèles, une petite erreur dans le choix du nombre de variables latentes n'a pas trop de conséquences. Ce n'est pas le cas du modèle classique de PLSR centré. Le meilleur $RMSEP$ est obtenu pour 7 variables latentes avec une valeur de 0.95. Mais pour 8 variables latentes, le $RMSEP$ monte à 1.25. Ici une petite erreur dans le choix du nombre de variables latentes peut compromettre le modèle. Enfin les deux plus mauvais modèles sont m_3, la PLSR non centrée, et m_4, la projection sur le spectre pur de l'éthanol.

Comparaison des b-coefficients des modèles NIPALS-PLSR $m3$ et VODKA-PLSR $m5$

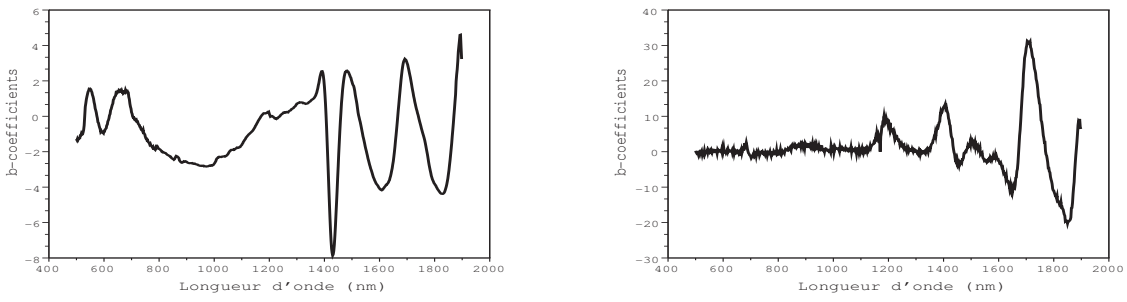
Il n'était pas possible de comparer les b-coefficients obtenus avec les modèles m_5 et m_6 puisqu'ils n'utilisent pas la même forme de données : m_5 utilise les spectres bruts, m_6 utilise les spectres centrés. Nous avons donc choisi de comparer $m3$, PLS sur données brutes, avec m_6 obtenu avec le NAS. Le choix de 7 variables latentes a été fait pour les 2 modèles : très près de

Données $\mathbf{r}$	m_1 $\mathbf{X}, \mathbf{y}$ $\mathbf{1}_P$	m_2 $\mathbf{X}, \mathbf{y}$ $\mathbf{X}'\mathbf{1}_N$	m_3 $\mathbf{X}, \mathbf{y}$ $\mathbf{X}'\mathbf{y}$	m_4 $\mathbf{X}, \mathbf{y}$ $\mathbf{k}$	m_5 $\mathbf{X}, \mathbf{y}$ **NAS**	m_6 $\mathbf{X}_c, \mathbf{y}_c$ $\mathbf{X}'_c\mathbf{y}_c$
LV1	10.2	5.93	6.22	3.14	1.06	4.11
LV2	2.25	3.80	3.74	3.07	1.04	3.23
LV3	2.06	5.06	2.81	3.27	**1.02**	1.86
LV4	2.09	3.16	1.81	2.26	1.04	1.23
LV5	2.30	2.22	1.26	1.93	**0.94**	1.05
LV6	2.94	2.50	1.04	2.42	**0.92**	**1.00**
LV7	1.43	2.23	1.03	1.88	**0.92**	**0.95**
LV8	1.12	1.46	1.34	1.21	**0.93**	1.25
LV9	1.09	**0.94**	**1.02**	**1.02**	**0.97**	**1.02**
LV10	1.08	**0.93**	1.38	**1.01**	**0.99**	1.40
LV11	**0.99**	**1.02**	1.19	**1.02**	**1.02**	1.20
LV12	**0.96**	**0.97**	1.08	1.03	1.04	1.11
LV13	**0.97**	**1.01**	1.19	1.03	1.04	1.23
LV14	**0.96**	**1.00**	1.18	1.02	1.01	1.22
LV15	1.22	1.11	1.16	1.17	1.28	1.21
LV16	1.22	1.05	1.27	1.19	1.29	1.33
LV17	1.22	1.11	1.34	1.19	1.28	1.35
LV18	1.13	1.20	1.54	1.13	1.21	1.44
LV19	1.12	1.19	1.58	1.13	1.22	1.58
LV20	1.08	1.16	1.63	1.09	1.19	1.71

TAB. 4.2 – Erreurs standard de prediction (RMSEP)

l'optimum de $RMSEP$ pour m_3 ; à l'optimum pour m_5.

Les b-coefficients de ces deux modèles, présentés figure 4.1, sont différents. En dessous de $1100nm$, les b-coefficients du modèle NAS sont proches de 0. Cela peut être expliqué par le spectre pur de l'éthanol, également nul dans cette plage. Dans la même plage, les coefficients de la PLSR ne sont pas nuls. Cela s'explique par le fait que $\mathbf{X}'\mathbf{y}$ n'est pas nul dans cette plage du fait d'autres composés du vin qui absorbent, en particulier les anthocyanes. Comme l'extraction des anthocyanes est contrôlée partiellement par la teneur en éthanol, une corrélation existe entre ces deux composés qui a été exploitée par la PLSR. En dessus de $1100nm$, des différences existent aussi dans l'intensité des pics, par exemple vers $1420nm$ et $1700nm$, ou dans la présence de pics vers $1200nm$ et $1600nm$ pour les b-coefficients de m_5.



(a) (b)

FIG. 4.1 – Vecteurs de b-coefficients pour 7 variables latentes obtenus par les modèles m3 NIPALS (a) et m5-VODKA-PLSR avec r=NAS (b)

4.4 Discussion

Trois propriétés découlent de la nouvelle écriture de PLSR.

4.4.1 Informations expérimentales et informations expertes dans le modèle PLSR

Jusqu'à présent, les modèles inverses utilisaient uniquement de l'information expérimentale sous forme d'un jeu d'étalonnage. Le modèle VODKA-PLSR permet d'utiliser d'autres sources d'informations, expertes ou expérimentales, au coeur même du calcul de l'algorithme de régression. Ainsi dans l'application le paramètre $\mathbf{r}$ est obtenu avec deux autres informations : (1) une

information experte constituée de spectres purs ; (2) une information expérimentale différente du jeu d'étalonnage. Le choix de $\mathbf{r}$ est essentiel pour les performances des modèles obtenus, toute l'information complémentaire utilisée pour définir $\mathbf{r}$ est de nature à les améliorer. Le champ des possibles est très large pour $\mathbf{r}$, une connaissance experte des produits analysés peut ainsi être exploitée pour s'orienter rapidement vers les meilleurs choix. Les modèles obtenus par VODKA-PLSR devraient dépasser les performances de NIPALS classique.

