Développements chimiométriques pour améliorer la robustesse des mesures spectrométriques appliquées aux agro-procédés

J.M. Roger

ACADÉMIE DE MONTPELLIER
Université de Montpellier II

Mémoire d'Habilitation à Diriger des Recherches

Section 26
Mathématiques Appliquées et Application des Mathématiques

présenté par

Jean-Michel ROGER

Développements chimiométriques pour améliorer la robustesse des mesures spectrométriques appliquées aux agro-procédés

soutenu le 28 Novembre 2005 devant la commission d'examen

Jean-marie	NAVARRO
Anna	DE JUAN
Nathalie	DUPUY
Ricardo	LEARDI
Douglas	RUTLEDGE
Robert	SABATIER
Véronique	BELLON-MAUREL

Remerciements

Il serait vain de remercier nominativement toutes les personnes ayant contribué à la rédaction de ce mémoire d'HDR, puisqu'il retrace plus de dix années de travail collaboratif.

Aussi, j'adresse globalement ma reconnaissance et mes remerciements les plus sincères à toutes ces personnes, avec une mention particulière pour celle qui est à l'origine de cette étape dans ma vie de chercheur et qui se reconnaîtra.

Quand les mystères sont malins,
ils se cachent dans la lumière.
JEAN GIONO

Table des matières

Publications et Travaux

Publications dans des revues scientifiques

[Blasco et al., 2002] Blasco, J., Aleixos, N., Roger, J. M., Rabatel, G., and Molto, E. (2002). Robotic weed control using machine vision. *Biosystems engineering*, 83(2) :149–157.

[Chauchard et al., 2004a] Chauchard, F., Cogdill, R., Sylvie, R., Roger, J. M., and Bellon-Maurel, V. (2004a). Application of ls-svm to non-linear phenomena in nir spectroscopy : development of a robust and portable sensor for acidity prediction in grapes. *Chemometrics and Intelligent Laboratory Systems*, 71(2) :141–150.

[Chauchard et al., 2004b] Chauchard, F., Roger, J. M., and Bellon-Maurel, V. (2004b). Correction of the temperature effect on near infrared calibration - application to soluble solid content prediction. *Journal of Near Infrared Spectroscopy*, 12(3) :199–205.

[Francois et al., 2003] Francois, J., Grandvalet, Y., Denoeux, T., and Roger, J. M. (2003). Resample and combine : An approach to improving uncertainty representation in evidential pattern classification. *Information Fusion*, 4(2) :75–85.

[Grenier et al., 2000] Grenier, P., Alvarez, I., Roger, J. M., Steinmetz, V., Barre, P., and Sablayrolles, J. M. (2000). Artificial intelligence in wine-making. *International Journal of Vine and Wine Making*, 34(2) :61–66.

[Nivière et al., 1994] Nivière, V., Grenier, P., Roger, J. M., Sevila, F., and Oussalah, C. (1994). Intelligent simulation of plant operation in the wine industry. *Food control*, 5 :91–95.

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

[Roger et al., 2003] Roger, J. M., Chauchard, F., and Bellon-Maurel, V. (2003). 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.

[Roger et al., 2005] Roger, J. M., Palagos, B., Guillaume, S., and Bellon-Maurel, V. (2005). Discriminating from highly multivariate data ; application to nir spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 79/1-2 : 31-41

[Roger et al., 2002] Roger, J. M., Sablayrolles, J. M., Steyer, J. P., and Bellon-Maurel, V. (2002). Pattern analysis techniques to process fermentation curves : Application to discrimination of enological alcoholic fermentations. *Biotechnology and Bioengineering*, 79(7) :804–815.

[Roussel et al., 2003a] Roussel, S., Bellon-Maurel, V., Roger, J. M., and Grenier, P. (2003a). Authenticating white grape must variety with classification models based on aroma sensors, ft-ir and uv spectrometry. *Journal of food engineering*, 60(4) :407–419.

[Roussel et al., 2003b] Roussel, S., Bellon-Maurel, V., Roger, J. M., and Grenier, P. (2003b). Fusion of aroma, ft-ir and uv sensor data based on the bayesian inference. application to the discrimination of white grape varieties. *Chemometrics and intelligent laboratory systems*, 65 :209–219.

[Sánchez et al., 2003] Sánchez, N. H., Lurol, S., Roger, J. M., and Bellon-Maurel, V. (2003). Ro-

bustness of models based on nir spectra for sugar content prediction in apples. *Journal of Near Infrared Spectroscopy*, 11 :97–107.

[Steinmetz et al., 1999] Steinmetz, V., Roger, J. M., Molto, E., and Blasco, J. (1999). On-line fusion of colour camera and spectrophotometer for sugar content prediction of apples. *Journal of agricultural engineering research*, 73, nř 2 :207–216.

[Zeaiter et al., 2004a] Zeaiter, M., Roger, J. M., and Bellon-Maurel, V. (2004a). Robustness of models developed by multivariate calibration. part ii : Improving the robustness. *Trends in Analytical Chemistry*, 24, 5 : 437-445.

[Zeaiter et al., 2005] Zeaiter, M., Roger, J. M., and Bellon-Maurel, V. (2005). Dynamic orthogonal projection : A new method to maintain the on-line robustness of multivariate calibrations. application to nir based monitoring of wine fermentations. *Chemometrics and Intelligent Laboratory Systems*, in press.

[Zeaiter et al., 2004b] Zeaiter, M., Roger, J. M., Bellon-Maurel, V., and Rutledge, D. N. (2004b). Robustness of models developed by multivariate calibration. part i : The assessment of robustness. *Trends in Analytical Chemistry*, 23, 2 :157–170.

Brevets

[Roger et al., 1997] Roger, J.-M., DeRudnicki, V., Bonicel, J.-F., Leroy, G., and Delencre, G. (1997). Dispositif de désherbage électrique, sélectif et dirigé. Brevet ; n^o de publication FR 2770969, n^o d'enregistrement national 9714315. 1997. 23 p.

[Roger et al., 1998] Roger, J.-M., Bellon, V., Fatou, J.-M., Steinmetz, V., and Crochon, M. (1998). Procédé et installation pour la mesure de la teneur, notamment en sucre, de fruits et légumes. Brevet ; n^o de publication FR 2775345. 1998. 26 p.

Organisation de congrès

[BIO-DECISION'98] From sensors to decision support systems in agriculture Participation au comité scientifique / Présidence de scéance

[SENSORAL'98] Capteurs pour la qualité en Agro-Alimentaire Participation au comité d'organisation

[CHIMIOMÉTRIE 2004] Conférencier Invité La robustesse des étalonnages multi-variés

Thèses co-encadrées

[Pierre THOMPSON, 1997] Méthodologie de commande pour un manipulateur avec mobilité perturbée travaillant "au vol" sur l'environnement. École Nationale du Génie Rural, des Eaux et des Forêts Dirigée par A. LIEGEOIS Taux d'encadrement : 20%

[Nahid KARCHENASSE, 1998] Étude d'un système de diagnostic à partir des cas - Application au diagnostic de défaillance d'un sécateur électronique. École Nationale du Génie Rural, des Eaux et des Forêts Dirigée par C. MILLIER Taux d'encadrement : 90%

[Sylvie ROUSSEL, 1998] Optimisation des capteurs d'arômes et fusion multisensorielle appliquée à la caractérisation des produits agro-alimentaires. École Nationale Supérieure Agronomique de Montpellier Dirigée par V. BELLON MAUREL Taux d'encadrement : 10%

[Jérémie FRANÇOIS, 2000] Fusion de connaissances expérimentales et expertes : une approche évolutive du diagnostic. Université Technologique de Compiègne Dirigée par Y. GRANDVALET Taux d'encadrement : 90%

[Serge GUILLAUME, 2001] Induction de règles floues interprétables. INSA Toulouse Dirigée par A. TITLI Taux d'encadrement : 20%

[Olivier NAUD, 2002] Modélisation hybride pour la supervision de systèmes mécatroniques : application à la stabilité en pente de machines mobiles INSA Toulouse Dirigée par J. AGUILAR MARTIN Taux d'encadrement : 30%

[Magida ZEAITER, 2004] Mesures robustes en ligne des solutés organiques par spectrométrie infrarouge et étalonnages multi-variés. Université Montpellier II ; ED SPBI Dirigée par V. BELLON MAUREL Taux d'encadrement : 90%

[Jean-Noël PAOLI, 2004] Fusion de données géoréférencées. Université Montpellier II ; ED SPBI Dirigée par F. SEVILA Taux d'encadrement : 40%

[Fabien CHAUCHARD, prévue en 2005] Mesure par spectrométrie dans les milieux diffusants. Application à la caractérisation de la qualité du raisin au champ. Université Montpellier II ; ED SPBI Dirigée par V. BELLON MAUREL Taux d'encadrement : 90%

Principaux projets

[MAGALI, 1986-1990] Récolte robotisée de pommes Partenaire : PELLENC, LIRMM, SAGEM Responsable : A. BOURÉLY
Développement de l'interface utilisateur.

[PATCHWORK, 1994-1997] Machine tractée de désherbage électrique Partenaires : SRI (GB), IVIA (E), DIPSS (D). Responsable : J.M. ROGER
Animation du projet ; Développements informatiques ; Contrôle / Commande du bras manipulateur ; Conception de l'actionneur désherbant (Cf Brevet).

[SHIVA, 1994-1998] Nouveau système de tri des fruits à la qualité Partenaires : PELLENC (F), IVIA (SP), FOMESA (SP), SYFA (I) Responsable : V. STEINMETZ
Développements chimiométriques. (Cf Brevet)

[GLOVE, 1998-2001] Gant instrumenté pour mesurer la qualité des fruits Partenaires : APOFRUIT (I), ATB (D), VERHAERT (B), KUL (B) Responsable : M. CROCHON
Conception du dispositif ; Développements chimiométriques.

[VISHNU, 2001-2003] Mesure en ligne de la qualité des fruits Partenaires : PELLENC (F), IVIA (SP), FOMESA (SP), MONTAGNE (F) Responsable : J.M. ROGER
Animation du projet ; Conception du dispositif ; Développements chimiométriques.

Avant propos

Ce rapport est dédié à la synthèse de ma carrière de chercheur. Cette dernière a commencé, il y a bientôt quinze ans, dans le domaine de la modélisation de connaissances, appliquée au diagnostic. En 1997, c'est à dire approximativement en milieu de parcours, j'ai inflechi ma thématique de recherche, pour répondre à un problème concret de mise en œuvre des capteurs à base de spectrométrie proche infrarouge. C'est uniquement cette deuxième phase que j'ai choisi de rapporter ici.

