Mesures robustes en ligne des solutés organiques par spectrométrie infrarouge et étalonnages multivariés

M. Zeaiter

UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DE LANGUEDOC

THÈSE

Pour obtentir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Discipline **: Mathématiques appliquées et Application des mathématiques**
Formation Doctorale **: Génie des Procédés**
Ecole Doctorale **: Sciences des Procédés Biologiques et Industriels (SPBI)**

présentée et soutenue publiquement

par

Magida ZEAITER
le 02 Décembre 2004

Titre :

Mesures Robustes en Ligne des Solutés Organiques par Spectrométrie Infrarouge et Étalonnages Multivariés

JURY

M. RUTLEDGE Douglas	Président
Mme. BELLON-MAUREL Véronique	Directeur de thèse
M. HUVENNE Jean-Pierre	Rapporteur
M. BERTRAND Dominique	Rapporteur
M. BARROS Antònio	Examinateur
M. BOULET Jean-Claude	Examinateur
M. STEYER Jean-Philippe	Examinateur

Remerciements

Le travail présenté de ma thèse a été réalisé dans l'équipe IODE de l'unité de recherche Information et Technologies pour les Agro-procédés (ITAP) du Cemagref de Montpellier.

Je tiens tout d'abord à remercier M. Douglas Rutledge, de l'INAP-G de PARIS, d'avoir accepté d'assurer la présidence de ce jury. Je lui dois mes premiers pas sur le chemin de la "*Chimiométrie*", il a su m'apporter toujours conseils et encouragements.

Ma profonde gratitude s'adresse surtout à ma directrice de thèse, Mme Véronique BELLON-MAUREL pour son encadrement, sa disponibilité, sa patience, ses conseils et surtout son encouragement permanent. De même, je tiens à exprimer ma profonde gratitude à M. Jean-Michel ROGER pour son encadrement et son attention permanente. Son enthousiasme et sa disponibilité ont été une immense source d'appui, de motivation et d'encouragement qui m'ont permis de mener à bien ce travail. Leurs pertinences et leurs exigeances scientifiques m'ont permis de prendre de bonnes leçons pour la suite de mon chemin.

Je remercie M. Jean-Pierre HUVENNE, de l'Université de LILLE - Laboratoire CNRS, et M. Dominique BERTRAND de l'INRA de NANTES de m'avoir fait l'honneur d'accepter de juger ce travail.

Je remercie vivement M. Antònio BARROS, de l'Université d'AVEIRO du PORTUGUAL, d'avoir accepté de faire partie du jury de ma thèse.

Je remercie M. Sylvie ROUSSEL, M. Gilles RABATEL et M. Bernard PALAGOS d'avoir fait partie de mon comité de suivi de thèse. Je vous remercie pour votre disponibilité, vos conseils et votre soutien pendant ces trois ans.

Je tiens à exprimer ma gratitude à Monsieur Jean-Philippe STEYER, de l'INRA de NARBONNE, Laboratoire de Biotechnologies et de l'Environnement (LBE) ; M. Jean-Claude BOULET ainsi que M. Jean-Marie Sabrayrole et Mme Evelyne AGUERA, de l'INRA de PECHROUGE. Cela fut un plaisir de collaborer, d'échanger ses connaissances et de travailler avec vous pour mettre vos procédés au service de mes outils chimiométriques.

Je remercie Jean-Luc Lablee, la personne la plus recherchée au Cemagref, pour toute son aide technique (montages, appareillages,...), informatique (l'imprimante pour commencer et ....), sa patience et sa disponibilité (malgré tout).

Je remercie Michèle Egea pour tout son aide dans les démarches administratives dès mon premier jour au Cemagref (et c'est pas encore fini ! ! !) avec infiniment de patience, sa bonne humeur et toute sa générosité.

Je ne saurais pas oublier l'aide et le soutient de Serge (SG), Michel, Serge(SC), Nathalie, Sylvie, Benoit, Fabien, Iana, Sébastien. Leur appui et leur amitié ont été essentiels.

Je dédie mes pensées à mes amis, toutes les personnes que j'aime ici ou ailleurs, qui m'ont accompagné pendant ces trois ans, chacun se reconnait, merci de vos encouragements.

Mes pensées vont particulièrement vers ma famille, merci pour tout votre soutient, je vous dois ma présence ici.

Je remercie l'ensemble des membres du Cemagref de Montpellier pour leur accueil et leur soutien. C'était un plaisir de passer trois ans parmi vous.

Publications

Liste des Publications

1. **Zeaiter Magida, Roger Jean-Michel, BELLON-MAUREL Véronique and RUTLEDGE N. Douglas**. Robustness of methods developed by multivariate calibration. Part I : Robustness assessment, Trends in Analytical Chemistry TrAC, Vol.23, No2, p.157-170, 2004.
2. **Zeaiter Magida, Roger Jean-Michel, BELLON-MAUREL Véronique**. Robustness of methods developed by multivariate calibration. Part II : Influence of preprocessing methods. Trends in Analytical Chemistry TrAC. Accepted in November 2004.
3. **Zeaiter Magida, Roger Jean-Michel, BELLON-MAUREL Véronique**. Dynamic Orthogonal projection (DOP) a method for on-line process monitoring to maintain IR-based measurements. Application on alcoholic wine fermentation monitoring using NIR spectroscopy. Submitted October 2004 to Journal of Chemometrics.

Liste des Posters

1. **Chimiométrie 2002** : Étude de la répétabilité des mesures spectrales des solutés organiques par spectrométrie ATR/IR-TF. 4 et 5 Décembre 2002 - CNAM-Paris.
2. **Chimiométrie 2003** : Effet du centrage sur la robustesse des modèles d'étalonnage multivariés. 2 et 3 Décembre 2003 - CNAM-Paris.
3. **Chemometrics in Analytical Chemistry - CAC 2004** : Dynamic Orthogonal Projection (DOP) : a new method to maintain the robustness of IR-based measurements for on-line process monitoring. Institut Technologique de Lisbonne (IST)- 20-23 Septembre 2004 - Lisbonne, Portugal.

Liste des Communications Orales

4. **Communication orale " Heliospir "** : 'Suivi en ligne des procedes par spectrométrie proche infrarouge'. Cemagref, Montpellier - 11 Mars 2004.

5. **Chimiométrie 2004 - Communication Orale** : Dynamic Orthogonal Projection (DOP) : a new method to maintain the robustness of IR-based measurements for on-line process monitoring. Institut National Agronomique de Paris (INAP-G) - 1er Décembre 2004 - Paris.

Table des matières

Table des figures

Liste des tableaux

Notations

$\mathbf{X}$	matrice de n spectres, par p longueurs d'onde	
p	nombre de colonnes de $\mathbf{X}$	
n	nombre de lignes de $\mathbf{X}$	
j	indice de colonne	
i	indice de ligne	
g	nombre de facteurs ou grandeurs d'influence	
λ	longueur d'onde	
μ	la moyenne	
υ	la vraie valeur	
f	fonction multivariée	
$\mathbf{y}$	vecteur des n valeurs de référence	
$\widehat{\mathbf{y}}$	vecteur des n valeurs de référence estimées	
$\mathbf{b}$	vecteur des p coefficients de la régression	
b_0	intercept	$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} + b_0$
$\mathbf{1}_n$	vecteur colonne contenant n 1	
$\mathbf{I}_p$	matrice identité de dimension p	
$\overline{\mathbf{x}}^T$	vecteur ligne des moyennes des colonnes de $\mathbf{X}$	
$\overline{\mathbf{X}}$	matrice contenant n fois $\overline{\mathbf{x}}^T$	$\overline{\mathbf{X}} = \mathbf{1}_n\overline{\mathbf{x}}^T$
$\overline{y}$	valeur moyenne de $\mathbf{y}$	
$\|\mathbf{x}\|$	norme euclidienne de $\mathbf{x}$, i.e. $(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}$	
SEC	erreur standard d'étalonnage	$\mathrm{SEC}^2 = \frac{1}{n-n_1-1}\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$
SEV	erreur standard de validation	$\mathrm{SEV}^2 = \frac{1}{n_1-1}\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$
SEP	erreur standard de test	$\mathrm{SEP}^2 = \frac{1}{n}\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$
R_C	critère de robustesse global	
$R_C(t)$	critère de robustesse dynamique en fonction du temps	
G	facteur ou grandeur d'influence	
$\delta\widehat{\mathbf{y}}$	erreur sur la valeur prédite $\widehat{\mathbf{y}}$	

$\delta\mathbf{x}$	vecteur du spectre du facteur d'influence sur $\mathbf{x}$
$\delta\mathbf{x}_1$	vecteur des variations systématiques structurés de $\delta\mathbf{x}$
$\delta\mathbf{x}_2$	vecteur des variations aléatoires de $\delta\mathbf{x}$
LV	nombre de variables latentes du PLS
k	nombre de composantes principales
$\mathbf{W}$	matrice de poids
$\mathbf{T}$	matrice des coordonnées factorielles (scores)
$\mathbf{P}$	matrice des vecteurs propres (loadings)
$\overrightarrow{\mathcal{S}}$	l'espace des spectres de dimensions p
$\overrightarrow{\mathcal{C}}$	l'espace des spectres contenant les paramètres d'intérêts
$\overrightarrow{\mathcal{N}}$	l'espace des spectres contenant le reste de l'information
$\overrightarrow{\mathcal{E}}$	l'espace des résidus
$\mathbf{X}^+$	matrice des spectres de $\overrightarrow{\mathcal{C}}$
$\mathbf{X}^-$	matrice des spectres de $\overrightarrow{\mathcal{N}}$
$\mathbf{Z}^-$	matrice de base de $\overrightarrow{\mathcal{N}}$
$\mathbf{E}$	matrice des résidus dans $\overrightarrow{\mathcal{E}}$
$Col(\mathbf{X})$	l'espace engendré par les colonnes de $\mathbf{X}$
$Col(\mathbf{X}^+)$	l'espace engendré par la matrice $\mathbf{X}^+$
$Col(\mathbf{X}^-)$	l'espace engendré par la matrice $\mathbf{X}^-$
$Col(\mathbf{E})$	l'espace engendré par la matrice $\mathbf{E}$
SNR_j	rapport signal sur bruit au variable j
τ	l'instant de recalage ou de contôle
n_τ	nombre de points de recalage à l'instant τ
K	fonction de noyau kernel
ε	largeur du noyau
$\mathbf{A}_{(n_\tau \times n_0)}$	matrice du noyau contenant les combinaisons linéaires de $\mathbf{y}_0$ qui estime $\mathbf{y}_\tau$
$\mathbf{a}_{(ij)}$	coefficients de la matrice $\mathbf{A}$

$\mathbf{X}_0$ matrice des spectres d'étalonnage

$\mathbf{y}_0$ vecteur des valeurs de référence d'étalonnage

$\mathbf{X}_\tau$ matrice des spectres correspondant au point de recalage à l'instant τ

$\mathbf{y}_\tau$ vecteur des valeurs de référence aux points de recalage à l'instant τ

$\mathbf{X}^*$ matrice des spectres corrigés

$\mathbf{X}_0^*$ matrice des spectres d'étalonnage corrigés

$\mathbf{D}$ matrice des spectres de différence $\mathbf{D} = \mathbf{X}_\tau - \widehat{\mathbf{X}}_\tau$

Abreviations

ACP	Analyse en composantes Principales (Principal Components Analysis)
$ANOVA$	Analyse de variance (ANalysis Of VAriance)
CLS	Régression classique aux moindres carrées (Classical Least Squares)
DOP	Projection orthogonale dynamique (Dynamic Orthogonal Projection)
DO	Projection orthogonale directe (Direct Orthogonal Projection)
DS	Standardisation directe (Direct Standardization)
DWT	Domaine de transformation des ondelettes (Wavelet Transform Domain)
DT	Correction de la tendance (De-Trend)
EPO	Projection orthogonale externe (External Orthogonal Projection)
$FT-IR$	Infrarouge à Transformée de Fourrier (Fourrier Transform-Infrared)
IIR	Réduction indépendante des interférences (Independent Interference Reduction)
ILS	Régression aux moindres carrées inverses (Inverse Least Squares)
KF	Filtre de Kalman (Kalman Filter)
$LMVC$	Etalonnage multivarié linéaire (Linear Multivariate calibration)
MLR	Régression linéaire multiple (Multiple Linear Regression)
MIR	Moyen Infrarouge (Near Infrared Spectroscopy)
MSC	Correction multiplicative de la dispersion (Multiplicative Scatter Correction)
$OPLS$	Projection orthogonale sur les structures latentes (Orthogonal Projection to Latent Structures)
OSC	Correction orthogonale du signal (Orthogonal Signal Correction)
PCR	Régression sur composantes principales (Principal Components Regression)
PLS	Régression aux moindres carrées partielles (Partial Least Squares)
PDS	Standardisation directe par morceaux (Piecewise Direct Standardization)
$RMSEC$	Racine carrée de la moyenne des carrés d'erreurs d'étalonnage
$RMSEP$	Racine carrée de la moyenne des carrés d'erreurs de test
RNV	Transformation normale standard robuste (Robust Standard Normal Variate transformation)

$SPIR$	Spectroscopie Proche Infrarouge (Near Infrared Spectroscopy)
SNV	Transformation normale standard (Standard Normal Variate transformation)
SEC	L'erreur standard d'étalonnage
$SECV$	L'erreur standard d'étalonnage en cross validation
SEV	L'erreur standard de validation
SEP	L'erreur standard de test
UVE	Élimination des variables non informatives

Chapitre 1

Introduction et problématique

Sommaire

1.1 Spectroscopie Infrarouge

1.1.1 Forces

Face aux besoins industriels de mesures de contrôle qualité fiables, rapides, économiques, avec le moins de préparation d'échantillons, de pollution et de risque sanitaires, ainsi qu'avec le développement des procédés industriels et l'augmentation de la production, de nombreux industriels s'orientent vers l'adoption de la spectroscopie infrarouge (SIR) comme technique de mesure de routine [148, 2, 50, 117, 91]. La spectroscopie proche infrarouge (PIR) nécessite moins de préparation d'échantillons que les autres méthodes analytiques. Les spectres PIR contiennent de l'information sur les propriétés chimiques et physiques du milieu analysé et conduisent à la mesure quantitative simultanée de plusieurs paramètres. Ces mesures quantitatives s'avèrent possibles grâce au développement des techniques d'étalonnage multivariés permettant de relier les propriétés d'un ensemble d'échantillons aux intensités spectrales ou absorbances à plusieurs longueurs d'onde [8]. Les premiers travaux qui ont été développés dans le domaine de la spectroscopie notamment de la PIR pour fournir de l'information sur les propriétés physiques et chimiques des produits analysés, sont ceux de Norris [107] en appliquant le modèle linéaire basé sur la loi de Beer-Lambert qui relie la concentration à la densité optique mesurée en une seule longueur d'onde. Les spectres dans le domaine du PIR sont constitués des bandes spectrales qui se recouvrent et qui sont très étalées dans lesquelles les intensités sont très corrélées. A part des cas simples, il n'existe pas un modèle ou une loi qui relie la concentration à l'absorbance en PIR. D'où l'utilisation des modèles d'étalonnage multivariés, tels que : la régression linéaire multiple (MLR), la régression aux moindres carrés partiels (PLS) et la régression par composantes principales (PCR). Une revue de toute ces méthodes se trouve dans [25].
C'est d'ailleurs grâce au développement de telles méthodes de traitement mathématique, et à l'amélioration conjointe des possibilités de calcul que la spectroscopie PIR a pris son essor.

Ces méthodes utilisées pour extraire l'information à partir des mesures spectrales relèvent de la '*chimiométrie*'. Massart et al. ont défini la chimiométrie comme étant :

"*la discipline qui utilise les mathématiques, les statistiques et la logique formelle (a) pour concevoir ou choisir des procédures expérimentales optimales ; (b) pour fournir le maximum d'information chimique d'intérêt en analysant les données chimiques ; et (c) pour obtenir des connaissances sur les systèmes chimiques*"[97]. La chimiométrie est apparue à la suite du développement des instruments de mesures chimiques rapides utilisés pour des analyses de routine. Ces instruments fournissent des mesures indirectes représentées par de grandes bases de données spectrales qui nécessitent d'être interprétées afin d'extraire l'information chimique souhaitée. La chimiométrie s'est développée récemment dans divers domaines d'optimisation, traitement de signal, classification, discrimination, reconnaissance des formes, diagnostic, contrôle des processus, intelligence artificielle [177]. Au delà des analyses de routine, la spectroscopie PIR présente un grand nombre d'atouts pour être utilisée pour le suivi en ligne des procédés [12, 16, 132, 134]. En particulier, l'utilisation des fibres optiques peu onéreuses [36], rend possible son application en ligne directement sur une chaîne de production [101, 149].

1.1.2 Faiblesses

Or, bien qu'ils présentent tous les critères pour les applications en temps réels [134, 181], les systèmes spectrométriques PIR sont rarement utilisés pour la caractérisation en ligne des solutés organiques en conditions industrielles [108].

En effet au delà des difficultés d'étalonnage inhérentes à la PIR (bandes d'absorption larges, fortement corrélées, spectre faiblement résolu et dominé par l'eau), la sensibilité des spectres PIR aux facteurs non contôlés rend cette technique délicate à utiliser en conditions industrielles, éminément variables. Les mesures spectrales sont fortement influencées par la température des échantillons, la température de l'appareil de mesure, le changement d'une pièce optique (usure de l'appareil, changement de lampe,...), l'hétérogenéité du milieu analysé, la lumière diffuse, l'apparition de produits de moindre importance absorbant dans la zone spectrale d'intérêt ainsi que par la présence de l'eau qui est le produit majoritaire en PIR. Ces mesures deviennent différentes de celles de la base d'étalonnage ou d'apprentissage du modèle, et par la suite elles conduisent à des valeurs érronées. Ainsi, pour les applications industrielles où les conditions de mesures

changent, il est nécessaire de redévelopper ou d'adapter le modèle à l'application et aux conditions de mesure, ce qui exige l'arrêt des mesures et la collection d'une autre base de données spectrales prises dans les conditions réels avec les mesures de référence correspondantes.

Cette sensibilité du modèle à l'action des facteurs d'influence que l'on rencontre en conditions industrielles traduit la non-robustesse du modèle d'étalonnage. Ainsi le développement de méthodes de mesure PIR robuste passe par une procédure d'étalonnage particulièrement soignée, voire par le développement de nouvelles méthodes d'étalonnage ciblées sur ce problème.

1.2 Étalonnages non robustes

1.2.1 La démarche recommandée pour étalonner

Dans le règlement standard ASTM (American Society for Testing and Materials)[8], l'étalonnage multivarié en spectroscopie est défini comme étant : '*le processus de création d'un modèle qui relie les propriétés d'un ensemble des échantillons de référence connus à leurs intensités ou absorbances à plus d'une longueur d'onde ou fréquence.*' Dans ce règlement pour l'étalonnage multivarié en spectroscopie PIR et MIR, la construction du modèle passe par plusieurs étapes :

1. Sélection des échantillons d'étalonnage ;
2. Acquisition des données spectrales et de référence ;
3. Construction d'un modèle d'étalonnage ;
4. Validation du modèle ;
5. Utilisation du modèle pour prédire des nouveaux échantillons ;
6. Contrôle de l'étalonnage ;
7. Maintenance du modèle d'étalonnage.

Généralement, un modèle d'étalonnage est construit en utilisant une grande base de données constituée des mesures spectrales et de référence des mêmes échantillons pris

pendant une longue période pour représenter une large gamme de variation dans le modèle. Le modèle linéaire, p.e. PLS, est utilisé pour établir le modèle d'étalonnage [178]. Cette étape comprend plusieurs parties séquentielles : détection des données aberrantes, prétraitement des spectres (centrage, réduction, normalisation, sélection de variables, méthodes d'orthogonalisation).

Après sa construction, le modèle est testé en utilisant un autre ensemble de données. Durant cette étape les performances du modèle sont évaluées, telles que justesse, sélectivité, robustesse [142]. Ensuite, le modèle peut être utilisé pour prédire des nouvelles mesures spectrales. Il reste les deux dernières étapes de contôle et de maintien de la robustesse du modèle lors des utilisations réelles qui sont les plus importantes à étudier surtout pour des applications industrielles.

La mauvaise gestion de ces deux étapes aggrave le problème de manque de robustesse du modèle d'étalonnage pour des applications en ligne. Ces deux dernières étapes sont abordées dans cette thèse.

1.2.2 Causes du manque de robustesse des modèles d'étalonnage

Pour comprendre comment les facteurs d'influence peuvent agir sur la qualité du modèle, il faut revenir à l'équation mathématique de l'étalonnage linéaire. En spectroscopie, un modèle d'étalonnage linéaire est de la forme :

$$\mathbf{y} = \mathbf{X}\mathbf{b} + b_0 + \mathbf{e} \tag{1.1}$$

$\mathbf{y}_{n,1}$ représente les valeurs de référence, $\mathbf{X}_{(n,p)}$ représente la matrice de n mesures spectrales et de p variables ou longueurs d'onde, $\mathbf{b}_{(p,1)}$ représente le coefficient de régression permettant de prédire les valeurs de références des nouvelles mesures et $\mathbf{e}$ est la matrice des erreurs résiduelles.

$\mathbf{X}_{(n,p)}$ étant la matrice des spectres, elle contient à la fois l'information chimique du milieu analysé, p.e. $\mathbf{X}_{chimique}$, l'information liée à son état physique (granulométrie

du produit, homogénéité,couleur,etc.), p.e. $\mathbf{X}_{physique}$ et celle provenant du milieu environnant (lumière parasite, variation de la température,...), p.e. $\mathbf{X}_{environnement}$, comme l'indique l'équation 1.2 :

$$\mathbf{X} = \mathbf{X}_{chimique} + \mathbf{X}_{physique} + \mathbf{X}_{environnement} \tag{1.2}$$

$\mathbf{X}_{chimique}$ correspond d'une part à la mesure de l'information chimique recherchée, notée $\mathbf{X}^{+}_{chimique}$, et d'autre part à la partie non utile notée $\mathbf{X}^{-}_{chimique}$. Pour la prédiction de la concentration du milieu analysé, c'est la partie chimique $\mathbf{X}^{+}_{chimique}$ qui est intéressante. Quant aux autres parties ($\mathbf{X}^{-}_{chimique}$, $\mathbf{X}_{physique}$ et $\mathbf{X}_{environnement}$), elles représentent l'erreur systématique due à des facteurs non contrôlés. Dans la suite, l'ensemble des termes de $\mathbf{X}$ qui sont dûs aux facteurs responsables de l'erreur systématique seront notés $\mathbf{X}^{-}$ et ceux correspondant à la propriété d'intérêt seront notés $\mathbf{X}^{+}$. D'où l'équation suivante :

$$\mathbf{X} = \mathbf{X}^{+} + \mathbf{X}^{-} + \mathbf{R} \tag{1.3}$$

avec $\mathbf{R}$ dus à des facteurs inconnus.

Les différents facteurs pouvant influencer la robustesse des modèles d'étalonnages sont des facteurs non contrôlés contenus dans $\mathbf{X}^{-}$ de l'équation 1.3. Pour des applications en ligne de la spectrométrie PIR, c'est l'effet de la variation de ces facteurs sur les mesures spectrales qui est à la base des problèmes de robustesse des modèles d'étalonnage (figure 1.1).

Le manque de robustesse peut être dû alors à la présence des :

- Facteurs intrinsèques, comme la présence du spectre de l'eau qui absorbe beaucoup dans le PIR et qui perturbe toute autre information, ou bien la nature et la composition complexe du milieu d'analyse.

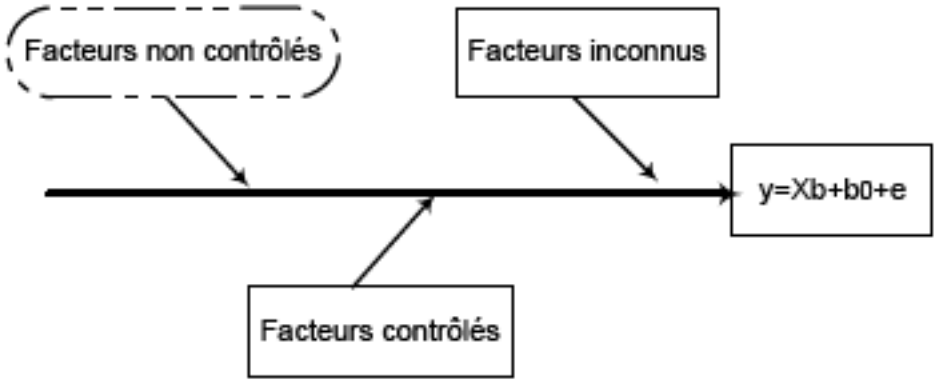

FIG. 1.1 – Diagramme de causes du manque de la robustesse des étalonnages multivariés

- Facteurs extrinsèques, comme la variation de la température, la variation de la lumière parasite.

- Facteurs instrumentaux, tels que la source, le détecteur, l'optique.

Ces facteurs viennent se fixer à un niveau particulier au cours de l'étalonnage, ensuite lors des mesures ultérieures, la variation de leurs niveaux a des effets significatifs sur la réponse du modèle d'étalonnage. Ces effets se manifestent par l'apparition d'une dérive ou d'un biais dans les réponses du modèle d'étalonnage. L'évaluation de la robustesse des modèles d'étalonnage multivariés revient alors à évaluer l'effet de ces facteurs sur la réponse du modèle. L'amélioration de la robustesse consistera à diminuer l'effet des facteurs sur le modèle.

1.3 Problématique de la thèse

Cette thèse a pour objet l'étude de la robustesse des modèles d'étalonnage multivariés avec, comme cas d'étude, les modèles développés pour des applications industrielles. Elle propose une méthodologie pour assurer la robustesse dans le cas de suivi d'un procédé continu, bénéficiant de mesures ponctuelles de contrôle.

En effet comme indiqué dans la démarche préconisée par l'ASTM, l'étalonnage doit être contrôlé (étape 7). Celà suppose donc que sont disponibles des données de contrôle

ponctuelles. Ainsi l'ensemble des données disponibles dans le cas d'une mesure PIR en ligne est le suivant : (i) des mesures spectrales rapides, non destructives, effectuées sur un procédé de mesure lent et continu ; (ii) en parallèle, des mesures de références moins fréquentes prises dans le cadre du contrôle du procédé.

La méthodologie développée dans cette thèse consiste à utiliser toutes les informations disponibles en ligne afin d'assurer une mesure robuste par spectrométrie SIR.

Pour celà cette thèse répondra à trois questions :

- "*Qu'est ce que la robustesse des étalonnages multivariés et comment la mesure-t-on ?*" est la première question posée pour analyser la robustesse. C'est un préalable indispensable à toute amélioration.
- Un modèle optimal développé dans les conditions d'étalonnage, peut présenter des problèmes de robustesse vis-à-vis de l'effet de la variation de certains facteurs d'influence connus. D'où, l'intérêt de l'amélioration intrinsèque de la robustesse de ce modèle. "*Quelles sont les méthodes disponibles pour améliorer la robustesse des étalonnages multivariés ?*" est la deuxième question posée dans l'analyse de la robustesse.
- La troisième question s'intéresse au développement de la méthodologie pour maintenir la robustesse du modèle d'étalonnage développé hors ligne, dès lors qu'il est utilisé en ligne pour effectuer des mesures en temps réel. Ce modèle doit conserver sa robustesse malgré les changements éventuels dans les conditions de mesure dus à l'apparition de nouveaux facteurs d'influence en ligne. A ce niveau de l'analyse de la robustesse, la question posée est : "*comment utiliser les informations de contrôle en ligne pour maintenir la robustesse de cet étalonnage en ligne ?*".

  Les informations de contrôle en ligne correspondent à quelques couples de mesures de références et de mesures spectrales correspondantes acquises en ligne. Le but est d'utiliser ces informations disponibles en ligne représentatives du processus de mesure, pour capitaliser la base d'étalonnage prise dans des conditions optimales, afin de maintenir et d'améliorer la robustesse des mesures ultérieures.

L'objectif de la thèse sera donc d'apporter des réponses originales aux

trois questions suivantes :

1. **Comment définir et caractériser la robustesse de l'étalonnage multivarié ?**
2. **Comment améliorer intrinsèquement la robustesse de l'étalonnage multivarié ?**
3. **Comment utiliser les informations de contrôle en ligne pour maintenir la robustesse de l'étalonnage multivarié ?**

Pour celà, la thèse se déroulera selon le plan suivant.

1.4 Plan de la thèse

1.4.1 Définition et évaluation de la robustesse

Dans le chapitre 2 la notion de "*robustesse*" est évoquée. Afin de pouvoir analyser la robustesse des modèles des étalonnages multivariés, il était nécessaire de commencer notre étude sur la robustesse des étalonnages par définir cette notion ,"*robustesse*", dans notre domaine d'application. La deuxième partie de ce chapitre s'intéresse à l'évaluation de la robustesse : méthodes et critères d'évaluation.

À la fin de cette partie un critère pour évaluer la robustesse des étalonnages multivariés est proposé relativement à la définition de la robustesse déjà donnée.

1.4.2 Amélioration de la robustesse

Après avoir défini la robustesse du modèle d'étalonnage, ainsi que son critère d'évaluation, la troisième étape de l'analyse de la robustesse consiste à étudier l'amélioration de la robustesse. Cette étude fait l'objet du chapitre 3. Dans ce chapitre toutes les méthodes de prétraitement utilisées pour améliorer la prédiction du modèle d'étalonnage sont discutées et les meilleures méthodes sont sélectionnées pour la suite de la démarche.

1.4.3 Maintenance de la robustesse en ligne

Le modèle d'étalonnage optimisé sera utilisé en ligne. Ce modèle doit conserver sa robustesse lors des mesures en continu. La maintenance de la robustesse du modèle en ligne fait l'objet du chapitre 4. Dans ce chapitre, une nouvelle méthode est proposée pour maintenir en ligne la robustesse du modèle d'étalonnage, lors du suivi des processus continus par spectrométrie IR.

1.4.4 Application

Deux exemples pratiques d'application de cette méthode font l'objet du chapitre 5. D'une part, une fermentation alcoolique du vin blanc sera suivie en temps réel par spectrométrie infrarouge pour démontrer l'intérêt de la méthode sur des facteurs d'influence de type physique (température). Les expériences ont été menées à l'INRA de PechRouge. D'autre part, une fermentation destinée à épurer les vinasses sera suivie pour illustrer l'intérêt de la méthode sur des facteurs d'influence de type chimique (ajout de composés parasites).

1.4.5 Conclusion et perspectives

L'ensemble de la méthodologie sera discutée en conclusion et les perspectives seront largement ouvertes pour l'application de cette méthodologie dans tous les secteurs industriels utilisant le PIR.

Chapitre 2

Robustesse des étalonnages multivariés : Définition et évaluation

Sommaire

2.1 Introduction

L'objectif de ce chapitre est de cerner la notion de robustesse afin de proposer une définition consensuelle et un critère pour la mesurer. C'est un préalable indispensable à toute procédure d'amélioration de la robustesse. Dans un premier temps, différentes définitions de la robustesse qui ont été trouvées dans la littérature, sont présentées. La définition la plus récente a été proposée dans les directives de la conférence internationale "Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH)" [75] au sujet de la validation des procédures analytiques, éditées pour des applications standard dans la Communauté Européenne, le Japon et les Etats-Unis. La robustesse d'un procédé analytique est définie comme étant une des caractéristiques de la méthode de validation liée à la fidélité et à la reproductibilité. Cette définition, la plus appliquée en chimie analytique, ne satisfait pas nos besoins. Par conséquent, une définition de la robustesse liée aux problèmes des étalonnages multivariés utilisés en spectroscopie (SPIR) est proposée en se basant sur les définitions déjà établies.
Après avoir défini la robustesse, les méthodes et les critères les plus utilisés pour son évaluation sont présentés. Une attention particulière sera accordée à ce qui concerne la spectroscopie PIR, dans le cadre de la prédiction et du transfert d'étalonnage. En conclusion, nous proposons un critère pour évaluer la robustesse relativement à notre définition. Ce chapitre a été le résultat d'un article publié dans le journal Trends in Analytical Chemistry TrACs ([185]).

2.2 Définitions

Cette section montre, via différentes définitions, comment est vue la robustesse selon le domaine d'application. Le terme "*robustness*", qui n'a pas encore été défini jusqu'à présent par un organisme officiel de contrôle de qualité ou de métrologie, est toujours considéré comme étant une propriété importante de la méthode de mesure. Dans la littérature, beaucoup de chercheurs en ont déjà discuté. Chacun propose une

définition adaptée à son domaine d'application qui peut même changer au sein d'un même domaine. De plus, il y a des confusions possibles entre les termes anglais et les termes français ce qui a justifié l'existence d'une norme AFNOR [4] pour préciser le vocabulaire métrologique dans les deux idiomes [1]. Par conséquent, différentes sources de définitions (avec leur équivalent en anglais) seront employées pour donner la définition la plus adaptée à nos besoins et la plus claire possible.

En métrologie, la qualité d'une méthode ou d'un instrument est habituellement caractérisée en utilisant différents termes comme : reproductibilité, répétabilité, exactitude, justesse et fidélité. La reproductibilité d'une méthode est souvent employée dans le contexte du laboratoire, alors que la robustesse est utilisée pour les applications industrielles. Selon la norme ISO [49], la ***reproductibilité*** d'une méthode de mesure est '*la fidélité des résultats issus des mesures inter-laboratoires*' tandis que la ***répétabilité*** des résultats est ' *l'étroitesse de l'accord entre les résultats des mesures successives d'un même mesurande réalisées dans les mêmes conditions de mesure (le même analyste, même instrument, même jour si possible)*'.
Dans la littérature, l'***exactitude*** et la ***précision*** des résultats sont très souvent confondus. Pour cela, la distinction entre ces deux termes s'avère importante. Selon la norme ISO 1994 [49], l'**exactitude** est caractérisée par un terme de *justesse* et un autre de *fidélité*. La justesse décrit la déviation systématique par rapport à la vraie valeur, qui est normalement un but inaccessible, puisqu'elle serait donnée par une mesure parfaite ; Alors que la fidélité (qui correspond à précision en anglais), décrit des erreurs aléatoires. La **fidélité** comme définie par ISO [49] est '*l'étroitesse de l'accord entre les résultats d'essai indépendants obtenus dans des conditions stipulées*'. Comme définie par la directive de l'ICH, [75] la fidélité peut être considérée à trois niveaux : la répétabilité, la fidélité intermédiaire et la reproductibilité. Elle est exprimée comme une variance, un écart type ou un coefficient de variation d'une série de mesures. Se rapportant à [4], la fidélité ne devrait pas être employée pour l'exactitude. Fearn et

[1] Attention à la confusion Français/Anglais. Les termes exactitude, fidélité et justesse sont traduits en anglais respectivement par accuracy, presision et trueness

al. [111] indiquent que la bonne fidélité n'assure pas la bonne exactitude si l'erreur systématique significative est présente ou si le terme de justesse n'est pas accompli. Osborne [109] considère que l'exactitude des mesures quantitatives par spectroscopie proche infrarouge PIR, est particulièrement influencée par l'incertitude des données de référence utilisées pour l'étalonnage. L'exactitude est habituellement évaluée sur la base de l'erreur standard de prédiction en validation croisée (SECV) et l'erreur standard de prédiction en test (SEP) obtenue en comparant, par exemple, les résultats du SPIR avec les mesures de référence sur les mêmes échantillons témoins. Il considère que l'évaluation de l'exactitude d'un étalonnage en SPIR pourrait être réalisée en décomposant la variance totale définie par le SEP selon l'équation suivante :

$$SEP^2 \quad = \quad S_r^2 + S_{PIR}^2 + S_e^2 \tag{2.1}$$

où S_r est la répétabilité de la méthode de référence, S_{SPIR} la répétabilité de la mesure du spectre SPIR et le S_e est le manque d'ajustement du modèle d'étalonnage. Pour déterminer l'exactitude vraie de la méthode d'analyse en SPIR, l'erreur de la méthode de référence devrait être éliminée. En outre, il est possible d'estimer la limite de l'exactitude en l'absence de l'erreur dans l'équation de régression, au delà de laquelle il n'est pas possible d'améliorer l'exactitude des étalonnages du SPIR.

Un autre terme caractérisant les résultats de mesure est l'*incertitude*, habituellement évaluée pendant la validation de la méthode [172, 15]. L'incertitude de la mesure est définie comme un paramètre associé au résultat d'une mesure qui caractérise la dispersion des valeurs qui pourraient raisonnablement être attribuées au mesurande [4]. Par conséquent, l'incertitude est liée à l'exactitude puisque toutes les deux décrivent l'étroitesse de l'accord entre la valeur expérimentale moyenne et la valeur vraie. L'incertitude, caractérisée par l'écart-type, correspond à la gamme de valeurs qui peuvent être données comme résultat de mesure à partir de la distribution statistique des résultats. Cependant, on ne doit pas confondre l'erreur et l'incertitude. L'*erreur* est la différence entre un résultat individuel et la valeur vraie ; elle peut être utilisée pour corriger le résultat, ce qui n'est pas valable pour l'incertitude [41]. En considérant les définitions

déjà mentionnées ci-dessus, nous pouvons conclure que tous ces différents termes, utilisés comme critères pour la validation des méthodes d'analyses se complètent.