4.4.2 Choix de l'algorithme NIPALS

Les algorithmes NIPALS-P et sa généralisation VODKA-PLSR calculent $\mathbf{\Sigma}$, une pseudo-inverse. Cette opération est coûteuse en temps. Pour réaliser une PLSR en utilisant $\mathbf{r}$ différent de $\mathbf{X}'\mathbf{y}$, l'algorithme proposé VODKA-PLSR est pour l'instant la seule solution. Pour réaliser une PLSR classique, l'algorithme NIPALS est nettement plus rapide. En pratique, NIPALS pourrait être utilisé au début de la construction d'un modèle, pour confirmer la possibilité d'obtenir un étalonnage robuste et pour définir le choix des prétraitements. Si les résultats sont concluants, l'utilisation de VODKA-PLSR dans une seconde étape conduirait à un meilleur modèle.

4.4.3 Présence d'une incohérence dans NIPALS ?

Une récente discussion a opposé les partisans d'une incohérence de NIPALS ([31], [32] aux partisans de sa parfaite cohérence [33]). Les coordonnées ou scores $\mathbf{T}$ de NIPALS introduisent $\mathbf{W}$ ce qui induit une incohérence apparente des b-coefficients de NIPALS. Mais leur réecriture selon l'équation 2.19 montre la parfaite cohérence avec le modèle général, lui-même fondé sur des règles basiques d'algèbre linéaire. Pour nous il n'y a donc pas d'incohérence dans NIPALS.

5

Discussion et conclusion

Sommaire

5.1 Place centrale des informations utiles et nuisibles 73

5.1.1 L'information utile, pour les étalonnages 75

5.1.2 L'information nuisible, pour les prétraitements 77

5.2 La notion de métrique introduite par $\mathbf{\Sigma}$ 78

5.2.1 Construction de $\mathbf{\Sigma}$ 78

5.2.2 Utilité fonctionnelle de $\mathbf{\Sigma}$ pour les étalonnages directs 79

5.2.3 Utilité fonctionnelle de $\mathbf{\Sigma}$ pour les étalonnages inverses 80

5.2.4 Perspectives d'une métrique $\mathbf{S}$ dans $\mathbb{R}^N$ 80

5.3 Combinaison de modèles d'étalonnage et de prétraitement 81

5.4 Le NAS, concentré d'information experte pour l'IDC et VODKA-PLSR 82

5.5 Gestion par les projections orthogonales de plusieurs informations nuisibles 83

5.6 Conclusion générale 83

Le coeur de ce mémoire concerne la gestion de l'information par les modèles d'étalonnage et de prétraitement. Les trois sections suivantes lui sont consacrées.

5.1 Place centrale des informations utiles et nuisibles

Le processus de construction d'un modèle d'étalonnage ou de prétraitement peut être décomposé en plusieurs étapes (voir figure 5.1). Les données à notre disposition ont deux origines. Elles

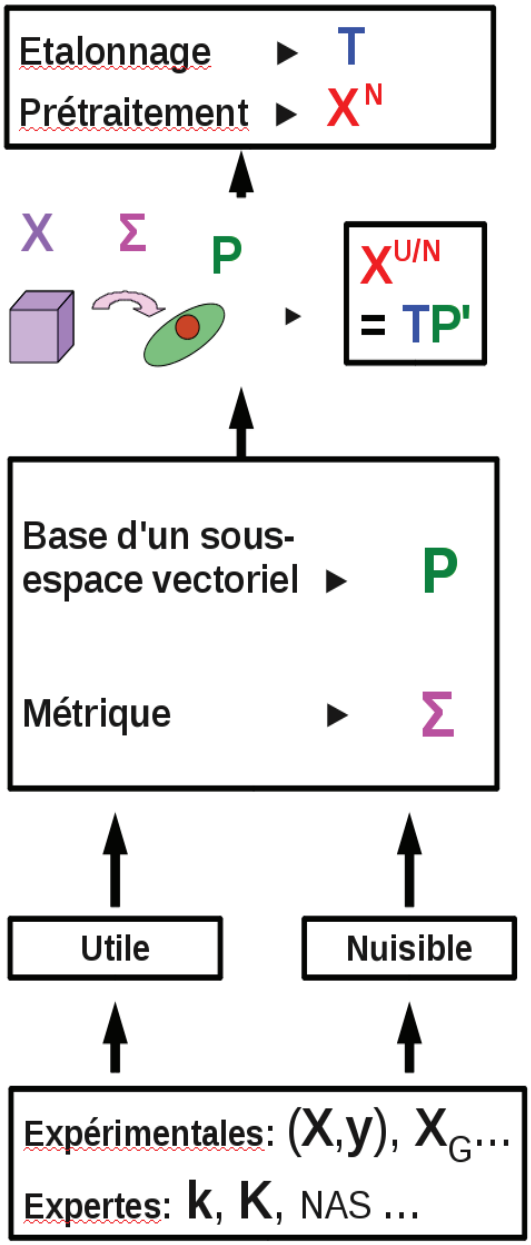


FIG. 5.1 – Modèle général : utilisation des informations expérimentales et expertes pour caractériser les informations utiles ou nuisibles, bases de la construction de modèles d'étalonnage et de prétraitement

peuvent être expérimentales, résulter d'acquisitions de données sur des échantillons. Elles peuvent aussi être expertes, résulter d'une connaissance universelle, par exemple des spectres purs. Ces données expérimentales et expertes servent à caractériser au mieux soit les informations utiles, soit les informations nuisibles, soit les deux. Ces informations utiles et nuisibles conduisent au calcul des deux paramètres $\mathbf{P}$ et $\mathbf{\Sigma}$, desquels un modèle d'étalonnage ou de prétraitement pourra être construit selon le modèle général. Ses performances dépendront à la fois de la qualité de l'information mise en oeuvre, ainsi que de la manière dont cette information sera utilisée.

La représentation des informations utiles, nuisibles et inutiles est donnée par la figure 1.1 page 24. Le choix d'une méthode est orientée par la nature de l'information identifiée, voir figure 5.2.

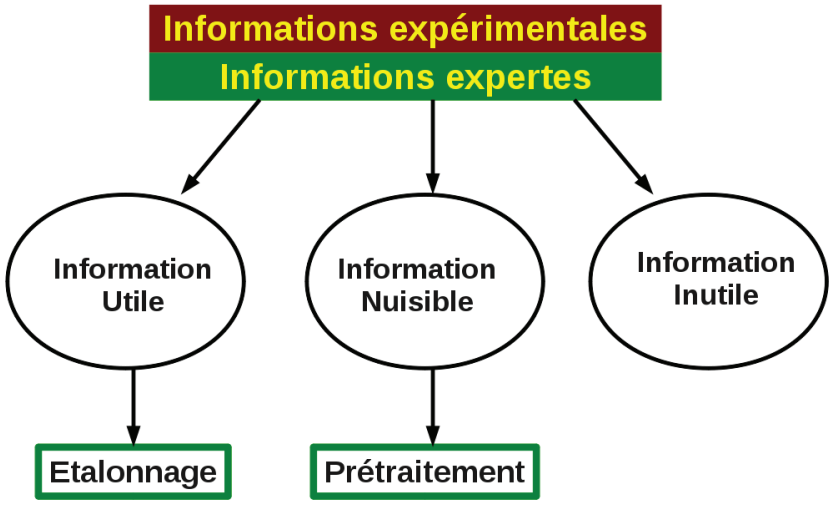


FIG. 5.2 – Modèle général : utilisation des informations expérimentales et expertes pour caractériser les informations utiles ou nuisibles en vue de construire des modèles d'étalonnage ou de régression

5.1.1 L'information utile, pour les étalonnages

Les étalonnages utilisent toujours une information utile, parfois corrigée d'une information nuisible (cas des étalonnages directs). Cependant la nature de l'information expérimentale ou

experte restreint le champ des possibilités.