Une première partie présente ma problématique de recherche. Puisque cette recherche a été initiée par un problème applicatif très pratique, j'y ai emprunté une démarche analytique, qui m'a amené à l'identification de trois voies de recherche. Chacune d'elles est donc ensuite exposée individuellement. Pour ce faire, j'utilise à chaque fois deux articles : Un premier, de portée générale, par exemple une revue, me permet de poser le cadre ; Le deuxième expose une contribution originale. Enfin, pour chacune de ces voies de recherche, des perspectives sont données.

Les 6 articles suivants ont donc été choisis comme représentatifs de ma recherche ; ils seront référencés spécialement dans le texte et sont livrés *in extenso* en annexe.

[Article I] Chauchard, F., Roger, J. M., and Bellon-Maurel, V. (2004b). Correction of the temperature effect on near infrared calibration - application to soluble solid content prediction. *Journal of Near Infrared Spectroscopy*, 12(3) :199–205.

[Article II] Roger, J. M., Chauchard, F., and Bellon-Maurel, V. (2003). 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.

[Article III] Zeaiter, M., Roger, J. M., and Bellon-Maurel, V. (2004a). Robustness of models developed by multivariate calibration. part ii : Improving the robustness. *TrAC Trends in Analytical Chemistry*, 24, 5 : 437-445.

[Article IV] Zeaiter, M., Roger, J. M., and Bellon-Maurel, V. (2005). Dynamic Orthogonal Projection. A new method to maintain the on-line robustness of multivariate calibrations. Application to NIR based monitoring of wine fermentations. *Chemometrics and Intelligent Laboratory Systems*, in press.

[Article V] Roger, J. M., Sablayrolles, J. M., Steyer, J. P., and Bellon-Maurel, V. (2002). Pattern analysis techniques to process fermentation curves : Application to discrimination of enological alcoholic fermentations. *Biotechnology and Bioengineering*, 79-7 :804–815.

[Article VI] Roger, J. M., Palagos, B., Guillaume, S., and Bellon-Maurel, V. (2005). Discriminating from highly multivariate data by focal eigen function discriminant analysis. application to nir spectra. *Chemometrics and Intelligent Laboratory Systems*, 79, 1-2 :31-41

Notations

Les lettres majuscules grasses sont employées pour désigner des matrices, p.e. $\mathbf{X}$; les lettres minuscules grasses désignent des vecteurs colonnes, p.e. $\mathbf{x}_j$ désigne la $j^{\text{ème}}$ colonne de $\mathbf{X}$; les vecteurs lignes sont désignés par l'opérateur de transposition, p.e. $\mathbf{x}_i^T$ désigne la $i^{\text{ème}}$ ligne de $\mathbf{X}$; les lettres minuscules non grasses désignent des scalaires, p.e. des éléments de matrice x_{ij} ou des indices i. En cas de besoin, la dimension des matrices peut être indiquée par un double indiçage entre parenthèses, p.e. $\mathbf{X}_{(np)}$ indique que la matrice $\mathbf{X}$ a n lignes et p colonnes.

Sauf indication contraire, les notations suivantes sont employées :

$\mathbf{X}$	Une matrice de n spectres, par p longueurs d'onde	
p	Le nombre de colonnes de $\mathbf{X}$	
n	Le nombre de lignes de $\mathbf{X}$	
j	Un indice de colonne	
i	Un indice de ligne	
n_{LV}	Nombre de variables (vraies ou latentes) du modèle	
$\mathbf{y}$	Le vecteur des n valeurs de référence	
$\widehat{\mathbf{y}}$	Le vecteur des n valeurs de référence estimées	
$\mathbf{b}$	Le vecteur des p coefficients de la régression	
b_0	L'intercept	$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b} + b_0$
$\mathbf{1}_n$	Le vecteur colonne contenant n 1	
$\mathbf{I}_p$	La matrice identité de dimension p	
$\overline{\mathbf{x}}^T$	Le vecteur ligne des moyennes des colonnes de $\mathbf{X}$	
$\overline{\mathbf{X}}$	La matrice contenant n fois $\overline{\mathbf{x}}^T$	$\overline{\mathbf{X}} = \mathbf{1}_n\overline{\mathbf{x}}^T$
$\overline{y}$	La valeur moyenne de $\mathbf{y}$	
$\|\mathbf{x}\|$	La norme euclidienne de $\mathbf{x}$, i.e. $(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}$	
$Col(\mathbf{X})$	L'espace engendré par les colonnes de $\mathbf{X}$	
rank$(\mathbf{X})$	Le rang de $\mathbf{X}$, i.e. la dimension de $Col(\mathbf{X})$	
SEC	L'erreur standard d'étalonnage	$SEC^2 = \frac{1}{n-n_{LV}-1}\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$
SEP	L'erreur standard de test	$SEP^2 = \frac{1}{n}\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$
BS	Le biais de prédiction	$BS = \overline{(\widehat{\mathbf{y}} - \mathbf{y})}$
SEP_C	Le SEP corrigé du biais	$SEP_C^2 = SEP^2 - BS^2$
RPD	L'erreur résiduelle de prédiction (Residual Prediction Deviation)	$RPD = SEP/\sigma(\mathbf{y})$

Chapitre 1

Introduction - La problématique de mon action de recherche

Le Cemagref étant particulièrement investi dans la recherche technologique, ma problématique a été générée par un problème pratique, comme le décrit la première partie de ce chapitre. Cette problématique très concrète est ensuite reprise de manière plus théorique, afin de mettre à jour les thématiques de recherches que j'ai choisi d'exposer dans ce mémoire.

1.1 Des problèmes pratiques aux questions scientifiques

1.1.1 Quelles applications ?

J'utiliserai le terme d'agro-procédé pour désigner tout procédé de production primaire exécuté dans "l'entreprise agricole élargie", à savoir touchant la filière agricole ou forestière, depuis la parcelle jusqu'aux opérations de post-récolte (tri, conditionnement, premières transformations impliquant l'agriculteur, telles que la vinification).

Après une période où le principal problème était d'assurer un volume de production suffisant, les agro-procédés sont maintenant fortement contraints par des objectifs de qualité, de sécurité et de préservation de l'environnement. La réalisation de ces objectifs, tout en maintenant un niveau de productivité économiquement rentable, requiert une réelle maîtrise des agro-procédés.

Tout comme dans l'industrie classique, l'opérateur maîtrise un procédé s'il est capable à tout instant d'en connaître l'état, de manière à déterminer les actions à mener pour se rapprocher d'un optimum souhaité. Or, contrairement aux procédés de l'industrie classique, les éléments constitutifs des agro-procédés sont biologiques, donc complexes et requièrent une instrumentation particulière. C'est ainsi que le Cemagref travaille à l'élaboration de capteurs dédiés à la détermination de cet état. Pour bien illustrer la difficulté de cette tâche, il convient de donner quelques exemples de paramètres qu'il est (ou qu'il serait) utile de mesurer :

- La bonne conduite d'un verger s'appuie entre autres sur la connaissance de l'état physiologique des fruits. Une connaissance parfaite de cet état nécessiterait la mesure de dizaines de paramètres physico-chimiques. Par exemple, pour la récolte, les experts ont néanmoins réussi à déterminer les quelques paramètres les plus importants : la taille des fruits, leur couleur, leur fermeté, leurs teneurs en pigments, dont la chlorophylle, en sucres[1], en amidon et en acides organiques. Si les mesures de la taille et de la couleur restent assez simples, celles des autres paramètres nécessitent des analyses chimiques coûteuses en temps et en argent, destructives et non accessibles à l'agriculteur.

[1] La plupart des fruits contiennent plusieurs sucres différents. Leur teneur totale est mesurée par réfractométrie et exprimée en degrés Brix.

– Pour diminuer les intrants chimiques dans les cultures, on peut envisager de moduler leur apport en fonction des besoins, au niveau intra-parcellaire. Pour les herbicides, une telle automatisation requiert de pouvoir déterminer, en temps réel, un taux de mauvaises herbes, par exemple en calculant un ratio de surfaces foliaires entre la culture et les adventices. Tout le problème réside alors dans la discrimination culture / sol / mauvaises herbes. Si on ajoute à cela que chaque espèce de mauvaise herbe réagit différemment aux herbicides, on arrive au problème encore plus complexe de la reconnaissance des espèces.
– Immédiatement après la récolte, les fruits tels que les pommes ou les pêches sont triés en classes de qualité, d'une part pour établir des classes de valeur marchande et d'autre part pour orienter la suite de leur traitement (conservation en chambre frigorifique, transport, transformation, etc.). Actuellement, cette qualité est incomplètement évaluée en ligne sur la base du calibre (ou du poids), de la couleur et des défauts d'aspect. La caractérisation serait bien meilleure avec la mesure de la fermeté et de la qualité organoleptique. La mesure de ce dernier critère peut être approchée par la détermination du taux de sucres au moyen de la spectrométrie infrarouge.

Le champ applicatif global de mes recherches est le développement de capteurs non destructifs pour la caractérisation d'objets biologiques. Plus précisément, je me suis focalisé sur la mesure par spectrométrie dans les agro-procédés.

1.1.2 Quelles solutions technologiques ?

Cette partie présente les principes technologiques des différents capteurs à base de spectrométrie que nous avons développés. Deux configurations majeures peuvent être distinguées : la mesure en extérieur et la mesure en ligne.

La mesure en extérieur

Cette catégorie intéresse les capteurs fonctionnant en conditions extérieures (au champ), qu'ils soient embarqués sur une machine ou utilisés par un opérateur piéton.

Dans le cadre du projet GLOVE (FAIR PL 97- 3399 ; 1998 - 2001), nous avons réalisé un capteur piéton, dont la partie spectrométrique permet de mesurer le taux de sucres et la teneur en chlorophylle interne des pommes et des pêches. Nous avons en projet actuellement le développement du même type de capteur, mais qui sera capable de mesurer, sur le cep de vigne, les teneurs en polyphénols, en acides organiques et en sucres ainsi que de détecter la présence de pourritures dans les grappes de raisin. Ce type de dispositif utilise une source lumineuse de faible intensité (quelques Watt), une optique de collimation sommaire (la mesure se fait au contact, ou presque) et un spectromètre à barrette de silicium opérant dans la gamme ultra-violet, visible et très proche infrarouge (UV - VIS - VNIR i.e. de 200 à 300 nm ; de 300 à 750 nm et de 750 à 1100 nm). Ce type de spectromètre présente l'avantage d'être compact, robuste et rapide.