En chimie analytique, plusieurs définitions pour la robustesse sont traditionnellement données. Mulholland [102] définit la robustesse d'un procédé analytique comme étant ' *sa résistance vis à vis des petites variations dans les conditions expérimentales, qui pourraient se produire en transférant la méthode : entre les laboratoires, les opérateurs ou les instruments*'. Cette définition se rapproche surtout de celle qui a été suggérée par le guide français pour la validation des méthodes d'analyse [27]. La "Pharmacopea XXII des Etats Unis (USP) [29]" définit la *ruggedness* comme suit : c'est pour une méthode analytique, sa capacité à reproduire les résultats obtenus par l'analyse du même échantillon sous une variété de conditions expérimentales normales telles que différents laboratoires, différents jours, sous différentes températures, etc. Ainsi, le terme *ruggedness* donne une indication sur l'exactitude prévue de la méthode. Il est considéré comme étant un des critères de performance utilisés pour l'évaluation de la fidélité lors de la validation d'une méthode de mesure [102, 153, 150]. Comme défini par l'ICH [76, 77] la robustesse d'un procédé analytique est la mesure de sa capacité à rester inchangé sous l'influence des petites variations délibérées dans les paramètres de la méthode et fournit une indication de sa fiabilité durant l'utilisation normale. Par conséquent, la robustesse évalue la capacité d'une méthode à résister à des changements qui peuvent survenir lors de l'utilisation normale ; c'est semblable à la définition donnée dans [102]. Rodriguez et al. [118] définissent la "ruggedness" et la "robustness" comme l'inertie du procédé de génération des résultats analytiques qui sont biaisés par des petites variations autour de la valeur standard des variables expérimentales. Cette valeur standard est celle donnée par une méthode analytique standard ou celle correspondant à la valeur optimale obtenue lorsque des méthodes internes sont développées.

Dans le **contrôle de qualité**, selon Taguchi [143], la robustesse d'un processus est sa capacité à produire des produits uniformément bons avec un effet minimal de l'influence dûe aux changements des facteurs incontrôlables pouvant agir sur la qua-

lité de la fabrication. Pour Feinberg [47], la robustesse d'une méthode est le niveau des effets observés lorsque des petites variations délibérées des conditions expérimentales surviennent. Par conséquent, la robustesse devrait être considérée comme une propriété de l'étalonnage utilisé dans un procédé analytique. Comme vu ci-dessus, le terme '*ruggedness*' est fréquemment employé comme synonyme de la '*robustesse*' [150, 118, 157]. Une distinction entre ***ruggedness*** et ***robustesse*** a été montrée à travers des définitions dans [166]. Le niveau de '*ruggedness*' obtenu à partir d'une épreuve entre les laboratoires (différents laboratoires, analystes, équipements analytiques, etc.), est employé notamment dans le cas des problèmes de transférabilité, alors que la '*robustesse*' est determinée à partir des mesures intra-laboratoire (différentes températures, concentrations, etc.) [60].

Dans l'**étalonnage multivarié**, la robustesse est considérée comme étant *la sensibilité des prédictions du modèle d'étalonnage aux changements des facteurs externes tels que des variations environnementales, instrumentales, et des conditions de mesure de l'échantillon sous lesquelles les spectres ont été mesurés pendant la phase d'étalonnage* [141]. Dans son étude sur la robustesse des modèles d'étalonnage utilisant les réseaux de neurones artificiels, Gemperline [54] considère qu'un étalonnage multivarié est dit robuste s'il est peu sensible aux faibles changements dans la réponse de l'appareil de mesure. Blanco et al. [17] considèrent la robustesse en SPIR comme étant la "stabilité du modèle d'étalonnage dans le temps". La ***stabilité*** comme définie dans [55] est "l'invariabilité d'une propriété spécifique d'une substance, d'un dispositif, ou d'un appareil au cours du temps, ou sous l'influence des facteurs en général extrinsèques". C'est le point auquel un appareil de mesure demeurera invariable, ayant des variations dans le temps de la température, de la puissance, etc. tandis que la ***fiabilité*** d'un système est sa capacité à exécuter ses fonctions dans des conditions indiquées pendant une période de temps bien définie [78]. Rodriguez et al. [118] considèrent la robustesse en tant que propriété de la géométrie de la surface de réponse à proximité de l'optimum étudié. Récemment, Roussel [122] a considéré le concept de la robustesse d'un étalonnage SPIR comme la capacité du modèle à rester stable sous de petites perturbations,

par l'analyse des surfaces de réponses.

En statistiques, Box [22] a présenté le mot "robuste" pour décrire les procédures qui donnent de bons résultats même en cas de violation des hypothèses sur lesquelles ces procédures sont basées. Comme décrit dans [89], la première approche de la robustesse a été donnée par Huber [72] et Hampel [80]. Huber [72] a présenté l'approche "minimax" pour spécifier les erreurs du modèle, afin d'étudier la sensibilité de la prédiction optimale . Dans un contexte qualitatif, comme cité dans [89], Hampel a défini la robustesse d'un estimateur en fonction du point de défaillance ou breakdown point (BP). Le BP mesure l'insensibilité d'un estimateur aux données aberrantes et donc la robustesse de l'analyse statistique. Le BP est défini comme la plus petite fraction de la contamination qui excentre l'estimateur de sa vraie valeur [186, 71]. Pour l'étalonnage multivarié, ce point peut représenter le pourcentage maximal de points aberrants qu'une méthode d'étalonnage peut supporter [163, 87]. D'ailleurs, la robustesse apparaît dans des statistiques robustes, pour améliorer la résistance des estimateurs à l'occurrence des données aberrantes, en employant la médiane au lieu de la moyenne [47]. En outre, la robustesse du modèle se rapporte à son insensibilité aux déviations liées à la non-conformité avec les hypothèses statistiques (linéarité, normalité, indépendance des variables) [137].

Dans cette étude, étant intéressés par la robustesse des modèles d'étalonnage multivariés pour des applications en ligne de la spectrométrie SPIR, et après avoir vu toutes les définitions sus-citées, nous adoptons la définition de la robustesse d'un modèle d'étalonnage multivarié comme étant *la stabilité de sa capacité prédictive vis-à-vis des perturbations appliquées au voisinage des conditions standard.* Par cette définition nous considérons qu'un modèle d'étalonnage robuste est un modèle stable vis à vis des perturbations dûes aux variations des facteurs d'influence.

2.3 Évaluation de robustesse

L'évaluation de la robustesse est une étape importante pour vérifier la qualité du modèle d'étalonnage multivarié. Youden en 1961 [182] était le premier à discuter la "ruggedness" de la méthode analytique, disant que *le processus analytique doit être*

"rugged". Le test de la robustesse d'une méthode de mesure est exigé par les directives de l'ICH comme étant une partie du développement de la méthode [77, 160]. Souvent, le modèle d'étalonnage est calculé en employant la technique de validation croisée ; le modèle ayant le moins d'erreur de prédiction est choisi et moins d'attention est accordée à sa robustesse. Dans une étude sur la sensibilité des modèles d'étalonnage multivariés, Swierenga et al [141] considèrent que la robustesse des modèles ne peut pas être jugée seulement en terme d'erreur de prédiction ; en effet, les modèles peuvent posséder une petite erreur de prédiction (parce que les conditions expérimentales ne diffèrent pas) et en même temps être très sensibles à des petites perturbations dans les conditions expérimentales. Par conséquent, la sensibilité des modèles d'étalonnage doit être estimée par l'intermédiaire d'une étude de robustesse. La plupart des stratégies d'évaluation de la robustesse essayent de trouver la gamme centrée sur les conditions standard, sous lesquelles le modèle continue à donner des réponses satisfaisantes c'est-à-dire semblables à celles produites dans les conditions standard. Habituellement, toutes les méthodes d'évaluation de la robustesse comportent l'utilisation d'un test de robustesse suivi de calcul d'indice de robustesse.

2.3.1 Tests de robustesse

Les tests de robustesse sont largement appliqués pour étudier le potentiel des sources de variabilité et leur influence sur les résultats des méthodes de mesures.

Ils exigent une identification rapide des effets significatifs (c'est à dire des effets qui peuvent affecter de manière significative le résultat) [43, 130]. La méthodologie du test de la robustesse donnée en chimiométrie et en statistique est en accord avec les directives de l'ICH qui définit le test de robustesse comme étant une "étude expérimentale dans laquelle on évalue l'influence de petits changements des conditions opératoires ou des conditions environnementales sur les réponses mesurées ou calculées"[157]. Les changements pris en compte dans le test de robustesse reflètent ceux qui peuvent se produire quand une méthode est transférée entre différents expérimentateurs, différents instruments, etc. [154, 151]. Pour les méthodes analytiques, deux définitions du test de robustesse sont trouvées. La première se rapporte à celle du guide français pour la

validation des méthodes d'analyse [27] et celle de l'ICH [75] vue précédement et qui est la plus fréquemment utilisée. La seconde définition, donnée par la Pharmacopeia XXII [29] des Etats Unis, est davantage associée au terme "ruggedness". Cette dernière décrit le test de ruggedness comme étant "l'évaluation du degré de reproductibilité des résultats du test effectué sous des conditions normales, des conditions opératoires prévues entre différents laboratoires et différents analystes". Ainsi, exécuter le test de robustesse à la fin de la phase d'étalonnage augmente le risque de devoir complètement développer le modèle de nouveau. Comme rapporté par les directives de l'ICH [77] et les directives hollandaises de pharmacie [150], l'évaluation de la robustesse devrait être exécutée pendant le développement du procédé analytique. Dans [102, 150] le test de robustesse est considéré comme partie de l'évaluation de l'exactitude dans la partie de validation de la méthode. Leeuwen et al. [152] considèrent un test de robustesse comme l'évaluation de la fiabilité qui complète le développement de la méthode. Les résultats d'un test de robustesse indiquent combien les facteurs expérimentaux devrait être étroitement contrôlés. Par conséquent, Faber [43] recommande d'inclure les résultats de cet essai dans le dossier d'enregistrement de contrôle de qualité.

En testant la robustesse d'une méthode analytique trois niveaux ont été considérés dans les directives du Food and Drug Administration (FDA) [27]. Le premier niveau se rapporte à la définition de l'ICH de la robustesse et devrait inclure la vérification de la reproductibilité en employant un deuxième analyste. Ce niveau est exigé pour toutes les méthodes. Le second niveau tient compte de l'effet des changements plus profonds des conditions opératoires quand une méthode est prévue pour être appliquée dans un laboratoire différent et avec un équipement différent. Le troisième niveau considère "une collaboration complète des essais de test" ce qui est rarement fait. De ces définitions deux approches ont pu être trouvées selon les facteurs à examiner. Habituellement, un test de robustesse exige un plan d'expérience. Pour les deux premiers niveaux du test de robustesse les facteurs sont choisis parmi ceux reliés aux conditions opérationnelles et environnementales connus sous le nom de *facteurs liés à la procédure*. Ils sont examinés en utilisant un plan d'expérience classique tandis les facteurs tels que : laboratoires, analystes, instruments, facteurs reliés à la procédure, etc. sont examinés en utilisant

un plan hiérarchisé ou Nested-design [65]. Swierenga et al. [141] ont appliqué le test de robustesse sur des modèles d'étalonnage de spectres Raman. Ce test comprenait les étapes suivantes :

1. Choix d'un sous-ensemble représentatif d'échantillons, avec leur paramètre de référence correspondant à prédire par le modèle ;
2. Choix des facteurs externes qui peuvent influencer les résultats selon leurs niveaux ;
3. Choix des niveaux pour que les facteurs soient examinés ;
4. Construction d'un plan d'expérience approprié comprenant des niveaux de variation des facteurs externes ;
5. Exécution des expériences dans des circonstances définies par le plan d'expérience, en utilisant un sous-ensemble d'échantillons choisis ;
6. Prédiction des paramètres désirés de l'échantillon en utilisant le modèle d'étalonnage à étudier et détermination de l'erreur de prédiction (RMSEP) à chaque point expérimental du plan ;
7. En conclusion, calcul des effets des facteurs externes sur l'erreur du modèle de prédiction : Analyse statistique des résultats et interprétation.

Chacune des étapes mentionnées ci-dessus sera détaillée dans les sections suivantes.

2.3.1.1 Le choix d'un sous-ensemble représentatif d'échantillons

Le sous-ensemble d'échantillons utilisé pour le test de robustesse est choisi pour être représentatif des mesures [142]. Le choix des échantillons à analyser parmi beaucoup d'échantillons est généralement une tâche difficile. Plusieurs stratégies sont déjà disponibles pour obtenir un sous-ensemble, dont la distribution est le plus possible semblable à la distribution originale [141]. Ces techniques seront discutées plus tard.

2.3.1.2 Le choix des facteurs externes

Le test de robustesse est habituellement appliqué à un procédé optimisé dans des conditions standard. En conditions réelles, des facteurs quantitatifs et qualitatifs af-

fectent les réponses de la méthode. Différentes méthodes existent pour identifier ces facteurs. Le diagramme d'Ishikawa pourrait être employé pour les identifier [40]. Ces facteurs choisis doivent représenter les changements qui sont le plus susceptibles de se produire quand une méthode est transférée entre les laboratoires, les analystes ou les instruments ou dans le temps, et qui pourraient avoir une influence sur la réponse de la méthode [65, 160]. Cependant, seulement un nombre limité de facteurs sont considérés. Leurs interactions ne sont pas prises en compte tant que leurs effets sont négligeables par rapport à ceux des facteurs principaux [161]. Ce nombre augmente en fonction des applications prévues du modèle : utilisation interne, sur différents emplacements, dans des études de collaboration [43].

2.3.1.3 Le choix des niveaux des facteurs

Une fois que les facteurs à examiner sont choisis, le nombre de niveaux examinés est fixé. Les niveaux correspondent à des valeurs des facteurs, dans un intervalle légèrement plus grand que les limites inférieures et supérieures de celles trouvées dans les situations pratiques sous lesquelles le modèle sera appliqué. Selon le degré du polynôme de la courbe de réponse prévue, deux, trois ou cinq niveaux sont considérés pour chaque facteur [112]. Quand les valeurs de niveau sont petites autour de la valeur standard, Nijhuis [105] recommande d'employer des plans d'expériences à deux niveaux simples correspondant à une réponse linéaire. Le choix de plus de deux niveaux augmente le nombre d'expériences, ce qui n'est pas recommandé dans un test de robustesse. Un plan de cinq niveaux est préférable pour une surface de réponse plus précise (une réponse non linéaire) [43]. Les niveaux sont normalement situés autour des niveaux standard avec un intervalle symétrique, toutefois un intervalle dissymétrique peut être employé dans des conditions de contraintes [162].

2.3.1.4 Le choix du plan d'expériences

Les facteurs choisis sont examinés dans un plan d'expériences. Seuls les plans d'expériences déjà utilisés dans des études de robustesse ont été décrits [102].

Les plans d'expériences à deux niveaux ou plans de criblages renferment les plans

factoriels complets, qui considèrent toutes les combinaisons possibles entre les G facteurs à deux niveaux, indiqués par le signe + pour le niveau élevé et − pour le niveau bas. Ce sont des plans du premier ordre parce que, dans la région d'intérêt, la réponse y peut être estimée par le modèle :

$$y = b_0 + b_1 G_1 + b_2 G_2 ... + b_g G_g + b_{12} G_1 G_2 ... + b_{g(g-1)} G_{g-1} G_g + e$$

avec g le nombre de facteurs G d'influences étudiés, et b le coefficient du modèle et e l'erreur du modèle. En général, le nombre d'expériences dans un plan factoriel complet à deux niveaux est de 2^g expériences. Ils ne sont pas souvent appliqués en raison du grand nombre d'expériences impliquées [43]. Ils sont le plus souvent employés dans des cas de simulation où le coût expérimental est peu élevé [141, 73, 130]. Les plans factoriels fractionnaires correspondent à une fraction du plan factoriel complet dans laquelle les interactions élevées sont considérées comme étant négligeables ; ils sont fortement économiques pour le criblage. Le nombre d'expériences est réduit par un nombre p selon 2^{g-h} avec h le nombre de facteurs qui ont servi au criblage. La plus petite fraction d'un plan s'appelle un plan factoriel partiel saturé. Une étude comparative de l'application du plan factoriel complet et partiel pour le test de robustesse a montré des résultats comparables entre les deux types de plans [116]. Les plans de Plackett-Burman (PB), également basés sur les matrices de Hadamard ont plus de flexibilité. Ils représentent une catégorie des plans saturés avec le degré le plus élevé de fractionnement. D'une façon générale, les plans de PB sont décrits pour un certain nombre d'expériences n = 4, 8, 12, 16,... p.e. multiples de quatre. Par conséquent, avec N expériences, ils permettent l'étude d'un nombre fixe des facteurs $(n-1)$. Les signes de la première ligne de chaque plan de n-expériences sont donnés par Plackett & Burman [115] ; les $n-2$ lignes suivants sont obtenues par la permutation cyclique des signes comparés aux rangées précédentes. Le nombre de facteurs pour le plan de PB est fixé par le nombre d'expériences ; après cette détermination, les facteurs potentiels restants dans le plan sont définis comme des

variables fictives. Un facteur fictif est un facteur imaginaire [65]. La condition que les interactions doivent être négligeables est en général supposée dans les tests de robustesse [74]. Pour les plans de PB, la présence des interactions dans le modèle vrai induit une structure complexe de biais sur les coefficients de premier ordre. Par conséquent, ils sont habituellement utilisés pour estimer seulement les effets principaux et ils ont été souvent employés pour des études rapides de robustesse et donc recommandés dans les méthodes officielles pour le test de la robustesse [43, 112, 130, 151, 74]. Des plans sursaturés ont été employées par Heyden et al. [159] pour le test de la robustesse d'une méthode de chromatographie liquide ; ils sont construits par l'intermédiaire des fractions des plans de PB. Une colonne du plan PB est choisie comme une branche qui signifie que les signes (+) et (-) de cette colonne sont employés pour dédoubler le plan de PB dans deux plans sursaturés. À partir d'un plan PB (n expériences de $n-1$ facteurs), deux plans sursaturés sont obtenus avec les $n/2$ expériences employées pour examiner $n-2$ facteurs. Ils permettent d'évaluer la robustesse de la méthode en estimant la variance totale de la réponse. Aucun plan sursaturé ne peut être créé à partir des plans PB avec n =16, 32, 40 et 56 ; ils sont créés par reploiement (chaque colonne présentée deux fois) [88].

Les plans de surface de réponses ou plans à trois niveaux appelés aussi les plans de second ordre. Ils sont employés afin de décrire le rapport entre les réponses et les facteurs quantitativement à l'aide des modèles polynômiaux. Pour G facteurs d'influences étudiés, la réponse dans la région d'intérêt peut être estimée par le modèle général du second ordre.

$$y = b_0 + (b_1 G_1 ... + b_g G_g) + (b_{11} G_1^2 ... + b_{gg} F_g^2) + (b_{12} G_1 G_2 ... + b_{g(g-1)} G_{g-1} G_g) + e$$

Ils utilisent un plan factoriel complet à trois niveaux qui ne sont pas fréquemment utilisés pour tester la robustesse à cause du grand nombre d'expériences relatif. Un autre type de plan à trois niveaux est le plan central composite (CCD) déduit d'une combinaison d'un plan factoriel complet et d'un plan en étoile, avec les

centres confondus. Le nombre d'expériences $n = 2^{g-h}+2g+n_0$ avec g le nombre de facteurs, h le nombre de facteurs supplémentaires à aliaser afin de réduire le plan complet et n_0 le nombre d'expériences au centre du plan. Le plan CCD permet d'évaluer non seulement les effets principaux, mais également les interactions et les effets quadratiques. D'autres avantages du CCD sont représentés par son approche séquentielle pour l'expérimentation, puisqu'il peut être établi en deux temps (plan factoriel et puis plan en étoile) et sa flexibilité en choisissant la distance des points étoiles du centre [65]. Ses inconvénients résultent du fait que les points étoiles sont en dehors de l'hypercube et le nombre de niveaux à ajuster à chaque facteur est réellement 5 au lieu de 3. Il exige également un grand nombre d'expériences même pour un faible nombre de facteurs [112]. Quelques auteurs les ont employés [180] pour le test de robustesse. Par exemple, dans son étude sur la robustesse de la chromatographie liquide (LC) et de la chromatographie d'électrophorèse (la EC), Fabre [43, 42] a employé le plan CCD après un plan de PB pour explorer la variation de la réponse à l'intérieur et légèrement en dehors du domaine étudié dans les expériences de criblage. Roussel [121] a utilisé un CCD après un plan factoriel pour comparer les effets de différents types de bruit sur les performances de différents modèles d'étalonnage multivariés. S'il est difficile de réaliser les ajustements pour établir des plans CCD, des plans alternatifs sont employés comme des plans de Box-Behnken (BB) [24]. Dans les plans de BB, les points expérimentaux se trouvent sur une sphère à l'intérieur de l'hypercube. Par conséquent, le nombre d'expériences est moindre que dans un plan CCD. Par exemple, le nombre d'expériences pour un plan à 3-facteurs de BB est de quinze dont douze sont placées sur les milieux des bords du cube et trois comme points centraux. Un inconvénient est la représentation graphique de la réponse en fonction des facteurs, parce que les mesures dans le coin du cube n'ont pas été calculées, ainsi l'effet de chaque facteur sur la réponse ne peut pas être évalué [112]. Ragonese [116] a employé ce type de plan d'expérience pour l'optimisation et l'examen de la robustesse d'une méthode de chromatographie électrophorèse. Il a employé le test de convenance "System Suitability Testing

(SST)", dérivé du test de robustesse pour fournir l'intervalle de confiance pour les valeurs des réponses prédites. Les plans pour surfaces de réponse sont d'un grand intérêt pour le transfert de méthode parce qu'ils donnent une image complète du comportement et des limites de la méthode [157].

Les plans reflétés ou les plans à 3 niveaux bien-équilibrés sont habituellement employés pour examiner des facteurs à trois niveaux [103, 153, 1, 158]. Le plan reflété ou "reflected design" est un plan à deux niveaux (qui peut être complet, partiel ou de PB) qui est exécuté deux fois. Le premier plan considère le premier extrême et le niveau standard et le second plan considère l'autre extrémité et le niveau standard. Il est employé pour avoir une interprétation plus précise de l'influence des facteurs sur les résultats.

Des plans asymétriques contiennent des facteurs examinés à différents niveaux. Sept plans principaux orthogonaux ont été proposés par [3]. Ce type de plan a été employé pour un essai de robustesse par Hund et al. [74]. Il permet l'examen simultané d'un facteur à quatre niveaux, et douze facteurs à 2 niveaux.

Le plan d'expérience de Taguchi a été conçu et développé par Dr. Genichi Taguchi au Japon. Il est basé sur les matrices orthogonales développées dans les années 1890s par Hadamard. Il est habituellement nécessaire de déterminer les valeurs optimales pour les divers facteurs étudiés pour un procédé donné et en même temps d'examiner leur robustesse. Les facteurs à optimiser s'appellent les facteurs de commande et ceux à examiner pour la robustesse s'appellent des facteurs de bruit, des variables environnementales ou les sources de bruit [18]. L'avantage de ce plan est que les facteurs de commande sont distingués des facteurs de bruit. En outre, certains facteurs de commande et de bruit peuvent être examinés à différents niveaux [65, 84]. Wortel et al. [179] ont appliqué cette approche. Ils ont employé le concept basé sur la construction d'une rangée interne (facteurs de commande) et d'une rangée externe (facteurs de bruit) mais ils ont employé un plan factoriel pour améliorer l'efficacité de l'étude.

Les plans hiérarchisés sont employés pour examiner la robustesse définie par la re-

productibilité. Ils tiennent compte de différentes sources de variations : laboratoires, instruments dans les laboratoires, jours, répétitions dans les jours ; chaque source de variation représente un facteur qui pourrait être étudié aux différents niveaux [55]. Les plans hiérarchisés sont les modèles des effets aléatoires ; le modèle correspondant est donné par :

$$y_{ij} = \mu + a_j + e_{ij} \qquad \text{i=1,..,n} \qquad \text{et} \qquad \text{j=1,...,g} \qquad (2.2)$$

où a_j est l'effet aléatoire d'un facteur qui est normalement distribué avec une moyenne nulle et une variance σ_a^2, y_{ij} est le résultat mesuré, μ est la moyenne de la population et e_{ij} est l'erreur aléatoire des résultats. Un plan hiérarchisé est recommandé pour étudier l'effet des sources de variabilité qui se manifestent au cours du temps. En outre, le calcul est basé non seulement sur le SEP mais également sur plusieurs composantes de variance qui pourraient également être comparées entre les méthodes ($s^2_{repetitions} + s^2_{days} + s^2_{instruments} + s^2_{laboratory}$). Ainsi, il permet d'obtenir la contribution relative de chaque facteur à la variance totale et donc d'évaluer la robustesse du procédé analytique vis-à-vis de chaque facteur [156]. Dans une étude comparative des méthodes d'étalonnage pour l'analyse des spectres SPIR en reflectance, Howard et al. [93] ont démontré que le plan hiérarchisé convenait bien.

Les plans D-Optimaux sont employés :

- quand la région expérimentale n'est pas régulière dans la forme due aux contraintes ;
- quand le nombre d'expériences choisies par un plan classique doit être réduit, ou
- pour les modèles qui dévient du premier ou du second ordre habituel.

Le modèle peut être présenté dans la notation de matrice comme : $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ où $\mathbf{y}_{(n\times 1)}$ est un vecteur colonne des valeurs de réponse et $\mathbf{X}_{(n\times g)}$ est la matrice des expériences, g le nombre de facteurs définis dans le plan, $\mathbf{b}_{(g\times 1)}$ est le vecteur colonne des coefficient d'estimation du modèle d'étalonnage et $\mathbf{e}_{(n\times 1)}$ est le vecteur colonne d'erreur. Le but du plan optimal est de choisir les n bonnes valeurs pour

$\mathbf{X}$ correspondant au minimum de $(\mathbf{X}^T\mathbf{X})^{-1}$. Par conséquent, un certain nombre de critères ont été élaborés :

- critère du A-optimalité : réduit au minimum la trace de $(\mathbf{X}^T\mathbf{X})^{-1}$.
- critère du D-optimalité : réduit au minimum $\det(\mathbf{X}^T\mathbf{X})^{-1}$.
- critère de E-optimalité : réduit au minimum la valeur propre maximale de $(\mathbf{X}^T\mathbf{X})^{-1}$.
- critère de G-optimalité : réduit au minimum la valeur maximum de la variance de la réponse prédite $\frac{1}{n}\left(\mathbf{y}^T\mathbf{y}\right)$.

Deux algorithmes sont employés pour établir le plan D-optimal : l'algorithme de Mitchell et l'algorithme de Federov [100, 34]. Ils visent à choisir "n" échantillons, dans le domaine expérimental, qui contiennent ensemble la quantité maximale de l'information. Cependant, aucune étude de cas n'est connue de nous concernant l'application de cette approche au test de robustesse des étalonnages multivariés.

2.3.1.5 L'exécution expérimentale des essais

Une fois que le plan d'expérience est établi, les expériences sont de préférence effectuées dans un ordre aléatoire pour éviter d'introduire du biais dans la réponse suite à la présence des facteurs non contrôlés. Quand il est difficile d'exécuter un grand nombre d'expériences dans des conditions identiques, il faut tenir compte des changements systématiques (effets de bloc) [112]. Cependant, quand le plan n'est pas saturé, il est recommandé d'arranger les expériences de telle manière que des effets de temps soient en grande partie expliqués par les facteurs ou les interactions factices ; les facteurs factices sont des variables imaginaires pour lesquelles le changement d'un niveau à l'autre dans le plan d'expérience n'a aucune signification physique [105].

2.3.1.6 Le calcul des réponses

A partir des expériences exécutées, un certain nombre de réponses a pu être déterminé. Un modèle d'étalonnage basé sur ces réponses est établi. Cependant, dans le cas de vérification de la dérive, les résultats des mesures sont corrigés avant étalonnage [160].

2.3.1.7 L'analyse statistique des résultats et interprétation

Lorsque les facteurs étudiés correspondent à un changement dans les conditions de mesure, l'analyse des résultats des plans d'expériences donne une idée des problèmes potentiels qui pourraient se produire quand la méthode est répétée dans différentes conditions ou transférée par exemple, dans un laboratoire différent [157]. Des méthodes statistiques et graphiques peuvent être employées.

Interprétation statistique : L'effet E du facteur sur la réponse y est défini comme :

$$E = \frac{\sum y_+}{\frac{n}{2}} - \frac{\sum y_-}{\frac{n}{2}}$$

avec n le nombre des essais du plan d'expériences. L'effet a pu être calculé aussi comme étant :

$$E = \frac{\sum y_+ - \sum y_-}{n} - \frac{\sum y_+ - \sum y_-}{2g}$$

avec n le nombre d'expériences à chaque niveau. C'est l'effet produit en changeant un facteur du niveau nominal à un niveau externe. D'une façon générale, l'effet d'un facteur sur une réponse est évalué en employant un test de Student. Un effet est considéré significatif si sa valeur absolue excède l'effet critique ($|E_{critical}|$), qui est la valeur pour $t_{critical}$ donné issu d'un test de Student à un niveau de significatif α= 0,05 (parfois 0,01 ou 0,1) et pour le nombre approprié de degrés de liberté :

$$\|E\| > E_{critical} = t_{critical}(SE)_e$$

où $(SE)_e$ est l'erreur standard estimée sur un effet et $t_{critical}$ est la valeur du tableau de Student pour α= 0.05 et un degré de liberté (ddl) = nombre d'expériences utilisées pour l'estimation de $(SE)_e$ [160]. L'intervalle de confiance des effets est défini comme : $[E - t_{critical}(SE)_e; E + t_{critical}(SE)_e]$.
En outre, il est possible d'évaluer les effets des facteurs examinés sur la réponse à

partir d'un test de convenance (SST), résultant du test de robustesse comme proposé par Mullholland [103, 102]. L'ICH [76] considère un SST comme partie intégrale de beaucoup de méthodes analytiques. Il permet de vérifier la pertinence et l'efficacité de la méthode de mesure [29]. Les directives de l'ICH recommandent après l'évaluation de la robustesse, de mener un SST afin de s'assurer que la validité du procédé analytique est maintenue lors de son utilisation. Les limites du SST sont habituellement obtenues, pendant l'optimisation et la validation de la méthode, à partir des résultats des essais du test de robustesse [160] ; ces limites sont celles de l'intervalle de confiance à 95% défini par :

$$\text{Niveau bas} = \left[\overline{y} - t_{\alpha,ddl}\left(\frac{(SE)_e}{\sqrt{n}}\right); \infty\right[$$
$$\text{Niveau haut} = \left]0; \overline{y} + t_{\alpha,ddl}\left(\frac{(SE)_e}{\sqrt{n}}\right)\right]$$

où $\overline{y}$ est la moyenne de la réponse ; $t_{\alpha,ddl}$ le t-Student pour α=0.05 avec le ddl le degré de liberté et n le nombre d'expériences pour estimer SE à chaque niveau ; SE = l'estimation de l'écart-type de l'effet de chaque facteur dans des conditions du test de robustesse. L'estimation de l'écart-type $(SE)_e$ peut être effectuée de différentes manières :

1. **À partir de l'erreur expérimentale sur les réponses :**

$$(SE)_e = \sqrt{\frac{s^2}{n}}$$

Si s est connue le calcul est direct de SE. Si s est inconnue, elle peut être déterminée en effectuant des expériences supplémentaires (répétitions au centre du domaine ou duplication du plan d'expériences). Mais, dans le cas où il est impossible d'exécuter des mesures supplémentaires, s peut être évaluée en considérant les effets des interactions d'ordre supérieur à trois [75, 74, 58].

2. **À partir de la distribution des effets**

Ceci est représenté par la méthode de Lenth et l'algorithme de Dong [74, 130].

Nijhuis et al. [105] ont discuté ces méthodes qui ont été employées pour analyser des effets sans estimation expérimentale de la variance. Ces méthodes consistent à estimer s en utilisant la médiane de la valeur absolue des effets multipliée par une constante appropriée pour une variable aléatoire qui suit une distribution normale pour une estimation de l'écart-type. Cette estimation est utilisée pour calculer des marges d'erreur (ME), qui seront utilisées comme critère de décision sur l'effet des facteurs étudiés dans un test de robustesse. Pour plus de détails sur ces algorithmes, vous pouvez avoir recours à [185, 160].

3. **À partir de l'analyse de variance (ANOVA)**

   L'analyse de la variance (ANOVA) peut être appliquée pour identifier des effets significatifs [156, 179, 118]. Elle est employée avec des plans d'expériences hiérarchisés [69, 151]. Elle ne peut pas être appliquée quand des plans saturés ou fortement fractionnés sont employés parce que le nombre de degrés de liberté est petit [105].

Interprétation graphique : L'interprétation graphique est également employée pour décider sur la signification des effets.
Dans les tests de robustesse, le plan d'expérience choisi est généralement celui avec le nombre minimum d'expériences par exemple un plan saturé ou fortement fractionné. Par conséquent, ils n'y a pas suffisament de degrés de liberté pour évaluer la signification d'effets en utilisant les essais statistiques. Le même cas se présente quand les plans d'expériences ne peuvent pas être répétés. Ces problèmes sont surmontés en appliquant les 'normal-probability' et les 'half-normal' plot comme outil graphique pour juger visuellement sur la signification des effets, comme proposé par Daniel et cité dans [162, 74, 160]. Les effets qui dévient de la ligne droite sont considérés comme significatifs.

La pente de la ligne passant par les effets supposés non significatifs donne une évaluation de l'écart type de l'erreur [105]. Différents algorithmes sont employés pour extraire l'interprétation quantitative à partir de ces représentations graphiques [86, 105]. Sanz et al. [130] ont appliqué la technique du half-normal plot avec l'algorithme de

Lenth pour l'étude de la robustesse d'une méthode polarographique en utilisant l'étalonnage multivarié. Les interprétations étaient semblables à celles trouvées par la méthode de répétition décrite ci-dessus.

Des figures de contour de robustesse (contour plots) [179] sont utilisées comme un moyen de combiner des informations sur la complexité du modèle, sur les performances du modèle et sur les facteurs externes. Une figure de contour est une technique graphique pour représenter une surface à trois dimensions en traçant les tranches constantes de z, appelée les *contours*, sur un format à deux dimensions. C'est-à-dire, étant donnée une valeur pour z, des lignes sont tracées pour relier tous les (x, y) où cette valeur de z se produit. La figure de contour est une alternative à une surface de réponse à trois dimensions et est utilisée dans le cas du modèle de prédiction.

2.3.2 Indices de robustesse

Les indices de robustesse utilisés pour l'évaluation de la robustesse des étalonnages multivariés sont peu nombreux. Ils sont tous basés sur la réduction au minimum d'une fonction objective de l'erreur de prédiction.

2.3.2.1 Critères basés sur le rapport signal/bruit (SNR)

En littérature, les critères de robustesse basés sur le travail et la philosophie de Taguchi pour l'amélioration et le contrôle de qualité, sont fréquemment trouvés [65]. Taguchi définit la non qualité d'un produit comme étant "la perte provoquée par le produit à la société du moment où le produit est embarqué". La forme d'une telle fonction, appelée "Loss function" ou "fonction de perte", dépend du processus ou du produit et est souvent difficile à établir. Mathématiquement, la fonction de perte L des performances du rendement y comme défini par Taguchi est :

$$L_{\Omega}(y) = a(y - \upsilon)^2$$

où y et υ sont respectivement les performances du système et la vraie valeur à atteindre avec a comme constante de coût. La valeur prévue de la fonction de perte du rendement y qui est une fonction des entrées X peut être décomposée comme suit :

$$EL_{\Omega}(y) = aE(y-\upsilon)^2 = a\sigma^2 - a(\upsilon-\mu)^2$$

où $\mu = E(y)$ et σ^2 sont respectivement la moyenne et la variance de y dans l'espace de bruit (Ω). υ est la valeur à atteindre.
L'objectif de Taguchi est que la réponse soit sur la cible avec le minimum de variance. Cet objectif est atteint en maximisant le rapport signal/bruit d'estimateur (SNR) donné par Taguchi [20].

$$\begin{aligned} \frac{S}{N} &= 10\log_{10}\frac{\mu^2}{\sigma^2} && \text{le plus grand est le meilleur} \\ \frac{S}{N} &= -10\log_{10}\frac{1}{n}\sum y_i^2 && \text{le plus petit est le meilleur} \\ \frac{S}{N} &= -10\log_{10}\frac{1}{n}\sum\frac{1}{y_i^2} && \text{le plus grand est le meilleur} \end{aligned}$$

Les approches de Taguchi ont été évoquées et discutées dans [65]. Il y a eu également une certaine critique de l'utilisation des rapports signal/bruit de Taguchi comme critères de qualité dans [23]. L'inconvénient de cette méthode est que le SNR ne contient pas toute l'information appropriée. En outre, cette définition de qualité n'assure pas que, pour L(y) minimal, la valeur prévue de la caractéristique de qualité soit sur la cible. Dans cet objectif, un bas σ^2 avec peu du biais peut être préférable si la perte attendue est inférieure à une situation sans biais avec un plus grand σ^2. Wortel et al. [179] a combiné la philosophie de Taguchi, des plans d'expériences et des spectres artificiels obtenus pour différents facteurs de perturbations artificiellements générées, pour évaluer et améliorer la robustesse de l'étalonnage du PIR. Basé sur l'approche de Taguchi, ils ont proposé d'employer le rapport signal/bruit $Z(\theta)$ pour exprimer les

performances de la prédiction donné par le RMSEP, comme critère de robustesse :

$$Z(\theta) = 10 \log \frac{\overline{RMSEP}^2}{\sigma_{RMSEP}} \; RMSEP = \sqrt{\frac{1}{n} \sum \left(\widehat{y}_i - \upsilon\right)^2}$$

où $\overline{RMSEP}$ et σ^2_{RMSEP} sont la moyenne et la variance du RMSEP données par le plan expérimental. n est le nombre d'échantillons d'essai et $\widehat{y}_i$ et υ sont la valeur prévue et vraie respectivement. Une comparaison est faite entre les résultats obtenus et ceux donnés par un modèle de base (ou référence) basé sur la moyenne de minimum du RMSEP. $Z(\theta)$ est une approximation du rapport de biais/dispersion dû à l'influence des facteurs G. Wortel définit ce critère comme rapport signal/bruit des performances de la prédiction. Ce choix est complètement défendable. Son inconvénient principal est que, étant une erreur, le RMSEP est inversement proportionnel aux performances des prédictions. Par conséquent maximisant $Z(\theta)$ peut être incorrect, menant à réduire au minimum la variance σ^2_{RMSEP}, alors que RMSEP a une valeur élevée.