L'étalonnage direct, en absence d'informations expérimentales

Les étalonnages inverses supposent de connaitre une base d'étalonnage. Cette base peut ne pas être accessible pour des raisons de coût (trop cher, trop lent,...) ou plus souvent à cause de contraintes techniques. Un exemple classique est donné par l'imagerie hyperspectrale. Un étalonnage inverse demanderait à ce qu'un certain nombre de pixels, l'unité de surface de base en spectroscopie, puissent être analysés de manière à quantifier la grandeur d'intérêt. Cela n'est pas possible, dans cette situation seul un étalonnage direct peut être appliqué. Les performances d'un modèle d'étalonnage direct dépendent directement de la qualité des informations à notre disposition. Le terme ˜qualité ˝recouvre certes un concept d'exhaustivité, cet aspect a été largement discuté au chapitre 3 et l'accent mis sur l'importance d'associer les informations expérimentales et expertes afin de caractériser toute l'information nuisible. Mais le terme ˜ qualité ˝recouvre aussi un concept de qualité d'acquisition. Chaque grandeur d'influence doit être caractérisée de la manière la plus indépendante possible. Et le spectre de la grandeur d'intérêt doit être aussi précis que possible. Nous avons vu que l'IDC est particulièrement intéressante pour l'analyse d'images hyperspectrales. Il est possible de sélectionner une partie de l'image dont les spectres seront pris comme spectres purs pour une grandeur d'influence ou pour la grandeur d'intérêt. Une conséquence non voulue peut être d'incorporer une grandeur non apparente. Si l'information extraite de l'image concerne une grandeur d'influence et que la grandeur non apparente est aussi une grandeur d'influence, ce n'est pas un probléme. Si l'information extraite de l'image concerne une grandeur d'influence et que la grandeur non apparente est la grandeur d'intérêt, le modèle perdra toute capacité de prédiction. Si l'information extraite de l'image concerne la grandeur d'intérêt et que la grandeur non apparente est une grandeur d'influence, le modèle obtenu peut être fortement perturbé. Ainsi une bonne pratique d'application d'un étalonnage direct en imagerie hyperspectrale consisterait à utiliser le maximum d'informations expertes collectées en conditions rigoureuses au laboratoire, et obligatoirement le spectre pur de la grandeur d'intérêt. Les informations expérimentales extraites de l'image sont utiles et utilisables mais doivent être soigneusement sélectionnées.

L'étalonnage inverse, en l'absence d'informations expertes

Généralement les grandeurs d'intétrêt sont des concentrations de molécules connues dont le spectre pur est connu. Toutefois il peut arriver que la grandeur d'intérêt n'ait pas de spectre pur. C'est par exemple le cas pour toutes les mesures rhéologiques, fermeté, viscosité,...Dans ce cas l'étalonnage direct est difficile voire impossible à appliquer, par contre l'étalonnage inverse reste une bonne alternative.

Cette situation est particulière puisque la prédiction n'est plus basée sur un signal, par exemple une empreinte de spectre pur, mais sur l'effet d'autres composés chimiques plus ou moins corrélés à la grandeur d'intérêt estimée. Par exemple la fermeté d'une baie de raisin dépend de son niveau de maturation, donc de sa teneur en sucres et en acides puisque les sucres augmentent, l'acidité baisse et la fermeté diminue au cours de la maturation. L'espace spectral utile pour la fermeté est apporté par les informations sucre et acidité complétées par d'autres contributions de grandeurs physiques, alors qu'*a priori* il s'agit d'informations nuisibles. Un modèle peut certes être construit et utilisé, mais des problèmes de robustesse peuvent être attendus.

5.1.2 L'information nuisible, pour les prétraitements

Les prétraitements s'intéressent principalement à l'information complémentaire à une information qui n'est pas une information utile. Il faut bien faire la distinction entre les informations nuisibles et les informations inutiles, deux notions très différentes, voir figure 1.1 page 24.

- L'information nuisible est une information spectrale apportée par une grandeur d'influence dont l'empreinte spectrale est partiellement collinéaire avec celle de la grandeur d'intérêt. Cela se traduit par le partage d'un sous-espace vectoriel commun noté $\mathcal{E}^U \cap \mathcal{E}^N$, voir figure 1.1 page 24. C'est cette propriété qui justifie le qualificatif de nuisible. La performance d'un étalonnage est directement liée à sa capacité à identifier la partie de l'information utile qui n'est pas influencée par l'information nuisible.
- L'information inutile est l'information spectrale orthogonale à l'information utile. Cette information n'est pas prise en compte par les étalonnages.

Ainsi l'élimination de l'information nuisible est un enjeu majeur, alors que l'élimination de l'information inutile n'offre pas d'intérêt pratique. Pourtant la confusion est parfois faite.

L'Orthogonal-PLSR ou OPLSR [7] est une PLSR précédée d'une OSC. Il a été démontré ([23],[24]) que cette OSC n'améliore pas les performances du modèle OPLSR lorsque celui-ci est

comparé à une PLSR classique. Lors de l'exécution de l'OPLSR, soient a et b les nombres d'axes respectivement éliminés par l'OSC et utilisés par la PLSR. Alors une simple PLSR classique avec $c = a + b$ axes donne exactement le même modèle ([24]). Cette constatation est expliquée par la confusion faite entre information nuisibles et informations inutiles. En effet les composantes $\mathbf{t}_{OSC}$ sont choisies selon un critère d'orthogonalité par rapport à $\mathbf{y}$. L'espace enlevé par l'OSC contient donc uniquement de l'information inutile. D'où l'inaptitude de l'OSC et de l'OPLSR à améliorer les performances (précision, robustesse) des étalonnages.

Cependant l'outil OSC n'est pas mis en cause. Avec la même routine OSC utilisée différemment, les conclusions sont différentes. Nous avons proposé d'appliquer le modèle OSC sur des informations autres que le jeu d'étalonnage : des informations expertes sous forme de spectres purs ([34]) ou des informations expérimentales sous forme de spectres issus d'un plan d'expérience ([35]). Dans ces deux cas, l'information nuisible était bien caractérisée et a été prise en compte par l'OSC, ce qui a permis une amélioration significative des modèles obtenus.