Dans le projet VISHNU (CRAFT 1999-70106 ; 2001 - 2003), nous avons développé un capteur spectrométrique pour mesurer en temps réel le taux de sucres et l'acidité du raisin dans une machine à vendanger. Un spectromètre VNIR, associé à une source lumineuse puissante (200 Watt) et à une optique spécifique permettent de mesurer à distance la teneur en sucres moyenne d'une couche de baies de raisin.

La mesure en ligne

Il s'agit ici des capteurs fixes, montés sur une ligne faisant défiler des produits (p.ex. des fruits) ou sur un procédé batch dans lequel le même produit évolue au cours du temps (p.ex. une cuve de vinification).

Dans le projet VISHNU, nous avons développé un capteur spectrométrique pour mesurer en temps réel le taux de sucres des pommes et des pêches, sur une ligne de tri, à la cadence de 10 fruits par seconde. Le défi technologique que nous avons relevé consistait à mesurer une information spectrométrique de bonne qualité, sur des objets en mouvement (les fruits avancent et tournent sur eux mêmes), en moins de 50 ms. Les mêmes principes technologiques ont été retenus que pour le capteur embarqué dans la machine à vendanger (Cf paragraphe précédent).

Dans le projet MELON (Projet de transfert financé par l'ANVAR), la cadence est moins élevée (3 fruits par seconde) et la mesure se fait au contact, mais les fruits possèdent un épiderme plus opaque, plus épais et de structure très variable. Nous utilisons une source lumineuse moyennement puissante (50 Watt), associée à un spectromètre VIS - VNIR.

Dans le projet IRVIN (Région LR COST TRIAL 0207), nous développons un capteur de caractérisation des moûts lors de la fermentation. En plus des composés majoritaires tels que sucres et alcool, nous avons l'ambition de mesurer quelques composés impliqués dans la qualité organoleptique du produit fini, tels que les polyphénols. Non seulement ces paramètres sont importants pour le bon déroulement de la fermentation, mais ils peuvent aussi être utilisés comme une "empreinte digitale" de la vendange, comme nous l'avons montré dans [Article V]. Le capteur utilise un spectromètre UV - VIS - VNIR couplé à une sonde à fibre optique, de manière à pouvoir multiplexer la mesure (un seul spectromètre pour plusieurs cuves).

1.2 Le problème à résoudre

Les mesures que nous réalisons reposent sur des capteurs indirects[2] multivariés, nécessitant un étalonnage. En pratique, la relation entre les valeurs mesurées $\mathbf{x}$ et la valeur y de la grandeur qui nous intéresse est représentée par un *modèle* du type $y = f(\mathbf{x})$. Quelle que soit la technique employée pour le construire, un problème majeur réside dans la *robustesse*[3] de ce modèle, c'est à dire dans sa capacité à rester opérationnel dans des conditions différentes de celles du laboratoire (en extérieur ou en ligne).

Les solutions technologiques présentées dans 1.1.2, les conditions opératoires et la particularité des produits mesurés recèlent nombre de causes potentielles de ce problème de robustesse :

- **La complexité du signal :** Les molécules organiques présentent des pics d'absorption principaux dans la gamme du MIR (2,5 à 25 μm). Les absorptions mesurées dans le VNIR sont donc des combinaisons des $n^{\text{ièmes}}$ harmoniques, de niveau beaucoup plus faible (une bonne présentation de ces phénomènes quantiques pourra être trouvée dans [Lachenal, 2000]). De plus, un très large pic d'absorption de l'eau masque l'information sur les autres molécules. Il en résulte que la relation entre les spectres mesurés et les concentrations ne suit plus les lois théoriques. En particulier, une variation de concentration ne se traduit plus seulement par une variation d'une absorption, mais aussi comme un changement de forme du spectre.
- **Présence de grandeurs d'influence :** Quand les conditions environnementales ne sont pas maîtrisées, certaines grandeurs physiques peuvent influer sur la mesure. Nous avons mis en évidence dans [Sánchez et al., 2003] que la température du produit et la température du spectromètre sont des grandeurs d'influence très perturbatrices. Les variations de l'intensité du rayonnement incident peuvent aussi causer des perturbations. Elles peuvent avoir principalement deux origines : Une "pollution" par la lumière naturelle ; Une variation du spectre de la source. Ce dernier point pose un problème particulier : Le spectre de la source est normalement pris en compte en opérant une référence optique (un blanc). Si cette opération ne peut pas être faite avant chaque mesure, mais seulement de manière épisodique (p.ex. à la mise en marche du capteur), la prise en compte de la référence peut poser des problèmes de robustesse. En effet,

[2] Les capteurs indirects ne mesurent pas directement la grandeur Y cherchée, mais une autre grandeur, influencée par Y.

[3] La définition de ce terme sera discutée lors de l'analyse théorique, en section 1.3

une mauvaise acquisition entraîne des conséquences pour l'ensemble des mesures à venir. Dans certains cas, nous avons donc préféré mettre en œuvre une stabilisation de la source lumineuse et utiliser des spectres d'intensité plutôt que d'absorbance (Cf 1.3 pour plus de détails sur ces notions). L'acquisition et la prise en compte de la référence optique sont donc des problèmes que nous traitons au cas par cas, que je n'évoquerai plus dans la suite de ce mémoire.

- **Présentation de l'échantillon :** Un problème spécifique apparaît dans le cas d'un capteur embarqué sur une machine ou installé en ligne. Quand la mesure se fait sans contact, le positionnement de l'échantillon vis à vis du capteur n'est pas parfaitement maîtrisé, ce qui entraîne une variabilité sensible du trajet optique.
- **Variabilité de l'échantillon :** Enfin, un dernier problème a trait à la variabilité chimique et physique des produits mesurés. Ceux ci sont tous des produits biologiques, de composition exacte inconnue. Nous cherchons à mesurer quelques composés chimiques, dans une solution qui en contient plusieurs centaines. Certes, ces composés "parasites" sont en concentration moindre que ceux qui nous intéressent, mais leur influence sur le spectre mesuré n'est pas toujours négligeable (c'est le cas notamment de l'amidon par rapport aux sucres dans une pomme). Outre cette variabilité chimique, la matrice physique des échantillons est elle aussi variable (texture des fruits, turbidité des moûts, etc.). Ceci entraîne une variabilité dans le comportement du rayonnement lumineux, notamment sur sa diffusion par la matière.

Parfois, certaines de ces causes peuvent être éliminées ou réduites (p.ex. la température du spectromètre peut être régulée).

Lorsque les causes ne peuvent pas être éliminées, le seul moyen d'améliorer la robustesse de la mesure réside dans la manière de construire le modèle.

La construction des modèles d'étalonnage pour les capteurs multivariés entre dans une discipline scientifique assez récente, la *chimiométrie*. Ce terme fut cité la première fois en 1972 par le suédois Swante Wold et l'américain Bruce R. Kowalski. De nombreux comités, congrès et journaux sont maintenant dédiés à cette discipline, que l'on peut définir comme *la discipline chimique qui utilise les mathématiques et les statistiques pour (i) concevoir ou sélectionner les procédures de mesures et les expériences optimales (ii) fournir le maximum d'informations chimiques par l'analyse des données* ([Otto, 1999]).

Les recherches exposées dans ce mémoire ont donc pour objectif de développer des méthodes chimiométriques pour améliorer la robustesse des mesures spectrométriques appliquées aux agro-procédés.

1.3 Analyse théorique de la problématique

Le but de cette section est de réaliser une analyse théorique du problème. Pour cela, je présenterai brièvement les principes théoriques de la spectrométrie et de l'étalonnage de la mesure en lui donnant une interprétation géométrique. Ce point de vue me servira à dégager les principales causes de non robustesse des étalonnages et ainsi à identifier des voies de recherche.

1.3.1 Les modèles théoriques de la spectrométrie

Un premier modèle simple

Le principe de base de la spectrométrie repose sur la mesure de l'interaction entre un rayonnement électromagnétique et la matière à différentes fréquences. Connaissant l'intensité du rayonnement incident, la mesure de celle du rayonnement transmis, rétro-diffusé ou réfléchi est chargée d'information

sur la matière exposée au rayonnement. Si le rayonnement utilisé est photonique : Ultra Violet (UV), Visible (VIS), Proche Infrarouge (NIR) ou Moyen Infrarouge (MIR), le terme exact est spectrophotométrie. L'application principale de la spectrométrie est la chimie analytique. En effet, dans la gamme photonique (essentiellement dans l'infrarouge) on observe des absorptions dues aux fréquences de résonance des liaisons moléculaires. Cette absorption est modélisée par loi de Beer-Lambert. Pour une solution d'un composé en concentration c, elle relie l'intensité incidente $I_0(\lambda)$, l'intensité transmise $I(\lambda)$ et la longueur du chemin optique l, par la relation suivante :

$$I = I_0 e^{-k(\lambda)lc}$$

où $k(\lambda)$ est un coefficient d'extinction, fonction de la longueur d'onde λ. L'absorption d'un photon étant un phénomène quantique, le spectre $k(\lambda)$ présente des pics centrés sur des longueurs d'onde fondamentales (correspondant à des nombres d'onde donnés), ainsi que sur leurs harmoniques (multiples des nombres d'onde) et leurs combinaisons ([Lachenal, 2000]). Ceci forme donc un spectre caractéristique du composé chimique ayant interagi avec le rayonnement. Dans des conditions idéales, c'est à dire dans le cas où un seul composé est responsable de l'absorption, en utilisant une gamme de longueurs d'onde appropriée et en maîtrisant le chemin optique, on peut déterminer la concentration c à partir du spectre d'absorbance, théoriquement proportionnel au spectre caractéristique $k(\lambda)$:

$$A(\lambda) = -\ln \frac{I(\lambda)}{I_0(\lambda)} = k(\lambda)lc$$

Notons que dans ce cas idéal, il n'est nul besoin d'utiliser un spectromètre, puisqu'il suffit de mesurer l'absorption à une longueur d'onde correctement choisie. Cette mesure simple est d'ailleurs utilisée, sous le nom de densité optique, dans certaines applications particulières.

Des modèles plus affinés

Même dans des conditions de laboratoire, le cas idéal décrit ci-dessus est rarement rencontré. Dès que plusieurs composés sont présents en mélange, les absorbances se combinent.