2.3.2.2 Critères robustes basés sur la méthodologie de surface de réponse.

1. La méthode de Jones [81] a été proposée comme alternative à la méthode de Taguchi. Elle est basée sur l'utilisation de la méthodologie de surface de réponse extraite à partir des plans de Taguchi, sans application des rapports de SNR. Cette méthode consiste à calculer un plan d'expérience pour tous les facteurs d'intérêt. Un modèle approprié est choisi pour représenter la réponse en fonction des facteurs. Puis, les facteurs avec effet significatif sur la réponse sont divisés en facteurs contrôlés G (de commande) et facteurs environnementaux N (de bruits). La réponse est graphiquement tracée en fonction des facteurs de bruit et de commande. Analysant cette surface de réponse, Jones a employé l'intégrale du carré de l'erreur (la fonction de perte) comme critère de performance :

$$L(x) = cte \int_{Rz} \left[\upsilon - \widehat{y}_{xz}\right]^2 dz$$

avec Rz la région d'intérêt des facteurs d'environnement ou de bruit z, $\widehat{y}_{xz}$ étant une valeur prévue de la réponse y pour une certaine combinaison donnée entre x et z, υ est la valeur à atteindre ou la réponse idéale et x est le facteur de contrôle. $L(x)$ a le même problème que le SNR qui contient deux objectifs expérimentaux (bias $\rightarrow 0$ et variance $\rightarrow 0$) avec une relation fixe entre eux. Par conséquent, $L(x)$ est séparé en :

$$M(x) = \int_{\mathbb{R}_F} \left[\upsilon - \widehat{y}_x\right]^2 dz$$
$$V(x) = \int_{\mathbb{R}_F} \left[\widehat{y}_{xz} - \widehat{y}_x\right]^2 dz$$

où $M(x)$ est le carré de la déviation de la réponse moyenne par rapport à la cible et $V(x)$ est le carré de la variation moyenne de la réponse moyenne.
Pour commander le rapport entre $M(x)$ et $V(x)$, un facteur de poids est introduit. Il consiste en :

$$L(x) = \lambda V(x) + (1-\lambda)M(x) \quad \text{avec} \quad 0 \leq \lambda \leq 1$$

Par l'ajustement de λ, l'importance relative du biais et de la variance est controlé.

2. La méthode pondérée de Jones est une variante de la méthode de Jones. Elle emploie la variance $V(x)$ avec un facteur de poids qui considère la distribution de probabilité des facteurs.

$$WJ_c = \int_{\mathbb{R}_\frown} \left[\widehat{y}_x - \widehat{y}_{xc}\right]^2 W_x dx$$

avec le c comme point d'intérêt (point central du domaine) ; $\mathbb{R}_\frown$ est la région elliptique autour de c qui renferme la majeure partie de la distribution de probabilité des erreurs contenu dans l'ensemble des facteurs contrôlés au point c ; W_x est la densité de probabilité de la distribution des facteurs contrôlés, qui agit en tant que poids.

3. La méthode de variance projetée (PV) [65, 165] décrit la variance totale ($\sigma_{t,c}$)

comme étant la somme de l'erreur pure dans les mesures de la réponse (σ_e) et de l'erreur propagée au point c d'intérêt (σ_c).

$$\sigma_{t,c}^2 = \sigma_e^2 + \sigma_c^2$$

Réduire au minimum cette variance maximise la robustesse de la méthode.

4. Le coefficient de robustesse, RC, représenté par la probabilité que, pour une variation connue de la variable indépendante x, la variable dépendante y appartienne à un intervalle prédéfini, représenté par la plus petite fraction symétrique autour du point d'intérêt c pour $\alpha = 0.05$. Plus le RC est grand, plus les résultats sont robustes. Si la vraie distribution des erreurs de x n'est pas connue, alors le RC est représenté par la distance de Mahalanobis entre c et un certain point a de la distribution. Ce critère RC est employé particulièrement pour des problèmes d'optimisation de plans de mélange [32].

5. Dans son étude sur la robustesse de la fusion des capteurs à l'égard du dysfonctionnement des capteurs. François [52] a évalué la robustesse en utilisant la fonction $\mathcal{I}$ qui représente la "vague d'erreur". Il s'agit de la surface représentant l'erreur de prédiction comme fonction des perturbations (p.e. la dérive) dues aux facteurs d'influences G responsables des erreurs systématiques et du bruit N :

$$I_0 = \mathcal{I}(N, G)$$

où $\mathcal{I}$ est la fonction d'erreur, N le bruit et G la dérive. Elle est semblable à la fonction donnée par un plan expérience. Par conséquent, l'évaluation de la robustesse exige un premier critère extrait à partir de la surface de vague d'erreur telle que le secteur de la surface de vague d'erreur prise au-dessous d'un seuil donné d'erreur. Puis, un deuxième critère ρ a été calculé avec une approche stochastique : le domaine d'incertitude ou d'erreur est défini comme un espace probabiliste où on suppose que les paramètres ou les perturbations incertains suivent une fonction commune donnée de densité de probabilité. Ce critère tient

compte du degré de la fonction $\mathcal{I}$ et de la fonction de loi de probabilité supposant que la probabilité pour avoir une grande erreur de prédiction augmente tout en s'éloignant de la position optimale.

$$\rho = \int \int_{\Omega} \|\mathcal{I}(N,G)\|^2 \mathcal{K}(N,G) dN dG$$

où Ω est l'espace des perturbations, $\|\mathcal{I}(N,G)\|$ la norme du gradient de la fonction d'erreur $\mathcal{I}$, $\mathcal{K}$ la loi de probabilité des perturbations pour un capteur donné. Dans le même contexte de la fusion de capteurs, Steinmetz [137] a proposé l'utilisation de deux critères mentionnés ci-dessus avec quelques variations et a inclus le critère I_{1i} de pertinence de chaque capteur i :

$$I_{1i} = I_{0_i} - I_0$$

où I_{0_i} est l'erreur obtenue sans employer le i^{eme} capteur et I_0 est obtenu quand tous les capteurs sont utilisés et leur biais maximal toléré est déterminé par le rapport :

$$I_2 = \frac{\int \int_{\Omega} dN dG}{2 N_{max} G_{max}}$$

où $\int \int_{\Omega} dN dG$ est la surface du domaine limité par N et G. N_{max}, G_{max} sont les valeurs maximales de N et de G. Il a combiné les trois critères mentionnés ci-dessus dans :

$$I_3 = \frac{\rho}{I_1 I_2}$$

afin d'évaluer la robustesse du système de fusion de capteurs. Plus l'index I_3 est petit, plus robuste est le système.

6. $(RMSEP)$ est le critère le plus utilisé pour comparer la robustesse des modèles d'étalonnage [118, 141]. Roussel [121] a comparé la robustesse des modèles d'étalonnage, en simulant des bruits N, pour des analyses de la composition des grains

par spectroscopie SPIR. Elle a proposé d'employer un critère quantitatif de robustesse pour chaque modèle. Ce critère R_N est la dérivée partielle du (SEP) par rapport à chaque facteur de bruit (N) :

$$R_N = \frac{\partial SEP}{\partial N}$$

qui est la pente de la fonction d'erreur à l'égard de chaque facteur de bruit, avec

$$SEP_c = \sqrt{\frac{1}{n_t - 1} \sum_{i=1}^{n_t} (\widehat{y}_i - y_i - biais)^2}$$

$$biais = \frac{1}{n_t} \sum_{i=1}^{n_t} (\widehat{y}_i - y_i)$$

où n_t est le nombre d'échantillons de test et $\widehat{y}_i$ est la valeur prédite et y_i la vraie valeur.

2.4 Synthèse : Définition du critère de la robustesse des étalonnages multivariés

Différents critères sont employés dans la littérature, quelquefois d'une façon confuse, pour déterminer la qualité de la méthode de la spectroscopie SPIR. Une distinction entre ces critères a rendu plus clair l'intérêt de l'utilisation de chacun pour évaluer les caractéristiques d'une méthode de mesure.

La définition de la robustesse que nous proposons est : *la stabilité de la capacité prédictive du modèle d'étalonnage vis-à-vis des perturbations appliquées au voisinage des conditions standard.* Le critère adopté pour évaluer cette robustesse des modèles d'étalonnage multivariés a été déterminé en considérant :

- la capacité prédictive du modèle d'étalonnage multivarié. Elle est représentée par le SEP défini par : $\mathrm{SEP}^2 = \frac{1}{n_t} \sum (\widehat{\mathbf{y}} - \mathbf{y})^2$ où n_t est le nombre d'échantillon de

test ;

- les perturbations qui sont à la base des problèmes de robustesse du modèle d'étalonnage. Elles sont représentées par les grandeurs d'influences G qui ont lieu en temps réels.
- la stabilité est évaluée par rapport à des conditions standard, c'est-à-dire le SEP calculé dans les conditions industrielles est évalué par rapport à SEP_0 correspondant aux conditions d'étalonnage du modèle ;
- la stabilité de la capacité prédictive du modèle d'étalonnage par rapport aux conditions standard est alors représentée par la stabilité du SEP par rapport à SEP_0. D'où, l'évaluation de cette stabilité est représentée par le rapport entre SEP_0 et SEP.

Par conséquence, le critère de robustesse R_C que nous proposons pour calculer la robustesse du modèle d'étalonnage multivarié est représenté par l'équation suivante :

$$\boxed{R_C \;=\; \frac{SEP_0}{SEP}} \tag{2.3}$$

D'où, si $R_C \in [0,1[\Longrightarrow$ le modèle est non robuste.
si $R_C \in [1,+\infty[\Longrightarrow$ le modèle est robuste.

Dans le cas de suivi en ligne, ce critère peut être calculé soit : - à la fin du procédé de mesure, permettant d'évaluer la *robustesse globale* du modèle. - en ligne, sur une fenêtre de largeur l représentant le passé proche du processus et permettant d'évaluer la *robustesse dynamique ou temporelle* du modèle tel que :

$$R_C(t) \;=\; \frac{SEP_0}{SEP(t)}$$

2.5 Conclusion

Le besoin industriel de méthodes robustes a rendu nécessaire l'évaluation de la robustesse, qui est devenue une étape importante dans le développement d'une méthode. Tout d'abord, nous avons commencé par montrer dans ce chapitre, la diversité de l'uti-

lisation du mot *robustesse* selon le domaine d'application. Après avoir analysé cette diversité, nous avons abouti à notre définition de la robustesse des modèles d'étalonnage multivariés. Ensuite, nous avons présenté plusieurs méthodes utilisant les plans d'expériences citées dans des études de robustesse et qui pourraient donc être employées pour tester la robustesse dépendant du nombre de facteurs considérés et du nombre d'expériences qu'on voudrait utiliser. Ce choix est relatif au but du test. Dans le cas de simulation, l'utilisation des plans complets factoriels et des plans centraux composites (D-optimaux) est recommandée. Dans le cas d'une application réelle, où nous cherchons à étudier beaucoup de facteurs avec peu d'expériences, les plans de Plackett Burman ou les plans asymétriques sont recommandés. Dans une troisième partie de ce chapitre, nous avons discuté les critères utilisés pour évaluer la robustesse d'une méthode. Ils font souvent l'hypothèse que les données sont normalement distribuées, ce qui n'est pas toujours le cas, et elles emploient la variance comme critère pour évaluer la robustesse. Le critère R_N proposé par Roussel semble être le plus intéressant sauf qu'il suppose que les facteurs d'influence sont connus.
Finalement, nous avons proposé un critère de robustesse R_C relatif à la définition de la robustesse donnée ci-dessus, pour évaluer la robustesse spécifiquement pour la spectroscopie infrarouge et les méthodes d'étalonnage multivariés.

Maintenant que nous avons défini notre critère d'évaluation de la robustesse, nous pouvons envisager de l'améliorer. C'est l'objet du chapitre suivant.

Chapitre 3

Amélioration de la robustesse des étalonnages multivariés

Sommaire

3.1 Introduction

Ce chapitre fait l'objet du second point de l'analyse de notre problématique. Il traite toutes les méthodes les plus généralement employées pour améliorer intrinsèquement la robustesse des étalonnages multivariés. Ce problème est surtout dû aux variations des conditions de mesures. Ces facteurs d'influence perturbent les mesures spectrales : une perturbation $\delta\mathbf{x}$ est additionnée au spectre. Cette perturbation se manifeste au niveau de la réponse du modèle de prédiction (équation 1.1) par :

$$\delta\widehat{y} \quad = \quad \delta\mathbf{x}^T\mathbf{b}$$

ce qui donne :

$$|\delta\widehat{y}| \quad = \quad \|\delta\mathbf{x}\|.\|\mathbf{b}\|\,|cos(\delta\mathbf{x},\mathbf{b})| \tag{3.1}$$

Pour minimiser $\delta\widehat{\mathbf{y}}$, il suffit de minimiser un ou plusieurs des trois termes de la partie droite de l'équation 3.1.

Un modèle d'étalonnage multivarié consiste en plusieurs étapes comme le montre la figure 3.1.

La première étape est relative à l'optimisation de la base d'étalonnage. La deuxième étape est reliée aux méthodes de prétraitement des spectres, employées pour corriger les effets additifs et multiplicatifs dus aux variations des conditions de mesures. Ces deux grandes classes de méthodes font l'objet des deux parties de ce chapitre. Pour chacune d'elle, une revue des principales méthodes est effectuée, et l'amélioration intrinsèque de la robustesse est discutée. La troisième phase "'l'étalonnage"' fera l'objet d'un chapitre particulier.

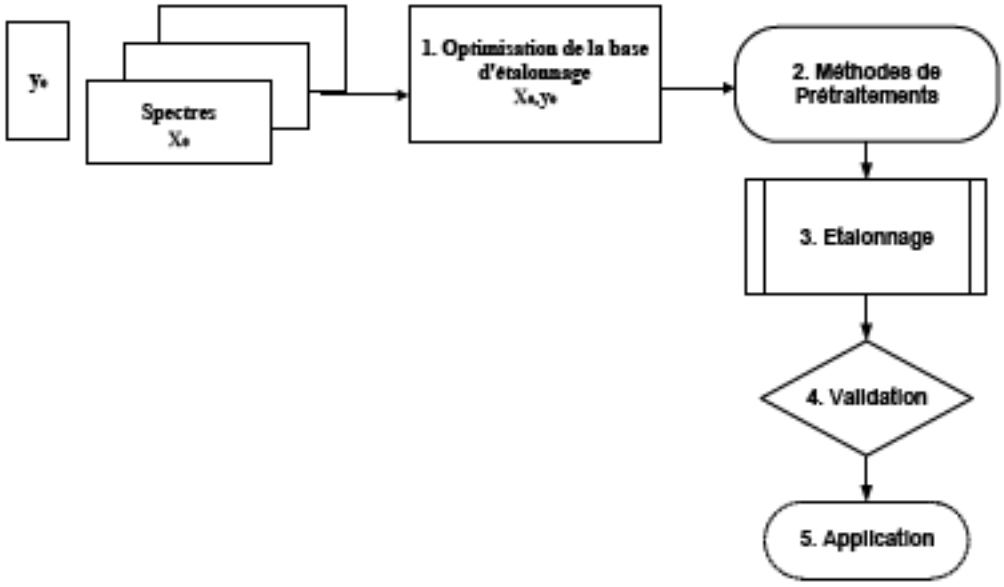

FIG. 3.1 – Étape d'amélioration de la robustesse du modèle d'étalonnage multivarié

3.2 Optimisation de la base d'étalonnage

La construction d'un modèle d'étalonnage passe par plusieurs étapes qui sont toutes basées sur l'utilisation d'une base de données dite base d'étalonnage. Cette base est constituée des mesures spectrales $\mathbf{X}$ et des mesures de référence $\mathbf{y}$ correspondant aux mêmes individus et prises dans les mêmes conditions. L'optimisation de la base d'étalonnage est donc une des premières phases de l'optimisation du modèle d'étalonnage.

3.2.1 Diagnostic des données aberrantes

Il y a fondamentalement deux approches différentes pour manipuler des données aberrantes; le premier est la détection et le rejet de ces données avant modélisation, le second est l'incorporation de leur processus d'élimination lors de la modélisation. Différentes stratégies de diagnostic peuvent être employées pour la détection des données aberrantes telles que : la distance de Mahalanobis, la distance robuste, distance de Cook [87]. La régression par composantes principales robustes (RPCR-Robust principal component regression) développée par [168].Elle est basée sur l'utilisation de la distance de Mahalannobis pour choisir le sous-ensemble d'observations ou d'individus [106].

3.2.2 Choix de la base d'étalonnage

L'étape suivante consiste à sélectionner les individus formant la base d'étalonnage optimale, en utilisant des algorithmes basés sur des principes statistiques (p.e. le maximum de variance) [68, 48]. Cette sélection est très utilisée dans le cas des produits naturels, où l'ensemble d'étalonnage ne peut pas être créé artificiellement. Dans ce cas, les mesures spectrales du lot sont effectuées et ensuite seuls les individus choisis par l'analyse de leur spectre comme étant les plus représentatifs du lot sont analysés par la méthode de référence. Elle est aussi utile dans le cas de très grandes bases de données, où un sous-ensemble choisi de façon à être représentatif suffira pour établir un modèle d'étalonnage. En spectroscopie, les méthodes classiques utilisées pour choisir le sous-ensemble des spectres sont représentées soit par l'algorithme de sélection aléatoire, soit par le choix basé sur l'ordre croissant des valeurs de $\mathbf{y}$. L'inconvénient de ces algorithmes réside dans le fait que les individus choisis ne sont pas forcément représentatifs du domaine expérimental.

D'autres algorithmes tels que l'algorithme de Federov et l'algorithme de Kennard et Stone (K&S) [31], [48], ont été développés et utilisés pour choisir le sous-ensemble optimal des spectres d'étalonnage. L'algorithme de Federov consiste à choisir les spectres qui couvrent de façon optimale le domaine d'intérêt, en maximisant le déterminant de $\mathbf{X}^T\mathbf{X}$ avec $\mathbf{X}$ centré *a priori* [65],[59]. Le choix est fait sur $\mathbf{X}$ indépendamment des réponses de $\mathbf{y}$. L'algorithme de K & S [83] choisit également l'ensemble des individus qui couvrent le domaine expérimental spectral global basé sur leur distance les uns des autres (distance Euclidienne ou de Mahalanobis) indépendamment de $\mathbf{y}$. Une variante de l'algorithme de K& S est l'algorithme DUPLEX ; il consiste à choisir alternativement les ensembles de données de l'étalonnage et du test. Il conduit évidemment au meilleur choix des individus d'étalonnage, parce que des individus d'essai sont pris sur la frontière des limites expérimentales ce qui n'est pas le cas avec K& S. Mais le problème de la représentativité du choix de la base d'étalonnage tenant compte de la réponse d'intérêt $\mathbf{y}$ demeure toujours. Isaksson et al. [80] ont comparé deux stratégies utilisées pour le choix des individus de la base d'étalonnage sur des mesures spectrales. Le meilleur

est basé sur l'analyse des classes. Il choisit les individus qui sont les plus éloignés du centre de chaque classe. Le second, basé sur la variance spectrale, choisit l'échantillon ayant la plus grande valeur absolue des absorbances. Puis, le procédé est répété jusqu'à ce que le nombre désiré de spectres soit choisi. Ferre et al. [48] ont comparé les algorithmes de Federov et de K & S avec l'algorithme de sélection aléatoire. Ils ont montré les avantages de choisir les sous-ensembles D-optimaux pour l'étalonnage avec l'algorithme de Federov. Tous les algorithmes mentionnés ci-dessus tendent à choisir les individus les plus représentatifs du domaine expérimental et ayant la distribution la plus uniforme possible sur tout le domaine. Ils essayent de tenir compte de toute la variabilité spectrale de l'espace d'étalonnage en utilisant la variance des spectres ou leur distance l'un de l'autre.

Conclusion

L'optimisation de la sélection des échantillons constituant la base d'étalonnage est une étape primordiale permettant d'améliorer sa représentativité. Toutefois, le choix des spectres les plus éloignés les uns des autres n'implique pas qu'ils correspondent aux valeurs de référence les plus éloignées. En outre, ces méthodes de sélection tendent à choisir des individus qui peuvent être aberrants. De plus, il y a plus de chance d'avoir des données aberrantes dans les mesures spectrales $\mathbf{X}$ que dans les mesures de référence du laboratoire $\mathbf{y}$. D'où notre idée d'appliquer ces algorithmes sur les valeurs de référence $\mathbf{y}$, et non sur $\mathbf{X}$, pour construire la base d'étalonnage optimale qui couvre tout le domaine d'intérêt avec moins de risques de présence des données aberrantes. Cette méthode de sélection est celle qu'on propose pour notre application.

3.2.3 Centrage

Comme mentionné dans [35], le centrage est l'une des étapes standard dans la plupart des régressions multivariées. Différents modes de centrage existent [26]. En spectroscopie, c'est le centrage selon le premier mode (en colonne) qui est utilisé. Suite au centrage, les différences entre les individus sont sensiblement améliorées en termes de réponses chimique et spectrale. Néanmoins, Seasholtz dans [133], recommande de

ne pas centrer les données d'étalonnage dans le cas où :

- La variation des spectres est linéaire en fonction de la concentration ;
- Les spectres ne présentent pas de variations dans la ligne de base (addition d'une constante) ;
- La somme des concentrations des composés n'est pas égale à une constante.

Ces conclusions ont été trouvées lors de l'examen de la propagation d'erreur dans le modèle. Seasholtz a montré que les incertitudes dans les coefficients des modèles résultent de celles qui se situent dans l'espace d'étalonnage. Dans le cas du centrage, l'incertitude reliée à la moyenne calculée est ajoutée à l'incertitude de la réponse. Ceci induit l'augmentation de l'incertitude dans les coefficients des modèles. Le tableau 3.1 montre une comparaison entre les prédictions des modèles utilisant des données brutes ou centrées dans le cas de la régression linéaire simple.

$No-centering$	$Centering$
$\mathbf{x}$	$\mathbf{x}-\overline{\mathbf{x}}$
y	$y-\overline{y}$
$\mathbf{b}=\mathbf{x}\backslash y$	$\mathbf{b}=(\mathbf{x}-\overline{\mathbf{x}})\backslash(y-\overline{y})$
$\widehat{y}=\mathbf{x}^T\mathbf{b}$	$\widehat{y}=\overline{y}+(\mathbf{x}-\overline{\mathbf{x}})^T\mathbf{b}$
$\delta\widehat{y}=\delta\mathbf{x}^T\mathbf{b}+\mathbf{x}^T\delta\mathbf{b}$	$\delta\widehat{y}=\left(\frac{\delta y}{\sqrt{n}}\right)+\delta(\mathbf{x}-\overline{\mathbf{x}})^T\mathbf{b}+(\mathbf{x}-\overline{\mathbf{x}})^T\delta\mathbf{b}$

TAB. 3.1 – Différence entre le centrage et le non-centrage des données quant à l'erreur de prédiction ([133]).

Seasholtz dans [133] a conclu que s'il n'y a pas de changement de ligne de base, le nombre de facteurs utilisés pour estimer le coefficient de régression $\mathbf{b}$ et l'erreur de prédiction sont les mêmes pour un modèle centré ou brut. Si le nombre optimal de facteurs est réduit après centrage, le modèle doit être établi avec les données centrées. Sinon, il est préférable de construire le modèle sans centrage.

Nous avons identifié que le centrage pouvait provoquer, dans certains cas, une non robutesse. Quand les données sont centrées, le modèle d'étalonnage est établi en considérant la moyenne comme étant le centre du domaine d'étalonnage $b_0 = \overline{y} - \overline{\mathbf{x}}^T\mathbf{b}$, par conséquent ce modèle est le plus représentatif de la base d'étalonnage. Le non centrage force le modèle à passer par le point (0,0) qui correspond au spectre de concentration

nulle. Ainsi, les modèles non-centrés avec $\mathbf{b}_0 = 0$ se rapprochent le plus du modèle théorique donné par la loi de Beer-Lambert pour l'analyse spectroscopique quantitative. De plus, l'extrapolation au-delà du domaine d'étalonnage est plus performante en utilisant les données brutes.

Conclusion

Donc, quand le centrage ne réduit pas le nombre de facteurs du modèle d'étalonnage, il vaut mieux ne pas centrer, car le centrage assurait alors une bonne prédiction à l'intérieur du domaine d'étalonnage mais pas à l'extérieur ; d'où la préférence, dans certains cas, et en particulier dans nos applications du non centrage.

3.2.4 Réduction

La réduction est utilisée pour réduire l'effet des faibles variations dans les données en donnant le même poids à toutes les intensités spectrales. Elle est calculée en divisant la réponse à chaque longueur d'onde j par l'écart-type des réponses de tous les spectres à cette longueur d'onde.

$$\sigma\left(\mathbf{x}\right)_j = \sqrt{\frac{1}{n-1}\left(\mathbf{x}_j - \overline{\mathbf{x}}_j\right)^T\left(\mathbf{x}_j - \overline{\mathbf{x}}_j\right)}$$

$$\mathbf{x}_{j,corrige} = \frac{\mathbf{x}_j}{\sigma(\mathbf{x})_j} \tag{3.2}$$

La réponse chimique est réduite d'une manière identique pour chaque composé. A noter que cette réduction de la variance est uniquement effectuée après centrage des données. Cette méthode est plus utilisée quand on analyse des composés chimiques concentrations faibles et dont des bandes spectrales sont masquées par celles des composés majeurs.

3.2.5 Conclusion : optimisation de la base d'étalonnage et robustesse

Examinons l'effet de ces méthodes d'optimisation de la base d'étalonnage sur l'amélioration de la robustesse du modèle. Bien qu'elle ne soit pas directement liée à la minimisation de l'un des termes de l'équation 3.1. La base d'étalonnage influence la représentativité du modèle. Optimiser la base d'étalonnage contribue intrinsèquement à la robustesse d'une façon indirecte en balayant l'espace le plus large du domaine expérimental. Pour des modèles linéaires, la mauvaise utilisation du centrage peut surtout affecter l'extrapolation du modèle en dehors du domaine expérimental. C'est un des problèmes que l'on peut rencontrer au cours des applications en ligne et que l'on peut considérer comme un manque de robustesse du modèle d'étalonnage. La réduction des données spectrales donne le même poids au bruit et aux structures informatives. C'est pour cela que ce prétraitement n'est pas de nature à améliorer la robustesse des modèles d'étalonnage, et devra être évité.

3.3 Méthodes de prétraitements des données spectrales

Les méthodes de prétraitements sont largement appliquées pour corriger les données spectrales. Elles peuvent être divisées en deux différentes catégories selon le type de correction réalisée. La première catégorie (prétraitement géométrique) modèle le spectre sans en changer la dimension, pour corriger le décalage et la dérive de la ligne de base, la curvilinéarité, les effets additifs et multiplicatifs principalement dus à la diffusion. La deuxiéme catégorie consiste à réduire la dimension spectrale.

3.3.1 Prétraitements géométriques

Ces méthodes n'utilisent que l'information sur les spectres $\mathbf{X}$ et ont vocation à améliorer le spectre, sans en changer la dimension. Elles incluent les méthodes de

lissage utilisées pour réduire le bruit et les méthodes de différenciation pour corriger le recouvrement des pics et les lignes de base affines.

3.3.1.1 Normalisation

Les méthodes de normalisation donnent le même poids aux absorbances des différentes longueurs d'onde. Elles sont appliquées spectre par spectre, sauf certaines qui nécessitent l'utilisation de toute la base de données pour déterminer les coefficients de correction. Dans ce qui suit, différentes méthodes utilisées pour la normalisation des spectres sont présentées.

Transformation normale standard des variables La diffusion de la lumière dûe aux interactions entre la radiation du rayon IR et les particules de l'échantillon analysé, crée généralement un décalage des longueurs d'onde absorbantes et une courbure de la ligne de base. Ceci affecte l'interprétation des spectres et l'étalonnage des spectres PIR en réflexion diffuse. Il s'agit d'une variation du chemin optique qui entraîne une variation de la ligne de base qui est différente selon les longueurs d'onde. L'effet de ces variations est à la base du décalage des longueurs d'onde et peut varier largement entre et à l'intérieur des spectres d'un même échantillon [79, 64, 5].
La transformation normale standard des variables (SNV) a été proposée par Barnes et al. [11] pour réduire les effets multiplicatifs de la lumière diffuse, la différence des dimensions des particules et le changement de multicollinéarité le long des spectres PIR. Chaque spectre est indépendamment centré et puis divisé par son écart-type. Son désavantage réside dans le fait que les effets sont supposés uniformes le long du spectre. Cette condition n'est pas toujours remplie ce qui augmente le risque d'introduire des artéfacts.

Transformation normale standard robuste Guo et al. [61] ont résolu ce problème des artéfacts ajoutés par l'utilisation de la transformation SNV en le traduisant par le problème de "*closure*". La "*closure*" est un terme statistique indiquant que la somme des données est nécessairement égale à une constante, de façon à ce que si les variables

changent dans une certaine direction, les autres variables doivent changer dans la direction opposée pour compenser le changement et assurer la somme constante. Guo et al. [61] ont alors surmonté ce problème en présentant une nouvelle méthode, la transformation normale standard robuste (RNV). Cette méthode modifie la SNV en employant le percentile au lieu de la moyenne pour procéder à la correction. Le r^{ieme} percentile de $\mathbf{x}$ est la valeur par rapport à laquelle $r\%$ des absorbances sont plus petites et (100-r)% des valeurs d'absorbance sont plus grandes. Ainsi, le $percentile(\mathbf{x})$ est le pourcentage des valeurs d'absorbances du spectre $\mathbf{x}$ qui correspondent au r^{ieme} percentile. Le spectre corrigé par RNV est :
$\mathbf{x}_{new} = \frac{\left(\mathbf{x} - percentile\,(\mathbf{x})\right)}{\sigma\left(\mathbf{x} \leq percentile\,(\mathbf{x})\right)}$
le $\sigma\left(\mathbf{x} \leq percentile\,(\mathbf{x})\right)$ signifie l'écart type des valeurs de $\mathbf{x}$ qui sont inférieures au percentile. L'inconvénient de cette méthode est l'optimisation du niveau de percentile à considérer. Son avantage est représenté par l'utilisation du percentile au lieu de la moyenne. Le percentile est moins sensible aux valeurs aberrantes, ce qui assure une correction plus efficace.

3.3.1.2 Correction de tendance : le De-Trend

Cette méthode corrige le décalage de la ligne de base et la curvilinearité, généralement trouvés dans les spectres de réflexion des échantillons en poudre ou emballés. De-Trend corrige la curvilinéarité de ligne de base en l'exprimant comme une fonction quadratique de la longueur d'onde λ [11, 134]. Le spectre corrigé résultant est donné par : $\mathbf{x}_{new} = \mathbf{x} - \left(a_0 + a_1\boldsymbol{\lambda} + a_2\boldsymbol{\lambda}^2\right)$, où a_i sont les coefficients du polynôme et $\boldsymbol{\lambda}$ le vecteur de longueur d'onde du spectre.

Tandis que SNV corrige le décalage linéaire de ligne de base, le De-Trend peut être employé après SNV pour éviter toute tendance curviligne comme expliqué par Barnes [11]. Cette méthode est intéressante pour éliminer les effets de lignes de base qui sont fonction de la longueur d'onde.

3.3.1.3 Correction de la dispersion multiplicative

Comme la SNV, la correction multiplicative de la dipersion (MSC) [79] se concentre sur les questions de dispersion de la lumière et de variation des dimensions des particules. Elle corrige simultanément les effets multiplicatifs (de pente) et additifs (de biais) [94]. L'approche de la MSC est basée sur deux hypothèses. D'abord, un spectre d'échantillon est considéré comme l'addition de deux spectres, le premier dû à la diffusion de la lumière ($\mathbf{d}$) et le second dû à des absorbances chimiques ($\mathbf{c}$). D'où, le spectre $\mathbf{x}$ de chaque échantillon est représenté par :

$$\mathbf{x} = \mathbf{d} + \mathbf{c}$$

La MSC consiste à corriger la partie diffuse $\mathbf{d}$ du spectre. Tout d'abord le spectre de diffusion ($\mathbf{d}$) est modélisé par la méthode des moindres carrés appliquée sur une gamme de longueurs d'onde d'un spectre $\mathbf{x}_{ref}$ de référence, qui est supposé exempt de toute variation chimique. Ce spectre de référence est généralement le spectre moyen $\overline{\mathbf{x}}$ de la base d'étalonnage.

$$\mathbf{d} = a + m\mathbf{x}_{ref} + \mathbf{e}_{cal}$$

avec a l'ordonnée à l'origine, m la pente et $\mathbf{e}_{cal}$ les résidus. Le spectre corrigé par MSC est par conséquent :

$$\mathbf{x}_{new} = \frac{(\mathbf{x} - a)}{m}$$

La MSC réduit l'effet de la réflexion spéculaire et rend les données plus linéaires. Cette méthode donne de bons résultats si le spectre de référence calculé pour déterminer les valeurs de l'ordonnée à l'origine et de la pente ne contient réellement pas de variation du composant d'intérêt. D'une façon générale le spectre de référence $\mathbf{x}_{ref}$ utilisé est le spectre moyen des données d'étalonnage $\overline{\mathbf{x}}_{cal}$, ce qui n'est pas le spectre

idéal pour corriger la pente et l'intercept. Par conséquent, la MSC risque d'enlever de $\mathbf{x}$ l'information corrélée avec $\mathbf{y}$ et donc être nocive au modèle de prédiction. Les méthodes "extended multiplicative scatter correction" (EMSC) et "spectral interference substraction" (SIS) sont deux méthodes de prétraitements utilisant les connaissances *a priori*, comme les spectres purs des composés et les effets des interférences pour calculer les facteurs de correction (a et m) [96]. EMSC est conçue pour améliorer la séparation entre la lumière due à la diffusion et celle due à l'absorbance ; SIS élimine les interférences dont les effets sur les spectres sont connus. Ces deux méthodes nécessitent une connaissance *a priori* des spectres purs (EMSC) et des effets des interférences (SIS). Plus de détails sur ces deux méthodes se trouvent dans [96]. Finalement, on peut dire que la MSC, EMSC et la SIS corrigent le spectre de diffusion.

3.3.1.4 Lissage

Le lissage est une méthode de prétraitement qui essaye de réduire le bruit représenté par les changements aléatoires de l'amplitude des points de mesure dans le signal. La méthode la plus simple consiste à utiliser une fenêtre glissante qui remplace le point du centre de la fenêtre par la moyenne m de la fenêtre.
L'algorithme le plus utilisé pour le lissage est l'algorithme Savitzky et Golay (SG) [56] ; la largeur de la fenêtre de lissage doit être choisie avec précaution. Plus cette fenêtre est large, plus grande est la réduction du bruit, mais également plus grande est la possibilité de la déformation du signal. Pour l'analyse quantitative, la déformation des pics est de moindre importance, parce que les mêmes opérations de prétraitement sont faites pour les échantillons d'étalonnage et de test. Rutledge et al.[125] ont appliqué l'algorithme de SG pour lisser les vecteurs propres obtenus à chaque étape de l'algorithme PLS-NIPALS [144]. Les capacités prédictives des modèles ont été améliorées en utilisant cette technique pour des niveaux de bruit de l'ordre de 10 à 20%. Pour optimiser le degré du polynôme utilisé pour le lissage, Barak [10] a présenté "'le filtre polynômial au degré adaptatif ou Adaptative Degree Polynomial Filter (ADPF)"' qui peut être considéré comme une extension du filtre de SG pour lequel le degré du polynôme n'est pas fixé *a priori*. Le filtre numérique adapte le degré du polynôme convenable au fur

et à mesure de la progression de la fenêtre le long du spectre. L'information disponible dans une fenêtre est employée pour déterminer le degré du polynôme qui devrait être appliqué dans la prochaine fenêtre. Cette technique utilisée pour lisser des données réduit la nécessité d'indiquer le degré du polynôme a priori pour la différenciation et améliore par la suite la réduction du bruit.

3.3.1.5 Différenciation

Dans les applications spectroscopiques, la différenciation est largement appliquée pour augmenter la résolution spectrale et pour corriger le bruit de fond [56], [10], [57]. La dérivée première permet de corriger les effets additifs. L'algorithme le plus appliqué pour la dérivation est l'algorithme de SG. Il utilise une fenêtre mobile de largeur prédéfinie dans laquelle les données sont ajustées par un polynôme de degré prédéfini [57]. L'algorithme SG utilise une fonction de convolution et par la suite le nombre de points utilisés pour définir la largeur de la fenêtre doit être correctement précisé pour s'assurer que la dérivée représente le comportement local du spectre. Des approches comprenant l'algorithme de SG [57] et la Transformation de Fourier (FT) sont employées pour le lissage et la différenciation des données [114]. Les filtres de lissage corrigent généralement les aspects très étroits (bruit), alors que les filtres de différenciation tendent à enlever à la fois les aspects très larges et très étroits [135]. La différenciation ajoute du bruit si le signal de base était déjà bruité. De même, quand des spectres dérivés (particulièrement ceux de la dérivée première) sont employés pour l'étalonnage, les vecteurs propres ne peuvent pas être facilement interprétés. En outre, la dérivée rend l'interprétation visuelle du spectre résiduel plus difficile, et localise de ce fait les absorptivités spectrales des impuretés dans les individus. .

3.3.1.6 Conclusion : prétraitements géométriques et robustesse

Étudions comment les méthodes géométriques influent sur l'équation (3.1) qui exprime la non robustesse.