5.2 La notion de métrique introduite par $\mathbf{\Sigma}$

Le modèle général fait apparaître une nouvelle matrice $\mathbf{\Sigma}$, à vocation de métrique ou pseudo-métrique dans $\mathbb{R}^P$. L'intérêt de $\mathbf{\Sigma}$ est de modifier l'espace lors des projections, de manière à favoriser certaines directions et d'en défavoriser d'autres. Les propriétés de $\mathbf{\Sigma}$ sont caractérisées en trois parties : (1) la construction de $\mathbf{\Sigma}$; (2) le noyau Ker de la fonction produit-scalaire ; (3) la relation entre $\mathbf{\Sigma}$ et $\mathbf{P}$. Une quatrième partie introduit rapidement la notion de métrique dans $\mathbb{R}^N$.

5.2.1 Construction de $\mathbf{\Sigma}$

Dans le modèle général, $\mathbf{\Sigma}$ est une matrice carrée symétrique de dimensions $(P \times P)$ ayant la signification d'une métrique, c'est à dire un objet mathématique permettant de mesurer des distances. En plus de $\mathbf{I}_P$, la métrique Euclidienne, deux autres formes de matrices $\mathbf{\Sigma}$ ont été rencontrées dans les implémentations du modèle général. Nous montrons qu'il s'agit bien au moins de pseudo-métriques.

- Dans la SBC, la matrice $\mathbf{\Sigma}$ est sous la forme : $(\mathbf{U}'\mathbf{U})^{-1}$. Dans la PLSR ou l'OSC, elle est sous la forme d'une pseudo-inverse de Moore-Penrose $(\mathbf{U}'\mathbf{U})^{+}$. Il s'agit d'une distance de Mahalanobis lorsque $\mathbf{U}$ est centrée et $\mathbf{U}'\mathbf{U}$ est inversible. Dans tous les cas, $\mathbf{\Sigma}$ est au

moins semi-définie positive. En effet, soit $\mathbf{ADB'}$ la SVD de $\mathbf{U}$. Alors $\mathbf{U'U} = \mathbf{BEB'}$ avec $\mathbf{E}$ obtenue en portant au carré chaque terme de $\mathbf{D}$. Dès lors $(\mathbf{U'U})^+ = \mathbf{BFB'}$ avec $\mathbf{F}$ obtenue en inversant tous les termes non nuls de la diagonale de $\mathbf{E}$. On remarque que $\mathbf{E}$ et $\mathbf{F}$ sont des matrices diagonales ne comportant que des termes positifs ou nuls. Soit maintenant un vecteur $\mathbf{x}$ quelconque non nul de dimension $(P \times 1)$. Il existe un vecteur $\mathbf{z}$ tel que $\mathbf{z} = \mathbf{B'x}$. Alors $\mathbf{x'}(\mathbf{U'U})^+\mathbf{x} = \mathbf{v'Fv}$. Cette valeur est évidemment positive ou nulle, d'où la propriété recherchée.

- La matrice $\mathbf{\Sigma}$ est sous la forme : $\mathbf{I}_P - \mathbf{U}(\mathbf{U'U})^{-1}\mathbf{U'}$ dans deux étalonnages directs : DC et IDC. Le produit $\mathbf{x'\Sigma x}$ peut sécrire $\mathbf{x'}(\mathbf{I}_P - \mathbf{U}(\mathbf{U'U})^{-1}\mathbf{U'})(\mathbf{I}_P - \mathbf{U}(\mathbf{U'U})^{-1}\mathbf{U'})\mathbf{x}$. Soit $\mathbf{v} = (\mathbf{I}_P - \mathbf{U}(\mathbf{U'U})^{-1}\mathbf{U'})\mathbf{x}$. Alors $\mathbf{x'\Sigma x} = \mathbf{v'v} \geqq 0$, soit la propriété recherchée.

5.2.2 Utilité fonctionnelle de $\mathbf{\Sigma}$ pour les étalonnages directs

Lors de la construction des étalonnages directs : DC, SBC, IDC, l'information expérimentale ou experte est utilisée dans le but d'incorporer de l'information nuisible dans $\mathbf{\Sigma}$. Deux formes de matrices $\mathbf{\Sigma}$ ont été observées : pondération (Soft Correction) ou projection orthogonale (Hard Correction). Nous étudions les performances théoriques de ces deux formes à la lumière des propriétés du noyau du produit scalaire qu'elles définissent.

Le produit scalaire selon $\mathbf{\Sigma}$ est la fonction $\mathcal{F}$ de $\mathbb{R}^P \times \mathbb{R}^P$ dans $\mathbb{R}$ telle que, pour tout couple de vecteurs $(\mathbf{u}, \mathbf{v})$, $\mathcal{F}(\mathbf{u}, \mathbf{v}) = \mathbf{u'\Sigma v}$. Le produit scalaire définit une norme : $\| \mathbf{u} \| = \sqrt{\mathbf{u'\Sigma u}}$ ainsi qu'une distance : $d(\mathbf{u}, \mathbf{v}) = \| \mathbf{u} - \mathbf{v} \|$.

Cas où $\mathbf{\Sigma}$ est de la forme $(\mathbf{U'U})^+$

Soit un vecteur $\mathbf{u}$ non nul de dimensions $(P \times 1)$ de l'espace nuisible, et $\mathbf{\Sigma} = (\mathbf{uu'})^+$ construite à partir de $\mathbf{u}$. Alors $\mathbf{u'\Sigma}$ est égal à $\mathbf{u'}(\mathbf{uu'})^+$, soit la pseudo-inverse de $\mathbf{u}$ qui donc ne peut pas être égal au vecteur nul. Donc $\mathbf{u'\Sigma u} \neq 0$. Ainsi le sous-espace vectoriel de l'information nuisible n'est pas inclus dans le noyau de cette fonction de produit scalaire. La correction de l'information nuisible n'est que partielle, d'où le qualificatif de Soft Correction.

Cas où $\mathbf{\Sigma}$ est une projection orthogonale

Supposons que les colonnes de $\mathbf{U}$ forment une base de l'espace nuisible. Soit $\mathbf{\Sigma}$ la projection orthogonale à $\mathbf{U}$. Alors tout vecteur-ligne $\mathbf{u}$ appartenant à l'espace nuisible vérifie : $\mathbf{u'\Sigma u} = 0$. Le sous-espace vectoriel de l'information nuisible est inclus dans le noyau $Ker(\mathcal{F})$. Toute

l'information nuisible est éliminée par la projection orthogonale à $\mathbf{U}$, d'où le qualificatif Hard Correction.