En première approximation, on peut considérer que l'absorbance mesurée est la somme des absorbances dues à chaque composé. Ainsi, si n_c est le nombre de composés, l'absorbance résultante sera :

$$A(\lambda) = l \sum_{i=1}^{i=n_c} k_i(\lambda) c_i \tag{1.1}$$

Soit p le nombre de longueurs d'onde mesurées. Supposons que les *spectres purs* $k_i(\lambda)$ sont connus et que la concentration cherchée est c_1. Si k_1 présente un pic d'absorption dans une zone où les autres composés n'absorbent pas, nous sommes ramenés au cas idéal précédent. Par contre, si les pics se superposent le calcul de c_1 nécessitera plusieurs variables. Or, dans la réalité de la spectrométrie NIR, les spectres purs ne présentent pas de zone d'absorption nulle et la détermination de c_1 à partir de l'équation 1.1 nécessite que $p \geq n_c$ (résolution d'un système de p équations à n_c inconnues). Cette première constatation simple justifie l'intérêt du capteur multi-varié. Cette démarche analytique pure est en fait difficile à mettre en œuvre, car l'additivité des absorbances est encore un cas idéal.

En deuxième approximation, il convient d'ajouter un spectre d'absorption $k_0(\lambda)$ dû à des phénomènes physiques tels que la diffusion dans le milieu :

$$A(\lambda) = l \left(k_0(\lambda) + \sum_{i=1}^{i=n_c} k_i(\lambda) c_i \right) \tag{1.2}$$

En troisième approximation, on peut modéliser l'absorbance d'un mélange en tenant compte des interactions d'ordre 1 entre les solutés, selon la relation suivante :

$$A(\lambda) = l\left(k_0(\lambda) + \sum_{i=1}^{i=n_c} k_i(\lambda)c_i + \sum_{i,j=1,i>j}^{i,j=n_c} k_{ij}(\lambda)c_i c_j\right) \tag{1.3}$$

Une démarche analytique devient alors complexe, car elle nécessite la connaissance des spectres d'interaction.

Tous ces modèles représentent une approximation, un cas idéal. Toutefois, ils traduisent bien le phénomène latent qui relie la grandeur mesurée (le spectre) à la grandeur cherchée (une ou plusieurs concentrations). Ils sont donc d'une grande utilité dans le cadre de la construction et de l'interprétation des modèles d'étalonnage.

1.3.2 L'étalonnage de la mesure par spectrométrie

Cette section expose brièvement les principes de l'étalonnage multivarié de la mesure par spectrométrie, en se limitant au cas linéaire (une description détaillée de ces techniques peut être trouvée dans [Martens and Naes, 1989]). L'angle de vue adopté dans cet exposé permettra de mieux aprréhender les problèmes de robustesse, abordés ensuite dans la partie 1.3.3. Pour plus de détails sur les notations choisies, le lecteur se reportera utilement au chapitre qui y est dédié, page 9.

Principe général

Un grand nombre de techniques d'étalonnage ont été (et sont encore) développées, qui ont en commun la notion d'apprentissage. Elles utilisent donc une base constituée d'un ensemble de spectres auxquels correspondent des valeurs de la grandeur à prédire, mesurées par des méthodes de référence.

Soient donc $\mathbf{X}$ une matrice de n spectres acquis sur p longueurs d'onde et $\mathbf{y}$ le vecteur des n valeurs de concentration correspondantes[4]. Le problème de l'étalonnage est de trouver une fonction f (un modèle) telle que $\mathbf{y} = f(\mathbf{X}) + \mathbf{e}$, avec $\mathbf{e}$ (le résidu) petit. Une fois cette fonction trouvée, la mesure de la concentration $\widehat{y}$ d'un nouvel individu de spectre $\mathbf{x}$ sera donnée par $\widehat{y} = f(\mathbf{x})$. Dans le cas particulier des modèles linéaires, le problème consiste à trouver un vecteur $\mathbf{b}$, de dimension p et un scalaire b_0, tels que :

$$\mathbf{y} = \mathbf{X}\mathbf{b} + b_0 + \mathbf{e} \tag{1.4}$$

On suppose généralement que les résidus sont dûs à des bruits de mesure, de moyenne nulle. On a alors : $b_0 = \overline{\mathbf{y}} - \overline{\mathbf{X}}\mathbf{b}$.

Les modèles linéaires sont les plus couramment utilisés en spectrométrie, pour plusieurs raisons : relation théorique linéaire entre concentration et absorbance ; interprétation aisée ; faculté de généralisation réputée meilleure et plus prévisible ; temps de calcul très brefs. De plus, si une relation non linéaire est soupçonnée, il est possible d'effectuer un prétraitement non linéaire sur les spectres, puis d'utiliser un modèle linéaire. Nous nous limiterons donc aux modèles linéaires.

Etalonnage direct

Cette approche s'appuie sur une démarche analytique, basée sur la connaissance des spectres purs impliqués. Repartons du modèle exprimé par la relation 1.2. Supposons que le composé dont la concentration doit être mesurée est le numéro 1. Supposons $\mathbf{k}_0$ constant et omettons le facteur l. On a :

[4] D'autres grandeurs que des concentrations peuvent être mesurées par spectrométrie, mais pour des raisons de clarté, nous n'en parlerons pas ici.

$$\mathbf{X} = \mathbf{1}_n\mathbf{k}_0^T + \mathbf{y}\mathbf{k}_1^T + \mathbf{c}_2\mathbf{k}_2^T + \cdots + \mathbf{c}_{n_c}\mathbf{k}_{n_c}^T + \mathbf{R} \tag{1.5}$$

où $\mathbf{c}_i$ est le vecteur contenant les n concentrations du produit i, $\mathbf{k}_i$ le vecteur des p absorbances du spectre pur $k_i(\lambda)$ et $\mathbf{R}$ une matrice de la même dimension que $\mathbf{X}$, recueillant les bruits de mesure que nous supposerons de moyenne nulle. Cette relation peut être mise sous la forme matricielle :

$$\mathbf{X} = \mathbf{Y}\mathbf{K}^T + \mathbf{R} \tag{1.6}$$

où : $\mathbf{Y}$ est la matrice $(n \times n_c + 1)$ contenant une colonne de 1 plus les concentrations des n_c composés chimiques dans les n individus ; $\mathbf{K}$ est la matrice $(p \times n_c + 1)$ contenant le spectre $\mathbf{k}_0$ plus les spectres purs des composés.

– Si on suppose $\mathbf{k}_1$ connu, on peut extraire $\mathbf{y}$ de l'équation 1.5 par :

$$\mathbf{y} = \mathbf{X}\frac{\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} - \left(\frac{\mathbf{k}_0^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} + \mathbf{c}_2\frac{\mathbf{k}_2^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} + \cdots + \mathbf{c}_{n_c}\frac{\mathbf{k}_{n_c}^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1}\right) + \mathbf{R}\frac{\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1}$$

Ce qui donne :

$$\mathbf{b} = \frac{\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} \tag{1.7}$$

$$b_0 = \overline{y} - \overline{\mathbf{x}}^T\mathbf{b} = -\left(\frac{\mathbf{k}_0^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} + \overline{c}_2\frac{\mathbf{k}_2^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} + \cdots + \overline{c}_{n_c}\frac{\mathbf{k}_{n_c}^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1}\right)$$

$$\mathbf{e} = (\mathbf{c}_2 - \overline{\mathbf{c}}_2)\frac{\mathbf{k}_2^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} + \cdots + (\mathbf{c}_{n_c} - \overline{\mathbf{c}}_{n_c})\frac{\mathbf{k}_{n_c}^T\mathbf{k}_1}{\mathbf{k}_1^T\mathbf{k}_1} - \mathbf{R}\mathbf{b} \tag{1.8}$$

Le vecteur $\mathbf{b}$, donné donné par l'équation 1.7 et représenté sur la figure 1.1, est colinéaire à $\mathbf{k}_1$. L'équation 1.8 montre que la variance du résidu est directement liée aux variances des concentrations $\mathbf{c}_2, \cdots, \mathbf{c}_{n_c}$ et aux covariances entre $\mathbf{k}_1$ et les autres spectres purs. On retrouve alors une explication à la remarque faite en 1.3.1, à savoir que si les spectres purs ne sont pas indépendants, c'est à dire que les produits scalaires $\mathbf{k}_2^T\mathbf{k}_1, \cdots, \mathbf{k}_{n_c}^T\mathbf{k}_1$ ne sont pas nuls, les concentrations $\mathbf{c}_2, \cdots, \mathbf{c}_{n_c}$ polluent d'autant plus la mesure spectrale qu'elle varient.

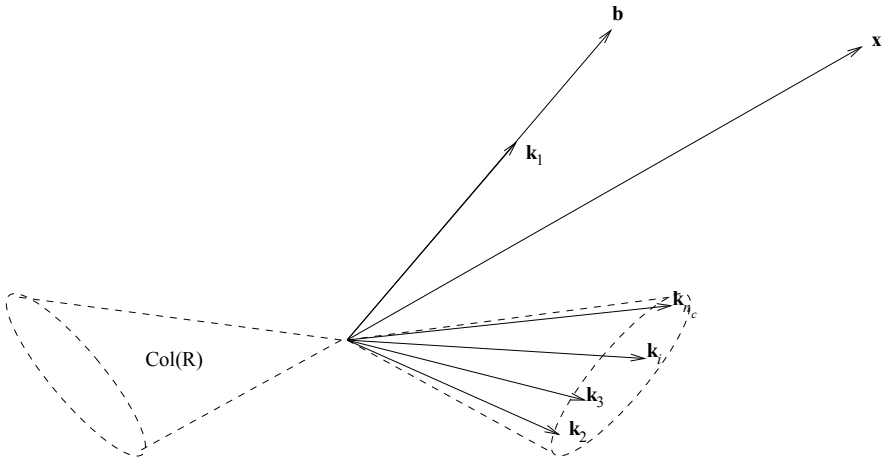

FIG. 1.1 – Etalonnage par projection sur le spectre pur $\mathbf{k}_1$ (fausse perspective dans $\mathbb{R}^p$).

– Si on connaît $\mathbf{K}$, c'est à dire tous les spectres purs, on peut utiliser l'équation 1.6 et obtenir :

$$\begin{aligned}\mathbf{Y} &= \mathbf{XK}(\mathbf{K}^T\mathbf{K})^{-1} - \mathbf{RK}(\mathbf{K}^T\mathbf{K})^{-1} = \mathbf{XB} + \mathbf{E}\\ \mathbf{b} &= \left[b_{21}b_{31}\cdots b_{(p+1)1}\right]^T\\ b_0 &= b_{11}\\ \mathbf{e} &= \mathbf{e}_1\end{aligned}$$

Dans cette approche, tous les spectres purs contribuent à $\mathbf{b}$, qui appartient au sous espace généré par $\mathbf{K}$, comme représenté sur la figure 1.2.