- **Effet sur $\|\delta\mathbf{x}\|$:**

  Supposons qu'un spectre de perturbation $\delta\mathbf{x}$ est constitué de la somme de deux

spectres : $\delta\mathbf{x}_1$ (le spectre structuré, p.e. spectre de diffusion, facteurs d'influence) et $\delta\mathbf{x}_2$ (spectre du bruit). $\delta\mathbf{x}_1$ et $\delta\mathbf{x}_2$ sont deux spectres orthogonaux dans l'espace $\mathbb{R}^p$ d'où,

$$\|\delta\mathbf{x}\| \quad = \quad \|\delta\mathbf{x}_1\| + \|\delta\mathbf{x}_2\| \tag{3.3}$$

Les méthodes de normalisation et de différenciation sont reliées à $\|\delta\mathbf{x}_1\|$, alors que le lissage est relié à $\|\delta\mathbf{x}_2\|$ comme suit :

1. **Réduction de $\|\delta\mathbf{x}_1\|$:**
   - La partie structurée de l'erreur $\delta\mathbf{x}_1$ représente l'effet des facteurs d'influence. Les méthodes de prétraitements éliminent une partie de $\delta\mathbf{x}_1$: décalage de la ligne de base (SNV, RNV, dérivée première), une rotation de la ligne de base (MSC, dérivation) et des variations quadratiques de la ligne de base (DT). EMSC est capable d'enlever des spectres plus complexes.
2. **Reduction de $\|\delta\mathbf{x}_2\|$:**
   - $\delta\mathbf{x}_2$ peut être à la base de l'hétéroscédasticité des données. L'héteroscédascticité résulte du fait que la variance tout le long du spectre n'est pas la même. Les méthodes de normalisation suppose cette variance identique pour toute les longueurs d'onde, ce qui n'est pas toujours le cas car le terme de l'erreur $\delta\mathbf{x}_2$ peut varier pour chaque longueur d'onde. L'hétéroscedasticité est donc une violation de cette hypothèse [26]. Ce qui en fait un facteur d'influence de la robustesse des modèles d'étalonnage.
   - $\delta\mathbf{x}_2$ est réduit par le lissage, par exemple en ajustant les absorbances à une fonction polynômiale.

- **Réduction de $\|\mathbf{b}\|$:**

  Le lissage contribue aussi à réduire $\|\mathbf{b}\|$ comme il a été mentionné dans [125].

3.3.2 Réduction des dimensions

Ces méthodes consistent à réduire les dimensions de l'espace des prédicteurs. Leur but est de trouver le sous espace contenant principalement les variations reliées à y. Ces méthodes regroupent : les méthodes de projection orthogonale et les méthodes de sélection de variables.

3.3.2.1 Projection orthogonale

Ces méthodes divisent l'espace des p variables spectrales $\overrightarrow{\mathcal{S}}$ en trois sous-espaces orthogonaux : (i) $\overrightarrow{\mathcal{C}}$ contenant principalement des effets dûs aux variations de y, (ii) $\overrightarrow{\mathcal{N}}$ contenant principalement des effets dûs aux variations systématiques structurées et (iii) $\overrightarrow{\mathcal{E}}$ contenant principalement des effets dûs aux variations du bruit aléatoire, tel que :

$$\overrightarrow{\mathcal{S}} = \overrightarrow{\mathcal{C}} \oplus \overrightarrow{\mathcal{N}} \oplus \overrightarrow{\mathcal{E}}$$

Le but est d'identifier une base orthonormale $\mathbf{P}^-$ représentant le maximum des variations systématiques $\overrightarrow{\mathcal{N}}$, et de projeter $\mathbf{X}$ dans le sous-espace orthogonal à $\mathbf{P}^-$, pour donner les spectres corrigés $\mathbf{X}^*$:

$$\mathbf{X}^* = \mathbf{X}\left[\mathbf{I}_p - \mathbf{P}^-\mathbf{P}^{-T}\right]$$

Ainsi l'identification de $\mathbf{P}^-$ est la clé pour effectuer la correction orthogonale.
Les n spectres de $\mathbf{X}$ constituent un nuage de points dans $\overrightarrow{\mathcal{S}}$. Ce nuage délimite un sous-espace de $\overrightarrow{\mathcal{S}}$, qui peut être divisé à son tour en différents sous-espaces de $\overrightarrow{\mathcal{C}}$, $\overrightarrow{\mathcal{N}}$ et $\overrightarrow{\mathcal{E}}$. Du point de vue matriciel, l'equation suivante est obtenue :

$$\mathbf{X} = \mathbf{X}^+ + \mathbf{X}^- + \mathbf{E} \tag{3.4}$$

$$\mathbf{X} = \mathbf{X}^- + \mathbf{R} \text{où } \mathbf{R} \text{ la matrice des résidus contient la majeure partie de } \mathbf{X}^+ + \mathbf{E}.$$

avec $Col(\mathbf{X}^+) \subset \overrightarrow{\mathcal{C}}$, $Col(\mathbf{X}^-) \subset \overrightarrow{\mathcal{N}}$ et $Col(\mathbf{E}) \subset \overrightarrow{\mathcal{E}}$.

$\mathbf{X}^+$ est la partie "*utile*" de $\mathbf{X}$ reliée à $\overrightarrow{\mathcal{C}}$, $\mathbf{X}^-$ est la partie "*inutile*" de $\mathbf{X}$ reliée à $\overrightarrow{\mathcal{N}}$ et $\mathbf{E}$ constitue les résidus reliés à $\overrightarrow{\mathcal{E}}$. Ainsi, $\mathbf{X}^*$ contient $\mathbf{X}^+$ avec une partie de $\mathbf{E}$. La décomposition de $\mathbf{X}$ peut être effectuée de différentes manières :

Identifier $\mathbf{P}^-$ directement à partir de $\mathbf{X}$: C'est le cas de la méthode **Orthogonal Signal Correction (OSC)** [176] developpée par Wold et al. qui commence par effectuer une ACP sur $\mathbf{X}$ pour déterminer les structures latentes correspondantes :

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

Ensuite, la matrice $\mathbf{T}$ est orthogonalisée par rapport à $\mathbf{y}$:

$$\mathbf{T}^- = \left[\mathbf{I}_n - \mathbf{y}\left(\mathbf{y}^T\mathbf{y}\right)^{-1}\mathbf{y}^T\right]\mathbf{T}$$

La matrice des structures latentes $\mathbf{P}^-$ correspondant à $\mathbf{T}^-$ est calculée en utilisant une PLS ou PCR :

$$\mathbf{P}^- = \left(\mathbf{T}^{-T}\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{X}$$

Identifier $\mathbf{P}^-$ à partir de $\mathbf{X}^-$: A partir de l'équation 3.4 on peut écrire :

$$\mathbf{X} = \mathbf{t}_1^-\mathbf{p}_1^{-T} + \mathbf{t}_2^-\mathbf{p}_2^{-T} + ... + \mathbf{t}_k^-\mathbf{p}_k^{-T} + \mathbf{R}$$
$$\mathbf{X} = \mathbf{T}^-\mathbf{P}^{-T} + \mathbf{R}$$

Cette identification de $\mathbf{X}^-$ peut être réalisée de différentes manières :

- ***Intrinsèquement*** : déduire $\mathbf{X}^-$ de $\mathbf{X}$.
- ***Extrinsèquement*** : déterminer $\mathbf{X}^-$ en utilisant un ensemble d'expériences qui

y sont dédiées, pour incorporer les variations systématiques.

A. Méthodes intrinsèques.

Ces méthodes visent à identifier $\mathbf{X}^{-}$, soit dans les structures latentes de $\mathbf{X}$ (approche indirecte), ou directement dans l'espace de prédiction contenant les variations communes de $\mathbf{X}$ et $\mathbf{y}$ (approche directe).

- Approche **indirecte** : Differents algorithmes de l'OSC ont été développés relativement à la version originale donnée par Wold et al. [176]. Ils diffèrent par la façon d'identifier $\mathbf{P}^{-}$, lors du calcul des composantes OSC. L'approche indirecte consiste à orthogonaliser $\mathbf{X}$ par rapport à $\mathbf{y}$ en utilisant la régression PLS ou PCR pour déterminer indirectement $\mathbf{X}^{-}$ tel que :

$$\mathbf{X}^{-} = \left[\mathbf{I}_n - \mathbf{y}\left(\mathbf{y}^T\mathbf{y}\right)^{-1}\mathbf{y}^T\right]\mathbf{X}$$

  Cette stratégie est utilisée dans la méthode de correction du signal par projection orthogonale ou Projected Orthogonal Signal Correction (POSC) proposée par [147], et dans la méthode de correction du signal par projection orthogonale directe ou Direct Orthogonal Signal Correction (DOSC) donnée par [174].
  Westerhuis et al. [174] a comparé cinq différents algorithmes OSC, parmi lesquels la DOSC qu'il propose, appliqués à des spectres PIR pour étudier l'humidité dans les maïs et la viscosité des fuels. Ils ont conclu qu'aucun de ces algorithmes n'est capable de réduire l'erreur de prédiction comparée à celle donnée par la PLS même si le nombre de variables latentes a été considérablement réduit. Les méthodes OSC aident à trouver le même sous-espace que celui trouvé par la PLS, mais en utilisant moins de variables latentes. Seule l'interprétabilité du modèle est améliorée par rapport à la PLS. Récemment, Azzouz et al. [9] ont trouvé que la capacité prédictive de l'OSC n'était pas optimale en utilisant l'OSC pour la prédiction des protéines. Comme il a été démontré dans [146] le désavantage de ces méthodes réside dans le risque de surestimation du nombre des composantes de l'OSC.
- Approche directe : pour éviter ce problème de surestimation, l'approche **directe**

a été adoptée. Il s'agit de trouver le sous-espace commun à $\mathbf{X}$ et $\mathbf{y}$ relatif à $\mathbf{X}^+$ tel que :

$$\mathbf{w}_{Xy} = \frac{\mathbf{y}^T\mathbf{X}}{\|\mathbf{y}^T\mathbf{X}\|}$$

Ensuite, une ACP est appliquée sur $\mathbf{w}_{Xy}$ pour déterminer la base orthonormée $\mathbf{P}^+$. Dans le cas où $\mathbf{y}$ est un seul vecteur, l'espace commun à $\mathbf{X}$ et $\mathbf{y}$ est de dimension unitaire, et la base orthonormée est représentée par le vecteur $\mathbf{w}$. D'où, $\mathbf{X}^-$ est calculée comme suit :

$$\mathbf{X}^- = \mathbf{X}\left[\mathbf{I}_p - \mathbf{w}^T\mathbf{w}\right]$$

Cette approche directe, la plus appropriée pour la régression, a été utilisée par Fearn et al. [46] et aussi par par Trygg et al. [147] pour développer la projection orthogonale sur les structures latentes ou "Orthogonal Projection to Latent Structures (OPLS)". C'est une modification de l'algorithme original NIPALS. OPLS consiste à éliminer les variations structurées de l'espace des premières composantes de la PLS qui contiennent le maximum de covariance entre $\mathbf{X}$ and $\mathbf{y}$. Ces variations sont celles qui affectent les prédictions de la PLS et qui sont à la base des problèmes de surestimation. Trygg dans [147] a appliqué cette méthode pour étudier les spectre PIR du bois.

Ces méthodes d'orthogonalisation possèdent quelques paramètres à régler tels que : le nombre de composantes OSC à choisir et le nombre de variables latentes pour la PLS. Toutefois ces méthodes intrinsèques de correction par projection orthogonale améliorent l'interprétabilité du modèle en réduisant sa complexité.

B. Méthodes extrinsèques :

Ces méthodes nécessitent d'avoir une matrice spéciale $\mathbf{Z}^-$, renfermant des spectres pris à différents niveaux des facteurs d'influences connus et contrôlés. $\mathbf{P}^-$ est identifiée par ACP sur $\mathbf{Z}^-$. Deux méthodes existent :

- **Réduction indépendante des interférences (IIR).** Dans [63], Hansen présente la méthode indépendante de réduction des interférences ou "**independent interference reduction**" (IIR). Elle élimine la majeure partie des effets des interférences et des dispersions du signal, avant l'étalonnage. La matrice $\mathbf{Z}^-$ est construite avec des échantillons dépourvus du composé d'intérêt et montrant des variations dans les facteurs d'influences. Les vecteurs propres sont divisés en ceux qui sont reliés aux interférences et ceux qui sont reliés aux composés d'intérêt. IIR réduit le nombre d'analyses de référence $\mathbf{y}$ exigées pour établir le modèle d'étalonnage. Le modèle résultant est plus simple à interpréter. Son inconvénient principal est qu'elle exige l'utilisation de deux jeux de données qui ne sont pas toujours disponibles.
- **Orthogonalisation par rapport aux paramètres externes (EPO).** L'EPO-PLS, développé par Roger et al. [120], utilise un ensemble d'échantillons mesurés à différents niveaux du facteur d'influence étudié ; Le spectre moyen à chaque niveau est calculé pour constituer la matrice $\mathbf{Z}^-$.

  L'EPO contribue à la réduction des effets des variations des facteurs d'influences externes prédéfinis ou bien une combinaison d'eux, en utilisant un petit ensemble des échantillons appropriés mesurés à différents niveaux des paramètres externes (facteurs d'influences). L'avantage de cette méthode est qu'elle ne nécessite pas de mesures de référence y. Le principe de l'EPO a été aussi appliquée pour la standardisation du modèle d'étalonnage entre les instruments [6]. Mais, seuls les facteurs d'influence connus sont considérés ; alors que les facteurs d'influence inconnus ne peuvent pas être pris en compte par cette méthode.

Conclusion : orthogonalisation et robustesse

Les méthodes d'orthogonalisation éliminent les $\delta\mathbf{x}$ dus aux variables concernées par la procédure d'orthogonalisation. Les méthodes intrinsèques améliorent le modèle PLS si des exemples de $\delta\mathbf{x}$ sont inclus dans la base d'étalonnage. Pour les méthodes extrinsèques, c'est surtout $|\cos(\delta\mathbf{x}, \mathbf{b})|$ qui est considérablement réduit dans le cas où $\delta\mathbf{x}$ est relié aux facteurs d'influence. De plus, ces méthodes réduisent la complexité

du modèle amenant à réduire $\|\mathbf{b}\|$. Par conséquent, les méthodes d'orthogonalisation tendent à améliorer la robustesse du modèle d'étalonnage.

3.3.2.2 Sélection de variables

Au lieu d'enlever les interférences modélisées en tant que spectres, les techniques de sélection de variables consistent à extraire les variables les plus liées aux composés recherchés. Ces méthodes ont été largement étudiées pour améliorer les modèles d'étalonnage multivariés. D'une façon générale, l'objectif principal du choix des variables est d'identifier un sous-ensemble de longueurs d'onde qui produit le moins possible d'erreur de prédiction de $\mathbf{y}$. Selon Rimbaud et al. [82] les méthodes de PLS et de PCR se comportent mieux après sélection de longueur d'onde. Ceci n'est pas toujours le cas car en sélectionnant les variables les plus corrélées, on peut éliminer celles qui corrigent les facteurs d'influence. En fait, Guyon et Elisseff [62] ont montré qu'une variable peut être parfaitement inutile en elle même mais utile lorsqu'elle est utilisée avec d'autres. Dans leur travaux sur la sélection de variables dans le domaine du *machine learning*, les auteurs ont groupé les méthodes de sélection de variables suivant trois stratégies différentes :

- les **filtres** qui choisissent les variables selon une fonction prédéfinie φ qui relie $\mathbf{X}$ à $\mathbf{y}$. Cette fonction est employée pour mesurer les variables d'intérêt j de la matrice $\mathbf{X}$, qui sont liées à $\mathbf{y}$. Ainsi, des variables j donnant la valeur optimale de $\varphi(\mathbf{x}_j, \mathbf{y})$ sont choisies.
- les méthodes **embarquées** qui effectuent la sélection des variables pendant l'étape d'apprentissage du modèle, afin de trouver la solution optimale. Elles diffèrent selon le modèle d'apprentissage utilisé.
- les méthodes dites **wrappers** qui utilisent le modèle d'étalonnage comme une boîte noire pour sélectionner des sous-ensembles de variables selon leurs capacités prédictives.

Méthodes de Filtres

Elles consistent à appliquer un seuil à la valeur d'une fonction objective φ qui relie

$\mathbf{X}$ à $\mathbf{y}$. En matière d'étalonnage multivarié, différentes modalités de φ sont trouvées dans la littérature :

- φ est indépendante de $\mathbf{y}$: Un exemple de cette méthode est basé sur $\varphi(\mathbf{x_j})$ étant le j^{ieme} poids des premières composantes d'une ACP appliquée sur $\mathbf{X}$. Les longueurs d'onde relatives à des vecteurs propres de valeur élevée sont retenues sur la base d'un ensemble de composantes choisies de l'ACP. Les longueurs d'onde choisies sont celles qui contribuent le plus à l'explication de la variance pour une composante principale donnée ([82]).
- φ dépend de $\mathbf{y}$: Le filtrage considère la relation entre la variable j et la réponse $\mathbf{y}$. $\varphi(\mathbf{x}_j, \mathbf{y})$ peut prendre les formes suivantes :

1. **le coefficient de corrélation** :

   $\varphi(\mathbf{x}_j, \mathbf{y}) = \mathbf{r}^2(\mathbf{x}_j, \mathbf{y})$.

   Les longueurs d'onde fortement corrélées avec $\mathbf{y}$ sont sélectionnées. Ce critère peut être également appliqué dans le cas de la sélection ou de l'élimination des variables ([96]). Une autre manière d'employer la fonction de corrélation est de choisir, au lieu de différentes variables, des intervalles qui ont une corrélation élevée avec $\mathbf{y}$; le but principal est de trouver le meilleur intervalle pour construire le modèle ([82]).

2. **la covariance** : $\varphi(\mathbf{x}_j, \mathbf{y}) = |Cov(\mathbf{x}_j, \mathbf{y})|$. Cette méthode choisit des longueurs d'onde ayant la plus grande valeur absolue de la covariance entre $\mathbf{x}_j$ et $\mathbf{y}$ [82].

3. le **rapport signal sur bruit SNR** : $\varphi(\mathbf{x}, \mathbf{y}) = \mathbf{SNR}$.

   Dans [99], Mc Shane et al. ont proposé de choisir les longueurs d'onde pour les employer dans la PLS tout en considérant les régions avec la plus grande variance spectrale. McShane et al. [98] ont défini le rapport du signal sur bruit comme suit :

$$\varphi(\mathbf{x}_j, \mathbf{y}) \quad = \quad SNR = \frac{\mathbf{b}_j}{\sigma(\mathbf{x}_j)}$$

   Le problème de la colinéarité est traité en employant la méthode de "la chaîne à rang multiple" ou " multiple ranking chain method" qui consiste à ranger les variables par ordre décroissant du SNR. La variable avec le rang le plus élevé est

employée dans la régression pour produire les spectres estimés et les spectres des résidus sont utilisés pour la régression et le calcul du second rang du SNR. Ce processus continue jusqu'à ce qu'un nombre prédéfini de chaînes soit produit. Le SNR mentionné ci-dessus n'est pas employé pour des données avec une grande quantité de données aberrantes. En effet, dans ce cas, $\sigma(\mathbf{x})$ tend à sous-estimer la véritable contribution du bruit, et par la suite à introduire du biais dans l'évaluation des coefficients de régression $\mathbf{b}$. Comme le bruit est supposé mieux modélisé à l'aide d'une distribution log-normale ou exponentielle que gaussienne, il est plus approprié d'employer les résidus qu'un écart type. Spiegelman et al. dans [136], ont proposé une méthode pour améliorer la sélection en rangeant les variables selon un SNR modifié qui tient compte de la quatrième puissance des résidus spectraux de la PLS et ceci à chaque longueur d'onde j au lieu de l'écart type. À chaque étape, la variable avec le rang le plus élevé est ajoutée au modèle et le $SECV$ est calculé ; l'ensemble des variables correpondant au $SECV$ minimum est retenu.

4. le coefficient de **fiabilité** c_j de b_j. Au lieu de choisir les variables d'intérêt, Centner et al. [28] ont proposé une nouvelle méthode (UVE) pour éliminer les variables non informatives (UVE). Les variables UVE sont celles qui ne contiennent pas plus d'information que les variables "aléatoires" (c.-à-d. le bruit). Centner a proposé le coefficient de fiabilité c pour chaque variable j ; calculé en utilisant le rapport entre la moyenne des j modèles obtenus par Jackknife et leur écart-type, tel que :

$$\varphi(\mathbf{x}_j, \mathbf{y}) = c_j \quad = \quad \frac{\text{Mean}(\mathbf{b}_j)}{\widehat{\sigma}(\mathbf{b}_j)}$$

où $\widehat{\sigma}(b_j)$ est une estimation de l'erreur standard des coefficients b_j, obtenu par jackknifing. Elle est basée sur une analogie avec la MLR pas à pas. La j^{ieme} variable à ajouter à la MLR doit avoir $c_j \geq c_{bruit}$, où c_{bruit} est obtenu avec des variables aléatoires artificielles (bruit) ajoutées aux données. Ce seuil est indicatif des valeurs du coefficient de fiabilité qui peuvent être atteintes par des variables

non informatives. L'erreur aléatoire produisant le seuil, ne devrait pas être trop grande pour qu'elle n'influence pas le modèle. Ce niveau devrait refléter le cas pratique qui pourrait être trouvé en réalité.

Cette méthode basée sur le coefficient de fiabilité améliore de manière significative la prédiction quand les données contiennent beaucoup de variables non informatives. Elle peut être considérée comme un procédé général de pré-sélection pour éviter des problèmes dans l'application de la MLR. Une alternative plus robuste (au sens statistiques) à la moyenne et l'écart-type est d'utiliser le rapport de la valeur médiane des coefficients de régression sur leur intervalle interquartile est également utilisé. Des variantes de c_j peuvent être trouvées dans [28]. L'UVE a été développée pour l'application de la PLS, mais peut être également utile pour la PCR ou les méthodes similaires.

Méthodes embarquées ou "embedded".

Les méthodes embarquées sont basées sur l'approche déterministe. Elles incorporent la sélection de variables comme étant une partie de l'étape d'apprentissage. Les méthodes de sélection pas à pas.

Ces méthodes de sélection commencent par ranger les longueurs d'onde selon le changement qu'elles apportent à une fonction particulière (fonction objective). Puis, les longueurs d'onde sont successivement ajoutées au modèle ou éliminées du modèle selon un seuil spécifique d'entrée/sortie ([53]). Ces méthodes présentent un problème de colinéarité quand des points voisins sont ajoutés consécutivement ([95]). Jouan-Rimbaud et al. [82] ont recommandé l'utilisation de la sélection pas à pas des longueurs d'onde ou bien de bandes des longueurs d'ondes pour fournir une meilleure stabilité du modèle, suivie de la régression linéaire multiple (MLR). L'efficacité du choix par étapes dépend du rang des variables qui les rend souvent sensibles aux distributions du bruit.

Méthodes enveloppées ou "Wrappers"

Ces méthodes emploient le modèle comme boîte noire et guident la sélection en utilisant ses performances. Différentes stratégies de sélection ont été développées basées

sur les approches probabilistes comme la méthode du recuit simulé (SA) ([67]) et les algorithmes génétiques (GA) ([90]). SA et GA sont deux techniques probabilistes d'optimisation bien connues pour éviter les solutions optimales locales. Dans ([142, 140]) Swierenga et al. ont utilisé l'algorithme SA pour la sélection de variables afin d'améliorer la transférabilité des modèles d'étalonnage ; cette technique recherche une solution optimale globale et utilise divers paramètres à optimiser. Cette méthode a montré des résultats semblables une fois appliquée à une base d'étalonnage à température standard et puis à un ensemble d'échantillons contenant des variations de température. Les algorithmes génétiques sont devenus assez connus et ont été appliquées pour des problèmes de sélection de variable par Roger et al. [119] ou des composantes principales par [12]. Mais l'algorithme GA présente des inconvénients, il représente un défi énorme de configuration pour l'utilisateur, à cause de nombreux facteurs à ajuster et qui affectent la sortie de l'algorithme : fonction de fitness, critère de convergence, la fréquence de mutation, le type de crossing over, le nombre des chromosomes considérés, la population initiale, et le nombre de générations ; Tous ces facteurs influencent les résultats. D'où, une sélection judicieuse de tous ces paramètres qui reste critique [97]. En particulier, ils exigent un niveau considérable d'attention de la part de l'utilisateur pour éviter les problèmes de sur ou de sous-estimation. De plus, les algorithmes d'optimisation tels que GA, SA sont lourds et gourmands en temps de calcul, comparés aux méthodes pas à pas, plus simples.

Conclusion : sélection de variable et la robustesse

Les méthodes de sélection de variables peuvent affecter la robustesse des prédictions du modèle d'étalonnage car :

1. quand le nombre des variables sélectionnées augmente, les chances de capter les variations des facteurs d'influence augmentent et par suite la valeur de $\|\delta\mathbf{x}\|$ augmente.
2. $\|\delta\mathbf{x}\|$ et $\|\mathbf{b}\|$ sont des sommes quadratiques, donc ils sont très sensibles à l'addition de termes même faibles.
3. quand des variables moins utiles pour le modèle sont sélectionnées, l'orthogonalité

de $\mathbf{b}$ et de $\delta\mathbf{x}$ est plus difficile à obtenir, ce qui amène une augmentation de la valeur de $|\cos(\delta\mathbf{x}, \mathbf{b})|$.

Donc les méthodes de sélection apparaissent favorables à la robustesse. Cependant, sélectionner les variables les plus corrélées à $\mathbf{y}$ n'assure pas que $\|\delta\mathbf{x}\|$ soit minimale car il peut y avoir quelques variations dans $\delta\mathbf{x}$ qui soient aussi reliées à $\mathbf{y}$ et qui risquent de contribuer à augmenter $\|\delta\widehat{y}\|$ de l'équation 3.1. La méthode UVE qui identifie les variables responsables de $\delta\mathbf{x}$ afin de les enlever réduit considérablement $\|\delta\mathbf{x}\|$ et par suite, contribue à l'amélioration de la robustesse du modèle.

3.4 Conclusion

Dans ce chapitre, les méthodes utilisées pour améliorer le modèle d'étalonnage multivarié ont été discutées, telles que les méthodes d'optimisation de la base d'étalonnage et les méthodes de prétraitement.

Alors que les algorithmes de sélection de la base d'étalonnage sont généralement appliqués sur les spectres, nous avons proposé de les appliquer sur les valeurs de référence $\mathbf{y}$ pour assurer le maximum de représentativité du domaine expérimental. Nous avons également expliqué l'effet néfaste du centrage sur la robustesse du modèle, surtout en ce qui concerne la sortie de gamme d'étalonnage.

L'effet des méthodes de prétraitement sur la robustesse du modèle linéaire d'étalonnage multivariés a été largement étudié. L'équation $|\delta\widehat{y}| = \|\delta\mathbf{x}\|\|\mathbf{b}\|\,|cos(\delta\mathbf{x}, \mathbf{b})|$ a été proposée comme base pour expliciter les causes des problèmes de robustesse.

Il a été conclu que les méthodes de prétraitements géométriques des spectres (SNV, RNV, De-trend, MSC, Lissage et Dérivation), corrigeant le décalage de la ligne de base, la curvilinéarité et les termes reliés aux bruits, réduisent en particulier le premier terme $\|\delta\mathbf{x}\|$. Il a été montré aussi que le lissage réduit le second terme $\|\mathbf{b}\|$, lui donnant un avantage par rapport aux autres méthodes de prétraitements géométriques. Mais,la largeur de la fenêtre de lissage doit être bien choisie.

Les méthodes de projection orthogonale (OSC, DOSC, OPLS, IIR, EPO) sont reliées plutôt au troisième terme, $|\cos(\delta\mathbf{x}, \mathbf{b})|$, qui représente l'angle entre les variations sys-

tématiques et le modèle. Ces méthodes visent à réduire $|\cos(\delta\mathbf{x}, \mathbf{b})|$, afin d'assurer une indépendance relative du modèle de $\delta\mathbf{x}$.

Les méthodes de sélection de variables (Wrappers, Filters, Embedded) influencent la robustesse via $\|\delta\mathbf{x}\|$ et $|\cos(\delta\mathbf{x}, \mathbf{b})|$. $\|\delta\mathbf{x}\|$ peut augmenter si les longueurs d'onde sélectionnées sont inutiles. Le terme $|\cos(\delta\mathbf{x}, \mathbf{b})|$ dépend des longueurs d'onde sélectionnées, et si le nombre des longueurs d'onde inutiles augmente, $|\cos(\delta\mathbf{x}, \mathbf{b})|$ augmente. Ainsi, la sélection des variables doit être effectuée avec précaution vis-à-vis de l'amélioration de la robustesse du modèle d'étalonnage.

Cette comparaison de la contribution de ces méthodes à l'amélioration de la robustesse a fait l'objet d'un article [184]. Pour le développement d'une méthode d'étalonnage robuste dans le cas des applications en ligne, nous retiendrons donc et mettrons en oeuvre les principes suivant :

- la sélection de la base d'étalonnage appliqués sur $\mathbf{y}$,
- ne pas utiliser le centrage systématiquement,
- la sélection de variables,
- réduction de la dimension par projection orthogonale

Ce développement fait l'objet du chapitre suivant (chapitre 4).

Chapitre 4

Méthodologie pour maintenir la robustesse des modèles d'étalonnage multivariés pour des applications en ligne

Sommaire

4.1 Introduction

Le problème traité dans ce chapitre correspond au troisième point de l'analyse de la problématique. Il consiste à étudier la maintenance de la robustesse des modèles d'étalonnage multivariés pour des applications industrielles en ligne et à proposer une méthodologie originale pour y parvenir. Le manque de robustesse, qui a été défini dans le chapitre 2, est un problème majeur pour des applications en ligne. Il est principalement dû aux variations de facteurs d'influence, tels que : la température, la complexité des échantillons, les perturbations instrumentales. Ainsi, un modèle d'étalonnage utilisé en ligne doit être régulièrement contrôlé et mis à jour pour améliorer sa robustesse comme indiqué dans les procédures ASTM décrites en introduction.

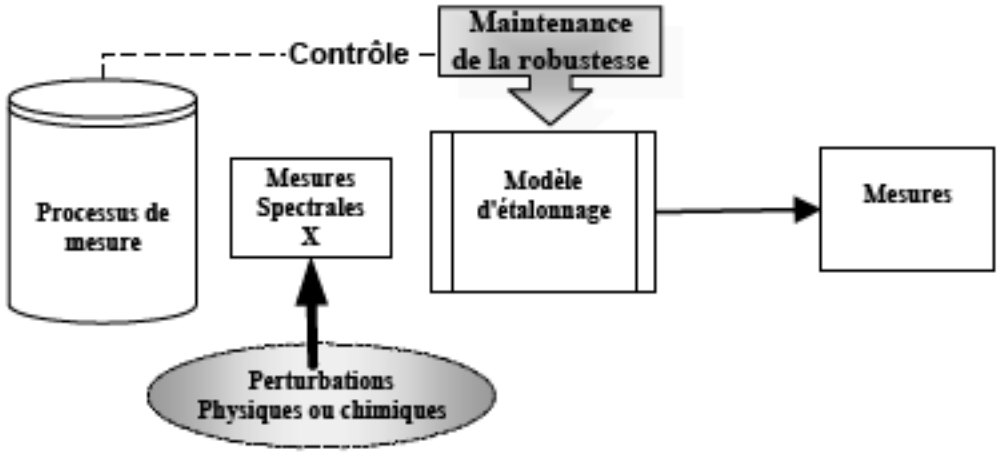

FIG. 4.1 – Maintenance en ligne de la robustesse du modèle d'étalonnage multivarié

Les données disponibles en ligne sont : (i) la base d'étalonnage, acquise dans des conditions contrôlées, p.e. au laboratoire, que l'on veut valoriser ; (ii) les mesures spectrales PIR fréquentes, prises continuellement en ligne pour le suivi du processus et (iii) les quelques mesures de référence en nombre restreint, p.e. une ou deux fois par jour, pour le contrôle du processus.

En ligne, le problème à résoudre et qui a généré la question de recherche, est le suivant :

"à un instant donné, comment utiliser au mieux la base d'étalonnage, les mesures de référence de contrôle passées et les mesures spectrales en ligne correspondantes, pour rendre l'étalonnage robuste pour les prochaines mesures."

Cette robustesse doit être maintenue vis-à-vis de l'apparition ou la disparition des effets durables des facteurs d'influences.

En utilisant les données disponibles en ligne, la correction des prédictions du modèle d'étalonnage peut être effectuée de différentes manières. La plus simple et la plus classique est la correction biais/pente des prédictions [110, 21, 51]. Cette méthode simple présente cependant un inconvénient majeur lorsque la perturbation disparaît et elle donne peu d'information sur les causes d'erreur. Une autre possibilité existe pour corriger les prédictions en ligne : elle consiste à reconstruire le modèle utilisant uniquement les spectres en ligne et les mesures de contôle. Ce sont par exemple les méthodes d'étalonnage local [171, 45]. Les méthodes locales sont simples à appliquer mais elles sont peu informatives. De plus, elles n'utilisent pas la base d'étalonnage initiale. L'ajout à la base d'étalonnage des mesures spectrales et de contôle acquises en ligne permet d'effectuer un étalonnage global ou exhaustif [85]. Les méthodes d'étalonnage global sont simples à réaliser, mais la différence de taille entre la base d'étalonnage très grande et les quelques mesures en ligne les rend difficiles à appliquer. De plus, elles sont peu informatives, car en ajoutant les quelques mesures en ligne on risque parfois de ne rajouter que de la confusion à la base d'étalonnage. Des méthodes de filtrage, comme le Filtre de Kalman (FK), peuvent aussi tirer parti des mesures disponibles en ligne. Celui-ci a été utilisé pour l'estimation en ligne du coefficient du modèle d'étalonnage [123, 145] et pour la correction de la dérive [124, 7]. Il présente cependant deux inconvénients. Le premier a trait au problème d'initialisation des paramètres du filtre, comme la matrice de transition, les matrices de variance et de covariance. Le second réside dans le fait que FK ne prend pas en compte la base d'étalonnage.

Ce chapitre présente une nouvelle méthode, la Projection Orthogonale Dynamique (DOP), qui permet de développer un modèle d'étalonnage pour le suivi en ligne par spectroscopie IR des processus continus, en présence de facteurs d'influences (physiques et chimiques). La méthode DOP utilise uniquement les bases de données qui sont disponibles en ligne.

La théorie de la méthode proposée (DOP) est présentée puis discutée vis-à-vis des principales approches qui peuvent être utilisées ainsi que de son implémentation et ses

paramètres à régler.

4.2 Théorie

4.2.1 Hypothèses

Considérons un processus continu, suivi en ligne par un capteur IR. Soit $\mathbf{X}$ la matrice des mesures spectrales et $\mathbf{y}$ le vecteur des valeurs de la propriété d'intérêt. Soit $\mathbf{X}_{0(n_0 \times p)}, \mathbf{y}_{0(n_0 \times 1)}$ la base d'étalonnage acquise dans les conditions contrôlées (p.e. au laboratoire), normalement utilisée pour le suivi du processus continu. Supposons que quelques mesures de référence sont prises épisodiquement, par exemple pour le contrôle du processus.

A un instant $t = \tau$ donné, posons $\mathbf{y}_{\tau(n_\tau \times 1)}$ les valeurs de référence acquises antérieurement au temps τ. Soit $\mathbf{X}_\tau$ la matrice des spectres acquis en ligne en même temps que ces mesures de référence.

La méthode DOP vise à établir un modèle d'étalonnage, robuste pour les futures prédictions, en tenant compte des effets des facteurs d'influences qui peuvent avoir eu lieu pendant le suivi du processus.

4.2.2 Principe général de DOP

L'idée maîtresse de DOP est basée sur : (i) l'estimation des spectres $\widehat{\mathbf{X}}_\tau$ qui auraient dû être mesurés en l'absence de perturbation ; (ii) le transfert du modèle d'étalonnage des conditions standard vers celles de la ligne, en utilisant les spectres $\mathbf{X}_\tau$ et $\widehat{\mathbf{X}}_\tau$ comme des *standards virtuels*[2].

4.2.2.1 Estimation de $\widehat{\mathbf{X}}_\tau$

L'estimation de $\widehat{\mathbf{X}}_\tau$ est la première originalité de notre approche. L'estimation de $\widehat{\mathbf{X}}_\tau$ doit être effectuée à partir des données disponibles en ligne $(\mathbf{X}_0, \mathbf{y}_0, \mathbf{y}_\tau)$. La méthode

[2]Un standard est un échantillon de produit mesuré dans les conditions d'étalonnage et dans les conditions perturbées et qui permet donc de corriger les perturbations spectrales.

la plus simple consiste à déterminer une combinaison linéaire $\mathbf{A}_{(n_\tau \times n_0)}$ de $\mathbf{y}_0$ qui estime $\mathbf{y}_\tau$ et à l'appliquer à $\mathbf{X}_0$. Le calcul de $\mathbf{A}$ peut être effectué de différentes manières, dont :

1. La méthode des plus proches voisins (kppv) consiste à calculer des distances entre le point à estimer et tous les points de la base d'étalonnage pour choisir les plus proches voisins afin de les utiliser pour l'estimation. La distance euclidienne est le plus souvent employée. Différentes variantes de cette méthode existent [155], par exemple en donnant plus de poids pour les plus proches voisins. Cette méthode est le plus souvent utilisée pour la classification. C'est une méthode simple et elle n'est basée sur aucune hypothèse statistique quant à la normalité de la distribution. Mais le nombre des voisins à choisir est un paramètre à optimiser. Cette méthode se comporte mal si la distribution de la base d'étalonnage présente des trous ou si on se trouve à l'extérieur des limites de la base.

2. Les méthodes à noyaux utilisent une fonction noyau pour estimer une fonction de densité. L'estimateur à noyau ou *kernel* est utilisé pour estimer, à partir d'un ensemble d'observations indépendantes et aléatoires, la valeur définie par $\widehat{y} = \frac{1}{n\varepsilon} \sum_{i=1}^{n} K(\frac{y-y_i}{\varepsilon})$ où K est la fonction de noyau et ε est la largeur du noyau. Différentes fonctions kernel existent : uniforme, triangulaire, gaussienne, cosinus, etc. [19, 70, 131, 66]. Les fonctions de noyau gaussiennes RBF "Radial Basis Functions" sont utilisées dans les méthodes de Support Vector Machine (SVM) pour la classification et pour la régression [139]. Le noyau gaussien K_g est souvent utilisé, c'est une fonction qui tend vers 0 quand la variable y tend vers l'infini. Les estimateurs à noyaux sont des estimateurs biaisés, dont la variance et le biais dépendent de ε. Des méthodes d'optimisation de ε existent [164]. Ces estimateurs, même s'ils sont biaisés, dont la variance et le biais dépendent de la largeur du noyau, présentent de bonnes propriétés de convergence une fois bien optimisés [66]. Un de leurs avantages tient au fait qu'il n'est pas nécessaire d'avoir des suppositions *a priori* sur la distribution et que des décisions probabilistes peuvent être plus facilement prises qu'avec la méthode kppv [155].