5.2.3 Utilité fonctionnelle de $\mathbf{\Sigma}$ pour les étalonnages inverses

Contrairement aux étalonnages directs qui incorporent de l'information nuisible dans $\mathbf{\Sigma}$, les étalonnages inverses utilisent deux formes de matrice sans information nuisible clairement identifiée : $\mathbf{I}_P$ et $(\mathbf{X}'\mathbf{X})^+$. L'identité $\mathbf{I}_P$ correspond à une métrique Euclidienne, elle n'entraine aucune déformation de l'espace lors des projections orthogonales ; alors que $(\mathbf{X}'\mathbf{X})^+$ déforme la projection. Bien que $(\mathbf{X}'\mathbf{X})^+$ ne contienne pas d'information nuisible clairement identifiée, nous pouvons malgré tout y voir un avantage pour l'information utile. Comparons la PCR et la PLSR.

Avec $(\mathbf{X}'\mathbf{X})^+$, cas de la PLSR, chaque variable spectrale est pondérée par des coefficients d'autant plus petits que cette variable et celles qui lui sont corrélées ont de la variabilité dans $\mathbf{X}$. Donc la PLSR attribue globalement des importances comparables à toutes les variables indépendemment de leur variabilité dans $\mathbf{X}$. Ce n'est pas le cas de la PCR où les variables ayant le plus d'importance sont celles ayant le plus de variabilité dans $\mathbf{X}$. Lorsque l'information utile se traduit par de fortes variations spectrales, les modèles PCR et PLSR auront des performances comparables. Mais lorsque l'information utile est apportée par des composés quantitativement minoritaires, donc se traduit par de faibles variations spectrales, la métrique de la PLSR a nettement plus de potentiel que la PCR pour prendre en compte cette information utile. C'est une raison expliquant l'avantage de la PLSR sur la PCR. Mais pour avoir encore plus de performances avec la PLSR, $\mathbf{\Sigma}$ peut être construite avec de l'information nuisible, par exemple une projection orthogonale à l'espace nuisible. Des premiers résultats, non présentés ici, sont très prometteurs.

5.2.4 Perspectives d'une métrique S dans $\mathbb{R}^N$

Le modèle général tel qu'il est présenté au chapitre 2 page 27 s'appuie implicitement sur deux espaces vectoriels : (1) l'espace $\mathbb{R}^N$ des individus muni de la métrique Euclidienne (voir l'équation 2.4) ; (2) l'espace $\mathbb{R}^P$ des variables muni de la métrique $\mathbf{\Sigma}$. Il est concevable que l'espace des individus soit muni d'une métrique non Euclidienne représentée par une matrice $\mathbf{S}$ de dimensions $(N \times N)$. Dès lors l'équation 2.4 page 29 serait remplacée par :

$$\widehat{\mathbf{y}} = \mathbf{T}(\mathbf{T}'\mathbf{S}\mathbf{T})^{-1}\mathbf{T}'\mathbf{S}\mathbf{y} \tag{5.1}$$

Dans le cas particulier où $\mathbf{T} = \mathbf{X}$, certains choix de $\mathbf{S}$ permettent de retrouver les méthodes WLSR et GLSR déjà décrites au chapitre 1. Au vu de ces applications, l'introduction d'une métrique $\mathbf{S}$ dans $\mathbb{R}^N$ n'est pas une idée nouvelle. Toutefois dans les deux cas WLSR et GLSR, la métrique $\mathbf{\Sigma}$ est Euclidienne. La question soulevée est la suivante : peut-on envisager un couple de métriques $(\mathbf{\Sigma}, \mathbf{S})$ indépendantes l'une de l'autre sans qu'une des deux soit Euclidienne ? Et si $\mathbf{\Sigma}$ et $\mathbf{S}$ peuvent être non Euclidiennes mais sont obligatoirement liées, quelle doit être la nature de leur relation ? Ce questionnement ouvrant potentiellement de nouvelles perspectives n'a pas été abordé.

5.3 Combinaison de modèles d'étalonnage et de prétraitement

Les étalonnages directs utilisent toujours de l'information experte, et éventuellement de l'information expérimentale. Au contraire, les étalonnages inverses ou régressions sont basés sur l'utilisation exclusive de l'information expérimentale, à l'exception de VODKA-PLSR qui combine informations expérimentales et expertes. Enfin les prétraitements performants de type EPO ou EMSC utilisent de l'information experte ou de l'information expérimentale différente de celle du jeu d'étalonnage. La question posée est de savoir comment les différentes méthodes d'étalonnage et prétraitement peuvent être combinées pour une utilisation optimale des informations expérimentales et expertes à notre disposition.

Une première stratégie consiste à intégrer l'information experte dans un prétraitement de type EPO ou EMSC par exemple, et à l'associer à une régression de type PLSR intégrant de l'information expérimentale. Il est reconnu que ces combinaisons améliorent significativement les performances de la PLSR puisque les informations mobilisées sont complémentaires.

Une autre stratégie est offerte par le modèle VODKA-PLSR puiqu'il permet aussi d'introduire de l'information experte ou une autre information expérimentale dans la régression. C'est le cas par exemple si le spectre pur de la grandeur d'intérêt est choisi comme paramétre $\mathbf{r}$. Cette approche s'est révélée aussi très performante, meilleure que la PLSR classique.

Il se pose donc la question de la gestion de l'information nuisible en complément ou avec les régressions. Est-il préférable de l'utiliser en deux étapes : prétraitement puis étalonnage inverse, comme c'est le cas actuellement ? Ou bien de l'incorporer directement dans le modèle de régression de VODKA-PLSR ? La réponse à cette question optimiserait l'utilisation conjointe des informations expertes et expérimentales.

En complément des considérations générales discutées dans les sections précédentes, les méthodes implémentées selon le modèle général confirment l'importance de deux concepts développés dans les sections suivantes.

5.4 Le NAS, concentré d'information experte pour l'IDC et VODKA-PLSR

Le NAS est l'information spectrale de la grandeur d'intérêt orthogonale à l'information spectrale des grandeurs d'influence. Cette définition postule que le NAS est défini dans un espace de dimension 1. Selon la figure 1.1 page 24, le NAS correspond à l'information de l'espace $\mathcal{E}^U$ qui n'appartient pas aussi à $\mathcal{E}^N$.

Toute la difficulté est d'estimer le NAS au mieux, de préférence en associant des informations expérimentales avec des informations expertes. Le NAS est donc un concentré d'information. Des étalonnages directs tels la DC ou l'IDC sont basés sur une définition du NAS. Les étalonnages indirects ont une relation plus complexe avec le NAS. Plusieurs grandeurs chimiques fortement corrélées à la grandeur d'intérêt peuvent être utilisées par un modèle de régression. Cette information n'est pas apportée par le NAS. Dans un cas extrême [36] nous avons pu observer que la quantification du gluten dans des farines de blé n'utilise pas l'information spectrale du gluten, mais celle de l'amidon qui est très fortement corrélé au gluten. Pour ce type d'étalonnage fondé uniquement sur des corrélations indirectes, le NAS n'est d'aucune utilité. Toutefois la plupart des modèles d'étalonnage s'appuient sur le signal spécifique de la grandeur d'intérêt. Dans ces situations, l'utilisation du NAS devrait donner des modèles plus robustes puisque insensibles aux influences des grandeurs corrélées plus ou moins avec la grandeur d'intérêt. Parmi les étalonnages indirects, la méthode VODKA-PLSR a la capacité d'utiliser directement le NAS.