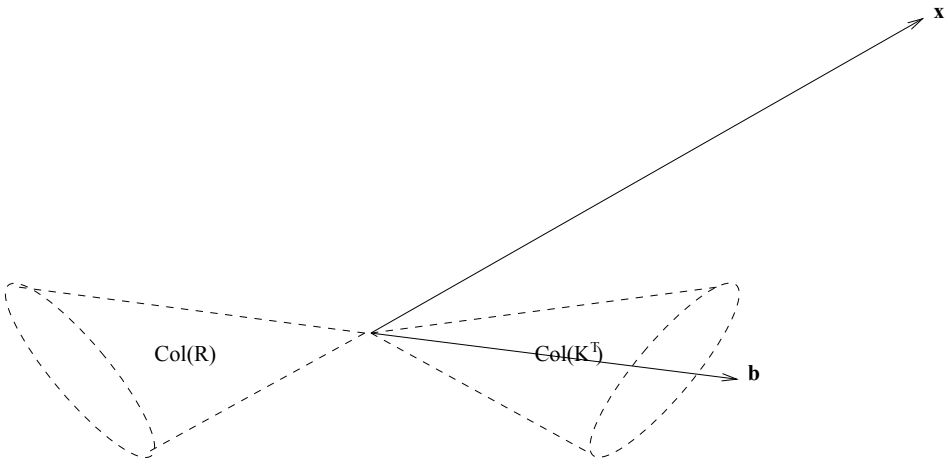

FIG. 1.2 – Etalonnage par projection sur le sous espace engendré par les spectres purs (fausse perspective dans $\mathbb{R}^p$).

– Si les spectres purs sont inconnus, mais que les concentrations de tous les composés sont données, on peut se ramener à la méthode précédente, en calculant une estimée de la matrice $\mathbf{K}$ par :

$$\widehat{\mathbf{K}} = \mathbf{X}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}$$

Toutes ces méthodes utilisent une décomposition de la matrice $\mathbf{X}$ sur le système de vecteurs $\{\mathbf{k}_1, \cdots, \mathbf{k}_{n_c}\}$ qui donne une bonne image de la réalité. Cependant, elles sont inadaptées aux cas où ni les spectres purs, ni les concentrations des composés annexes ne sont connus.

Étalonnage indirect

L'étalonnage indirect se propose d'apprendre la relation entre $\mathbf{X}$ et $\mathbf{y}$ sans connaissance *a priori*, comme les spectres purs.

Si les données sont préalablement centrées, on a $b_0 = 0$. Dans la suite, pour clarifier l'exposé, nous supposerons, soit que ce prétraitement est réalisé, soit que cette valeur de b_0 est recherchée. Cela veut dire que notre modèle est sans intercept : $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$.

Cette équation 1.4 est ordinairement résolue par la méthode de régression aux moindres carrés (MLR), qui donne :

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

Dans le cas de la spectrométrie, l'application de cette méthode pose deux problèmes :

– Comme le nombre de variables est élevé, on a souvent moins d'individus que de variables ($n < p$). Dans ce cas, $\mathbf{X}^T\mathbf{X}$ n'est pas inversible, car son rang est inférieur à sa dimension.
– Même si le nombre d'individus est suffisant pour éviter le problème calculatoire précédent, un autre écueil réside dans l'inter-corrélation des variables. Cette situation, connue sous le nom de *problème mal conditionné*, aboutit à un modèle très incertain. En effet, la variance des composantes de $\mathbf{b}$ est directement reliée aux valeurs propres de $(\mathbf{X}^T\mathbf{X})^{-1}$ et plus les variables de $\mathbf{X}$ sont corrélées, plus ces valeurs propres sont grandes. Dans notre cas, les variables sont par nature très corrélées car nous mesurons un spectre qui possède une certaine continuité : le signal mesuré à une longueur d'onde donnée est très dépendant de ceux mesurés aux longueurs d'onde voisines. La figure 1.3 illustre cette intercorrélation en montrant l'évolution du coefficient de détermination moyen entre $\mathbf{x}_i$ et $\mathbf{x}_j$ en fonction de la distance entre les deux longueurs d'onde λ_i et λ_j, calculée sur une base de $n = 2495$ spectres en transmission VIS - VNIR de grains de raisins. Deux variables distantes de 3.2 nm, c'est à dire voisines dans $\mathbf{X}$, sont liées par un R^2 de 0.998 ; deux variables distantes de 32 nm, c'est à dire de 10 colonnes dans $\mathbf{X}$ sont encore liées par un R^2 de 0.927.

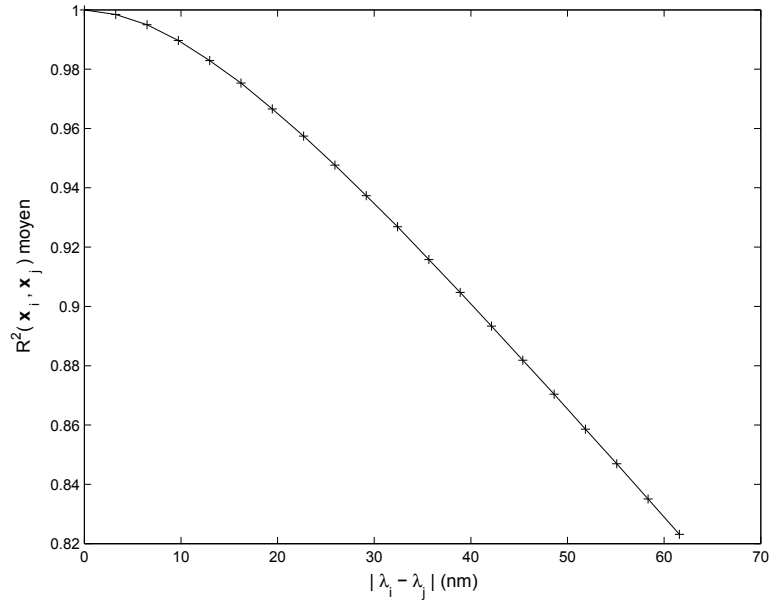

FIG. 1.3 – Coefficient de détermination moyen entre le signal mesuré à λ_i et celui mesuré à λ_j, en fonction de $|\lambda_i - \lambda_j|$. Calculs réalisés sur une base de $n = 2495$ spectres de transmission VIS - VNIR de grains de raisins.

C'est pour résoudre ces problèmes qu'ont été développées les méthodes de régression factorielle, dont une bonne description peut être trouvée dans [Martens and Naes, 1989]. Dans la suite de ce paragraphe, ne sera détaillée que la méthode la plus connue : la régression aux moindres carrés partiels (PLSR).

Les régressions factorielles reposent sur la décomposition de $\mathbf{X}$ selon une base de $Col(\mathbf{X})$. Les vecteurs de cette base sont appelés des *variables latentes*. Ces nouvelles variables, de par leur construction, pourront être utilisées pour établir une régression stable (bien conditionnée) avec $\mathbf{y}$.

Soit r le rang de la matrice $\mathbf{X}$; $r = \dim(Col(\mathbf{X}))$. L'équation 1.9 montre la décomposition de $\mathbf{X}$ sur les variables latentes $\{\mathbf{u}_1, \cdots, \mathbf{u}_r\}$ (vecteurs de dimension p), faisant apparaître les scores $\{\mathbf{t}_1, \cdots, \mathbf{t}_r\}$ (vecteurs de dimension n) :

$$\mathbf{X} = \mathbf{t}_1\mathbf{u}_1^T + \mathbf{t}_2\mathbf{u}_2^T + \cdots + \mathbf{t}_r\mathbf{u}_r^T \tag{1.9}$$

Les variables latentes sont obtenues par un processus itératif qui maximise la valeur absolue de la

covariance entre $\mathbf{Xu}$ et $\mathbf{y}$. Les scores sont orthogonaux dans R^n, c'est à dire que les variables latentes sont non corrélées. Les variables latentes sont ordonnées par ordre d'intérêt décroissant, c'est à dire que $|\text{Cov}(\mathbf{t}_i, \mathbf{y})| > |\text{Cov}(\mathbf{t}_{i+1}, \mathbf{y})|$.

On sélectionne ensuite les k premières variables latentes, séparant ainsi $Col(\mathbf{X})$ en deux sous espaces supplémentaires : U_k et U_k^+. Soit $\mathbf{T}_{(n,k)}$ la projection de $\mathbf{X}$ sur U_k ; on réalise une MLR entre $\mathbf{T}$ et $\mathbf{y}$. En appelant $\mathbf{V}_{(p,k)}$ la matrice de changement de base de $\mathbb{R}^p$ vers U_k, on a donc :

$$\begin{aligned} \mathbf{T} &= \mathbf{XV} \\ \widehat{\mathbf{y}} &= \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y} \\ \widehat{\mathbf{y}} &= \mathbf{XV}(\mathbf{V}^T\mathbf{X}^T\mathbf{XV})^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{y} \end{aligned}$$

Donc, finalement,

$$\mathbf{b} = \mathbf{V}(\mathbf{V}^T\mathbf{X}^T\mathbf{XV})^{-1}\mathbf{V}^T\mathbf{X}^T\mathbf{y} \quad \text{avec} \quad \mathbf{V} = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}$$

La décomposition de $\mathbf{X}$ en variables latentes, de l'équation 1.9 est plus efficace que la décomposition sur les spectres purs, de l'équation 1.5 en ce que les k scores retenus sont orthogonaux deux à deux, assurant une régression linéaire stable.

Par contre, alors que dans l'équation 1.5, les concentrations ($\mathbf{y}$) étaient bien isolées, dans l'équation 1.9, elles sont mêlées aux autres concentrations dans un ensemble de composantes. C'est pour cette raison que plusieurs variables latentes sont nécessaires pour reconstituer l'information. Le choix de k est d'ailleurs une étape importante de la réalisation des modèles par régression factorielle.

Un modèle réalisé par régression factorielle doit donc opérer une bonne séparation de l'espace $\mathbb{R}^p$ en deux sous espaces : un premier, U_k porteur d'information utile, généré par les k variables latentes et l'autre, U_k^+ recueillant l'information parasite due aux autres composés, aux phénomènes physiques et aux bruits, généré par le reste des variables latentes. La figure 1.4 illustre cette décomposition.