4.2.2.2 Transfert de l'étalonnage

La deuxième originalité de l'approche consiste à considérer chaque point de contrôle de la mesure en ligne comme un cas de transfert d'étalonnage entre les conditions de laboratoire et les conditions de la ligne. Le transfert de l'étalonnage a été largement étudié et différentes méthodes existent dans la littérature :

1. Les méthodes de standardisation optique visent à établir une matrice $\mathbf{F}_{(p\times p)}$ qui modélise la transformation entre les spectres perturbés $\mathbf{X}$ et les spectres qui auraient dû être acquis $\widetilde{\mathbf{X}}$ telle que : $\widetilde{\mathbf{X}} = \mathbf{F}\mathbf{X}$. Pour la détermination de $\mathbf{F}$, différentes méthodes ont été développées. Une bonne revue de ces méthodes existe dans [33][21][44]. La *standardisation directe (DS)* [170] qui détermine $\mathbf{F}$ par : $\mathbf{F} = \mathbf{X}^{-}\widetilde{\mathbf{X}}$, où $\mathbf{X}^{-}$ est la pseudo-inverse de $\mathbf{X}$; la méthode de standardisation directe par morceaux *Piecewise Direct Standardization (PDS)* développée par Wang et al. [169], calcule $\mathbf{F}$ en utilisant une fenêtre glissante, de façon à rendre $\mathbf{F}$ *quasi-diagonale*. A citer aussi la standardisation après transformation en ondelettes (wavelet transform) [167]. L'avantage principal des méthodes de standardisation optique tient au fait qu'elles donnent de l'information sur les facteurs d'influence corrigés. Néanmoins, si la perturbation disparaît, la correction agit dans la mauvaise direction et fausse le modèle.
2. Les méthodes d'étalonnage étendu sont basées sur le principe d'incrémentation de la base d'étalonnage en utilisant les spectres des standards. Ainsi, la méthode *repeatability file* développée par Westerhaus et al. [173], consiste à effectuer des répétitions des mesures spectrales des échantillons standard sur différents instruments. Ensuite, les spectres de différence, avec les valeurs de référence correspondantes à zéro, sont ajoutés à la base d'étalonnage. Dans notre cas les spectres de différence $\mathbf{D} = \mathbf{X}_{\tau} - \widehat{\mathbf{X}}_{\tau}$ peuvent être utilisés. Toutefois, pour être efficace, cette méthode nécessite l'utilisation de beaucoup de standards. De plus, cette méthode ne donne pas d'information sur la perturbation ayant eu lieu et donc sur les causes du manque de robustesse du modèle.
3. Les méthodes de projections orthogonales, expliquées au chapitre 3, consistent

à trouver le sous-espace $\overrightarrow{\mathcal{N}}$ de $\mathbb{R}^p$ qui contient la majeure partie des distorsions spectrales puis à projeter $\mathbf{X}_0$ sur l'orthogonal à $\overrightarrow{\mathcal{N}}$ pour donner $\mathbf{X}_0^*$. Parmi ces méthodes il y a la méthode "external parameter orthogonalization" (EPO)[120] développée à l'UMR ITAP (voir 3.3.2.1).

Le principal avantage des méthodes d'orthogonalisation est de rendre les données indépendantes des facteurs d'influences considérés. De plus, la partie des spectres éliminée par projection orthogonale donne de l'information sur les perturbations.

4.2.3 Conclusion

Après l'analyse des avantages et des inconvénients des méthodes qui peuvent être appliquées sur les deux étapes du principe de la méthode DOP proposée, nous avons choisi la fonction de noyau pour l'estimation, à $t = \tau$, des "standards virtuels" à partir de la base d'étalonnage et des mesures de contrôle prises jusqu'à cet instant. Ensuite, le transfert du modèle d'étalonnage des conditions standard vers celles en ligne est effectué en utilisant la projection orthogonale extrinsèque basée sur le principe de l'EPO (voir paragraphe 3.3.2.1). Cette dernière étape utilise la matrice de différence entre les spectres standards ($\widehat{\mathbf{X}}_\tau$) et les spectres en ligne ($\mathbf{X}_\tau$) comme la matrice contenant la majeur partie des variations dûes aux facteurs d'influence.

4.2.4 Implémentation mathématique de DOP

4.2.4.1 Étapes de l'implémentation

En considérant la conclusion du paragraphe précédent, nous avons construit la méthode DOP comme suit :

1. **Les spectres des standards virtuels** sont créés par estimation de $\widehat{\mathbf{X}}_\tau$ comme étant une combinaison linéaire de $\mathbf{X}_0$. Cette combinaison linéaire est donnée par des fonctions de noyaux centrées sur les éléments de $\mathbf{y}_\tau$ et appliquées

sur $\mathbf{y}_0$:

$$\begin{aligned} \widehat{\mathbf{X}}_\tau &= \mathbf{A}\mathbf{X}_0 \\ a_{ij} &= K_{y_{\tau_i}}(y_{0_j}) \quad \text{où } K_{y_{\tau i}} \text{ est une fonction du noyau centrée sur } y_{\tau i} \end{aligned}$$

2. **La matrice de différence D** entre $\widehat{\mathbf{X}}_\tau$ et $\mathbf{X}_\tau$ est calculée :

$$\mathbf{D} = \mathbf{X}_\tau - \widehat{\mathbf{X}}_\tau$$

   D contient l'ensemble des spectres de perturbation.

3. **Une base orthonormée P** de l'espace représenté par **D** est estimée par application d'une ACP :

$$\mathbf{D} = \mathbf{T}\mathbf{P}^T + \mathbf{R} \tag{4.1}$$

   où **T** sont les coordonnées factorielles de **D** et **P** les vecteurs propres correspondants.

   Cette base permet d'expliciter l'espace des perturbations au temps τ.

4. **Les spectres de la base d'étalonnage** sont corrigés par projection sur l'orthogonal à l'espace engendré par **P** :

$$\mathbf{X}_0^* = \mathbf{X}_0 \left(\mathbf{I}_p - \mathbf{P}\mathbf{P}^T\right) \tag{4.2}$$

   Cette base d'étalonnage $\mathbf{X}_0^*$ devient indépendante des variations des effets des facteurs d'influence dans **P**.

5. **Un nouveau modèle d'étalonnage** est établi en utilisant $\mathbf{X}_0^*$ et $\mathbf{y}_0$. Comme la base d'étalonnage initiale est corrigée par projection orthogonale, la correction est intégrée au modèle d'étalonnage. Par conséquent, il n'est pas nécessaire de corriger les nouveaux spectres quand le modèle est utilisé.

4.2.4.2 Réglage des paramètres

Chacune des deux étapes de la méthode DOP dépend d'un paramètre qui doit être réglé :

1. Le premier paramètre de la méthode DOP est la fonction de noyau utilisée pour estimer $\mathbf{y}_\tau$ à partir de $\mathbf{y}_0$. Généralement, un *noyau* sert à estimer la densité d'une loi inconnue d'un échantillon aléatoire d'une population. C'est une fonction réelle, continue, centrée sur $\mathbf{y}_{\tau i}$ et d'intégrale unitaire [66]. Dans notre cas, cette dernière propriété est en fait remplacée par : $\sum_j a_{ij} = 1$. Le noyau gaussien, vérifiant ces hypothèses, est celui adopté pour notre application. Le réglage du noyau doit être effectué en tenant compte de la distribution de $\mathbf{y}_0$.
2. Le second paramètre à régler est la dimension k du sous-espace décrit par $\mathbf{P}$, i.e. le nombre de composantes principales retenues pour l'ACP. Différents tests existent dans la littérature [104] et peuvent être utilisés pour choisir le nombre de composantes k à retenir qui décrivent le maximum de variance de $\mathbf{D}$, p.e. les statistiques de Wilks comme suggéré dans l'application de l'EPO [120], ou le test de Durbin-Watson utilisé par Barros et al. [12]. Le taux de variance expliquée est aussi utilisé avec des valeurs empiriques, comme 90%, 95% ou 99%.

4.2.4.3 Discussion

Les deux principaux paramètres de DOP sont la fonction du noyau et la dimension k de l'ACP.

Pour la dimension k, le taux de variance est souvent utilisé. C'est surtout la largeur ε de la fonction de noyau gaussien qui est le plus critique. L'influence de ce facteur sur les performances de DOP peut être étudié : (i) soit par explication théorique, (ii) soit par calcul du R_C ou du SEP en fonction de ε sur des données expérimentales ou simulées (ce qui sera fait au chapire 5). L'explication théorique se traduit par l'étude de l'erreur commise sur l'utilisation de la fonction de noyau qui dépend en partie de la distribution de $\mathbf{y}_0$. Deux types d'erreur peuvent surgir de l'utilisation de ces fonctions de noyau pour l'estimation des valeurs $\mathbf{y}_\tau$ à partir de $\mathbf{y}_0$.

1. L'erreur d'approximation : c'est le cas où $\mathbf{y}_\tau$ recherchée se trouve en dehors de la gamme couverte par $\mathbf{y}_0$; ainsi la solution est considérée comme étant une approximation de toutes les valeurs de $\mathbf{y}_0$. Ceci donne une large erreur d'approximation. Généralement, dans ce cas, la moyenne de $\mathbf{y}_0$ est choisie.
2. L'erreur d'estimation : c'est le cas où $\mathbf{y}_\tau$ est située à l'intérieur de l'espace de la base d'étalonnage. Cette erreur est dûe à la différence entre la valeur recherchée et celle estimée.

Etant dans le cas d'une base d'étalonnage optimale, représentative du domaine expérimental, c'est surtout la cas de l'erreur d'estimation qui nous concerne en utilisant l'estimateur à noyau. Cette erreur d'estimation dépend surtout de la distribution de la base d'étalonnage qui constitue l'espace d'estimation. D'où l'importance d'étudier l'effet de la distribution de $\mathbf{y}_0$ sur l'erreur d'estimation du noyau. Utilisant une fonction de noyau $N(y_\tau, y_0)$, l'estimation $\widehat{\mathbf{y}}_\tau$, à partir de la distribution $g(y_0)$ de y_0, est réalisée par :

$$\widehat{y}_\tau = \frac{\int_{\mathbb{R}} y_0 g(y_0) N(y_\tau, y_0) d(y_0)}{\int_{\mathbb{R}} g(y_0) N(y_\tau, y_0) d(y_0)} \tag{4.3}$$

Cette estimation est fonction de la distribution $g(y_0)$ de y_0. Son influence sur l'erreur d'estimation est étudiée pour deux types de distribution de $\mathbf{y}_0$: uniforme (figure 4.2-(a1)) et gaussienne (figure 4.2-(b1)). Considérons une fonction $N(y_\tau)$ de noyau gaussien.

Si la distribution de y_0 est gaussienne, la fonction $g(y_0)$ est de la forme :

$$g(y_0) = e^{-\frac{(y0-\mu)^2}{2\sigma^2}} \qquad \text{et} \qquad N(y_\tau, y_0) = e^{-\frac{(y0-y_\tau)^2}{2\varepsilon^2\sigma^2}} \tag{4.4}$$

En remplaçant les termes $g(y_0)$ et $N(y_\tau, y_0)$ dans l'équation 4.3, l'estimation $\widehat{\mathbf{y}}_\tau$ est alors de la forme :

$$(4.3) = \frac{\int_{-\infty}^{+\infty} y_0 e^{-\frac{(y_0-\mu)^2}{2\sigma^2}} e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0)}{\int_{-\infty}^{+\infty} e^{-\frac{(y_0-\mu)^2}{2\sigma^2}} e^{-\frac{(y_0-y_1)^2}{2\varepsilon^2\sigma^2}} d(y_0)} = \frac{\mathrm{I}(y_\tau)}{\mathrm{S}(y_\tau)} \tag{4.5}$$

$$\Leftrightarrow \mathrm{I}(y_\tau) = \int_{-\infty}^{+\infty} (y_0 - \mu)\, e^{-\frac{(y_0-\mu)^2}{2\sigma^2}} e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0) + \mu \mathrm{S}(y_\tau)$$

Soit

$$\mathrm{I}(y_\tau) \quad = \quad \mathrm{J}(y_\tau) + \mu S(y_\tau)$$

En appliquant l'intégration par partie sur $\mathrm{J}(y_\tau)$, on pose :

$$\begin{aligned} u &= -\sigma^2 e^{\frac{-(y_0-\mu)^2}{2\sigma^2}} \\ v &= e^{\frac{-(y_0-\mu)^2}{2\varepsilon^2\sigma^2}} \end{aligned}$$

D'où,

$$\mathrm{J}(y_\tau) \quad = \quad [uv]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} u dv$$

$$\Leftrightarrow \mathrm{J}(y_\tau) = \left[-\sigma^2 e^{-\frac{(y_0-\mu)^2}{2\sigma^2}} e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} \right]_{-\infty}^{+\infty} - \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} (y_0 - y_\tau)\, e^{-\frac{(y_0-\mu)^2}{2\sigma^2}} e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0)$$

$$\left.\begin{array}{ll} & \mathrm{J}(y_\tau) = -\frac{1}{\varepsilon^2} I(y_\tau) + \frac{y_\tau}{\varepsilon^2} S(y_\tau) \\ \Leftrightarrow & \text{et} \\ & \mathrm{J}(y_\tau) = I(y_\tau) - \mu S(y_\tau) \end{array}\right] \Rightarrow I(y_\tau) = \frac{y_\tau + \mu\varepsilon^2}{1+\varepsilon^2} S(y_\tau) \qquad (4.6)$$

Des équations (4.5) et (4.6), on déduit :

$$(4.5), (4.6) \Rightarrow \hat{y}_\tau = \frac{1}{1+\varepsilon^2} y_\tau + \frac{1}{1+\varepsilon^2}\mu \Rightarrow y_\tau - \hat{y}_\tau = \frac{\varepsilon^2}{1+\varepsilon^2}(y_\tau - \mu) \qquad (4.7)$$

Ainsi, la valeur estimée $\widehat{y}_\tau$ donnée par l'équation 4.7 est fonction de ε , la constante du noyau permettant de définir sa largeur et de μ, le centre de y_0. Les figures 4.2-b1 et b2 montrent les résultats d'une simulation de cette application théorique en utilisant un noyau gaussien de $\varepsilon = 0.05$ et une distribution gaussienne de y_0 d'écart-type $\sigma = 0.5$. Nous remarquons que l'erreur d'estimation de la figure 4.2-b2 est une fonction croissante qui passe par zéro au centre μ. Plus y_τ s'éloigne de μ, plus l'erreur d'estimation augmente. Ceci explique l'inconvénient de l'utilisation d'une base d'étalonnage de distribution normale avec le noyau gaussien utilisé pour notre application.

Si la distribution de y_0 est uniforme, avec les extrêmités définies par a et b

$$g(y_0) = 1 \qquad \text{et} \qquad N(y_\tau, y_0) = e^{-\frac{(y0-y_\tau)^2}{2\varepsilon^2\sigma^2}} \qquad (4.8)$$

De la même manière le calcul intégral nous permet d'obtenir :

$$(4.3) = \frac{\int_a^b y_0 e^{-\frac{(y0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0)}{\int_a^b e^{-\frac{(y0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0)} = \frac{\mathrm{I}(y_\tau)}{\mathrm{S}(y_\tau)}$$

$$\Leftrightarrow \mathrm{I}(y_\tau) = \int_a^b y_0 e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0) = \int_a^b (y_0 - y_\tau)\, e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0) + y_\tau \int_a^b e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0)$$

$$\Leftrightarrow \mathrm{I}(y_\tau) \quad = \quad \left[-\varepsilon^2\sigma^2 \left(e^{-\frac{(y_0-y_\tau)^2}{2\varepsilon^2\sigma^2}}\right)\right]_a^b + y_\tau \int_a^b e^{-\frac{(y0-y_\tau)^2}{2\varepsilon^2\sigma^2}} d(y_0)$$

Or,

$$\mathrm{I}(y_\tau) \quad = \quad \mathrm{J}(y_\tau) + y_\tau S(y_\tau)$$

Par suite, le calcul de $\mathrm{I}(y_\tau)$ sera déduit de :

$$\left.\begin{array}{l} \mathrm{J}(y_\tau) = -\varepsilon^2\sigma^2 \left(e^{-\frac{(b-y_\tau)^2}{2\varepsilon^2\sigma^2}} - e^{-\frac{(a-y_\tau)^2}{2\varepsilon^2\sigma^2}}\right) \\ \mathrm{J}(y_\tau) = I(y_\tau) - y_\tau S(y_\tau) \end{array}\right] \Rightarrow I(y_\tau) = -\varepsilon^2\sigma^2 \left(e^{-\frac{(b-y_\tau)^2}{2\varepsilon^2\sigma^2}} e^{-\frac{(a-y_1)^2}{2\varepsilon^2\sigma^2}}\right) + y_\tau S(y_\tau) \quad (4.9)$$

D'où,

$$(4.3), (4.9) \Rightarrow \hat{y}_\tau \quad = \quad y_\tau - \frac{\varepsilon^2\sigma^2 \left(e^{-\frac{(b-y_\tau)^2}{2\varepsilon^2\sigma^2}} - e^{-\frac{(a-y_\tau)^2}{2\varepsilon^2\sigma^2}}\right)}{S(y_\tau)}$$

L'erreur d'estimation est alors représentée par :

$$\Rightarrow y_\tau - \hat{y}_\tau \quad = \quad \frac{\varepsilon^2\sigma^2 \left(e^{-\frac{(b-y_\tau)^2}{2\varepsilon^2\sigma^2}} - e^{-\frac{(a-y_\tau)^2}{2\varepsilon^2\sigma^2}}\right)}{S(y_\tau)} \qquad (4.10)$$

L'équation 4.10 montre que la valeur estimée est fonction de la largeur du noyau et des limites de la distribution. La figure 4.2-a montre les résultats d'une simulation de cette application théorique en utilisant le noyau gaussien de $\varepsilon = 0.05$ et une distribution uniforme de y_0 de bornes $a = -1$ et $b = 1$. Nous remarquons que l'erreur d'estimation

représentée dans la figure 4.2-a2 est nulle tant que y_τ est à l'intérieur de l'intervalle [-1 ;1]. Dès que y_τ sort de cet intervalle, l'erreur d'estimation s'accroît. Par conséquent, la gamme couverte par $\mathbf{y}_0$ doit être la plus large possible, pour que les limites normales de la base d'étalonnage soient loin des limites rencontrées en pratique.

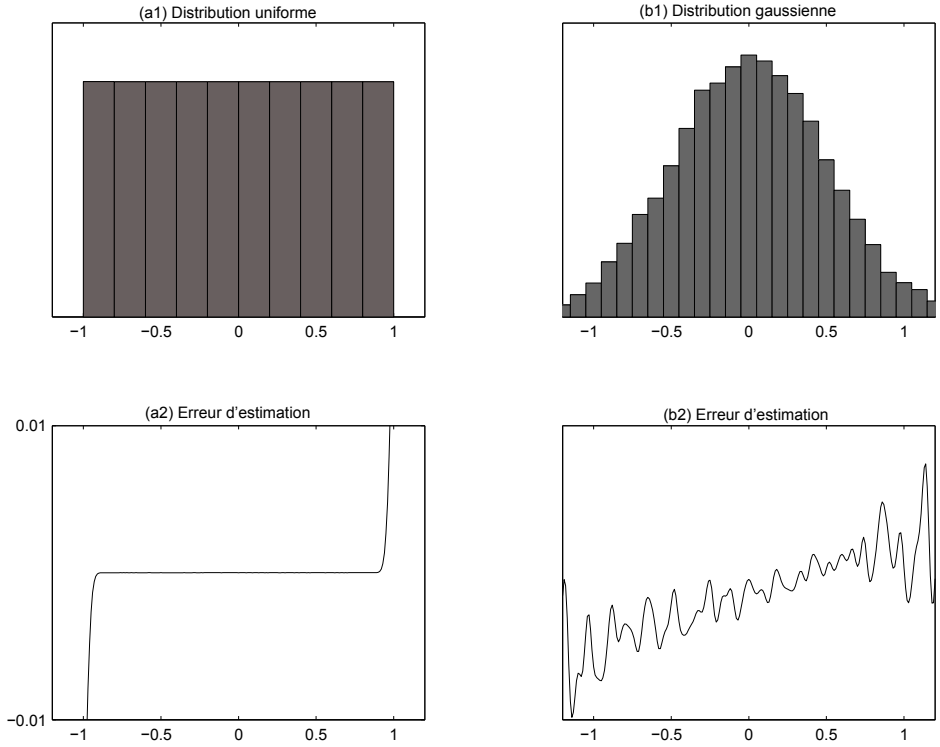

FIG. 4.2 – Influence de la distribution de y_0 sur l'erreur d'estimation par un noyau gaussien. a1 distribution uniforme ; b1 distribution normale ; a2 variation de l'erreur d'estimation en cas de distribution uniforme ; b2 variation de l'erreur d'estimation en cas de distribution gaussienne.

Donc une distribution uniforme de la base d'étalonnage est plus favorable pour l'utilisation de DOP avec une fonction de noyau gaussienne. En conséquence de quoi, nous avons choisi pour notre application une distribution uniforme de la base $\mathbf{y}_0$. Ceci est tout à fait conforme avec le choix d'utiliser les algorithmes de sélection de base d'étalonnage sur y (paragraphe 3.2.2) pour l'optimisation de la base d'étalonnage.

4.3 Conclusions

Dans ce chapitre une nouvelle méthode, baptisée DOP pour *Dynamic Orthogonal Projection*, a été proposée pour maintenir la robustesse en ligne des mesures effectuées par spectrométrie IR pour le suivi des processus continus. Cette méthode utilise :

- la base d'étalonnage ;
- des standards virtuels, estimés en ligne à partir de quelques mesures de recalage ;
- un transfert d'étalonnage continu par projection orthogonale.

Les paramètres de cette méthode ont été présentés. Elle nécessite l'utilisation d'un noyau, fonction de la distribution de la base d'étalonnage. L'influence de la nature de cette distribution sur l'estimation a été montrée. Une distribution uniforme semble être la mieux adaptée dans le cas de noyau gaussien, car l'erreur d'estimation dépend uniquement des extrémités de la distribution. Ainsi, il suffit de prendre une distribution avec des limites très larges par rapport aux limites attendues en ligne, ce qui est ordinairement réalisé. C'est une méthode simple à implémenter, elle possède uniquement deux paramètres à régler en fonction de la base d'étalonnage utilisée, la largeur du noyau utilisé et le nombre de composantes principales de l'ACP. DOP assure une mesure robuste en ligne en rendant la base d'étalonnage indépendante des facteurs d'influence ayant eu lieu. De plus, elle peut fournir un diagnostic des facteurs d'influence, par l'analyse de la matrice des résidus éliminés par la projection orthogonale. Ainsi elle permet d'avoir en temps réel des informations utiles à une meilleure compréhension du processus de mesure.

Dans le chapitre suivant, cette méthode est appliquée sur des cas réels.

Chapitre 5

Application de DOP pour le suivi en ligne des processus continus par spectroscopie IR

Sommaire

Ce chapitre présente deux applications de la méthode DOP au suivi en temps réel de processus continus. La première concerne le suivi par spectrométrie PIR d'une fermentation œnologique, soumise à des variations d'un facteur d'influence physique (la température). La deuxième s'intéresse au suivi par spectrométrie MIR d'un bioréacteur de dépollution des effluents viticoles, soumis à des variations d'un facteur d'influence chimique (la concentrations d'un composé parasite, l'ammoniaque). Au préalable, nous indiquons de manière générale comment fixer les paramètres de DOP et comment mesurer l'effet de la méthode DOP sur la robustesse.

5.1 Mises au point préliminaires de la méthodologie

5.1.1 Configuration de DOP

La méthode DOP développée possède deux paramètres à définir : la fonction de noyau utilisée pour l'estimation des "'standards virtuels"' et k le nombre de composantes principales pour déterminer les dimensions du sous-espace décrit par $\mathbf{P}$ (voir paragraphe 4.2.4.2).

5.1.1.1 Choix du noyau A

Plusieurs méthodes existent pour le calcul de $\mathbf{A}$. La plus appropriée consiste en une interpolation de $\mathbf{y}_\tau$ à partir de la base d'étalonnage $\mathbf{y}_0$. Dans notre cas, c'est la fonction de noyau gaussien qui est utilisée pour estimer par combinaison linéaire la mesure de référence obtenue en ligne à partir de l'ensemble des mesures d'étalonnage. Pour tenir compte de la distribution de $\mathbf{y}_0$, on relie la largeur du noyau à celle de la distribution par un coefficient ε tel que : $\sigma_A = \varepsilon\sigma(\mathbf{y}_0)$. Le noyau $\mathbf{A}$ est défini par l'équation suivante :

$$a_{ij} = \frac{1}{\sigma_A\sqrt{2\pi}} \exp\left(\frac{-(y_{\tau i} - y_{0j})^2}{2\sigma_A^2}\right)$$
$$a_{ij} = \frac{1}{\varepsilon\sigma(\mathbf{y}_0)\sqrt{2\pi}} \exp\left(\frac{-(y_{\tau i} - y_{0j})^2}{2\varepsilon^2\sigma^2(\mathbf{y}_0)}\right)$$
$$\text{avec} \qquad i = 1 : n_\tau \quad \text{et} \quad j = 1 : n_0 \tag{5.1}$$

avec ε le coefficient définissant la largeur du noyau gaussien, n_0 et n_τ sont respectivement le nombre de mesures de la base d'étalonnage et des mesures de recalage en ligne. La matrice du noyau $\mathbf{A}$ représente alors les coefficients utilisés pour estimer les valeurs de référence périodiques $\mathbf{y}_\tau$, obtenues en ligne, par combinaison linéaire des valeurs de référence de la base d'étalonnage $\mathbf{y}_0$. Ceci permet d'estimer les valeurs de $\mathbf{y}_\tau$ dans l'espace de plus grande dimension $\mathbf{y}_0$ en utilisant le noyau approprié défini dans l'équation 5.1. Le seul paramètre à régler est le paramètre ε.

Dans le où la distribution de $\mathbf{y}_0$ est uniforme, une approche théorique peut être développée pour savoir comment fixer ε.

Soient a et b les deux bornes de la distribution uniforme de $\mathbf{y}_0$. Soit Δ l'écart moyen entre deux valeurs consécutives. On a :

$$\begin{aligned} a &= \min(\mathbf{y}_0) \\ b &= \max(\mathbf{y}_0) \\ \Delta &= \frac{b-a}{n_0-1} \end{aligned}$$

La fonction du noyau est caractérisée par son centre c et sa largeur L autour du centre du noyau. Dans notre cas, pour un noyau gaussien, cette largeur peut être définie selon le poucentage de la surface de la distribution à considérer (figure5.1). Par exemple pour 95%, l'intervalle de confiance est compris entre $\pm 2\varepsilon\sigma(\mathbf{y0})$ et $L = 4\varepsilon\sigma(\mathbf{y0})$, pour 99.9%, l'intervalle de confiance est compris entre $\pm 3\varepsilon\sigma(\mathbf{y0})$ et $L = 6\varepsilon\sigma(\mathbf{y0})$, etc. Pour notre exemple de calcul théorique on considère le cas de 95% donc on a :

$$L = 4\varepsilon\sigma(\mathbf{y0})$$

Pour que l'estimation reste correcte, la plus petite largeur du noyau doit être au moins égale à l'intervalle entre deux valeurs de $\mathbf{y0}$. Sa plus grande valeur doit rester bien

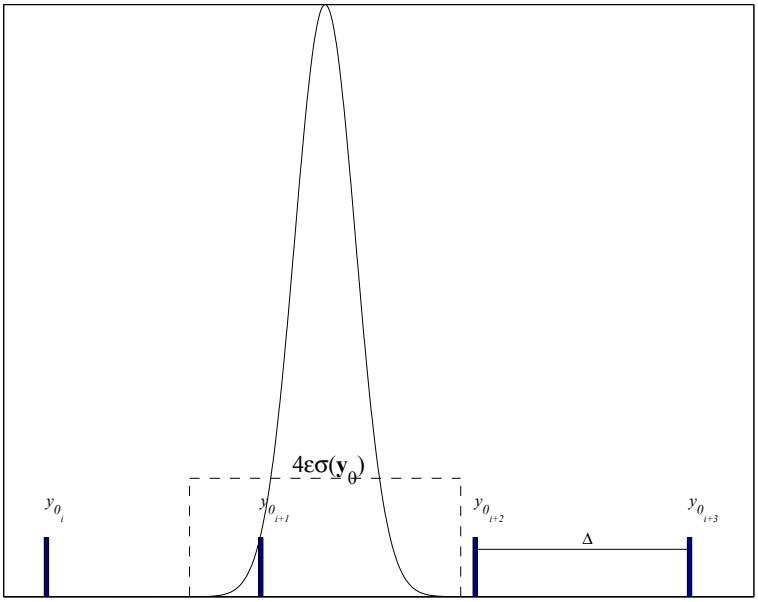

FIG. 5.1 – Estimation utilisant une fonction de noyau gaussien de largeur $4\varepsilon\sigma(\mathbf{y}_0)$ à partir d'une distribution uniforme.

inférieure à la largeur de la distribution $(b-a)$. Ces deux conditions nous donnent :

$$\Delta \leq L << b-a$$

$$\frac{b-a}{n_0-1} \leq 4\varepsilon\sigma(\mathbf{y0}) << b-a$$

Or, pour une distribution uniforme on a par théorie $\sigma(\mathbf{y0}) = (b-a)/2\sqrt{3}$, donc :

$$\frac{b-a}{n_0-1} \leq \frac{2\varepsilon(b-a)}{\sqrt{3}} << b-a$$

Soit, finalement :

$$\frac{1}{n_0-1} \leq \frac{2\varepsilon}{\sqrt{3}} << 1$$

C'est à dire, si n_0 est grand devant 1 :

$$\frac{\sqrt{3}}{2n_0} \leq \varepsilon << \frac{\sqrt{3}}{2}$$

Et, en ne considérant que les ordres de grandeur ;

$$\frac{1}{n_0} \leq \varepsilon << 1$$

On déduit de ce calcul théorique qu'une valeur de ε de l'ordre de $1/n_0$ est la plus appropriée pour l'application de DOP, dans le cas d'un noyau gaussien sur une base d'étalonnage de distribution uniforme.

5.1.1.2 Choix de k

Pour le paramètre k, généralement des valeurs empiriques de pourcentage de variance entre 90% et 99% sont utilisées (voir paragraphe 4.2.4.2).

5.1.1.3 Conclusion

Dans la première application, on étudiera l'influence de ces deux paramètres et on vérifiera la théorie développée ici. Dans la deuxième, on fixera *a priori* la valeur des paramètres.

5.1.2 Évaluation de la robustesse

La robustesse globale du modèle d'étalonnage a été évaluée, *a posteriori*, sur la base du critère de robustesse $R_C = \frac{SEP_0}{SEP}$ de l'équation 2.3 établie au chapitre 2, où SEP_0 est l'erreur de prédiction observée lors de l'étalonnage et SEP est l'erreur standard observée sur tout le modèle. La robustesse en temps réel a été évaluée en utilisant le critère de *robustesse dynamique* $R_C(t)$ exprimé par le rapport $R_C(t) = \frac{SEP_0}{SEP(t)}$. L'erreur de prédiction en ligne $SEP(t)$ est calculée, *a posteriori*, sur une fenêtre mobile centrée sur t. Ce critère est calculé pour le modèle avant et après correction par DOP pour évaluer l'amélioration de sa robustesse. Plus il est petit, moins la mesure est robuste. Il tend vers 1 lorsque le SEP se rapproche de SEP_0.

5.2 Première application : Cas des facteurs d'influence physiques

5.2.1 Introduction

La méthode DOP proposée dans le chapitre précédent est appliquée pour le suivi en ligne d'une fermentation alcoolique du vin blanc par spectrométrie PIR. La fermentation alcoolique a été choisie pour valider la méthode DOP car c'est un procédé lent et continu.

Des expériences permettant d'avoir en continu à la fois des mesures de référence et des mesures en PIR ont été réalisées à l'INRA - unité expérimentale de Pech Rouge. Cette application fait partie du projet IRVIN qui vise à développer un capteur PIR pour le suivi de fermentation alcoolique du moût directement sur les fermenteurs en utilisant des fibres optiques. La méthode de suivi de la fermentation alcoolique utilisée jusqu'à présent en ligne consiste à suivre la vitesse de dégagement du CO_2 à l'aide d'un débimètre massique. Ce dispositif est encore en phase de développement et n'est pas disponible sur le marché industriel [126]. L'avantage d'utiliser des capteurs PIR réside surtout dans le fait qu'ils rendent possible la caractérisation de la composition du milieu mesuré.

La méthode DOP a été appliquée pour le suivi de fermentation du vin blanc où la température a été choisie comme facteur d'influence à corriger. Des variations de la température ont été provoquées lors de la fermentation. Une première fermentation isotherme a permis de constituer une base d'étalonnage ; une deuxième anisotherme a fourni un jeu de test pour valider la méthode DOP.

5.2.2 Aperçu bibliographique sur la fermentation alcoolique

5.2.2.1 Description d'un cycle fermentaire

La fermentation alcoolique est une réaction exothermique [175]. Elle correspond à la transformation, par les levures, des sucres fermentescibles du moût (glucose et fructose)

en alcool (éthanol).

$$C_6H_{12}O_6 \overset{levures}{\rightarrow} 2C_2H_5OH + 2CO_2\nearrow + 24Kcal \tag{5.2}$$

Cette fermentation diffère selon le type de vinification. Dans le cas de vinification en blanc, ce qui est notre cas, elle se fait sur un milieu liquide. La figure 5.2 illustre

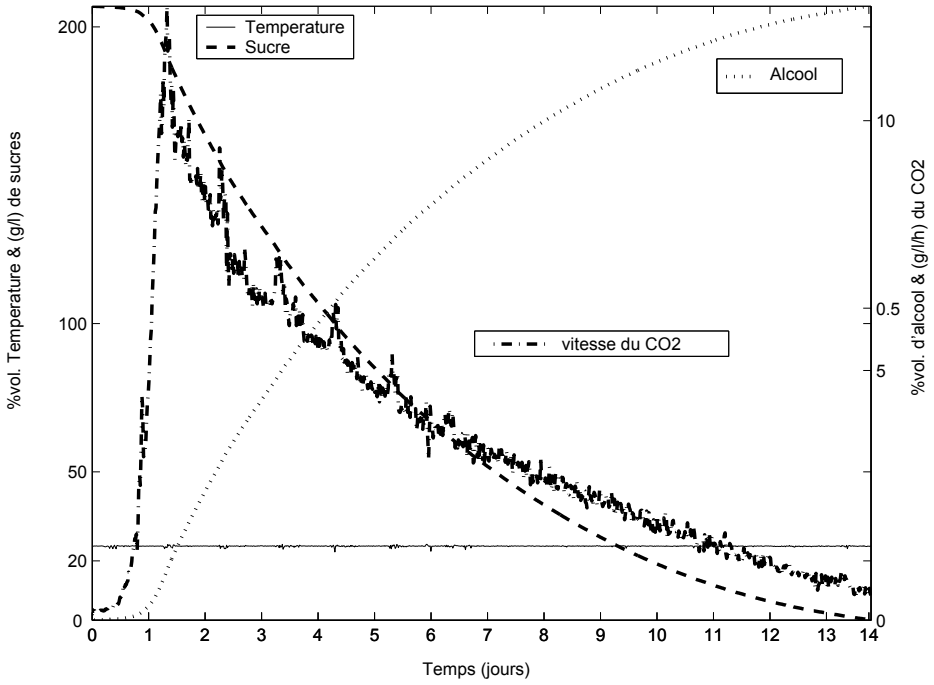

FIG. 5.2 – Description d'un cycle fermentaire en conditions isothermes 25^{o}C

la cinétique d'une fermentation alcoolique à température constante, représentée par le suivi en ligne du CO_2 dégagé. La vitesse instantanée du dégagement du CO_2 est proportionnelle à la vitesse de la fermentation (courbe d'évolution du taux d'alcool ou du sucre) [14]. Une fermentation est constituée de trois phases :

- Phase de latence où le milieu est saturé de CO_2. Sa durée, n'excédant pas les 24h généralement, est surtout fonction de la température.
- Phase de croissance cellulaire et de fermentation. Elle s'étend jusqu'à la vitesse maximale du dégagement du CO_2. Cette valeur maximale de la vitesse est fonction de la teneur initiale du moût en azote et de la température.

– Phase stationnaire. Elle correspond à la diminution progressive de l'activité des levures jusqu'à épuisement du milieu de sucres résiduels où la vitesse chute puis s'annule. Durant cette phase la fermentation se poursuit (consommation de 60 à 80% du sucre).

5.2.2.2 Facteurs d'influence

La cinétique de la fermentation alcoolique est influencée par trois facteurs principaux : la souche de levures, la composition du moût et le pilotage de la fermentation. L'effet de chacun des facteurs mentionnés, a fait l'objet de plusieurs travaux de recherche notamment ceux de Sablayrolles et al. [129, 126, 13].
Pour la conduite de fermentation, la législation n'autorise qu'un certain nombre limité d'interventions : la température, les ajouts d'activateurs (notamment l'azote), le remontage et la mise sous pression.
Dans cette application c'est surtout le facteur température lié à la conduite de la fermentation qui est considéré.

Effet de la température :

La température est un des paramètres sur lequel les oenologues peuvent agir en temps réel pour le pilotage de la fermentation. Elle permet l'activation de la fermentation, ce qui se traduit par une augmentation de sa vitesse.

L'effet de la température sur la cinétique de la fermentation alcoolique du vin ainsi que sur la qualité du produit final a été le sujet de plusieurs travaux de recherche [113, 37, 30, 127, 14, 128, 129].

La température a un effet sur les mesures spectrales en PIR. Elle agit sur les liaisons hydrogènes OH de l'eau qui absorbent fortement dans le PIR [92].

Vu l'intérêt de la variation de la température dans la conduite de fermentation et vu son effet sur les mesures spectrales en PIR, elle a été considérée comme facteur d'influence modèle sur la robustesse de l'étalonnage utilisé pour effectuer des mesures en PIR.