Mais le concept de NAS a ses limites puisque l'information utile peut aussi appartenir à un sous-espace vectoriel de dimension supérieure à 1. Les étalonnages directs ne sont pas conçus pour ce cas. Les étalonnages inverses sont bien mieux adaptés.

5.5 Gestion par les projections orthogonales de plusieurs informations nuisibles

L'utilisation conjointe d'informations expérimentales et expertes peut conduire à des données de différentes origines réparties dans des matrices différentes. Nous avons vu avec l'IDC qu'il est possible d'obtenir une matrice $\mathbf{R}$ concaténant différentes informations. De manière plus générale, ce concept peut être élargi aux projections orthogonales.

Supposons que nous ayons deux informations complémentaires sur les grandeurs d'influence, contenues dans les matrices $\mathbf{P}_1$ et $\mathbf{P}_2$. Ces informations peuvent être expérimentales et/ou expertes. L'objectif de la projection orthogonale est d'enlever les informations des sous-espaces vectoriels engendrés par les vecteurs-colonne de $\mathbf{P}_1$ et de $\mathbf{P}_2$. Cette opération n'est pas possible par deux projections orthogonales successives si les vecteurs de $\mathbf{P}_1$ et ceux de $\mathbf{P}_2$ ne sont pas orthogonaux entre eux. La meilleure solution est de construire une matrice $\mathbf{R}$ par concaténation de $\mathbf{P}_1$ et de $\mathbf{P}_2$, puis de projeter orthogonalement à $\mathbf{R}$.

5.6 Conclusion générale

L'étude de la gestion des informations expérimentales et expertes nous a conduit à un modèle général appliquable aux méthodes d'étalonnage comme de régression. Ce modèle est très informatif sur la manière dont les informations expérimentales et expertes sont gérées par deux matrices $\mathbf{P}$ et $\mathbf{\Sigma}$. Il montre bien que l'objectif premier est d'obtenir le sous-espace vectoriel le plus petit possible contenant l'information spectrale utile. Les étalonnages identifient directement l'information utile. Les prétraitements identifient puis enlèvent l'information nuisible. L'utilisation conjointe d'informations expérimentales et expertes n'est pas un dogme ni une obligation, mais un moyen pratique d'accéder à l'information utile et/ou nuisible la plus exhaustive possible, garante des meilleurs modèles prédictifs.

Basées sur ce modèle général, deux nouvelles méthodes sont proposées. La première méthode est l'IDC-Improved Direct Calibration, une méthode d'étalonnage direct. Elle intervient sur la construction de $\mathbf{\Sigma}$ en associant informations expérimentales et expertes. La deuxième méthode est VODKA-PLSR, l'identification d'une famille de modèles de régression parmi lesquels se trouve la PLSR classique. Un vecteur $\mathbf{r}$ est déterminé à partir d'informations expérimentales et/ou expertes. Il est utilisé pour définir l'espace utile dont les colonnes de $\mathbf{P}$ forment une base. Ces deux

méthodes élargissent l'utilisation des informations expérimentales et expertes par les étalonnages et prétraitements. Elles sont directement utilisables et devraient conduire à des modèles plus performants.

Le modèle général et ses deux implémentations IDC et VODKA-PLSR ouvrent un grand nombre de possibilités. Les possibilités pour déterminer l'information utile, représenté par $\mathbf{P}$, sont élargies *via* le paramètre $\mathbf{r}$ et l'algorithme proposé dans VODKA-PLSR. En ce qui concerne $\mathbf{\Sigma}$, sa signification est à approfondir. Dans les étalonnages directs, $\mathbf{\Sigma}$ est directement construite à partir d'informations nuisibles. Nous voyons alors se dessiner une symétrie d'association entre information utile dans $\mathbf{P}$ et information nuisible dans $\mathbf{\Sigma}$. La construction de matrices $\mathbf{\Sigma}$ contenant de l'information nuisible et utilisées dans les étalonnages inverses est une option prometteuse. Cela ne restreint en rien la conception de nouvelles matrices $\mathbf{\Sigma}$ dont le calcul serait basé sur des notions de distance totalement différentes à celles qui sont décrites dans ce mémoire.

Ce mémoire est axé sur la prédiction d'une seule grandeur d'intérêt. Des généralisations sont envisageables dans deux directions. En premier lieu l'application de ce modèle à la gestion de plusieurs grandeurs d'intérêt n'est pas abordée. Certes une solution pratique est de construire un modèle pour chaque grandeur d'intérêt ; mais pourquoi pas une gestion globale ? En second lieu, l'application de ce modèle à plusieurs tableaux de données spectrales pourrait constituer une passerelle vers l'analyse multitableaux.

A

Script Matlab et Scilab de la fonction VODKA-PLSR

Avertissement :

Ce script permet à chacun de réaliser facilement une VODKA-PLSR à l'aide de Scilab ou Matlab. L'objectif est une lisibilité maximum associée à un minimum de lignes de commandes. C'est pourquoi l'environnement est réduit : aucun prétraitement des données, pas de validation croisée ni de calcul d'écart-type de prédiction par exemple. Une utilisation de VODKA-PLSR en routine impliquerait une programmation plus complète.

En choisissant $\mathbf{r} = \mathbf{X}'\mathbf{y}$, les résultats ainsi obtenus pourront être comparés avec ceux de l'algorithme NIPALS de la PLSR1 d'une suite logicielle validée (Saisir, PLS-Toolbox, Sigma, Unscrambler par exemple). Les éléments de comparaison sont les suivants :

- l'identité des b-coefficients obtenus ;
- la colinéarité entre les $\mathbf{p}_i$;
- la vérification que pour les matrices $\mathbf{P}$: $\mathbf{P}'(\mathbf{X}'\mathbf{X})^{+}\mathbf{P}$ est une matrice diagonale.