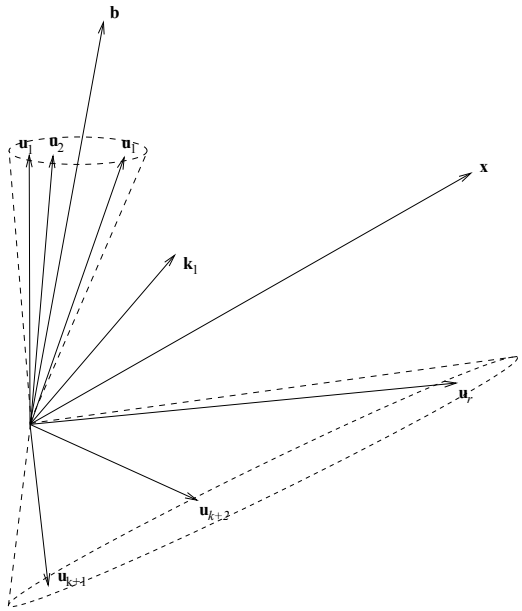

FIG. 1.4 – Etalonnage par projection sur des variables latentes (fausse perspective dans $\mathbb{R}^p$).

Le nombre k de variables latentes définit la dimension du sous espace dans lequel on opére la régression, que j'appellerai *dimension du modèle*. Le réglage de k peut s'opérer en observant la faculté de prédiction du modèle. On calcule l'erreur standard d'étalonnage (SEC), par application du modèle

sur l'ensemble d'apprentissage et l'erreur standard de validation (SEV), par application du modèle sur un ensemble de validation, pour des valeurs croissantes de k. On peut aussi, si l'on ne dispose pas de deux ensembles indépendants, procéder par validation croisée ([Wold, 1978]) sur l'ensemble d'apprentissage, et observer l'évolution de l'erreur standard de validation croisée (SECV) en fonction de k[5]. Alors que le SEC décroît constamment, on doit observer un minimum pour le SEV (ou le SECV). La valeur correspondante (k_0) est alors proche de la dimension "idéale" du modèle.

Il convient alors d'examiner le vecteur $\mathbf{b}$ pour des valeurs de k voisines de k_0. Ce vecteur, qui a la dimension d'un spectre, peut être tracé comme tel. On peut ainsi tenter d'interpréter ses formes (les pics, les alternances, etc.), qui doivent être reliées à des longueurs d'onde d'absorption (fondamentales, harmoniques ou combinaisons) des composés chimiques cherchés. Il est d'autre part d'usage de préférer des coefficients $\mathbf{b}$ peu "chahutés" et de norme faible.

Une autre méthode de réglage de k est d'ailleurs basée sur ce principe : Elle consiste à observer l'évolution de la norme de $\mathbf{b}$ en fonction de k ; La valeur du nombre k pour lequel cette norme croît brusquement est jugée comme la limite inférieure du sur-apprentissage.

La figure 1.5 montre une comparaison des deux méthodes de réglage de k, sur un exemple réel d'étalonnage de la mesure du taux de sucres des pommes par spectrométrie NIR. L'échantillon est composé de $n = 80$ individus, les spectres contiennent $p = 90$ absorbances (de 795 à 1080 nm) et la régression est effectuée par PLSR. La validation croisée est effectuée avec 80 blocs (leave one out). On y constate une assez bonne concordance entre les deux méthodes de réglage de k, qui donnent toutes deux une valeur de k_0 voisine de 9. La figure 1.6 montre l'évolution de $\mathbf{b}$ pour des valeurs de k allant de 5 à 12. On y voit clairement que pour $k \geq 9$, la forme de $\mathbf{b}$ devient très chahutée.

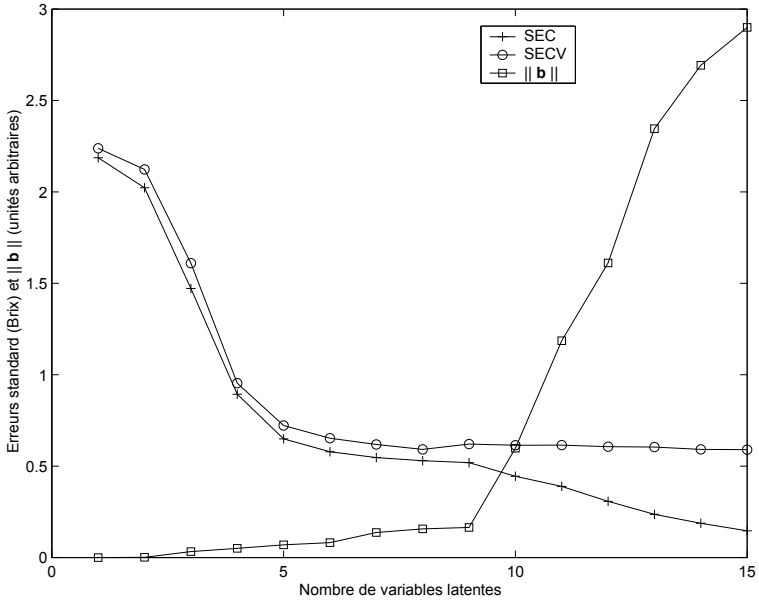

FIG. 1.5 – Exemple d'évolution comparée des erreurs standard (SEC et SECV) et de la norme de $\mathbf{b}$ en fonction du nombre k de variables latentes.

1.3.3 Les causes de non robustesse des étalonnages multivariés

Cette section se propose d'expliciter le problème de la robustesse des étalonnages multivariés. A partir de l'interprétation faite dans la partie 1.3.2, je dégagerai les principales causes de non robustesse.

[5]La validation croisée apparaît bien comme une méthode de réglage d'un modèle, mais ne dispense en rien de vrais tests.

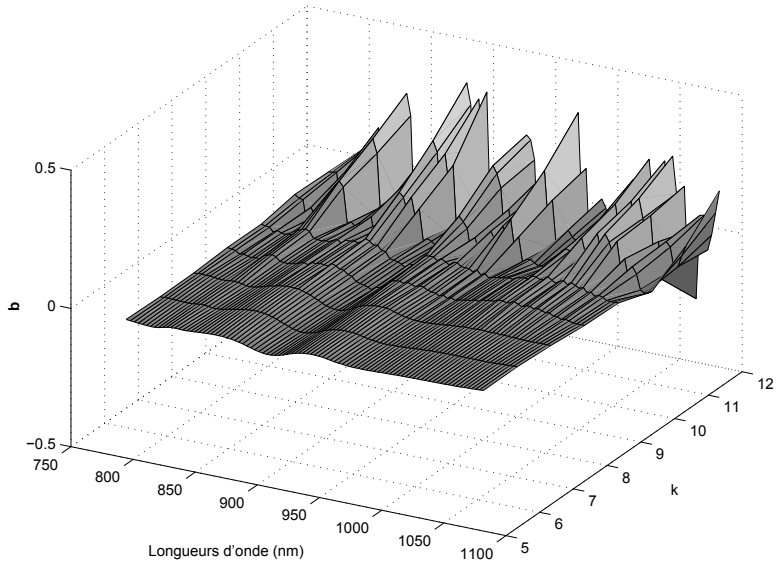

FIG. 1.6 – Exemple d'évolution de $\mathbf{b}$ en fonction du nombre k de variables latentes.

Définition et évaluation de la robustesse

Le terme de robustesse n'est pas défini clairement par la métrologie, où l'on trouve plutôt ceux de répétabilité, reproductibilité, justesse, fidélité, etc. Dans [Zeaiter et al., 2004], nous avons réalisé une recherche bibliographique et proposé la définition suivante :

La robustesse d'un modèle d'étalonnage multivarié traduit la stabilité de sa capacité prédictive au regard de perturbations au voisinage des conditions standard.

Nous nous intéresserons donc, dans la suite de cette section à l'effet d'une perturbation $\delta\mathbf{x}$ qui s'ajoute au spectre mesuré $\mathbf{x}$. L'analyse du problème de la robustesse du modèle revient alors à étudier comment la perturbation $\delta\mathbf{x}$ se propage à travers le modèle, pour produire une erreur $\delta\widehat{y}$, donnée par :

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

Ce qui donne :

$$|\delta\widehat{y}| = \|\delta\mathbf{x}\| \cdot \|\mathbf{b}\| \cdot |\cos(\delta\mathbf{x}, \mathbf{b})| \tag{1.10}$$

Dans la suite de cette analyse, j'examinerai différentes causes de non robustesse au travers de cette équation.

La dimension du modèle

L'étalonnage par régression factorielle nécessite de choisir un nombre k de variables latentes, qui définit la dimension de l'espace dans lequel on projette le spectre $\mathbf{x}$. Le choix de ce paramètre influence fortement la qualité du modèle.

Reprenons le modèle 1.3. Si le bruit de mesure et les variations de $k_0(\lambda)$ sont suffisamment faibles, les premières variables latentes seront construites principalement sur les spectres purs et les spectres d'interaction et les suivantes sur le bruit résiduel. Si tous les spectres sont dépendants, l'extraction de l'information relative à un composé nécessite autant de variables latentes que de spectres (purs et d'interaction), c'est à dire $k_0 = n_c(n_c + 1)/2$. Si la valeur de k est trop faible, on risque d'omettre des phénomènes d'interaction ou des dépendances de spectres et d'obtenir un modèle avec un faible

pouvoir explicatif. Par contre, si la valeur de k est trop forte, on risque d'inclure dans le modèle des variables latentes faiblement significatives, basées sur du bruit. Ce sur-apprentissage produit des modèles peu robustes. En effet, cela revient à inclure dans le modèle la partie du bruit de mesure qui est corrélée avec $\mathbf{y}$. Cette corrélation, par nature fortuite, ne se représente pas lors de la prédiction d'un individu test. Les variables latentes en surnombre provoquent alors une augmentation de l'erreur de prédiction.

Ce principe de minimalité de la dimension du modèle, dit *principe de parcimonie*, a été théoriquement explicité dans [Seasholtz and Kowalski, 1993]. On peut aussi lui donner une interprétation géométrique, grâce à l'équation 1.10. En effet, comme on l'a vu en 1.3.2, la norme de $\mathbf{b}$ augmente avec k. Un mauvais dimensionnement du modèle peut donc être une source de non robustesse.