5.2.3 Matériels et méthodes

5.2.3.1 Milieu de fermentation

1. **Moûts**

   Le moût utilisé est issu du cépage MACCABEU, il parvient de l'UEPechRouge. Ce moût est pasteurisé et conservé à 3^{o}C dans des cuves de 10hl à 20hl. Pour nos fermentations, un prélèvement de 100l est effectué sous pression en utilisant un mélange de gaz (CO_2 et N_2), ceci pour garder le moût stérile sans risque d'oxydation. Une dilution au 1/10 ème est effectuée sur le moût initial afin de diminuer la quantité du sucre de départ. En plus, un faible ajout de 30 mg/l d'N_2 ou 15g/hl de diammonium hydrogenophosphate (DAP), est effectué pour avoir des fermentations de courte durée.

   Ce moût est gardé à la température ambiante pendant la nuit pour monter en température. Le lendemain, la cuve est thermostatée à 25°C.

2. **Souche de Levures**

   Après avoir atteint la température consigne, le démarrage de la fermentation se fait au moment de l'innoculation. L'innoculum est préparé en pesant 10g de levures (Saccharomyces cerevisiae, K1 INRA) auxquels sont ajoutés 5g de sucre D(+)glucose, le tout est mélangé à 100ml d'eau chaude et maintenu à 30^{o}C à l'étuve pendant 20 minutes. Ensuite, ces levures sont diluées avec un peu de moût et ajoutées à la cuve.

5.2.3.2 Instrumentation

1. **Fermenteurs**

   Pour l'étude de validation de la méthode de suivi en ligne, les expériences ont eu lieu, à l'échelle pilote, dans des cuves de 100 litres en acier inoxydable 5.4.

2. **Système de régulation de la température**

   La température du fermenteur est réglée à l'aide de deux tuyaux en inox en forme de U qui sont intégrés dans le capot de la cuve. Ces tuyaux sont remplis d'eau et reliés à des vannes d'eau chaude (40^{o}C) ou froide (3^{o}C) pour effectuer le réglage.

Cette température est mesurée en continu à l'aide d'une sonde plongée dans la cuve. Suite à la commande de la température consigne par le logiciel, il y aura ouverture des vannes d'eau chaude ou froide qui sont responsables du réglage de la température. Un programme de variation de la température au cours de la fermentation peut être envisagé suite à l'utilisation d'un logiciel créé sous Labview version 5.1.

La température de la cellule de mesure de l'échantillon en transmission du spectromètre PIR a été réglée d'une façon équivalente à celle du fermenteur. Un montage PID a été mis en oeuvre (alimentation 220V, résistance de 100 ohms et une sonde Pt100 pour mesurer la température). Ce montage règle uniquement la température de la cellule de mesure (figure 5.3).

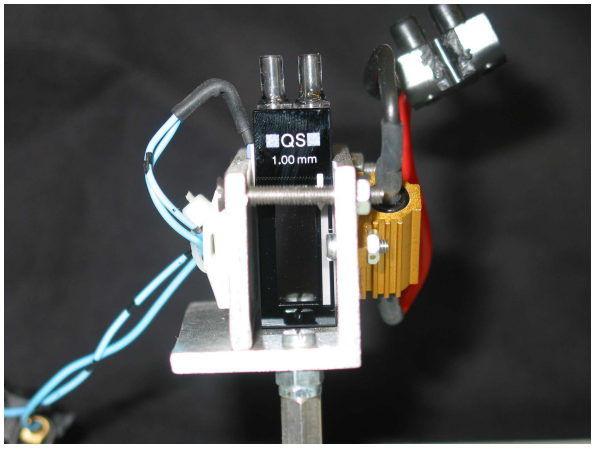

FIG. 5.3 – Montage de réglage de la température de la cellule de mesure de l'échantillon [J-L LABLEE04]

3. **Débimètre massique**

Le débimètre massique (type Brooks 5850) est branché en ligne sur la cuve de fermentation. Il permet d'enregistrer la vitesse instantanée de la fermentation exprimée en fonction de CO_2 dégagé $\frac{d(CO_2)}{dt}$. L'acquisition a lieu toutes les 20 minutes. Le gaz est desséché et la mesure de la quantité du CO_2 sec dégagé est estimée par intégration simple de la vitesse qui est directement donnée par le débimètre avec une précision de 3%.

D'après l'équation stœchiométrique 5.2 de la fermentation alcoolique, et connais-

sant le volume initial du moût et la teneur en sucre, on peut en déduire la quantité de CO_2 dégagé. Par la suite on atteint celle du sucre consommé et de l'éthanol produit d'après les équations de transformations suivantes [38] :

$$1\text{g } CO_2 \text{ dégagé} = 2.17\text{g de sucres consommés}$$
$$16.83 \text{ g de sucres} = 1\% \text{ vol. d'alcool}$$

L'erreur de la mesure de l'éthanol à partir du CO_2 dégagé est de $\mp$2,73g/l, le maximum d'erreur observé n'excède pas les 3 g/l ou 0.38 % vol. d'alcool, celle des sucres résiduels est de l'ordre de 4g/l [39], ce qui est acceptable pour les industriels dans le but du suivi en continu de la fermentation. La concentration initiale en sucre du moût était de 207g/l.

4. **Logiciel de mesure du débit massique**

   Ce logiciel est utilisé avec le débimètre massique pour le suivi et le contrôle des fermentations. Il a été conçu pour le calcul de la mesure du débit de dégagement de CO_2 effectuée avec le débimètre massique en continu. Le débit est mesuré avec une fréquence réglable. Sa valeur est moyennée pendant la durée comprise entre 2 stockages. Ainsi, le débit est mesuré chaque 20s et si les données sont enregistrées chaque 20 min, la valeur dans le fichier est la moyenne de 60 mesures instantanées.

5. **Spectromètre Proche Infrarouge**

   Le spectromètre UV-VIS-PIR Jasco V-570 est utilisé pour effectuer les mesures en ligne par transmission en utilisant une cellule à circulation de Quartz de 1mm de trajet optique. La plage de longueur d'onde balayée s'étend de 200nm à 2500nm, l'acquisition est faite à raison de 4000nm/min avec un pas de 2nm.

   La ligne de base est effectuée à 25^{o}C au début de chaque fermentation en utilisant l'eau distillée. Le spectromètre étant à double faisceau, les mesures sont effectuées par la suite par rapport à l'eau distillée comme référence.

   Des tuyaux de prélèvement d'échantillon émanant de la cuve sont reliés à la cellule de mesure du spectromètre JASCO, de façon à ce que le circuit reste étanche et ne présente aucune fuite (figure-5.4). Le prélèvement se fait automatiquement à

l'aide d'une pompe péristaltique (durée de pompage = 1 minute), l'échantillon passe dans la cellule de mesure où il reste 3 minutes afin de se stabiliser en température. Cette dernière est réglée grâce à une température consigne fixée manuellement, *a priori*, et un système PID qui permet le réglage (figure-5.3). Ensuite, la prise de spectre, d'une durée de 6 minutes, se déclenche et ainsi de suite toutes les 10 minutes il y a acquisition d'un spectre en UV/VIS/PIR entre 200nm-2500nm.

FIG. 5.4 – Montage du suivi en ligne par spectrométrie SPIR (JASCO-V570) de la fermentation alcoolique du vin blanc (échelle pilote - 100 L)

5.2.3.3 Conduite et suivi des fermentations

Deux fermentations ont été réalisées dans différentes conditions :

Fermentation isotherme

La première fermentation à été conduite dans des conditions isothermes à 25°C pendant une durée de douze jours.

Fermentation anisotherme

La deuxième fermentation a été réalisée sur le même moût dans des conditions anisothermes. Le profil de la variation de température montré en figure 5.5 est une simulation des variations qui peuvent avoir lieu sur les cuves industrielles.

Au début la température est maintenue à 25°C, ensuite elle augmente de 25°C à 35°C par pas de 0.2°C/h avec un palier la nuit. Donc il y a 5 paliers avec une pente de

0.2°C/h pendant 5 jours et puis une diminution de la température de 35°C à 25°C est réalisée pendant les 5 jours suivants.

Mesures de recalage aux points de contrôle

Cinq points de recalage ont été pris sur la fermentation anisotherme, comme montré par la figure 5.5. Les mesures de contrôle et les spectres correspondant à ces instants seront utilisés par la méthode DOP.

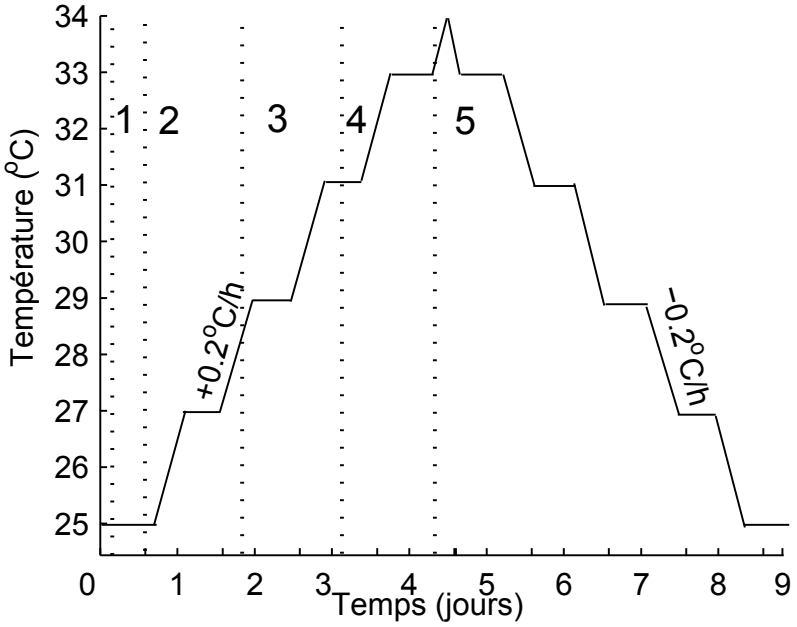

FIG. 5.5 – Programme de variation de la température qui a été imposé au cours de la fermentation anisotherme. Les barres verticales indiquent les moments de recalage utilisés par DOP.

5.2.4 Modélisation

5.2.4.1 Bases de données

Nous avons limité notre étude à la mesure du degré d'alcool. Les bases de données correspondant aux deux fermentations sont :

- $\mathbf{X}_{1(849\times1151)}$, $\mathbf{y}_{1(849\times1)}$, constituée des données spectrales et des valeurs de référence (degré d'alcool) relatives au suivi de la fermention isotherme ;

– Une base de données spectrales $\mathbf{X}$, constituée des spectres acquis en ligne à raison d'1 spectre/10 minutes durant la fermention anisotherme ;
– A chaque temps $t = \tau$, $\mathbf{X}_\tau$ et $\mathbf{y}_\tau$ contiennent les spectres et les mesures de référence des points de recalage antérieurs à τ.

5.2.4.2 Prétraitement des données

Tout d'abord, nous avons limité la plage de longueur d'onde à partir des spectres 200-2500nm (figure 5.6). Bien que l'alcool et les sucres absorbent surtout dans le PIR entre 2350nm et 2070nm respectivement, nous avons choisi de travailler dans une gamme $<$ 2000nm car le projet IRVIN prévoit l'utilisation de fibres optiques. De plus,

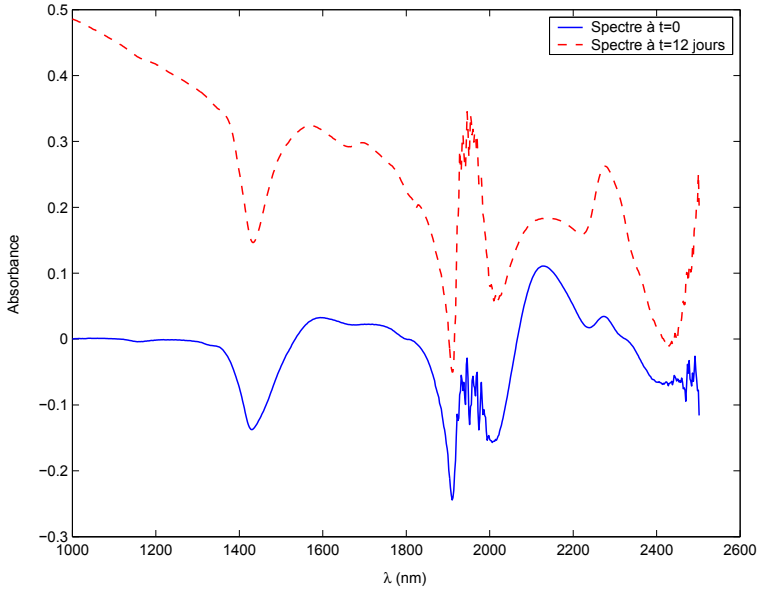

FIG. 5.6 – Spectres PIR de suivi de fermentation isotherme (25°C). Le spectre bleu du début de la fermentation et le spectre rouge de la fin de fermentation.

notre but méthodologique est d'avoir un modèle avec une zone de longueur d'onde sensible aux variations de température pour pouvoir tester notre nouvelle méthode DOP. Finalement, l'intervalle [1404nm-1546nm], renfermant les absorptions caractéristiques des fonctions alcooliques (R-OH) [111], a été retenu (figure 5.7).

Les données brutes ont été utilisées pour la modélisation. Ayant utilisé un spectromètre double faisceau, les mesures spectrales ont été corrigées par rapport à la référence

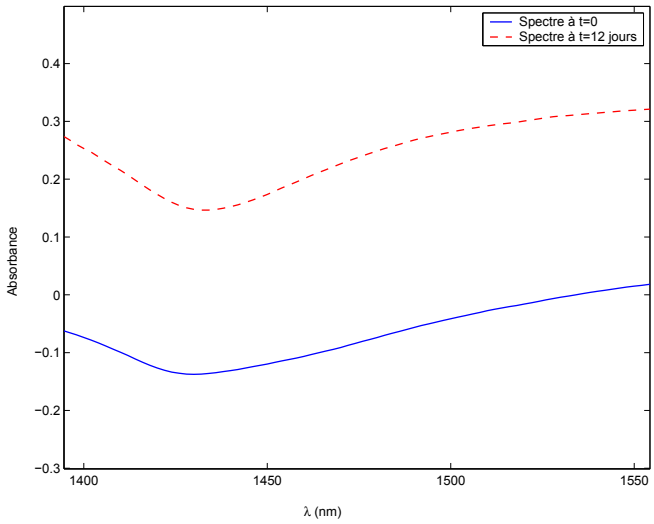

FIG. 5.7 – La plage sélectionnée de longeur d'onde pour le suivi de fermentation isotherme (25^oC)

(eau). Les données ne sont pas centrées.

5.2.4.3 Sélection de la base d'étalonnage

D'après l'étude effectuée sur les algorithmes de sélection de la base d'étalonnage et leur effet sur la robustesse du modèle d'étalonnage (paragraphe 3.2.2), l'algorithme Duplex a été appliqué sur $\mathbf{y}_1$ pour le choix des spectres les plus représentatifs de $\mathbf{X}_1$. L'algorithme Duplex choisit alternativement des échantillons pour l'étalonnage (les 2 premiers les plus éloignés) et d'autres pour la validation (les deux les plus éloignés des deux précédents) et ainsi de suite. Sur les 849 mesures de $\mathbf{y}_{01}$, 300 mesures ont été choisies pour l'étalonnage $\mathbf{X}_{01(300\times171)}, \mathbf{y}_{01(300\times1)}$ et le reste des données sera utilisé pour le test $\mathbf{X}_{test1(549\times171)}, \mathbf{y}_{test1(549\times1)}$.

Le nombre d'observations sélectionnées par l'algorithme a été établi en observant l'histogramme de $\mathbf{y}_{01}$, de façon à obtenir une distribution la plus uniforme possible en conformité avec la justification théorique du paragraphe 4.2.4.3.

5.2.4.4 Modèle PLS

Un modèle PLS a été calculé sur $\mathbf{X}_{01}, \mathbf{y}_{01}$. Une validation croisée leave-one-out (LOO) a permis de choisir le nombre de variables latentes. Le modèle est ensuite testé sur $\mathbf{X}_{test1}, \mathbf{y}_{test1}$, permettant de calculer l'erreur standard de prédiction du modèle d'étalonnage notée SEP_0.

5.2.5 Résultats

5.2.5.1 Modèle d'étalonnage dans les conditions isothermes

Le modèle d'étalonnage PLS construit sur la base d'étalonnage $\mathbf{X}_{01(300\times71)}, \mathbf{y}_{01(300\times1)}$ par validation croisée LOO, a permis d'obtenir l'erreur standard de validation croisée (SECV) et l'erreur standard d'étalonnage (SEC), reportées sur la figure 5.8. Le nombre de variables latentes choisies est celui qui correspond au minimum du SECV le plus proche du SEC. Ainsi le modèle résultant est construit avec 6 variables latentes expliquant $V_y = 99.84\%$ de la variance de y. Quand ce modèle est testé sur la base $\mathbf{X}_{test(549\times171)}, \mathbf{y}_{test(549\times1)}$, il donne une erreur de prédiction $SEP_0 = 0.18\%$vol. d'alcool. Cette prédiction est représentée par la figure 5.9. Les b-coefficients du modèle sont représentés dans la figure 5.10. Ils montrent un comportement chaotique entre 1430nm et 1450nm, ce qui est peut être dû au fait que le spectre de l'eau est soustrait du spectre de l'échantillon. Cette zone typique de l'absorption de l'eau est donc sujette à erreur.

5.2.5.2 Application du modèle d'étalonnage brut pour le suivi de la fermentation anisotherme

Le modèle d'étalonnage développé à 25°C est appliqué pour suivre en ligne une fermentation alcoolique en présence de variations de la température. Les résultats des prédictions de ce modèle sont données par la figure 5.11.

Sur cette figure, le modèle montre des problèmes de robustesse.

Décalage en début de fermentation : Dès le début du processus de fermentation (jour 1 à T=25°C), un biais de -0.5% est observé par rapport à la référence. Le critère de robustesse dynamique $R_C(t)$ correspondant, représenté dans la figure 5.14, varie

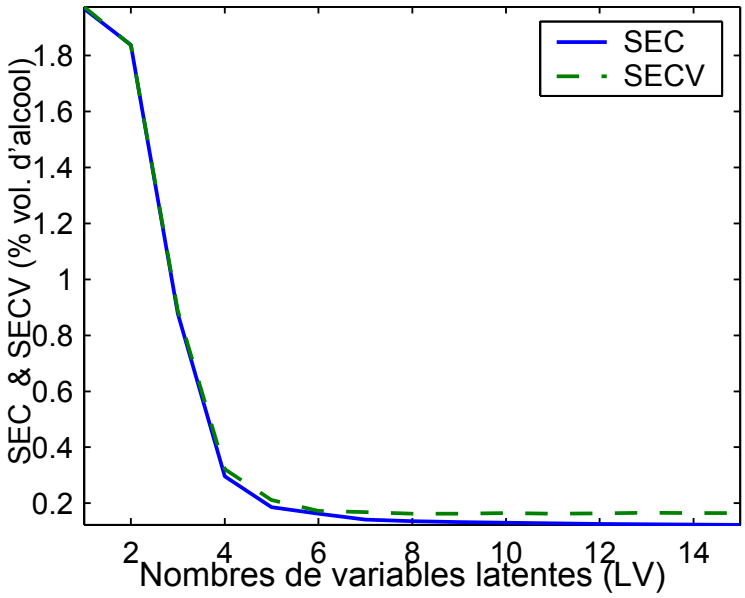

FIG. 5.8 – Variation du SEC et du SECV du modèle en fonction du nombre de variables latentes.

de 0.3 à 0.6, ce qui montre la non robustesse. Durant cette période, les conditions de température sont cependant les mêmes que lors de l'étalonnage. Ce biais est donc probablement dû à des facteurs non contrôlés, tels que la composition des moûts, la turbidité du milieu de mesure,... ou d'autres effets qui sont à la base de la différence entre les deux batches de fermentation.

Dérive due à la température : La figure 5.11 montre aussi que plus la température augmente (entre jour 2 et 5), plus le modèle présente une dérive marquée par rapport à la référence. La valeur de l'erreur est maximale à 34°C (jour 4-5), sa valeur absolue est approximativement de 1% vol. d'alcool. Dans le même temps, le critère de robustesse diminue jusqu'à moins de 0.2, ce qui signifie que le $SEP(t)$ est devenu 5 fois plus grand que le SEP_0 initial. A la fin de la fermentation, (jours 8 et 9), cette erreur disparaît en même temps que la température redevient égale à celle de l'étalonnage (25°C). Le critère de robustesse $R_C(t)$ augmente, pour atteindre des valeurs proches de 1. On remarque que l'effet du batch sur le modèle qui était présent au début de la fermentation disparait à la fin. Ceci peut être dû à deux causes : soit la présence d'un produit chimique non présent dans première fermentation et qui disparait à la fin ; soit la variation de la turbidité (ou la non homogénéité du moût au début de la

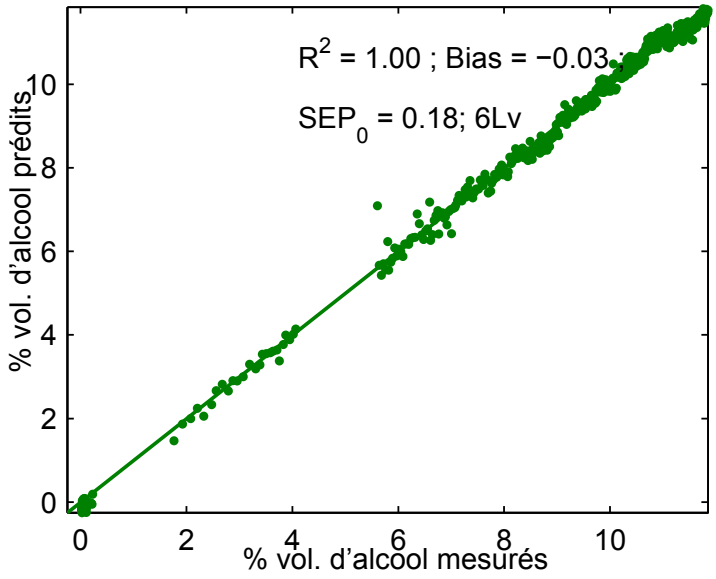

FIG. 5.9 – Test du modèle d'étalonnage dans les conditions isothermes (T=25°C)

fermentation) qui s'accroît au cours de la fermentation jusqu'au pic de vitesse de fermentation et qui diminue ensuite jusqu'à disparaître à la fin du fait du phénomène de clarification (les microorganismes cessent leur activité fermentaire et précipitent à la fin de la fermentation).

5.2.5.3 Application de la méthode DOP

Vérification expérimentale de l'influence des paramètres ε et k Pour verifier l'influence du coefficient de largeur ε de la fonction du noyau utilisée dans l'équation 5.1 sur les performances de DOP, une séquence de valeurs de ε, autour de la valeur théorique trouvée $\varepsilon = 1/n_0 = 3\ 10^{-3}$ (paragraphe 5.1.1.1) a été testée : $\varepsilon \in [10^{-4}, 10^{-3.8}, ..., 10^{-0.2}, 1]$.

Le nombre k de composantes principales pour déterminer les dimensions de la matrice $\mathbf{P}$ a été choisi en fonction du pourcentage de variance v expliqué par l'ACP. Différentes valeurs de v ont été testées : $v \in 80\%, 90\%, 95\%, 99.9\%, 100\%$.

L'optimisation du choix de ε et de k a été effectuée en même temps en appliquant l'algorithme DOP pour différentes valeurs des deux paramètres. Le critère de robustesse globale de prédiction (R_C) et le SEP ont été évalués en fonction des différentes valeurs

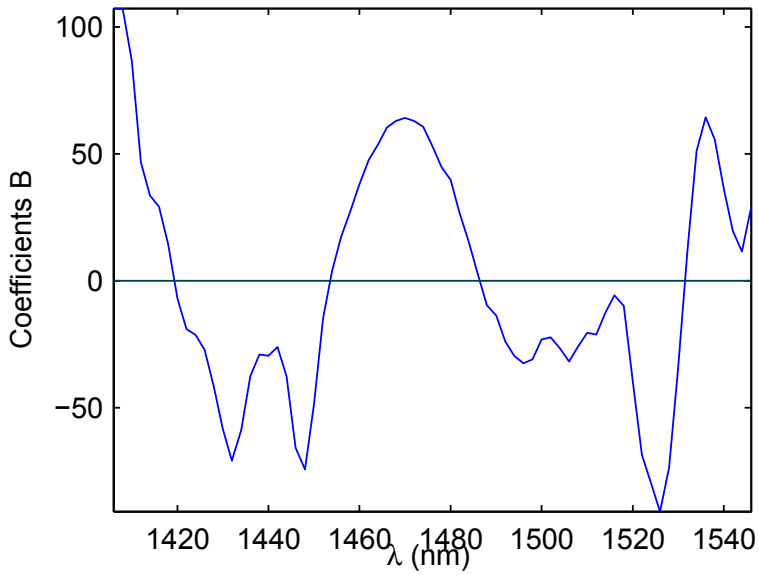

FIG. 5.10 – b-coefficients du modèle PLS établi dans les conditions isothermes avec 6LV

de ε et de v. La figure 5.12 et le tableau 5.1 reportent les résultats de cette simulation. Pour des valeurs très faibles de ε, quelle que soit la valeur de v, R_C est très petit, i.e. le SEP est très grand. Ceci est dû au fait que la largeur du noyau d'estimation est trop faible. A partir de $\varepsilon \simeq 2.5\ 10^{-4}$ l'erreur SEP chute brusquement, produisant une augmentation de R_C qui prend alors des valeurs comprises entre 0.7 et 1.1, selon la valeur de v. Ensuite, pour des valeurs de ε allant de $2.5\ 10^{-4}$ à 10^{-2}, on observe un plateau de robustesse optimale. A partir de 10^{-2}, R_C diminue de nouveau, car le noyau devient trop large. On remarque que la forme globale de l'évolution R_C en fonction de ε ne dépend pas de v, ce qui nous permettera de régler ces deux paramètres indépendemment. D'autre part, l'expérience valide empiriquement le niveau de ε que nous avions fixé par une approche théorique (paragraphe 5.1.1.1), i.e. $\varepsilon = 3.\ 10^{-3}$.

Effet de DOP sur le maintien de la robustesse Dans la figure 5.13, les prédictions sont tracées pour le modèle brut et le modèle corrigé par la méthode DOP, avec $\varepsilon = 3.\ 10^{-3}$, $v = 99\%$ et $V_y = 99.84\%$. Il apparaît nettement que la méthode DOP permet des prédictions beaucoup plus proches des valeurs de référence. Ceci est confirmé par la figure 5.14 qui montre l'évolution du critère de la robustesse dynamique $R_C(t)$.

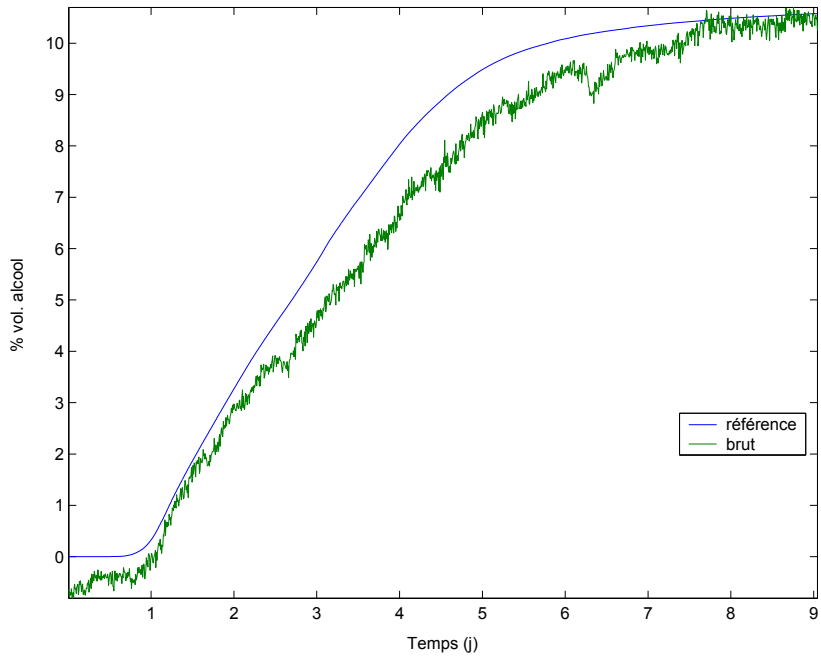

FIG. 5.11 – Prédiction du degré d'alcool lors de la fermentation anisotherme.

Pour les points de recalage de 1 à 3, l'espace de variation est de dimension 1 (k=1) et de dimension 2 (k=2) pour les points 4 et 5 (table 5.2).

Examinons l'effet de DOP pour les deux problèmes de robustesse relevés précédemment : le décalage en début de fermentation (effet batch) et la dérive due à la température.

Correction de l'effet batch : En début de fermentation, avant de correction par DOP, il est normal que les deux prédictions soient identiques figure (5.15). Dès la première correction, le modèle corrigé par DOP se comporte mieux, comme on peut le voir sur la figure 5.15. Les variations instantanées ne sont en rien altérées par la correction, dont l'effet est une pure translation. La première correction, apportée par le point 1 ($t = 2h30$), annule donc une grande partie du biais. Le critère $R_C(t)$ passe de 0.3 à plus de 1.2 [3]. Lors de ce premier recalage, comme on ne dispose que d'une seule mesure de référence, DOP n'utilise qu'un seul standard virtuel et par conséquent, le sous espace corrigé est de dimension 1 (tableau 5.2). La deuxième correction renforce la première en réduisant encore

[3]Sur la figure 5.14, cette augmentation instantanée est amortie par le lissage.

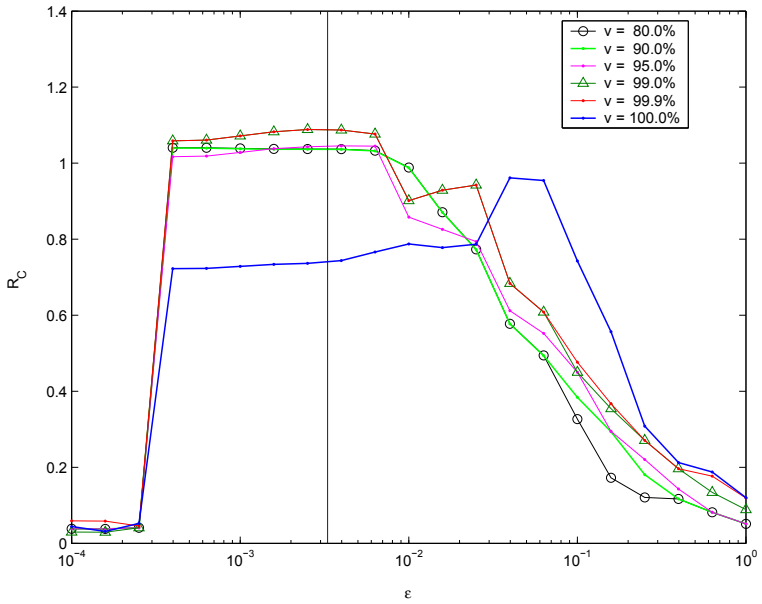

FIG. 5.12 – Variations de R_C en fonction de ε et de k. La barre verticale indique la valeur de $\varepsilon_{optimal} = 3.\ 10^{-3}$ que nous avons trouvé par la théorie.

le biais. Comme reporté par le tableau 5.2, une seule composante est encore suffisante pour corriger 99% de la différence entre les conditions standard et de la ligne. Ceci laisse à penser que le principal effet corrigé est un décalage global du spectre.

Correction de l'effet dû à la variation de la température :

Après le jour 1, alors que la température augmente, le modèle corrigé par DOP est beaucoup plus robuste que le modèle brut, comme le montre la figure 5.13. Un léger biais positif subsiste jusqu'au jour 3, ce qui cause la décroissance du critère $R_C(t)$ jusqu'à des valeurs inférieures à 1 mais plus proches de 1 que celles du modèle brut. Il est intéressant de constater que la correction apprise au point 5, c'est à dire pendant la phase d'augmentation de la température continue à être efficace lors de la phase de refroidissement. De même, le modèle est parfaitement opérationnel en fin de fermentation, alors que la température est redevenue conforme aux conditions standard, c'est à dire que la source des problèmes de robustesse a disparu. C'est un des avantages revendiqués de DOP.

Comme le montre le tableau 5.2, la dimension de l'espace corrigé par DOP augmente

ε	v% pour choisir k					
	80%	90%	95%	99%	99.9%	100%
0.0001	4.7	4.70	4.70	6.10	3.04	4.07
0.00016	4.7	4.73	4.73	6.12	3.08	5.72
0.00026	4.4	4.41	4.41	4.41	3.9	3.43
0.00039	0.17	0.17	0.18	0.17	0.17	0.24
0.00063	0.17	0.17	0.17	0.17	0.17	0.24
0.001	0.17	0.17	0.17	0.17	0.17	0.24
0.0016	0.17	0.17	0.17	0.16	0.16	0.24
0.0025	0.17	0.17	0.17	0.16	0.16	0.24
0.0039	0.17	0.17	0.17	0.16	0.16	0.24
0.0063	0.1	0.174	0.17	0.16	0.16	0.23
0.001	0.18	0.18	0.20	0.19	0.19	0.22
0.015	0.2	0.20	0.21	0.19	0.19	0.23
0.026	0.2	0.23	0.22	0.19	0.19	0.22
0.039	0.3	0.31	0.29	0.26	0.26	0.18
0.063	0.4	0.36	0.32	0.29	0.29	0.18
0.01	0.5	0.46	0.40	0.40	0.37	0.24
0.15	1.0	0.61	0.61	0.50	0.49	0.32
0.26	1.5	0.99	0.81	0.66	0.66	0.58
0.39	1.5	1.54	1.25	0.91	0.91	0.84
0.63	2.2	2.20	2.20	1.34	1.01	0.95
1	3.52	3.51	3.51	2.03	1.51	1.49

TAB. 5.1 – Variation du SEP global en fonction de ε et du pourcentage de variance v pour sélectionner k.

Point de contrôle	1	2	3	4	5
k	1	1	1	2	2

TAB. 5.2 – Dimension du sous espace corrigé (k) pour chaque point de recalage.

avec le temps. Au point 5, 2 composantes sont nécessaires pour corriger, ce qui montre que l'effet à éliminer est différent du précédent. Cet effet est visible sur la figure 5.16.

- Les vignettes a1 et a2 montrent les spectres d'étalonnage respectivement brut et après application de DOP au point 5 de contrôle. Ces spectres pris en conditions isothermes à 25°C, à différents moments de la fermentation correspondent à des degré d'alcool variant entre 0.3% et 11.7% vol. d'alcool.
- Les vignettes b1 et b2 montrent les spectres en ligne respectivement bruts et après application de DOP au point de contrôle 5. Ces spectres sont pris dans les conditions anisothermes (T=25°C-34°C).

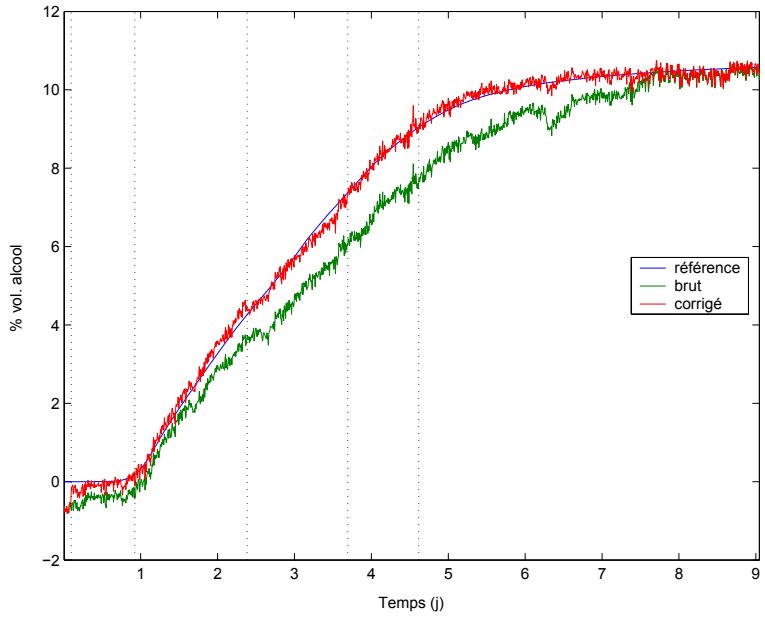

FIG. 5.13 – Prédictions du modèle brut (vert) et du modèle corrigé par DOP (rouge), sur la fermentation anisotherme ; température variant de 25^{o}C à 34^{o}C.

La comparaison des spectres de a1 et b1 du début de la fermentation, (0.2% en a1 et 0.3% en a2) montre que ces deux spectres mesurés presque à la même température (25°C et 26°C resp.) ont des profils différents. Les spectres de b1 sont beaucoup plus "'écrasés"'. Ceci explique les problèmes de robustesse. La différence entre les spectres est cependant difficile à interpréter car les moûts utilisés pour les deux fermentations sont issus de la même solution-mère. La figure 5.16(a1), montre que le premier facteur de variation des spectres n'est pas dû à un effet chimique (p.e. la teneur en alcool) de la fermentation, mais plutôt à l'effet de la variation d'un facteur physique sans doute lié à la turbidité du milieu de fermentation. Cette turbidité est fonction de l'activité des levures comme déjà expliqué dans le paragraphe 5.2.5.2. Ceci explique qu'à la fin de la fermentation (à partir de 8.5% vol. d'alcool, soit vers 5 jours) ce biais ou décalage des spectres commence à diminuer. On remarque aussi que le minimum du spectre présente un très faible déplacement vers les grandes longueurs d'onde, conjoint avec le décalage vertical. Ce déplacement traduit donc la conséquence "chimique" (sur les fréquences de vibration des OH) d'un phénomène lié à l'activité des microorganismes.