Ce programme ne contient aucun prétraitement, ceux-ci devront avoir été faits au préalable. Attention, les calculs sont un peu longs (2 à 3 minutes minimum) à cause du calcul de la pseudo-inverse de $\mathbf{X}'\mathbf{X}$. Les données en entrée sont :

x	matrice (n,v)	(données, n spectres, v variables spectrales)
y	vecteur (n,1)	(valeurs de la grandeur d'intérêt à prédire)
r	vecteur (v,1)	(vecteur arbitraire)
a	entier non nul	(nombre maximum de variables latentes)

```
function[res]=vodka _ plsr(x,y,r,a)
[n,v] = size(x) ;
p=zeros(v,a) ;
b=zeros(v,a) ;
xx=x'*x ;
s=pinv(xx) ;
p( :,1)=xx*r ;
p( :,1)=p( :,1)/sqrt(p( :,1)'*s*p( :,1)) ;
b( :,1)=s*p( :,1)*inv(p( :,1)'*s*xx*s*p( :,1))*p( :,1)'*s*x'*y ;
for i=2 :a ;
POP=eye(v,v) - s*p( :,1 :i-1)*p( :,1 :i-1)' ;
r2=POP'*r ;
xx2=xx*POP ;
p( :,i)=xx2'*r2 ;
p( :,i)=p( :,i)/sqrt(p( :,i)'*s*p( :,i)) ;
b( :,i)=s*p( :,1 :i)*inv(p( :,1 :i)'*s*xx*s*p( :,1 :i))*p( :,1 :i)'*s*x'*y ;
end
res.p_loads=p ;
res.b_coeff=b ;
endfunction
```

Glossaire

ACP : Analyse en Composantes Principales

DC : Direct Calibration

DOP : Dynamic Orthogonal Projection

DVS : Decomposition en Valeurs Singulieres

EROS : Error Removal by Orthogonal Substraction

EMSC : Extended Multiplicative Signal Correction

EPO : External Parameter Orthogonalisation

GLSR : Generalised Least Square Regression

IDC : Improved Direct Calibration

NAP : Net Analyte Preprocessing

IDC : Improved Direct calibration

LMM : Linear Mixture Model

MSC : Multiplicative Signal Correction

NAP : Net Analyte Preprocessing

NAS : Net Analyte Signal

NIPALS : Non-Linear Iterative Partial Least Square

OLSR : Ordinary Least Square Regression

OSC : Orthogonal Signal Correction

OPLSR : Orthogonal-PLSR

PLSR : Partial Least Square Regression

PCA : Principal Component Analysis

PCR : Principal Component Regression

RM-CPCA : Regression Models through Constrained PCA

RR : Ridge Regression

SBC : Science-Based Calibration

SIMPLS : Straightforward Implementation of a Statistically-Inspired Modification of the PLS method

SNV : Standard Normal Variate

SVD : Singular Value Decomposition

TOP : Transfer Orthogonal Projection

VODKA-PLSR : Vector Orientation Decided through Knowledge Assessment PLSR

Index

étalonnage, 6, 8, 28, 37
étalonnage direct, 8, 29–31, 37, 55, 81, 82
étalonnage inverse, 8, 12, 29, 37, 81

centrage, 8, 20

Direct Calibration, 8, 31, 37, 61, 82
Dynamic Orthogonal Projection, 16, 31, 37

Error Removal by Orthogonal Substraction, 31
Error Removal by Orthogonal Subtraction, 17, 37
Extended Multiplicative Signal Correction, 21, 32
External Parameter Orthogonalisation, 16, 31, 37

Generalised Least Square Regression, 13
grandeur d'influence, 7

Improved Direct Calibration, 41, 42, 55, 61, 82
Independant Interference Reduction, 16, 31, 37
information expérimentale, 6, 8, 12, 15, 17, 20, 29, 31, 37, 66, 81
information experte, 6, 8, 18, 20, 29, 37, 66, 81
information nuisible, 23, 28, 30
information utile, 8, 23, 28, 30

métrique, pseudo-métrique, 28, 61, 78
modèle général, 27, 28, 30, 31, 34, 37, 55, 78
modèle linéaire de mélange, 2, 10, 42

Net Analyte Preprocessing, 17, 32, 37
Net Analyte Signal, 55, 60, 82
NIPALS, 32, 63, 68, 71
NIPALS-P, 64, 66, 71

Ordinary Least Square Regression, 12, 30, 37
ordonnée à l'origine, 59
Orthogonal Signal Correction, 17, 35, 37
Orthogonal-PLSR, 77

Partial Least Square Regression, 14, 32, 37, 70
pente, 59
plan d'expérience, 31
prétraitement, 6, 15, 17, 18, 20, 28, 30, 31, 37, 39
Principal Component Regression, 14, 30, 37, 80
projection orthogonale, 28, 79
pseudo-inverse, 32
pseudo-inverse (Moore-Penrose), 34, 63

Science-Based Calibration, 11, 31, 37
sous-espace vectoriel nuisible, 37
sous-espace vectoriel utile, 37
Standard Normal Variate, 20, 32

Transfer Orthogonal Projection, 16, 31, 37

Vector Orientation Decided through Knowledge
Assessment-PLSR, 63, 65, 66, 68, 71,
81, 82

Weighted Least Square Regression, 13

Bibliographie

[1] H.Martens, T.Naes, Multivariate Calibration, Wiley, 1989.

[2] H.Martens, J.P.Nielsen, S.B.Engelsen, Light scattering and light absorbance separated by extended multiplicative signal correction, application to near infra-red transmission analysis of powder mixtures, Analytical Chemistry 75(3) (2003) 394–404.

[3] R.Marbach, A new method for multivariate calibration, Journal of Near Infrared Spectroscopy 13 (2005) 241–254.

[4] A.Lorber, K.Faber, B.R.Kowalski, Net analyte signal calculation in multivariate calibration, Analytical Chemistry 69(8) (1997) 1620–1626.

[5] J.Sun, A correlation principal component regression analysis of nir data, Journal of Chemometrics 9 (1995) 21–29.

[6] S.Wold, A.Ruhe, H.Wold, W. D. III, The collinearity problem in linear regression, the partial least square (pls) approch to generalized inverses, Journal of Science and Statistical Computations 5 (1984) 735–743.

[7] J.Trygg, Parsimonious multivariate models., Ph.D. thesis, Umea University, Sweden (2001).

[8] P.W.Hansen, Pre-processing method minimizing the need for reference analyses, Journal of Chemometrics 15 (2001) 123–131.

[9] J.M.Roger, F.Chauchard, V.Bellon-Maurel, Epo-pls external parameter orthogonalisation of pls, application to temperature-independant measurement of sugar contents in fruits, Chemometrics ant Intelligent Laboratory Systems 66 (2003) 191–204.

[10] A.Andrew, T.Fearn, Transfer by orthogonal projection : making near infra-red calibrations robust to between-instrument variation, Chemometrics ant Intelligent Laboratory Systems 72 (2004) 51–56.

[11] M.Zeaiter, J.M.Roger, V.Bellon-Maurel, Dynamic orthogonal projection, a new method to maintain the on-line robustness of multivariate calibration, application to nir-based monitoring of wine fermentations., Chemometrics ant Intelligent Laboratory Systems 80 (2006) 227–235.

[12] Y.Zhu, T.Fearn, D.Samuel, A.Dhar, O.Hameed, S.G.Brown, L.B.Lovat, Error removal by orthogonal substraction (eros) : a customised pre-treatment for spectroscopic data, Journal of Chemometrics 22 (2008) 130–134.

[13] S. Wold, H. Antti, F. Lindgren, J. Ohman, Orthogonal signal correction of near infra-red spectra., Chemometrics ant Intelligent Laboratory Systems 44 (1998) 175–185.