Le nombre de variables

Nous avons vu que, sauf dans le cas idéal d'un composé unique, mesuré par une gamme spectrale appropriée, la mesure multivariée est nécessaire. Cependant, la prise en compte dans le modèle d'un trop grand nombre de variables peut aussi s'avérer néfaste :

- On augmente la probabilité de capter des perturbations, qui viendront grossir $\delta\mathbf{x}$.
- On complexifie le modèle, ce qui induit une augmentation du nombre de variables latentes.
- On dilue l'information dans des spectres sans intérêt, rendant la construction des variables latentes moins pertinente.
- On augmente artificiellement et inutilement les termes $\|\delta\mathbf{x}\|$ et $\|\mathbf{b}\|$ de l'équation 1.10. Théoriquement, une zone spectrale qui n'est pas reliée à la concentration recherchée doit correspondre à des valeurs de $\mathbf{b}$ nulles. En pratique, les valeurs calculées sont non nulles (bien que faibles) et viennent grossir de manière quadratique la norme de $\mathbf{b}$.
- On risque d'augmenter la colinéarité entre $\delta\mathbf{x}$ et $\mathbf{b}$. Supposons que, avec un ensemble de variables minimal, la perturbation $\delta\mathbf{x}$ se trouve normalement dans l'espace orthogonal à $\mathbf{b}$, c'est à dire que la régression l'a bien identifié comme un bruit. L'adjonction de variables inutiles risque d'amener des perturbations qui modifient la direction de $\delta\mathbf{x}$ et le feront sortir de l'espace orthogonal à $\mathbf{b}$ et donc augmenter le terme $|\cos(\delta\mathbf{x}, \mathbf{b})|$ de l'équation 1.10.

Toutes ces raisons, par ailleurs partiellement redondantes, montrent bien que l'augmentation du nombre de variables est de nature à diminuer la robustesse du modèle.

La corrélation entre les perturbations et le modèle

L'équation 1.10 montre de manière évidente que la perturbation $\delta\mathbf{x}$ aura d'autant plus d'effet qu'elle sera colinéaire à $\mathbf{b}$. Le modèle ne sera parfaitement robuste que si $\delta\mathbf{x}$ est orthogonal à $\mathbf{b}$. Une non orthogonalité peut avoir plusieurs causes :

- Soit $\delta\mathbf{x}$ est l'expression d'une variation qui n'était pas présente dans les exemples ayant servi à l'apprentissage. La construction des variables latentes n'en a donc pas tenu compte. Par exemple, $\delta\mathbf{x}$ est dû à la variation d'une concentration d'un composé secondaire, qui était constante sur tous les individus de la base d'étalonnage ou à l'effet d'une grandeur d'influence non prise en compte dans l'étalonnage.
- Soit $\delta\mathbf{x}$ est de nature aléatoire et, bien que présent dans la base d'étalonnage, il a une direction variable. Cette hypothèse correspond par exemple au modèle 1.2 avec un terme $k_0(\lambda)$ de forme changeante. Il devient alors plus difficile pour la régression d'identifier un sous espace orthogonal à $\delta\mathbf{x}$.

1.4 Récapitulatif de ma problématique - Identification des voies de recherche

Le but de cette section est de lister les voies de recherche que j'ai suivies pour résoudre ma problématique. En croisant les problèmes listés dans la section 1.2 avec l'analyse de la section 1.3, je mettrai à jour les hypothèses qui ont guidé mes travaux de recherches.

Reprenons chacun des problèmes identifiés :

Complexité du signal

Les spectres purs des composés recherchés sont dominés par le spectre de l'eau toujours majoritaire dans les produits qui nous intéressent et par un effet de matrice (structure physique, suspensions, etc.). De ce fait, la norme de $\mathbf{k}_1$ est faible par rapport à celle de $\mathbf{k}_2$ (si l'eau est le composé numéro 2) et celle de $\mathbf{k}_0$. Cela se traduit donc par une dimension importante du modèle, qu'il faudra réduire par la **sélection des variables**, afin de concentrer la mesure spectrale sur les zones essentielles. Des **prétraitements géométriques** des spectres devront aussi être recherchés pour diminuer la norme de $\mathbf{k}_0$.

Grandeurs d'influence

Les grandeurs d'influence telles que la température du produit, ou la température de couleur de la source lumineuse déforment les spectres mesurés. Ces déformations peuvent être décomposées en deux termes. Le premier est une déformation verticale simple du spectre (translation, ajout d'une ligne de base, etc.), s'apparentant à un terme $\mathbf{k}_0$, qui pourra être pris en charge par les **prétraitements géométriques**. Le deuxième correspond à une déformation plus complexe, comme par exemple une translation horizontale, une dilatation (ou contraction) horizontale, ou le décalage de certains pics. Un traitement spécifique de **prise en compte de la grandeur d'influence** doit alors être mis en oeuvre. Dans le cas du suivi en ligne de procédés longs (plusieurs jours), se pose le problème de la dérive des conditions opératoires. Ce problème de **maintenance de la robustesse des modèles utilisés en ligne** mérite d'être traité séparément, car le nombre et la nature des grandeurs d'influence est difficile à appréhender.

Présentation de l'échantillon

La mauvaise présentation de l'échantillon peut provoquer des effets récurrents s'apparentant à ceux d'une grandeur d'influence et pourront donc être traités comme tels. Par exemple, si nous considérons la mesure en ligne sur des fruits, dans le cas d'une mesure sans contact, la distance entre le capteur et l'échantillon peut être vue comme une grandeur d'influence. Cependant, ces effets peuvent rapidement atteindre des dimensions très importantes. C'est le cas notamment lorsque le spectromètre capte du rayonnement issu d'une réflexion spéculaire sur la peau d'un fruit. D'autres effets, moins continus, correspondent à des incidents de mesure. C'est le cas par exemple lorsque le capteur se positionne sur le pédoncule ou sur le calice d'un fruit. Il est donc nécessaire de mettre au point un système de **discrimination des spectres** qui permette de rejeter les cas aberrants. Cet aspect, bien qu'à la limite de la problématique de la robustesse telle que définie en 1.3.3, est fondamental dans notre cadre applicatif.

Variabilité de l'échantillon

Les problèmes posés par la variabilité de l'échantillon sont complexes. Non pas que les effets sur le spectre mesuré soient particuliers, mais surtout parce que la variabilité elle même est parfois difficile à apprécier. La structure physique d'un fruit, par exemple, ne peut être approchée que par des mesures mécaniques peu fiables, comme la fermeté (peu répétable), l'élasticité (peu indicative),

etc. L'analyse complète de sa composition chimique fait appel à des techniques de mesure lourdes et coûteuses. Le parti a donc été pris de traiter les sources de variabilité identifiables (variété, origine, maturité, etc.) comme des grandeurs d'influence.

Voies de recherche identifiées

Les pistes identifiées ci-dessus peuvent être regroupées en quatre voies de recherche :

1. L'optimisation intrinsèque du modèle, visant la réduction de sa complexité, au moyen de la sélection des variables ou/et de prétraitements géométriques
2. La prise en compte des grandeurs d'influence
3. La maintenance de la robustesse des modèles utilisés en ligne
4. La discrimination à partir des spectres.

L'optimisation intrinsèque des modèles d'étalonnage se réfère aux méthodes susceptibles d'en améliorer la robustesse sans observer les effets des perturbations. C'est, bien naturellement, un sujet très étudié et de nombreuses méthodes existent. Dans [Article III], nous en avons réalisé une revue en discutant de leur effet sur la robustesse (Cf 3.2). Je n'ai pas produit de méthodes nouvelles dans cette voie de recherche. Je me suis contenté d'appliquer certaines méthodes existantes, sans lesquelles les capteurs développés dans les projets cités en 1.1.2 n'auraient certainement pas été opérationnels. J'ai accordé une attention toute particulière à la sélection de variables, comme en témoignent deux publications :

- La première ([Roger and Bellon-Maurel, 2000]) décrit l'utisation des algorithmes génétiques comme méthode de recherche stochastique de la meilleure sélection de variables, dans le cadre de la prédiction du taux de sucres des cerises par spectrométrie VNIR.
- La seconde ([Article V]) avait pour application la discrimination de cépage à partir de la courbe de cinétique fermentaire de vinification. Même si la mesure n'est pas spectrométrique, la problématique du traitement de l'infomation est similaire. Une comparaison de différentes techniques de discrimination et de sélection de variables associée y a été décrite (Cf 4.2).

Les autres voies de recherche sont reprises individuellement dans les chapitres suivants, car elles ont fait l'objet d'une production méthodologique.

Chapitre 2

La prise en compte des grandeurs d'influence

Ce chapitre traite de la deuxième voie de recherche identifiée en 1.4, c'est à dire l'amélioration de la robustesse vis à vis des grandeurs d'influence. Une première partie dresse un panorama des méthodes, en s'appuyant sur l'article [Article I] et propose une méthodologie générale de choix d'une classe de méthodes en fonction de certaines contraintes. La seconde partie expose une contribution originale, publiée dans [Article II], qui permet de rendre l'étalonnage moins sensible à une grandeur d'influence donnée. Enfin, la troisième partie propose des perspectives de recherche.

2.1 Introduction - Notations

Le terme de grandeur d'influence est emprunté à la métrologie, où il désigne toute grandeur qui, appliquée de l'extérieur, est susceptible de modifier les caractéristiques de la mesure. Toute grandeur physique ou chimique qui, s'appliquant sur l'échantillon ou sur le spectromètre, ajoute une composante $\delta\mathbf{x}$ au spectre est de nature à altérer la robustesse du modèle d'étalonnage. Nous proposons en outre d'étendre cette notion à toute condition extérieure stable qui influence la mesure spectrale, comme, par exemple un changement de spectromètre. Nous parlerons dans ce cas de grandeur discrète. Dans la mesure où, pour des raisons techniques ou économiques, il n'est pas toujours possible d'éliminer l'application de la grandeur d'influence, il nous faut trouver un moyen de diminuer son effet, c'est à dire de rendre le modèle moins sensible à $\delta\mathbf{x}$.

Soit G une grandeur d'influence dont l'effet est indésirable et soit g sa valeur. Dans le cas d'une grandeur physique ou chimique, g est un scalaire à valeur quasi-continue. Dans le cas d'une grandeur discrète, g prend sa valeur dans un ensemble de symboles dans lequel il n'existe pas de relations d'ordre. Par exemple, g peut coder la variété d'une pomme.