La figure 5.16-(b1) montre les spectres acquis en ligne dans les conditions aniso-

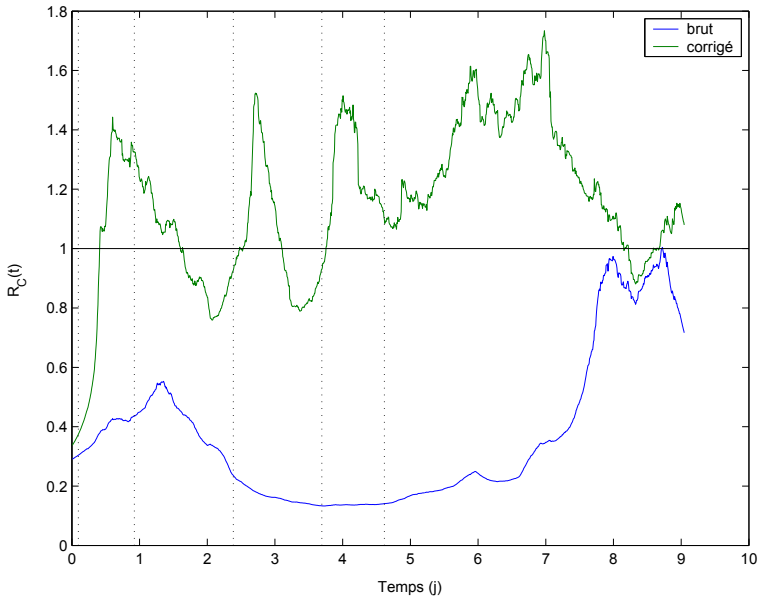

FIG. 5.14 – Comparaison de la robustesse du modèle brut (ligne bleu) et du modèle corrigé par DOP (ligne en vert), en utilisant le critère de robustesse dynamique $R_C(t)$.

thermes (25°C à 34°C), pris à différents moments de la fermentation. Dans cette figure on voit l'évolution ordonnée des spectres en fonction de la température mais pas en fonction du degré d'alcool. De plus, le minimum des spectres présente un shift qui évolue en fonction de la température : entre 26°C et 33°C il y a un shift de 25nm qui s'annule quand la température revient à 25°C. Ces spectres montrent donc un double phénomène : un décalage de la ligne de base et un déplacement du minimum qui sont à la base du problème de robustesse du modèle. Ce double phénomène explique qu'à partir du point de contrôle 4, on ait besoin de 2 dimensions pour corriger. En effet, c'est entre les points de contrôle 3 et 4 que se situe l'apparition du décalage.

Dans la figure (a2) on a les spectres de l'étalonnage corrigé par DOP au point 5, ce qui correspond à $\mathbf{X}_0^*$. Le minima de ces spectres est bien fixe à 1462nm. A gauche, la figure (b2) représente les spectres acquis en ligne corrigés par DOP au point 5 ; leur minimum est aussi centré à 1462nm et ils ressemblent au spectre d'étalonnage, ceci est expliqué par le fait que DOP transforme les spectres en ligne pour qu'ils ressemblent à ceux de la base d'étalonnage.

Les b-coefficients, calculés après application de DOP, sont montrés en figure 5.17. On remarque qu'ils sont beaucoup plus réguliers que ceux obtenus sur les spectres

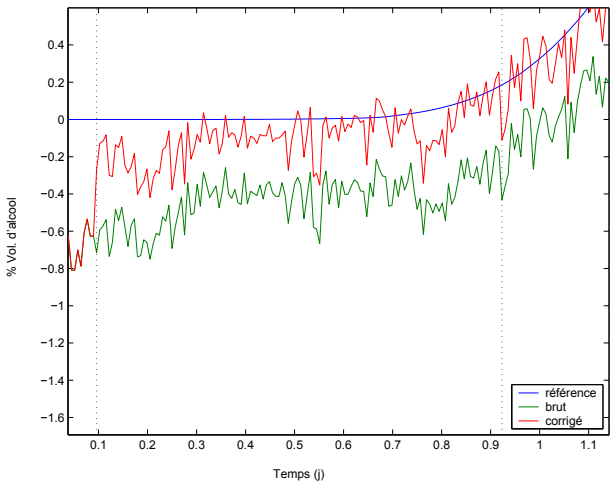

FIG. 5.15 – Prédictions du modèle brut et du modèle corrigé par DOP, sur le début du process : mesures de référence (bleu), prédictions par le modèle brut (vert) ; prédictions par le modèle corrigé par DOP (rouge).

bruts 5.10. Les points caractéristiques de ces b-coefficients sont 1410nm et 1440nm. Le point 1410nm correspond à une zone d'écartement maximal des spectres. Cette zone est caractéristique des alcools [111]. Le pic négatif à 1440 nm est également intéressant : il correspond à un épaulement qui marque la disparition des sucres.

Conclusion

Cette application nous a permis de montrer la capacité de DOP à éliminer les effets des variations des facteurs d'influence physiques : variation de température et effet de batch (turbidité).

De plus, cette application a permis de révéler les potentialités offertes par DOP pour aider les spectroscopistes à analyser le processus et en particulier l'effet de la grandeur d'influence. Elle a permis de confirmer la facilité d'utilisation de la méthode.

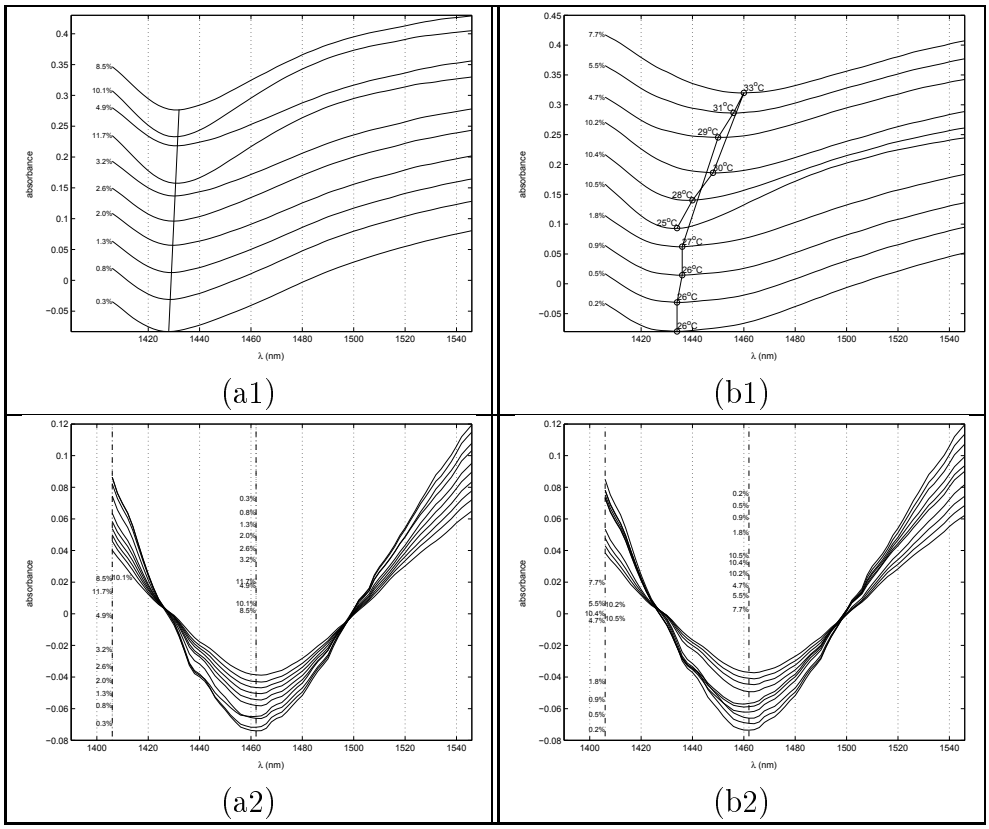

FIG. 5.16 – (a1 et a2) spectres d'étalonnage dans les conditions isothermes (T=25°C) : (a1) spectres brut, (a2) après application de DOP au point 5 de contrôle. (b1 et b2) spectres en ligne dans les conditions anisothermes (T=25°C-34°C) : (b1) spectres en ligne bruts, (b2) après application de DOP du point 5 de contrôle. La ligne verticale des figures a1 et b1 indique la variation de l'abscisse des minima.

5.3 Deuxième application : Cas des facteurs d'influence chimiques

Une deuxième application a été réalisée, permettant d'étudier les performances de la méthode DOP en cas de présence des facteurs d'influence de nature chimique. Il s'agit de suivre en ligne un processus continu de fermentation de vinasses par spectrométrie MIR. Durant ce processus plusieurs ajouts d'azote ont eu lieu, sous forme d'ammoniaque.

5.3.1 Matériels et méthodes

La base de données utilisée provient du contrôle en ligne par spectrométrie MIR de l'activité d'un bioréacteur de traitement des effluents d'une cave viticole. Cette

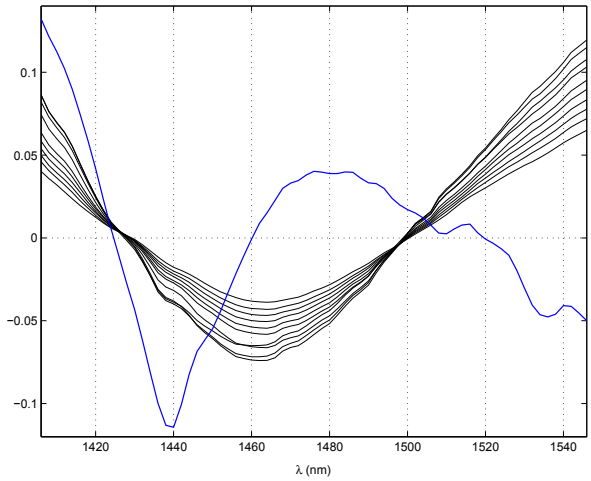

FIG. 5.17 – Comparaison entre les spectres corrigés $\mathbf{X}_0^*$ et le vecteur du b-coefficient du modèle DOP-corrigé au point 5 avec 2 LV

base a été fournie par le laboratoire de biotechnologie et de l'environement (LBE) de l'INRA de Narbonne. Les mesures spectrales sont acquises en ligne, à raison de 1 mesure/30minutes, à l'aide d'un spectromètre FT-IR du type Nicolet (Avatar 380FT-IR), en mode de transmission utilisant une cellule de mesure du type CaF_2 de 50 μm de trajet optique, dans le domaine du moyen infrarouge ($3000cm^{-1}$ à $1000cm^{-1}$) par intervalle de $3.8cm^{-1}$ (figure-5.18). L'objectif est de suivre par spectrométrie MIR la variation de la teneur en solutés organiques du bioréacteur durant tout le processus du traitement [138]. Pour notre application, seule la teneur du bioréacteur en acide gras volatil (AGV) est suivie.

Un modèle d'étalonnage PLS, intégré au logiciel du spectromètre, est utilisé en ligne pour prédire les AGV ; en parallèle une autre mesure de référence par titrimétrie est prise en ligne. Durant le processus, différents évènements ont eu lieu, ils sont résumés dans le tableau-5.3.

L'ammoniaque est considéré comme le facteur d'influence chimique qui affecte les prédictions des AGV car il absorbe dans la même plage de longueur d'onde que celle sélectionnée pour prédire les AGV, d'où l'intérêt d'étudier l'application de DOP pour réduire son effet.

Temps (jours)	Évènements	Effet sur la prédiction
1	Addition de NH_4OH (C_1)	faible biais
2		Dérive
4	Ajout de $NaOH$	Biais
6	Addition de NH_4OH ($C_2 > C_1$)	Dérive + biais
9	Alimentation en eau	Dérive + biais
10	Ré-alimentation en vinasse	
11		Biais
12	Arrêt du bioréacteur	
13	Redémarrage	Biais
14	Addition de NH_4OH ($C_3 > C_2$)	Dérive + biais

TAB. 5.3 – Bilan des évènements qui ont eu lieu en ligne au cours du processus. C_1, C_2 et C_3 représentent la quantité de composé ajoutée.

5.3.2 Modélisation

5.3.2.1 Base de données

La base de données acquise en ligne contient d'une part les mesures spectrales MIR : $\mathbf{X_2}_{(616\times520)}$ et d'autre part les mesures de référence du taux des AGV correspondantes : $\mathbf{y}_{2_{(616\times1)}}$. Deux points de recalage ont été choisis, comme représentés sur la figure 5.21. Le premier point a été pris au début du procédé, pendant la première addition de NH_4OH ; le second juste après l'ajout (plus important) de NH_4OH du jour 6.

5.3.2.2 Prétraitements

Les données brutes ont été utilisées sans centrage car les spectres en transmission ont été corrigés de la ligne de base par rapport au spectre de l'eau. La plage $1002cm^{-1} - 1576cm^{-1}$ a été sélectionnée pour construire le modèle d'étalonnage, conformément à l'expertise des personnes responsables du capteur MIR.

5.3.2.3 Sélection de la base de données

La base de données fournie pour cette application ne contient pas de réelle base d'étalonnage. On dispose uniquement de la base de données $\mathbf{X}_{2(616\times520)}, \mathbf{y}_{2(616\times1)}$ de suivi en ligne. Pour pouvoir appliquer la méthode DOP on a besoin d'une base d'éta-

lonnage représentative du procédé. Comme les données acquises pendant les jours 10 et 11 sont exemptes de toute variation de grandeur d'influence, elles vont nous servir de base d'étalonnage $\mathbf{X}_{02(100\times150)}$, $\mathbf{y}_{02(100\times1)}$. Les spectres de la figure 5.18 confirment bien l'abscence d'évènements majeurs[4] Cette base correspond à la période de réalimentation du processus en vinasse, pendant laquelle la solution du bioréacteur est exempte de NH_4OH. De plus, cette plage renferme une grande variation de concentration en AGV [276 mg/l - 1570 mg/l], alors qu'au cours du procédé complet, la teneur en AGV varie entre 250 et 2100 mg/l, comme l'indique la figure 5.21.

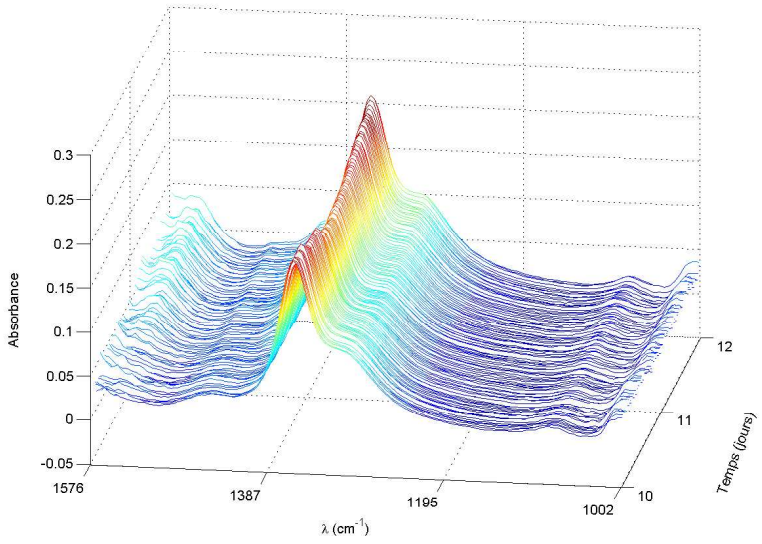

FIG. 5.18 – Spectres Moyen infrarouge de la base d'étalonnage, acquis en ligne pour le suivi du processus de mesures pendant la période du 10-11 ^{eme} jours $n_{02} = 100$.

5.3.2.4 Étalonnage

L'étalonnage est réalisé par une PLS, sans centrage. Une validation croisée leave-one-out a été réalisée sur $\mathbf{X}_{02}$, $\mathbf{y}_{02}$. L'examen de l'évolution des erreurs de validation croisée et d'étalonnage a permis de choisir le nombre de variables latentes. A ce nombre de variables latentes, correspond un pourcentage de variance de y expliquée par le modèle V_y, qui a été ensuite utilisé en ligne pour régler le ré-étalonnage du modèle PLS.

[4]Bien entendu, un tel choix est démonstratif et est dû aux données disponibles. Dans un cas réel la base d'étalonnage est constituée avant toute expérience.

5.3.2.5 Réglage de DOP

Les paramètres adoptés pour DOP sont :

- $\varepsilon = 0.05$. La valeur donnée par le calcul théorique, pour une distribution uniforme de $\mathbf{y}_0$ (ce qui n'est pas notre cas ici) est $1/100 = 10^{-2}$. Nous avons donc volontairement adopté une valeur supérieure (tout en conservant l'ordre de grandeur).
- $v = 90\%$.

Comme aucun jeu de test n'est disponible, la valeur de SEP_0, utilisée dans le calcul du critère d'évaluation de la robustesse R_C, est inconnue. Elle sera remplacée par la valeur de l'erreur de cross validation. Le critère de robustesse temporelle a été calculé sur une fenêtre glissante de largeur égale à un jour.

5.3.3 Résultats

5.3.3.1 Modèle d'étalonnage

La figure 5.19 montre le résultat de la validation croisée. Quatre variables latentes ont été sélectionnées pour construire le modèle. La prédiction de la validation croisée correspondante est donnée par la figure 5.20. On adopte donc $SEP_0 = 67.9 mg/l$.

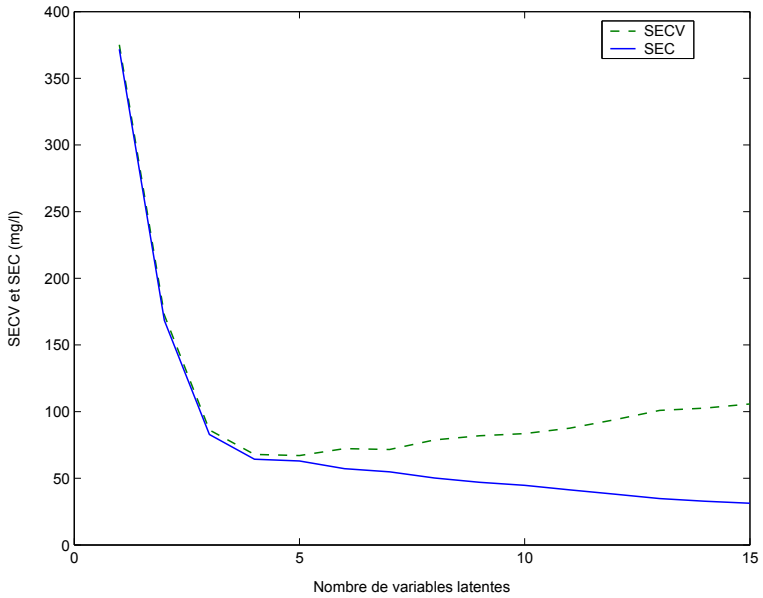

FIG. 5.19 – Evolution des erreurs d'étalonnage et de validation croisée.

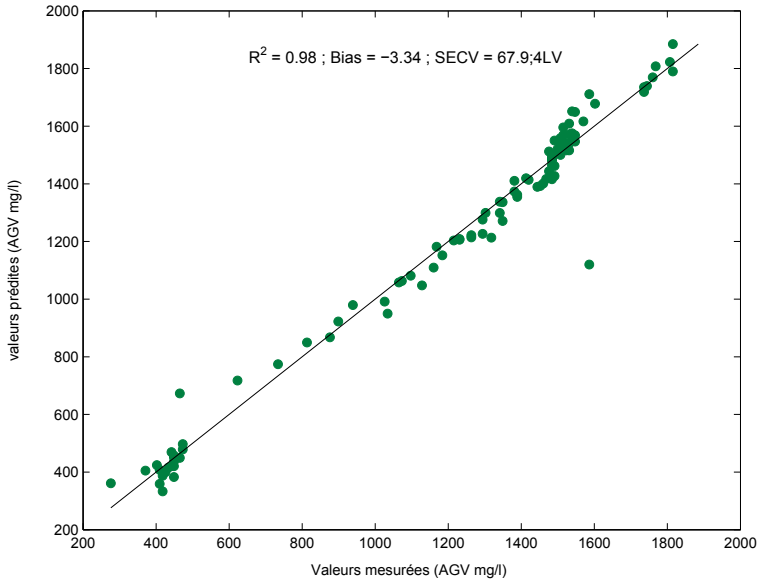

FIG. 5.20 – Prédiction de la teneur en AGV par validation croisée (4 variables latentes).

5.3.4 Application de DOP pour le suivi en ligne

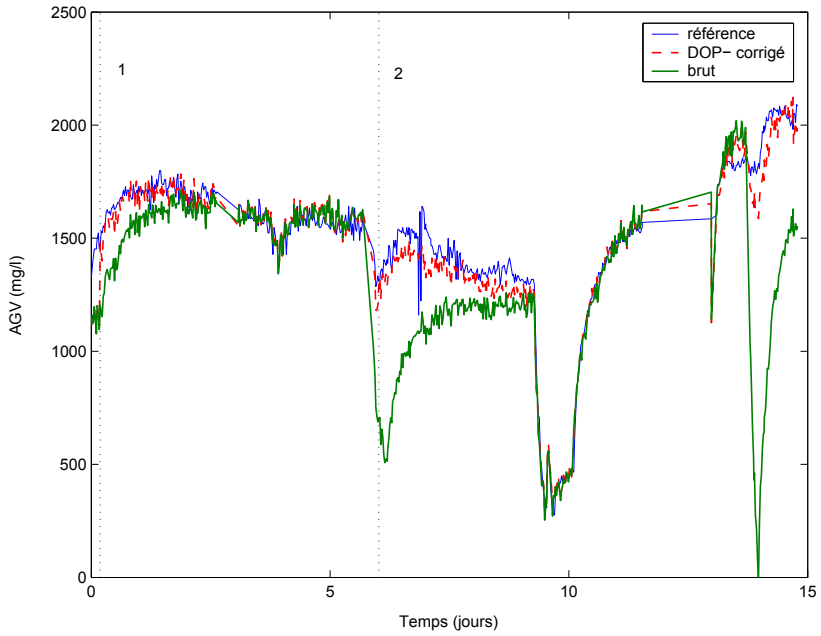

FIG. 5.21 – Prédictions du modèle brut et du modèle corrigé par DOP sur 15 jours de procédé, soumis à des additions de NH_4OH.

Le modèle brut présente des problèmes de robustesse vis-à-vis de l'addition de NH_4OH, comme le montre la figure 5.21. Au début du processus, un biais dû à la

présence de faibles teneurs en ammoniaque apparaît. Ceci se traduit sur la figure 5.22 par une faible valeur du critère de robustesse R_C, de l'ordre de 0.25 (i.e. $SEP = 4 \times SEP_0$). Lors du deuxième ajout d'ammoniaque, plus conséquent que le premier, à $t = 6$ jours, la dérive est très forte. Un biais de près de $-1000mg/l$ apparaît. Le critère de robustesse plonge vers des valeurs proches 0.1. Pendant la période des jours 10 et 11, l'erreur est quasi nulle, car il s'agit de l'intervalle ayant servi à l'étalonnage. Ensuite, au jour 14, lors de l'ajout massif d'ammoniaque, le même genre de décrochage négatif survient, avec une amplitude encore plus importante. Le biais atteint $-1500mg/l$ et le critère de robustesse atteint 0.05.

Le modèle corrigé par DOP est beaucoup plus robuste. Dès le premier point de recalage, au bout de $t = 5$h, l'influence spectrale de l'ammoniaque est apprise et éliminée. Jusqu'au deuxième ajout d'ammoniaque, au jour 6, la prédiction reste proche de la référence, avec un indice de robustesse supérieur à 1 [5]. Le deuxième point de recalage survient au moment de la deuxième addition d'ammoniaque, qui n'a qu'un très léger effet sur les prédictions. Un biais négatif subsiste et R_C passe sous la barre de 1, pour descendre jusqu'à environ 0.55. Comme pour le modèle brut, la période ayant servi à l'étalonnage est exempte d'erreur. Lors du troisième ajout d'ammoniaque (jour 14), on retrouve de nouveau un léger biais négatif et une valeur de R_C voisine de 0.55.

Diagnostic en ligne, par l'examen des résidus L'application de DOP, à chaque point de recalage, a pour effet d'éliminer un sous espace jugé nuisible, par projection orthogonale. Le projecteur sur ce sous espace peut être utilisé en chaque point du procédé, pour visualiser en temps réel le spectre non pris en compte par le modèle. Cela permet de visualiser les effets spectraux éliminés, sous forme d'un spectre à chaque point de mesure.

Ceci constitue donc une surface, comme visualisée dans la figure 5.23. Les deux fronts blancs correspondent aux points de recalage, instants de rédéfinition du projecteur, donc de discontinuité. Il est intéressant de noter le pic, apparaissant deux fois à $t = 6$ et $t = 14$ jours dans la zone de 1460 cm^{-1}, typique de l'ammoniaque.

[5] Rappelons ici que ce critère est optimiste, puisqu'il est calculé par rapport au $SECV_0$ et non au SEP_0.

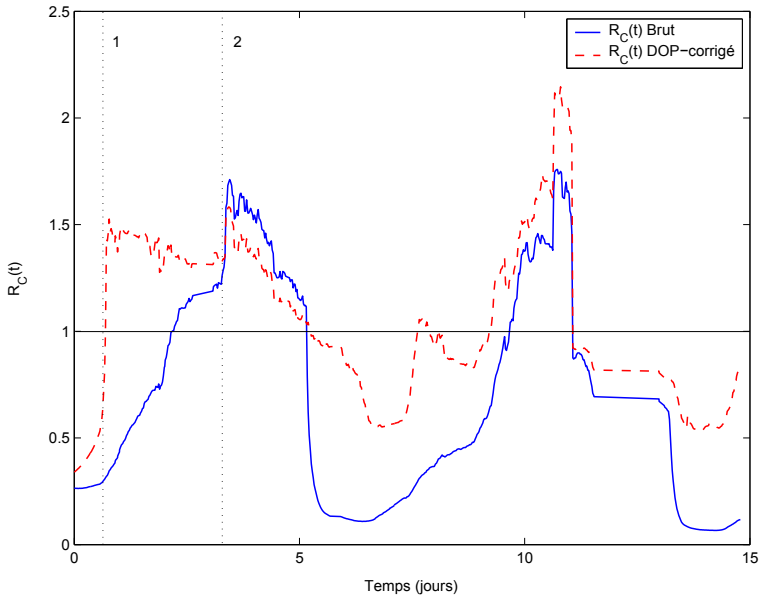

FIG. 5.22 – Évolution du critère de robustesse en fonction du temps.

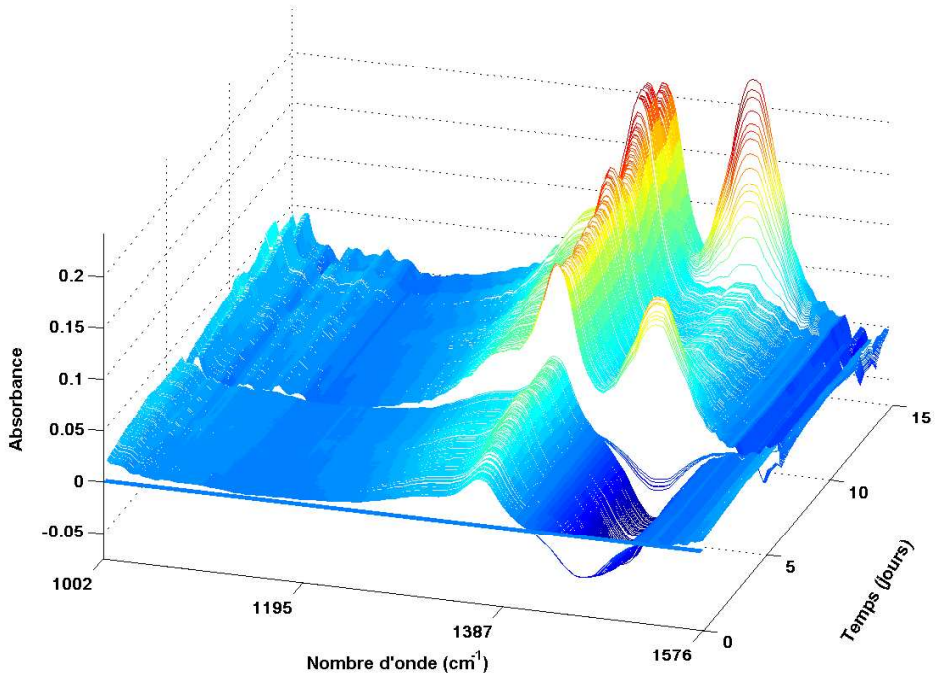

FIG. 5.23 – Spectres résidus, éliminés par la projection orthogonale de DOP.

5.3.4.1 Conclusion

Cette deuxième application a permis de confirmer la généricité de la méthode vis-à-vis (i) d'un autre type de spectroscopie (MIR au lieu de NIR), (ii) d'un autre type

de perturbation (chimique au lieu de physique). La facilité d'utilisation est également confirmée.

5.4 Conclusion

Ce chapitre a présenté deux applications de la méthode DOP à des cas de suivi en ligne réels. La première s'est intéressée à une grandeur d'influence physique : la température de la solution dans un procédé de fermentation alcoolique. Dans, la seconde les grandeurs d'influence chimiques ont été abordées, à travers le cas de l'addition d'ammoniaque dans un réacteur de dépollution anaérobie. Les différentes étapes de construction du modèle d'étalonnage ont été décrites pour chaque application ainsi que le choix des paramètres de la méthode DOP. L'intérêt de disposer d'une base d'étalonnage de distribution uniforme a été souligné et dans ce cadre, une règle de choix du paramètre de largeur de la fonction de noyau a été démontrée d'un point de vue théorique et validé par les tests sur la première application. La sensibilité de DOP au deuxième paramètre de recalage, à savoir le pourcentage de variance capturée par l'ACP lors de l'identification de la base du sous espace à éliminer, a été évaluée au travers de tests. Il semble que cette sensibilité soit faible et que donc, une grande plage de valeurs, de 90 à 99%, convient.

A travers ces deux applications, nous avons démontré de manière spectaculaire les avantages de la méthode DOP. En effet, dans la première application, la robustesse a été améliorée par un facteur 5 sur l'ensemble du procédé.

Dans la première application nous avons illustré comment, en réduisant les dimensions, DOP pouvait rendre un ensemble de spectres plus lisibles pour le spectroscopiste et les b-coefficients plus lisses et faciles à interpréter. Dans la deuxième, toutes les erreurs n'ont pas pu être complètement éliminées (un léger biais subsistait), mais la robustesse a malgré tout été améliorée par un facteur 4.

L'examen, pour chaque point du process, des résidus éliminés par la méthode DOP, obtenus par projection des spectres sur $\mathbf{P}$, permet d'autre part d'interpréter les grandeurs d'influence éliminées.

Toutes ces qualités de DOP permettent son utilisation pour différentes applications dans différents domaines.

Chapitre 6

Conclusion Générale et Perspectives

Cette thèse est une contribution à la maintenance d'une mesure robuste en ligne par spectrométrie infrarouge et étalonnages multivariés. Elle propose une nouvelle méthodologie qui rend possibles les mesures en ligne en conditions variables et en particulier l'utilisation de la spectroscopie PIR couplée à des fibres optiques pour le suivi en ligne des processus continus.

Un bilan ...

Pour aboutir à cet objectif nous avons traité trois questions scientifiques dont chacune a fait l'objet d'un chapitre de cette thèse. Dans le chapitre 2, nous avons répondu à la question : 'qu'est-ce que la robustesse des étalonnages multivariés et comment la mesure-t-on ?'. Nous avons trouvé une diversité dans les définitions données au terme "robustesse" selon les domaines d'application. Nous sommes dans le cas des mesures en ligne par SPIR où les effets des variations des facteurs d'influence inconnus et non contrôlés se manifestent par une déformation des mesures spectrales et donc par une erreur sur les prédictions du modèle d'étalonnage. Dans ce contexte, nous avons ainsi défini la robustesse des modèles d'étalonnage multivariés par "la stabilité de sa capacité prédictive (SEP) vis-à-vis des perturbations appliquées au voisinage des conditions standards". En se basant sur cette définition nous avons défini un critère de robustesse

et sa variante temporelle, qui permet d'évaluer la robustesse du modèle d'étalonnage utilisé en ligne. Ce chapitre a fait l'objet d'une publication sur la '*définition et l'évaluation de la robustesse des méthodes d'étalonnage multivariés*', publié dans le journal Trends in Analytical Chemistry (TrACs) [185].

Dans le chapitre 3, nous nous sommes intéressés à l'étude des différentes méthodes d'amélioration du modèle d'étalonnage multivarié pour répondre à la deuxième question : "Comment améliorer la robustesse du modèle d'étalonnage multivarié ?". Les méthodes étudiées sont surtout celles utilisées, lors du développement du modèle, pour améliorer sa performance dans des conditions contrôlées. Elles consistent en l'optimisation de la base d'étalonnage (élimination des données aberrantes, sélection du lot des échantillons d'étalonnage représentatifs, centrage et réduction), le prétraitement géométrique des données spectrales (normalisation, lissage et différenciation), la réduction des dimensions de l'espace d'étalonnage (sélection de variable et projection orthogonale). La part de la contribution de chacune de ces méthodes à l'amélioration de la robustesse a été étudiée en se basant sur une expression géométrique des problèmes de robustesse. Cette étude nous a montré qu'une base d'étalonnage optimale, représentative du domaine expérimental, était nécessaire pour le développement d'un modèle d'étalonnage optimal. Nous avons proposé d'appliquer l'algorithme DUPLEX ou Kennard et Stone pour effectuer le choix du lot optimal en se basant sur les mesures de référence (au lieu des spectres), car elles ont l'avantage d'avoir moins de risque de données aberrantes. Nous avons conclu que les méthodes de sélection de variables et d'orthogonalisation contribuent le plus à l'amélioration de la robustesse des prédictions d'un modèle d'étalonnage en sélectionnant l'espace le plus approprié à l'étalonnage indépendant de ceux contenant les variations des facteurs d'influence. Ce travail a été traduit par une publication, faisant suite à la première, soumise au journal Trends in Analytical Chemistry TrACs [184].

Dans le chapitre 4, pour répondre à la troisième question posée : "Comment maintenir en ligne la robustesse du modèle d'étalonnage multivarié ?", nous avons proposé une nouvelle méthode intitulée DOP, pour Dynamic Orthogonal Projection. La méthodologie de DOP nécessite seulement que quelques points de contrôle soient disponibles.

Cette méthode consiste, tout d'abord, à utiliser les données disponibles en ligne (base d'étalonnage, quelques mesures de contrôle) pour estimer des "*standards virtuels*", utilisant une fonction de noyau gaussien. Ces standards sont les spectres que l'on aurait dû avoir en abscence des perturbations. Ils sont utilisés avec les spectres correspondants pour effectuer le transfert de l'étalonnage des conditions standards vers celles en ligne. Les performances de la méthode DOP vis-à-vis de l'élimination des facteurs d'influence (physiques et chimiques), ont été testées sur deux applications différentes de suivi de procédés continus dans le chapitre 5 : les résultats de l'utilisation de DOP pour maintenir la robustesse du modèle d'étalonnage en présence des variations des facteurs physiques (température) et chimiques (addition d'ammoniaque) ont été présentés. Nous avons conclu sur la simplicité, l'efficacité et la rapidité de la méthode DOP. Dans le premier exemple l'effet de différence entre les batches a été éliminé en utilisant DOP, de même que l'effet des variations de température. Ceci a montré la capacité de DOP à gérer plusieurs effets en même temps. Le second exemple nous a montré la capacité de DOP à éliminer l'effet d'un facteur d'influence chimique, absorbant dans la même plage de longueur d'onde que le composé d'intérêt. De plus, DOP fournit de l'information en ligne sur les causes d'erreur, permettant ainsi de faire le diagnostic et de mieux comprendre le processus de mesure afin de l'améliorer en temps réel. Dans ce chapitre, nous avons aussi proposé un développement théorique permettant de définir la largeur du noyau utilisé en fonction du nombre d'échantillons dans la base d'étalonnage, ce qui facilite l'implémentation de DOP. Cette méthode, validée par l'expérience, a été proposée dans une publication qui a été soumise récemment à Journal of Chemometrics [183].

En résumé, les avantages de DOP sont :

- l'utilisation continuelle de la base d'étalonnage pour les futures prédictions en ligne ;
- la correction embarquée du modèle d'étalonnage vis-à-vis des différentes sources inconnues de perturbations d'ordre physique ou chimique ;
- sa capacité de gérer l'effet de différents facteurs d'influence en même temps ;

– l'information fournie, qui permet de faire un diagnostic en temps réel sur le processus ;
– son implémentation facile et pratique (seulement deux), avec un temps de calcul de l'ordre de quelques secondes ;

... et des perspectives

Les résultats obtenus en utilisant DOP ouvrent des perspectives intéressantes :

– Dans cette thèse nous avons choisi de gérer très simplement l'historique des mesures en conservant tous les points de contrôle. Ceci est acceptable pour les procédés "courts". En revanche, une stratégie plus complexe devra être mise au point pour les procédés plus "longs" à échelle industrielle : doit-on garder tous les points de contrôle ?, si oui, quels poids leur donner ?,...
– Des travaux complémentaires peuvent être menés sur les points de recalage. Pour l'instant nous avons défini les instants des points de recalage en ligne *a priori*, en considérant que ce sont des points réguliers peu fréquents. Mais, il sera intéressant de développer une méthode pour alerter l'expérimentateur sur la nécessité d'une mesure de référence, afin de pouvoir intervenir le plus tôt possible sur le recalage pour éviter la dérive. Faire appel à des outils automatiques semble être très utile à ce niveau là.
– Il sera intéressant d'étudier l'application de DOP aux problèmes de traitement d'autres types de données telles que les images multi-canales, voire hyperspectrales, pour lesquelles certains pixels peuvent être utilisés comme points de recalage (application en télédétection).
– La méthode DOP peut être étendue à la résolution des problèmes de transférabilité des modèles d'étalonnage multivariés entre différents instruments de mesures, différents batches, utilisant différentes variétés d'échantillons, différentes campagnes de mesures ou pendant différentes années. Elle permet de corriger le modèle d'étalonnage et de le rendre indépendant des conditions ayant eu lieu en temps réel. Il n'est plus nécessaire de refaire un modèle d'étalonnage, il suffit de quelques mesures de référence pour corriger la base d'étalonnage initiale afin de pouvoir l'utiliser en temps réel.