[14] H.C.Goicoechea, A.C.Olivieri, A comparison of orthogonal signal correction and net analyte preprocessing methods, theorical and experimental study, Chemometrics ant Intelligent Laboratory Systems 56 (2001) 73–81.

[15] T.Fearn, On orthogonal signal correction., Chemometrics ant Intelligent Laboratory Systems 50 (2000) 47–52.

[16] A.Savitsky, M.Golay, Smoothing and differenciation of data by simplified least square procedures, Analytical Chemistry 36 (1964) 1627–1639.

[17] R. DeSerio, Savitsky-golay filters, www.compadre.org (2008).

[18] J.Luo, K.Ying, J.Bai, Savitsky-golay and differenciation filter for even number data, Signal Processing 85 (2005) 1429–1434.

[19] R.J.Barnes, M.S.Dhanoa, S.J.Lister, Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra, Applied Spectroscopy 43 (1989) 772–777.

[20] J.Badia, Algèbre matricielle, Vol. FPS01, INRA, 1990.

[21] A.Kohler, C.Kirschner, A.Out, H.Martens, Extended multiplicative signal correction as a tool for separation and characterization of physical and chemical information in fourier transform infrared microscopy images of cryo-sections of beef loin, Applied Spectroscopy 59(6) (2005) 707–716.

[22] B.M.Wise, N.B.Gallagher, J.M.Shaver, M.A.Rasmussen, R.Bro, A guide to the orthogonalisation filter smorgasbord, in : Afrodata-Rabat, 2010.

[23] T.Verron, R.Sabatier, R.Joffre, Some theoretical properties of the o-pls method, Journal of Chemometrics 18 (2004) 62–68.

[24] E.K.Kemsley, H.S.Tapp, Opls filtered data can be obtained directly from non-orthogonalized pls1, Journal of Chemometrics 23 (2009) 518–529.

[25] H.Tenenhaus, La régression PLS, Technip, 1998.

[26] J.F.Durand, Eléments de calcul matriciel et d'analyse factorielle de données, Université Montpellier II, 2002.

[27] B.G.Osborne, T.Fearn, Near infrared spectroscopy in food analysis, Wiley, N.Y., 1986.

[28] F.Guillon, O.Tranquet, L.Quillien, J.P.Utille, J.J.Ordaz-Ortiz, L.Saulnier, Generation of polyclonal and monoclonal antibodies against arabinoxylans and their use for immunocytochemical location of arabinoxylans in cell walls of endosperm of wheat, Journal of Cereal Sciences 40 (2004) 167–182.

[29] S.Philippe, O.Tranquet, J.P.Utille, L.Saulnier, F.Guillon, Investigation of ferulate deposition in endosperm cell walls of mature and developing wheat grains by using a polyclonal antibody, Planta 225 (2007) 1287–1299.

[30] H.Mark, R.Rubinovitz, Chemometric calibration without matrices (almost), in : Pittcon-Chicago, 2009.

[31] R.J.Pell, L.S.Ramos, R.Manne, The model space in partial least squares regression, Journal of Chemometrics 21 (2007) 165–172.

[32] R.Ergon, Re-interpretation of nipals results solves plsr inconsistency problem, Journal of Chemometrics 23 (2009) 72–75.

[33] S.Wold, M.Hoy, H.Martens, J.Trygg, F.Westad, J.MacGregor, B.M.Wise, The pls model space revisited, Journal of chemometrics 23 (2009) 67–68.

[34] J.C.Boulet, T.Doco, J.M.Roger, Improvement of calibration models using two successive orthogonal projection methods, application to quantification of wine mannoproteins., Chemometrics ant Intelligent Laboratory Systems 87 (2007) 295–302.

[35] S.Preys, J.M.Roger, J.C.Boulet, Robust calibration using orthogonal projection and experimental design, application to the correction of the light scattering effect on turbid nir spectra., Chemometrics ant Intelligent Laboratory Systems 91 (2006) 28–33.

[36] J.C.Boulet, J.M.Roger, A new direct calibration method : Idc-improved direct calibration, in : Chimiometrie-Paris, 2009.

Résumé

Les spectres contiennent de l'information sur la composition d'échantillons. Cette information est extraite au moyen d'une première famille d'outils chimiométriques, les étalonnages. Une deuxième famille d'outils, les prétraitements, est destinée à enlever une information spectrale nuisible. Etalonnages et prétraitements sont construits à partir de deux types d'informations : (1) les informations expérimentales basées sur l'expérience ; (2) les informations expertes basées sur la connaissance *a priori*. L'objectif de la thèse est d'étudier les complémentarités et synergies entre ces deux types d'informations. Après une étude bibliographique, un modèle général commun aux étalonnages et prétraitements est proposé. L'information utile ou nuisible contenue dans un spectre est obtenue par projection orthogonale de ce spectre (selon un métrique $\mathbf{\Sigma}$) sur une matrice $\mathbf{P}$ dont les colonnes constituent une base de l'espace vectoriel associé à l'information utile ou nuisible. Selon les cas, l'information utile est conservée alors que l'information nuisible est éliminée. Le modèle général est ensuite implémenté par deux nouvelles méthodes. L'IDC-Improved Direct Calibration est une méthode d'étalonnage direct utilisant conjointement des informations expérimentales et expertes. Ensuite VODKA-PLSR est une généralisation de PLSR. Un vecteur $\mathbf{r}$ est mis en évidence, il permet d'inclure de l'information experte dans le modèle. En conclusion ce travail permet une vision plus synthétique des modèles existants, propose deux nouveaux modèles d'étalonnage et ouvre de nombreuses possibilités pour créer de nouveaux modèles d'étalonnage et de prétraitement.

Mots-clés: information, expérimental, expert, utile, nuisible, modèle, général, étalonnage, direct, inverse, régression, prétraitement, PLSR, IDC, VODKA-PLSR

Abstract

Spectra contain informations about the composition of samples. This information is obtained using calibration. Harmful spectral information can be previoulsy withdrawn using pretraitments. Both calibration and pretraitment models are based on two types of informations : (1) experimental information based on measurements onto samples ; (2) expert information based

on a previous knowledge. The aim of this thesis is to study the links between those two types of information. After a biography review, a general model including both calibrations and pretraitments is proposed. The useful or harmful spectral information is obtained after spectra have been orthogonaly projected (with a $\mathbf{\Sigma}$ metrix) onto a $\mathbf{P}$ matrix whose columns define a basis of the vectorial subspace described by the useful or harmful information. Thus useful information is kept whereas harmful information is withdrawn. Two new methods are proposed. First IDC-Improved Direct Calibration is a direct calibration method using both experimental and expert informations. Then VODKA-PLSR is a generalisation of PLSR. A vector $\mathbf{r}$ permits the use of expert information by the regression model. To conclude, this work allows a global view of existing tools, proposes two new models and offers new possibilities for building new models.

Keywords: information, experimental, expert, useful, harmful, model, general, calibration, direct, inverse, regression, pretraitment, PLSR, IDC, VODKA-PLSR