Ainsi les prolongements scientifiques de cette thèse, ainsi que les applications potentielles sont multiples. Nous ne pouvons que souhaiter qu'elle permettra d'encourager le développement de la spectrométrie infrarouge dans des applications industrielles.

Bibliographie

[1] L. Abdel-Malek. Test de robustesse. *Comett Euro Training Course In Advanced HPLC and Capillary Electrophoresis*, 1993. 43

[2] V. Acha, M. Meurens, H. Naveau, and S. N. Agathos. Atr-ftir sensor development for continuous on-line monitoring of chlorinated aliphatic hydrocarbons in a fixed-bed bioreactor. *Biotechnology and Bioengineering*, 68(5) :473–487, 2000. 20

[3] S. Addelman. *Technometrics*, 4 :21–46, 1962. 43

[4] AFNOR. In *Normes fondamentales : Vocabulaire international des termes fondamentaux et généraux de métrologie*, volume NFX07-001. 1994. 31, 32

[5] C. A. Andersson. Direct orthogonalization. *Chemometrics and Intelligent Laboratory System*, 47(1) :51–63, 1999. 66

[6] A. Andrew and T. Fearn. Transfer by orthogonal projection : making near-infrared calibrations robust to between instrument variation. *Chemometrics and Intelligent Laboratory Systems*, 72 :51–56, 2004. 76

[7] K. N. Andrew, S. C. Rutan, and P. J. Worsfold. Application of kalman filtering to multivariate calibration and drift correction. *Analytica Chimica Acta*, 388(3) :315–325, 1999. 86

[8] ASTM98. *Standard practices for infrared, multivariate, quantitative analysis (E 1655-97)*. ASTM Annual Book of Standards Vol.03.06, ASTM, west Conshohoken, 1998. 20, 22

[9] T. Azzouz, A. Puigdomenech, M. Aragay, and R. Tauler. Comparison between different data pre-treatment methods in the analysis of forage samples using near-

infrared diffuse reflectance spectroscopy and partial least-squares multivariate calibration method. *Analytica Chimica Acta*, 484(1) :121–134, 2003. 74

[10] P. Barak. Smoothing and differentiation by an adaptive-degree polynomial filter. *Analytical Chemistry*, 67 :2758–2762, 1995. 69, 70

[11] R.J Barnes, M.S. Dhanoa, and J. Lister Susan. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5) :772–777, 1989. 66, 67

[12] A. Barros. *Contribution à la sélection et la comparaison de variables caractéristiques*. PhD thesis, Institut National Agronomique Paris-Grignon Labratoire de chimie analytique, 1999. 21, 81, 92

[13] M. Bely. *Detection automatique et correction des carences en azote assimilable des fermentations alcooliques en conditions oenologiques : Etude cinétique et approche physiologique*. PhD thesis, Université de MontpellierII, 1990. 107

[14] M. Bely, J-M. Sablayrolles, and Barre P. Description of the alcoholic fermentation kinetics :its variability and significance. *Am. J. Enol.Vitic.*, 41 :319–324, 1990. 106, 107

[15] S. Bettencourt, J. N. Ricardo, M. Camoes, G. F. C. Filomena, and B. J. Seabra. Validation and quality control schemes based on the expression of results with uncertainty. *Analytica Chimica Acta*, 393(1-3) :167–175, 1999. 32

[16] M. Blanco and D. Serrano. On-line monitoring and quantification of a process reaction by near-infrared spectroscopy. catalysed esterification of butan-1-ol by acetic acid. *Analyst*, 125 :2059–2064, 2000. 21

[17] M. Blanco and D. Serrano. On-line monitoring and quantification of a process reaction by near-infrared spectroscopy. catalysed esterification of butan- 1-ol by acetic acid. *Analyst*, 125(11) :2059–2064, 2000. 34

[18] de J.H. Boer. *Chemometrical Aspects of Quality in Pharmaceutical Technology. The Application of Robustness Criteria and Multicriteria Decision making in Optimization Procedures for pharmaceutical formulations*. PhD thesis, Groningen, 1992. 43

[19] D. Bosq and J. P. Lecoutre. Theorie de l'estimation fonctionnelle. *Economica*, 1987. 88

[20] M. Bounou, S. Lefebvre, and X. D. Do. Improving the quality of an optimal power flow solution by taguchi method. *International Journal of Electrical Power & Energy Systems*, 17(2) :113–118, 1995. 50

[21] E. Bouveresse and D. L. Massart. Standardisation of near-infrared spectrometric instruments : A review. *Chemometrics and Intelligent Laboratory Systems*, 11 :3–15, 1996. 86, 89

[22] G.E.P. Box. Non-normality and tests on variances. *Biometrika*, 40 :318–335, 1953. 35

[23] G.E.P. Box. Signal-to-noise ratios, performance criteria and transformations. *Technometrics*, 30 :1–17, 1988. 50

[24] G.E.P. Box and D.W. Behnken. Some new three-level designs for the study of quantitative variables. *Technometrics*, 2 :455–475, 1960. 42

[25] R. G. Brereton. Introduction to multivariate calibration in analytical chemistry. *Analyst*, 125(11) :2125–2154, 2000. 20

[26] R. Bro and A. K. Smilde. Centereing and scaling in component analysis. *Journal of chemometrics*, 17 :16–33, 2003. 62, 71

[27] J. Caporal-Gautier, J.M. Nivet, P. Algranti, M. Guilloteau, m. Lallier, J.J. N'Guyen-Huu, and R. Russotto. Guide de validation analytique, rapport d'une commission sfstp. *STP Pharma Pratiques*, 2 :205–239, 1992. 33, 37

[28] V. Centner, D. L. Massart, and O. E. de Noord. Detection of inhomogeneities in sets of nir spectra. *Analytica Chimica Acta*, 330(1) :1–17, 1996. 79, 80

[29] United States Pharmacopeial Convention. The united states pharmacopeia xxii, the national formulary xvii. Technical report, The United States Pharmacopeia XXII, 1990. 33, 37, 47

[30] T.H.E. Cottrell and M.R. Mc Lellan. The effect of fermentation temperature on chemical and sensory characteristics of wines from seven white grape cultivars grown in new york state. *Am. J. Enol. Vitic.*, 37 :190–194, 1986. 107

[31] P. F. de Aguiar, B. Bourguignon, M. S. Khots, D. L. Massart, and R. Phan-Than-Luu. D-optimal designs. *Chemometrics and Intelligent Laboratory Systems*, 30(2) :199–210, 1995. 61

[32] J.H. de Boer, A.K. Smilde, and D.A. Doornbos. Introduction of a robustness coefficient in optimization procedures : Implementation in mixture design problems. part i : Theory. *Chemometrics and Intelligent Laboratory Systems*, 7 :223–236, 1990. 53

[33] O. E. de Noord. The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*, 23 :65–70, 1994. 89

[34] P. F. deAguiar, Y. VanderHeyden, and D. L. Massart. Study of different criteria for the selection of a rugged optimum in high performance liquid chromatography optimisation. *Anal. Chim. Acta*, 348(1-3) :223–235, 1997. 45

[35] N. R. Draper and H. Smith. Applied regression analysis. Wiley Series in Probability and Mathematical Statistics, pages 257–265. John Wiley and Sons, New York, second edition, 1980. 62

[36] R. D. Driver, J. N. Downing, M. L. Brubaker, Stark J. D., Vacha L., and T. L. Wilbourn. The influence of data preprocessing on the robustness and parsimony of multivariate calibration models. *Optically Based Methods for Process Analysis*, 1681 :236–249, 1992. 21

[37] C.S. Du Plessis. Influence de la température d'élaboration et de conservation sur les caratéristiques physicochimiques et organoleptiques des vins. *Bull. off. Int. Vin*, 624 :105–115, 1983. 107

[38] N ElHaloui, G. Corrieu, and D. Picque. Method for on-line prediction of kinetics of alcoholic fermentation in wine making. *J. Ferment. Bioeng.*, 68 :131–135, 1989. 110

[39] N.E. ElHaloui, Y. Cleran, J. M. Sablayrolles, P. Grenier, P. Barre, and G. Corrieu. Suivi et contrôle de la fermentation alcoolique en oenologie. *Revue Francaise d'Oenologie*, 115 :12–17, 1988. 110

[40] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikström, and S. Wold. Book review : Design of experiments, principles and applications. *Journal of chemometrics*, 15 :495–496, 2001. 39

[41] EURACHEM. Quantifying uncertainty in analytical measurement. Technical report, EURACHEM/CITAC, 2000. 32

[42] H. Faber and N. Mesplet. Robustness testing for a capillary electrophoresis method using the "short-end injection" technique. *Journal of Chromatography A*, 897(1-2) :329–338, 2000. 42

[43] K. Faber. Robustness testing in liquid chromatography and capillary electrophoresis. *Journal of Pharmaceutical and Biomedical Analysis*, 14 :1125–1132, 1996. 36, 37, 39, 40, 41, 42

[44] T. Fearn. Standardisation and calibration transfer for near infrared instruments : a review. *J. of Near Infrared Spectrosc.*, 9 :299–244, 2001. 89

[45] T. Fearn and A.M.C. Davies. Locally biased regression. *J. of Near Infrared Spectrosc.*, 11 :467–478, 2003. 86

[46] Tom Fearn. On orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 50(1) :47–52, 2000. 75

[47] M. Feinberg. *La validation des méthodes d'analyse : Une approche chimiométrique de l'assurance qualité au laboratoire*. Masson, Paris, 1996. 34, 35

[48] J. Ferre and F. X. Rius. Constructing d-optimal designs from a list of candidate samples. *Trends in Analytical Chemistry*, 16(2) :70–73, 1997. 61, 62

[49] International Organisation for Standardisation (ISO). Accuracy (trueness and precision) of measurement methods and results- part1 :general principles and definitions. *International Organisation for Standardisation*, ISO 5725-1, 1994. 31

[50] R. A. Forbes, M. Z. Luo, and D. R. Smith. Measurement of potency and lipids in monensin fermentation broth by near-infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 25(2) :239–256, 2001. 20

[51] M. Forina, M. C. Casolino, and C. de la Pezuela Martinez. Multivariate calibration : applications to pharmaceutical analysis. *Journal of Pharmaceutical and Biomedical Analysis*, 18(1-2) :21–33, 1998. 86

[52] J. François. *Système multi-capteurs*. PhD thesis, Université de Technologie de Compiègne, Génie Informatique Département : Cemagref - Montpellier, 1996. 53

[53] J. P. Gauchi and P. Chagnon. Comparison of selection methods of explanatory variables in pls regression with application to manufacturing process data. *Journal of Chemometrics and Intelligent Laboratory Systems*, 58(2) :171–193, 2001. 80

[54] P. J. Gemperline, C. JungHwan, B. Baker, B. Batchelor, and D. S. Walker. Determination of multicomponent dissolution profiles of pharmaceutical products by in situ fiber-optic uv measurements. *Analytica Chimica Acta*, 345(1-3) :155–159, 1997. 34

[55] Telecom Glossary. Federal standard 1037c. *http ://www.its.bldrdoc.gov/fs-1037/fs-1037c.htm*, 2000. 34, 44

[56] Golay and Savitzky. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(1627), 1964. 69, 70

[57] P.A. Gorry. General least-squares smoothing and differentiation by the convolution (savitzky-golay) method. *Analytical Chemistry*, 62 :570–573, 1990. 70

[58] J. Goupy. *La Méthode des Plans d'Expériences. Optimisation des choix des essais & de l'interprétation des résultats*. Masson, Paris, 1988. 47

[59] J. Goupy, editor. *Plans d'Expériences pour Surfaces de Réponse*. Technique et Ingénieurie. Dunod, Paris, 1999. 61

[60] J.M. Green. *Anal. Chem.*, 68 :305A–309A, 1996. 34

[61] Q. Guo, W. Wu, and D. L. Massart. The robust normal variate transform for pattern recognition with near-infrared data. *Analytica Chimica Acta*, 382(1-2) :87–103, 1999. 66, 67

[62] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003. 77

[63] P. W. Hansen. Pre-processing method minimizing the need for reference analyses. *Journal of chemometrics*, 15 :123–131, 2001. 76

[64] I. S. Helland, T. Naes, and T. Isaksson. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 29(2) :233–241, 1995. 66

[65] M. W. B. Hendriks, J. H. De Boer, and A. K. Smilde. *Robustness of Analytical Chemical Methods and Pharmaceutical Technological Products*, volume 19 of *Data Handling In Science and Technology*. Elseiver, Amsterdam, 1996. 38, 39, 41, 42, 43, 49, 50, 52, 61

[66] N. Hilgert. *Identification et contrôle de processus autorégressifs non linéaires incertains. Application à des procédés biotechnologiques.* PhD thesis, Université Paris-Sud centre d'orsay, 1997. 88, 92

[67] U.W.E. Hochner and J.H. Kalivas. Simulated-annealing-based optimization algorithms : Fundamentals and wavelength selection applications. *Journal of Chemometrics*, 9 :283–308, 1995. 81

[68] D. E. Honigs, G. M. Hieftje, and T. Hirshfeld. Number of samples and wavelengths required for the training set in near-infrared reflectance spectroscopy. *Applied spectroscopy*, 38(6) :844–847, 1984. 61

[69] M. Howard. Comparative study of calibration methods for near-infrared reflectance analysis using nested experimental design. *Anal. Chem.*, 58 :2814–2819, 1986. 48

[70] W. Härdle. Applied nonparametric regression. *Econometric Society Monographs Cambridge University Press*, 19, 1990. 88

[71] F. Hu and J. Hu. A note on breakdown theory for bootstrap methods. *Statistics and Probability Letters*, 50(1) :49–53, 2000. 35

[72] P.J. Huber. *Robustness and Designs.* John Wiley, New york : North-Holland, 1975. 35

[73] E. Hund, D. L. Massart, and J. Smeyers-Verbeke. Inter-laboratory studies in analytical chemistry. *Analytica Chimica Acta*, 423(2) :145–165, 2000. 40

[74] E. Hund, Y. Vander Heyden, M. Haustein, D. L. Massart, and J. Smeyers-Verbeke. Comparison of several criteria to decide on the significance of effects in a robustness test with an asymmetrical factorial design. *Analytica Chimica Acta*, 404(2) :257–271, 2000. 41, 43, 47, 48

[75] ICH. Harmonised tripartite guideline prepared with in the third international conference on harmmonisation of technical requirements for registration of pharmaceuticals for human use. text on validation of analytical procedures, 1994. 30, 31, 37, 47

[76] ICH. Harmonised tripartite guideline prepared with in the third international conference on harmmonisation of technical requirements for registration of pharmaceuticals for human use. validation of analytical procedures : Methodology, 1996. 33, 47

[77] ICH. Harmonised tripartite guideline prepared with in the third international conference on harmmonisation of technical requirements for registration of pharmaceuticals for human use. guidance for industry :analytical procedures and methods validation, 2000. 33, 36, 37

[78] IEEE. *IEEE Standard Computer Dictionary : A Compilation of IEEE Standard Computer Glossaries*. Institute of Electrical and Electronics Engineers, New York, 1990. 34

[79] T. Isaksson and T. Naes. The effect of multiplicative scatter correction (msc) and linearity improvement in nir spectroscopy. *Applied Spactroscopy*, 42(7) :1273 –1284, 1988. 66, 68

[80] T. Isaksson and T. Naes. Selection of samples for calibration in near-infrared spectroscopy. ii. selection based on spectral measurements. *Applied Spectroscopy*, 44(7) :1152–1158, 1990. 35, 61

[81] S.P. Jones. *Designs for minimizing the effect of environmental variables*. PhD thesis, University of Wisconsin-Madison, 1990. 51

[82] D. Jouan-Rimbaud, B. Walczak, D. L. Massart, I. R. Last, and K. A. Prebble. Comparison of multivariate methods based on latent vectors and methods ba-

sed on wavelength selection for the analysis of near-infrared spectroscopic data. *Analytica Chimica Acta*, 304(3) :285–295, 1995. 77, 78, 80

[83] R.W. Kennard and L.A. Stone. Computer aided design of experiments. *Technometrics*, 11(1) :137–148, 1969. 61

[84] R. Komanduri and M. Jiang. Application of taguchi method for optimization of finishing conditions in magnetic float polishing (mfp). *Wear*, 213(1-2) :59–71, 1997. 43

[85] M.S. Larrechi and M.P. Callao. Strategy for introducing nir spectroscopy and multivariate calibration techniques in industry. *Trends in Analytical Chemistry*, 22 :634–640, 2003. 86

[86] John Lawson, Scott Grimshaw, and Jason Burt. A quantitative method for identifying active contrasts in unreplicated factorial designs based on the half-normal plot. *Computational Statistics & Data Analysis*, 26(4) :425–436, 1998. 48

[87] Y. Z. Liang and O. M. Kvalheim. Robust methods for multivariate analysis : a tutorial review. *Chemometrics and Intelligent Laboratory Systems*, 32(1) :1–10, 1996. 35, 60

[88] D. K. J. Lin. Generating systematic supersaturated designs. *Technometrics*, 37(2) :213–225, 1995. 41

[89] Luan. *Méthodes robustes en analyse en composantes principales*. PhD thesis, Conservatoire national des arts et métiers, 1992. 35

[90] C.B. Lucasius and G. Keateman. Tutorial : Understanding and using genetic algorithms. part2. representation, configuration and hybridization. *Chemometrics and Intelligent Laboratory Systems*, 25 :99–145, 1994. 81

[91] S. Macho and M. S. Larrechi. Near-infrared spectroscopy and multivariate calibration for the quantitative determination of certain properties in the petrochemical industry. *Trends in Analytical Chemistry*, 21(12) :799–805, 2002. 20

[92] H. Maeda and Y. Osaki. Near infrared spectroscopy and chemometrics studies of temperature-dependent spectral variations of water : relationship between spec-

tral changes and hydrogen bonds. *J. Near Infrared Spectroscopy*, 3 :191–201, 1995. 107

[93] H. Mark and J. Workman. A new approach to generating transferable calibrations for quantitative near-infrared spectroscopy. *Spectroscopy*, 3(11) :28–36, 1988. 44

[94] H. Martens, S. Jensen, and P. Geladi. Proc. nordic symposium. *Applied Statistics*, page 325, 1983. 68

[95] H. Martens and T. Naes. *Multivariate calibration.* John Wiley and Sons, 1998. 80

[96] H Martens and E Stark. Extended multiplicative signal correction and spectral interference subtraction : new preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 9(8) :625–635, 1991. 69, 78

[97] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometics and Qualimetrics : Part A.* Elseiver, 1997. 21, 81

[98] M. J. McShane, B. D. Cameron, G. L. Cote, M. Motamedi, and C. H. Spiegelman. A novel peak-hopping stepwise feature selection method with application to raman spectroscopy. *Analytica Chimica Acta*, 388(3) :251–264, 1999. 78

[99] M. J. McShane, G. L. Cote, and C. H. Spiegelman. Variable selection in multivariate calibration of a spectroscopic glucose sensor. *Applied Spectroscopy*, 51(10) :15591564, 1997. 78

[100] T. J. Mitchell. An algorithm for the construction of "d-optimal" experimental designs. *Technometrics*, 16(2), 1974. 45

[101] G. Montague. Monitoring and control of fermenters. *Institute of Chemical Engineers*, page Warwickshire, 1997. 21

[102] M. Mulholland. Ruggedness testing in analytical chemistry. *Trends in Analytical Chemistry*, 7 :383–389, 1988. 33, 37, 39, 47

[103] M. Mulholland and J. Waterhouse. Development and evaluation of an automated procedure for ruggedness testing of chromatographic conditions in high-

performance liquid chromatography. *J. Chromatogr.*, 395 :539–551, 1987. 43, 47

[104] T. Naes and H. Martens. Principal component regression in nir analysis : viewpoints, background details and selection of components. *Journal of Chemometrics*, 2 :155–167, 1988. 92

[105] A. Nijhuis, H. C. M. van der Knaap, S. de Jong, and B. G. M. Vandeginste. Strategy for ruggedness tests in chromatographic method validation. *Anal. Chim. Acta*, 391(2) :187–202, 1999. 39, 45, 48

[106] D. Nolan. Asymptotics for multivariate trimming. *Stochastic Processes and their Applications*, 42(1) :157–169, 1992. 60

[107] K. H. Norris and J. R. Hart. Direct spectrophotometric determination of moisture content of grain and seeds. *Proceedings of the 1963 International Symposium on Humidity and Moisture, Priciples and methods of measuring moisture in liquids and solidsd.*, 4 :19–25, 1963. 20

[108] L. Olsson, U. Schulze, and J. Nielsen. On-line bioprocess monitoring - an academic discipline or an industrial tool ? *Tracs-Trends Anal. Chem.*, 17(2) :88–95, 1998. 21

[109] B.G. Osborne. Monitoring the accuracy of nir instruments. *Aspects of Applied Biology /Cereal Quality*, 15 :515–521, 1987. 32

[110] B.G. Osborne and T. Fearn. Collaborative evaluation of universal calibrations for the measurement of protein and moisture in flour by near reflectance. *J. Food Technology*, 18 :453–460, 1983. 86

[111] B.G. Osborne, T. Fearn, and P.H. Hindle. *Practical NIR spectroscopy with applications in food and beverage analysis.* John Wiley and Sons, second edition, 1993. 32, 113, 124

[112] M. Otto. *Chemometrics : statistics and computer application in analytical chemistry.* Wiley-VCH, New York, 1999. 39, 41, 42, 45

[113] C.S. Ough and M.A. Amerine. Studies with controlles fermentations. effects

of temperature fermentation on some volatile compounds in wine. *Am. J. Enol.Vitic.*, 18 :157–164, 1967. 107

[114] L. Pasti, D. Jouan-Rimbaud, D. L. Massart, and O.E. de Noord. Application of fourier transform to multivariate calibration of near-infrared data. *Analytica Chimica Acta*, 364(1-3) :253–263, 1998. 70

[115] R.L. Plackett and J.P. Burman. The design of optimum multifactorial experiments. *Biometrika*, 33 :305–325, 1946. 40

[116] R. Ragonese, M. Mulholland, and J. Kalman. Full and fractionated experimental designs for robustness testing in the high-performance liquid chromatographic analysis of codeine phosphate, pseudoephedrine hydrochloride and chlorpheniramine maleate in a pharmaceutical preparation. *Journal of Chromatography A*, 870(1-2) :45–51, 2000. 40, 42

[117] J. Rantanen, E. Rasanen, J. Tenhunen, M. Kansakoski, J-P. Mannermaa, and J. Yliruusi. In-line moisture measurement during granulation with a four-wavelength near infrared sensor : an evaluation of particle size and binder effects. *Chemometrics and Intelligent Laboratory Systems*, 56 :51–58, 2001. 20

[118] L. C. Rodriguez, B.G. Rosario, C. Garcia, M. Ana, S. Bosque, and M. Juan. A new approach to a complete robustness test of experimental nominal conditions of chemical testing procedures for internal analytical quality assessment. *Chemometrics and Intelligent Laboratory Systems*, 41(1) :57–68, 1998. 33, 34, 48, 54

[119] J. M. Roger and V. Bellon-Maurel. Using genetic algorithms to select wavelengths in near-infrared spectra : Application to sugar content prediction in cherries. *Applied Spectroscopy*, 54(9) :1313–1320, 2000. 81

[120] J.M. Roger, F. Chauchard, and V. Bellon-Maurel. Epo-pls external parameter orthogonalisation of pls application to temperature-independent measurement of sugar content of intact fruits. *Chemometrics and Intelligent Laboratory Systems*, 66(2) :191–204, 2003. 76, 90, 92

[121] S. Roussel. Calibration of near infrared instruments. *Uncorrected proof*, 1999. 42, 54

[122] S. Roussel, D. Funk, C.R. Hurburgh, and J.M. Roger. Noise comparison of multivariate calibration models for grain composition prediction based on nir spectroscopy. *Uncorrected proof.* 34

[123] S. C. Rutan. Fast on-line digital filtering. *Chemometrics and Intelligent Laboratory Systems*, 6 :191–201, 1989. 86

[124] Sarah C. Rutan, Eric Bouveresse, Kevin N. Andrew, Paul J. Worsfold, and D. L. Massart. Correction for drift in multivariate systems using the kalman filter. *Chemometrics and Intelligent Laboratory Systems*, 35(2) :199–211, 1996. 86

[125] D. N. Rutledge, A. Barros, and I. Delgadillo. Polish smoothed partial least squares regression. *Analityca chimica Acta*, 446 :281–296, 2001. 69, 71

[126] J-M. Sablayrolles. Conduite des fermentations en oenologie : les spécificités aux enjeux technologiques, 2003. 105, 107

[127] J-M. Sablayrolles and Barre P. Effect of anisothermal conditions on the kinetics of alcoholic fermentation by saccharomyces cerevisiae. *Bioproc. Eng.*, 4 :139–143, 1989. 107

[128] J-M. Sablayrolles and Barre P. Kinetics of alcoholic fermentation under anisothermal conditions. ii. prediction from kinetics under isothermal conditions. *Am. J. Enol.Vitic.*, 44 :134–138, 1993. 107

[129] J-M. Sablayrolles and Barre P. Kinetics of alcoholic fermentation under anisothermal enological conditions. i. influence of temperature evolution on the instantaneous rate of fermentation. *Am. J. Enol.Vitic.*, 44 :127–133, 1993. 107

[130] M. B. Sanz, L. A. Sarabia, A. Herrero, and M. C. Ortiz. A study of robustness with multivariate calibration. application to the polarographic determination of benzaldehyde. *Talanta*, 56(6) :1039–1048, 2002. 36, 40, 41, 47, 48

[131] D. W. Scott. *Multivariate Density Estimation : Theory, Practice, and Visualization.* John Wiley & Sons New York Chichester, 1992. 88

[132] M. B. Seasholtz. Making money with chemometrics. *Chemometrics Intell. Lab. Syst.*, 45(1-2) :55–63, 1999. 21

[133] M. B. Seasholtz and B. R. Kowalski. The effect of mean centering on prediction in multivariate calibration. *Journal of Chemometrics*, 6(2) :103–111, 1992. 13, 62, 63

[134] S. S. Sekulic, J. Wakeman, P. Doherty, and P. A. Hailey. Automated system for the on-line monitoring of powder blending processes using near-infrared spectroscopy ; part ii. qualitative approaches to blend evaluation. *Journal of Pharmaceutical and Biomedical Analysis*, 17(8) :1285–1309, 1998. 21, 67

[135] R. E. Shaffer and R. J. Combs. Comparison of spectral and interferogram processing methods using simulated passive fourier transform infrared remote sensing data. *Applied Spectroscopy*, 55(10) :1404–1413, 2001. 70

[136] C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, and G. L. Cote. Theoretical justification of wavelength selection in pls calibration development of a new algorithm. *Analytical Chemistry*, 70(1) :35–44, 1998. 79

[137] V. Steinmetz. *Fusion multisensorielle appliquée en temps réelaux décisions qualitatives sur les produits agro-alimentaires.* Génie des procédés bio-industriels, Ecole Nationale du Génie Rural, des Eaux et des Forêts, 1997. 35, 54

[138] P Steyer, J. C. Bouvier, T. Conte, P. Gras, J. Harmand, and J. P. Delgenes. On-line measurements of cod, toc, vfa, total and partial alkalinity in anaerobic digestion processes using infra-red spectrometry. *Water Science And Technology : a Journal Of The International Association On Water Pollution Research*, 45(10) :133–138, 2002. 126

[139] J.A.K. Suykens, T. Van Gestel, J. De Barbanter, B. De Moor, and J. Vandewalle. *Least squares Support vector Machines.* World Scientific, 2002. 88

[140] H. Swierenga, P. J. de Groot, A. P. de Weijer, M. W. J. Derksen, and L. M. C. Buydens. Improvement of pls model transferability by robust wavelength selection. *Chemometrics and Intelligent Laboratory Systems*, 41(2) :237–248, 1998. 81

[141] H. Swierenga, A. P. de Weijer, R. J. van Wijk, and L. M. C. Buydens. Strategy for constructing robust multivariate calibration models. *Chemometrics and Intelligent Laboratory Systems*, 49(1) :1–17, 1999. 34, 36, 38, 40, 54

[142] H. Swierenga, F. Wulfert, O. E. de Noord, A. P. de Weijer, A. K. Smilde, and L. M. C. Buydens. Development of robust calibration models in near infra-red spectrometric applications. *Analytica Chimica Acta*, 411(1-2) :121–135, 2000. 23, 38, 81

[143] G. Taguchi. Quality engineering (taguchi methods) for the development of electronic circuit technology. *Microelectronics and Reliability*, 37(3) :534, 1997. 33

[144] M. Tenenhaus. *La Regression PLS :Théorie et Pratique*. TECHNIP, Paris, 1998. 69

[145] Pekka Teppola, Satu-Pia Mujunen, and Pentti Minkkinen. Kalman filter for updating the coefficients of regression models. a case study from an activated sludge waste-water treatment plant. *Chemometrics and Intelligent Laboratory Systems*, 45(1-2) :371–384, 1999. 86

[146] J. Trygg. *Parsimonious Multivariate Models*. umea university, 2001. 74

[147] J. Trygg and S. Wold. Orthogonal projections to latent structures (o-pls). *Journal of Chemometrics*, 16 :119–128, 2002. 74, 75

[148] G. Vaccari, E. Dosi, A.L. Campi, R.A. Gonzalez-Vara y, D. Matteuzzi, and G. Mantovani. A near-infrared spectroscopy technique for the control of fermentation processes : An application to lactic acid fermentation. *Biotechnology and Bioengineering*, 43(10) :913–917, 1994. 20

[149] S. Vaidyanathan, G. Macaloney, L. M. Harvey, and B. McNeil. Assessment of the structure and predictive ability of models developed for monitoring key analytes in a submerged fungal bioprocess using near-infrared spectroscopy. *Applied Spectroscopy*, 55 :444–453, 2001. 21

[150] F.J. van de Vaart. Study group " quality in pharmaceutical analysis' of the work group " quality in analytical chemistry", section analytical chemistry of the kncv,

validation in pharmaceutical and biopharmaceutical analysis. *Het Pharmaceutisch Weekblad*, 127 :1992, 1992. 33, 34, 37

[151] Y. Van der Heyden. The ruggedness of analytical methods. *Analusis Magazine*, 22(5) :M27–M29, 1994. 36, 41, 48

[152] J. A. van Leeuwen, L. M. C. Buydens, B. G. M. Vandeginste, G. Kateman, A. Cleland, M. Mulholland, C. Jansen, F. A. Maris, P. H. Hoogkamer, and J. H. M. van den Berg. Res, an expert system for the set-up and interpretation of a ruggedness test in hplc method validation ; part 3 : The evaluation. *Chemometrics and Intelligent Laboratory Systems*, 11(2) :161–174, 1991. 37

[153] J. A. van Leeuwen, L. M. C. Buydens, B. G. M. Vandeginste, G. Kateman, P. J. Schoenmakers, and M. Mulholland. Res, an expert system for the set-up and interpretation of a ruggedness test in hplc method validation ; part 1 : The ruggedness test in hplc method validation. *Chemometrics and Intelligent Laboratory Systems*, 10(3) :337–347, 1991. 33, 43

[154] J. A. van Leeuwen, L. M. C. Buydens, B. G. M. Vandeginste, G. Kateman, P. J. Schoenmakers, and M. Mulholland. Res, an expert system for the set-up and interpretation of a ruggedness test in hplc method validation ; part 2 : The ruggedness expert system. *Chemometrics and Intelligent Laboratory Systems*, 11(1) :37–55, 1991. 36

[155] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, and J. Smeyers-Verbeke. *Handbook of Chemometics and Qualimetrics : Part B.* Elseiver, 1998. 88

[156] Y. Vander Heyden, K. De Braekeleer, Y. Zhu, E. Roets, J. Hoogmartens, J. De Beer, and D. L. Massart. Nested designs in ruggedness testing. *Journal of Pharmaceutical and Biomedical Analysis*, 20(6) :875–887, 1999. 44, 48

[157] Y. Vander Heyden, M. Jimidar, E. Hund, N. Niemeijer, R. Peeters, J. Smeyers-Verbeke, D. L. Massart, and J. Hoogmartens. Determination of system suitability limits with a robustness test. *Journal of Chromatography A*, 845(1-2) :145–154, 1999. 34, 36, 43, 46

[158] Y. Vander Heyden, M. S. Khots, and D. L. Massart. Three-level screening designs for the optimisation or the ruggedness testing of analytical procedures. *Analytica Chimica Acta*, 276(1) :189–195, 1993. 43

[159] Y. Vander Heyden, S. Kuttatharmmakul, J. Smeyers-Verbeke, and D. L. Massart. Supersaturated designs for robustness testing. *Analytical Chemistry*, 72(13) :2869–2874, 2000. 41

[160] Y. Vander Heyden, A. Nijhuis, J. Smeyers-Verbeke, B. G. M. Vandeginste, and D. L. Massart. Guidance for robustness/ruggedness tests in method validation. *Journal of Pharmaceutical and Biomedical Analysis*, 24(5-6) :723–753, 2001. 36, 39, 45, 46, 47, 48

[161] Y. Vander Heyden, F. Questier, and D. L. Massart. A ruggedness test strategy for procedure related factors : experimental set-up and interpretation. *Journal of Pharmaceutical and Biomedical Analysis*, 17(1) :153–168, 1998. 39

[162] Y. Vander Heyden, F. Questier, and D. L. Massart. Ruggedness testing of chromatographic methods : selection of factors and levels. *Journal of Pharmaceutical and Biomedical Analysis*, 18(1-2) :43–56, 1998. 39, 48

[163] P. Vankeerberghen, J. Smeyers-Verbeke, R. Leardi, C. L. Karr, and D. L. Massart. Robust regression and outlier detection for non-linear models using genetic algorithms. *Chemometrics and Intelligent Laboratory Systems*, 28(1) :73–87, 1995. 35

[164] P. Vieu. Non parametric regression :optimal local bandwith choice. *J.R. Statist. Soc. B*, 53 :453–464, 1991. 88

[165] I.N. Vuchkov and L.N. Boyadjieva. the robustness against tolerances of performance characteristics described by second order polynomials. In *First international conference-work-shop on optimal design and analysis of experiments*, Neuchatel, Switzerland, 1988. 52

[166] J.C. Wahlich and G.P. Carr. Chromatographic system suitability tests - what should we be using. *Journal of Pharmaceutical and Biomedical Analysis*, 8 :619–623, 1990. 34

[167] B. Walczak, E. Bouveresse, and D. L. Massart. Standardization of near-infrared spectra in the wavelet domain. *Chemometrics and Intelligent Laboratory Systems*, 36 :41–51, 1995. 89

[168] B. Walczak and D. L. Massart. Robust principal components regression as a detection tool for outliers. *Chemometrics and Intelligent Laboratory Systems*, 27(1) :41–54, 1995. 60

[169] Y. Wang, M.J. Lysaght, and B.R. Kowalski. Improvement of multivariate calibration through instrument standardisation. *Analytical Chemistry*, 64 :562–565, 1992. 89

[170] Y. Wang, D.J. VeltKamp, and B.R. Kowalski. Multivariate instrument standardisation. *Analytical Chemistry*, 63 :2750–2756, 1991. 89

[171] M.R. Warnes, G. Jarmila, G.A. Montague, and B. Kara. Application of radial basis function and feedforward artificial neural networks to the escherichia coli fermentation process. *Neurocomputing*, 20 :67–82, 1998. 86

[172] W. Wegscheider, H. J. Zeiler, R. Heindl, and J. Mosser. Quantifying uncertainty in sampling and analytical measurement. *Ann. Chim.*, 87(3-4) :273–283, 1997. 32

[173] M.O. Westerhaus. *Improving repeatability of calibrations across instruments.* Agriculture Research Center, Gembloux, Belgium, proc. 3rd int. conf. on near infrared spectroscopy edition, 1991. 89

[174] J. A. Westerhuis, S. de Jong, and A. K. Smilde. Direct orthogonal signal correction. *Chemometrics and Intelligent Laboratory Systems*, 56(1) :13–25, 2001. 74

[175] L. A. Williams. Heat release in alcoholic fermentation : a critical reappraisal. *Am. J. Enol.Vitic.*, 34 :234–242, 1982. 105

[176] S. Wold, H. Antti, F. Lindgren, and J. Ohman. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2) :175–185, 1998. 73, 74

[177] S. Wold and M. Sjostrom. Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems*, 44(1-2) :3–14, 1998. 21

[178] S. Wold, M. Sjostrom, and L. Eriksson. Pls-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2) :109–130, 2001. 23

[179] V.A.L. Wortel, W.G. Hansen, and S.C.C. Wiedemann. Optimising multivariate calibration by robustness criteria. *J. Near Infrared Spectroscopy*, 9 :141–151, 2001. 43, 48, 49, 50

[180] G. Wynia, P. Post, J. Broersen, and F. Maris. Ruggedness testing of a gas chromatographic method for residual solvents in pharmaceutical substances. *Chromatographia*, 39(5/6) :335–362, 1994. 42

[181] G. T. Xu, H. F. Yuan, and W. Z. Lu. Development of modern near infrared spectroscopic techniques and its applications. *Spectrosc. Spectr. Anal.*, 20(2) :134–142, 2000. 21

[182] W.J. Youden. *Mater. Res. Stand.*, 1 :863–865, 1961. 35

[183] M. Zeaiter, J.M.R. Roger, and V. Bellon-Maurel. Dop - dynamic orthogonal projection method. a new method to maintain the robustness of on-line used calibration models. application to monitor alcoholic fermentation using nir spectroscopy. *Journal of Chemometrics*, submited :xxxx, 2004. 137

[184] M. Zeaiter, J.M.R. Roger, and V. Bellon-Maurel. Robustness of models developped by multivariate calibration. part2.influence of preprocessing methods. *Trends in Analytical Chemistry*, submited :xxxx, 2004. 83, 136

[185] M. Zeaiter, M. Roger, V. Bellon-Maurel, and D.N. Rutledge. The robustness of models developed by multivariate calibration methods. part1 : the asssessment of robustness. *Trends in Analytical Chemistry*, 23(2) :157–170, 2004. 30, 48, 136

[186] J. Zhang. The sample breakdown points of tests. *Journal of Statistical Planning and Inference*, 52(2) :161–181, 1996. 35