2.2 Les grandes classes de méthodes - Stratégie de choix

Cette section est reliée à un article publié dans *Journal of Near Infra Red Spectroscopy* ([Article I]). Pour clarifier la rédaction, j'ai choisi de ne traiter que des grandeurs continues. La transposition au cas des grandeurs discrètes ne change en rien le fond de ce qui va suivre.

2.2.1 Théorie et stratégie

Comme cela a été exposé en 1.1, nos recherches visent à développer des capteurs appelés à fonctionner dans des conditions "hostiles". Le problème des grandeurs d'influence y est primordial, au contraire des laboratoires, où elles peuvent être contrôlées. La manière dont ces grandeurs vont être prises en compte contraint fortement le cahier des charges du projet.

Prenons l'exemple d'un appareil portable destiné à mesurer au champ les caractéristiques internes d'un fruit, utilisant un spectromètre simplifié, construit avec un source et quelques filtres. Si pour combattre l'effet de la lumière ambiante, il est nécessaire de normaliser le spectre, il faudra prévoir lors de la conception une ou plusieurs longueurs d'onde dédiées à cette opération.

Or, de nombreuses méthodes peuvent être utilisées pour combattre les effets d'une grandeur d'influence. Il nous est donc apparu nécessaire de développer une stratégie pour guider le choix de la méthode à mettre en œuvre pour chaque grandeur d'influence identifiée. La figure 2.1 illustre cette stratégie.

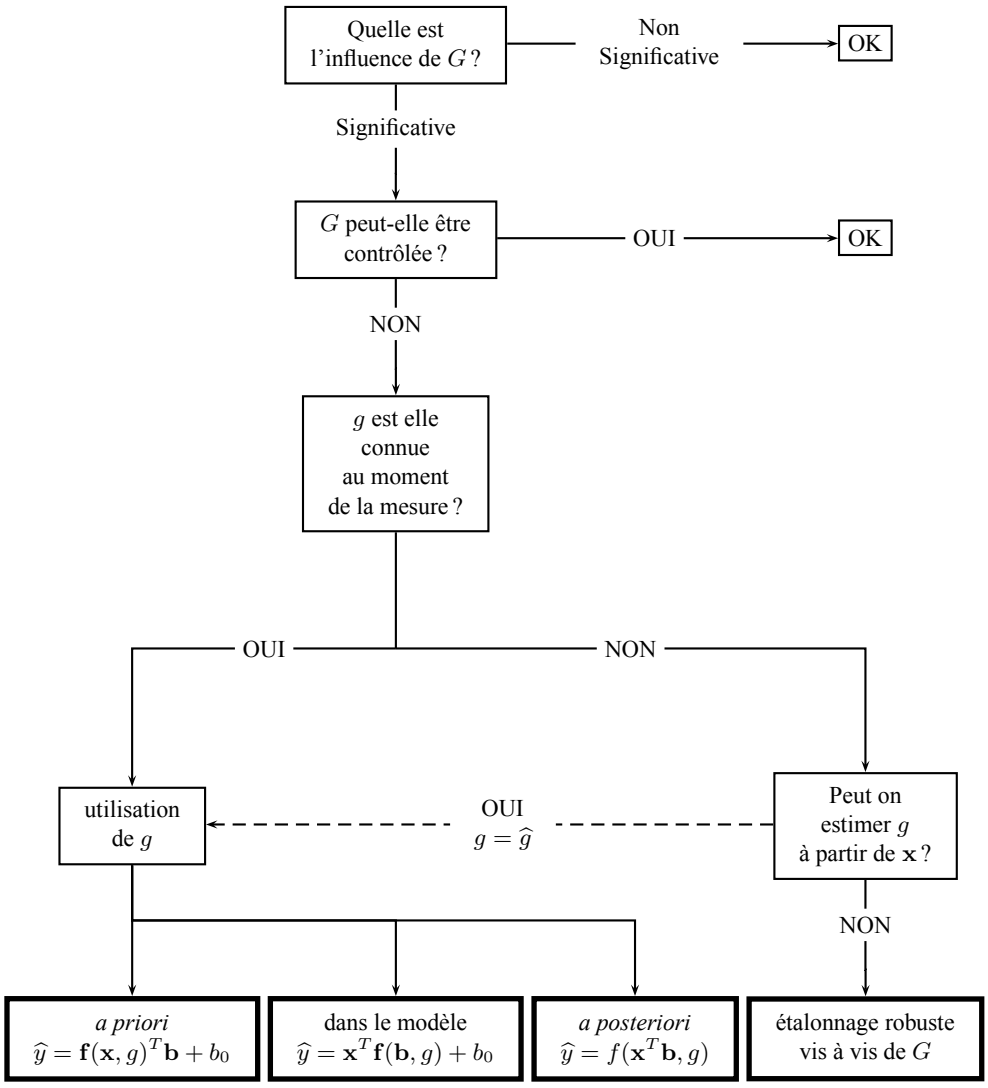

FIG. 2.1 – Stratégie de prise en compte d'une grandeur d'influence, selon que sa valeur est disponible ou non, au moment de l'utilisation du modèle.

La première étape consiste à évaluer l'influence de G sur la mesure spectrale et donc sur le modèle d'étalonnage, par exemple au moyen d'un plan d'expérience. Le résultat de ce test fournit un indice de l'influence de G, par exemple, si la grandeur d'influence est continue, sous la forme de la variation du SEP en fonction de g :

$$I_G = \frac{\partial(SEP)}{\partial g}$$

Prenons le cas d'une seule grandeur d'influence. Le test précédent nous permet alors de décider

de la significativité de l'influence de G : Si SEP_0 est le SEP du modèle constaté en l'absence de grandeur d'influence ; I_G l'indice d'influence de G ; si SEP_{max} est le SEP du cahier des charges du capteur ; et si ΔG représente la plage de variation admissible de G, on décidera de la significativité de G si $(I_G \Delta G)^2 + SEP_0^2 > SEP_{max}^2$.

Dans le cas (courant) de plusieurs grandeurs d'influences $(G_1, \cdots, G_k)$, on tirera profit de la méthode des plans d'expérience pour exprimer leurs interactions. A partir des plages de variation de ces grandeurs $(\Delta G_1, \cdots, \Delta G_k)$, on tracera la surface d'erreur, qui nous donne le SEP en fonction de toutes les valeurs $g_1, \cdots, g_k$, sur le domaine $\Delta G_1 \times \cdots \times \Delta G_k$. Le même raisonnement est alors tenu pour chaque grandeur d'influence G_i, mais en tenant compte des autres grandeurs d'influence. Cette méthode est en fait très complexe à mettre en pratique dans le cas où les effets des grandeurs d'influence ne sont pas indépendants. C'est pourquoi, dans une première approche, il est commode de considérer chaque grandeur séparément, comme nous l'avons décrit dans :

- [Sánchez et al., 2003], où nous avons étudié les influences combinées de : la température du fruit ; la température du spectromètre ; l'intensité de la lumière ambiante sur le prototype GLOVE (appareil portable de caractérisation interne des fruits)
- [Chauchard et al., 2004], où nous avons testé la technique de modélisation non paramétrique LS-SVM (pour Least Square Support Vector Machine) pour la prédiction des acides organiques dans le raisin de cuve par spectrométrie NIR. La robustesse du modèle a été testée vis à vis de différents bruits de mesure simulés, tels que : bruit gaussien, décalage en longueurs d'onde, diverses autres déformations du spectre.

La deuxième étape dépend du contexte. A partir du moment où il a été prouvé que l'influence de G est significative, il est très naturel d'essayer d'en éliminer les variations. Par exemple, la régulation en température du spectromètre, qui limite les déformations de ses pièces optiques, est une solution souvent employée en industrie. Cependant, la mise en œuvre de ce contrôle n'est pas toujours techniquement ou économiquement opportune. C'est le cas du contrôle en température du spectromètre pour un appareil portable. De même, s'agissant de la température des fruits sur une ligne de tri, imposer que tous les fruits soient à la même température peut induire une organisation du travail économiquement non viable.

C'est l'arborescence qui se déroule en dessous de la **troisième étape** qui nous intéresse plus particulièrement. De façon tout aussi pragmatique que précédemment, elle commence par une distinction sur la disponibilité de g au moment de l'utilisation du modèle.

Les méthodes suivantes ont été identifiées :

- Cas où g est connue :

  méthode 1.1 - Correction *a priori* **:** Il s'agit de modifier le spectre acquis $\mathbf{x}$ grâce à la valeur g et ainsi construire un nouveau vecteur $\mathbf{x}^*$. Ceci peut s'opérer de plusieurs manières :
  - Si on connait la relation entre g et $\delta\mathbf{x}$, on peut envisager d'opérer une réelle correction du spectre, pour le rétablir dans sa valeur "normale". On utilise ensuite le modèle non modifié, avec $\mathbf{x}^*$. En fait, cette opération est très délicate car, soit elle nécessite une connaissance experte des phénomènes liant g et $\delta\mathbf{x}$ qui sont souvent complexes, soit elle procède par apprentissage où la variable à prédire est de dimension très supérieure à la variable prédictive.
  - Plutôt que d'essayer de rétablir $\mathbf{x}$ dans sa forme originale, une autre solution consiste à ajouter aux variables prédictives du modèle des nouvelles variables dépendant de g. Par exemple, on réalise un étalonnage avec une colonne supplémentaire dans $\mathbf{X}$, contenant la variable g pour chaque individu.

  méthode 1.2 - Correction du modèle : Les coefficients de la régression $\mathbf{b}$ sont modifiés grâce à la valeur g pour donner un nouveau vecteur $\mathbf{b}^*$. Comme précédemment, l'apprentissage direct de cette correction s'avère difficile. On lui préfèrera plutôt une approche discrète,

sous forme de régression locale. Il s'agit de sélectionner un modèle dans une bibliothèque de modèles étalonnés pour différentes valeurs $(g_1, \cdots, g_k)$ de g. On pourra adopter le modèle correspondant à la valeur g_i la plus proche de g, voire calculer une interpolation entre les deux valeurs g_i et g_{i+1} encadrant g. Cette approche est tout à fait adaptée au cas des grandeurs discrètes. Notons que b_0 peut aussi être modifié par ce processus.

méthode 1.3 - Correction *a posteriori* : Cette méthode consiste à appliquer un deuxième modèle de correction, à la suite du modèle originel. Classiquement, un modèle linéaire est utilisé, sous le nom de correction "biais - pente". D'autres formes de fonction, plus complexes, peuvent être envisagées, par exemple pour prendre en compte des effets non linéaires.

- Cas où g est estimable à partir de $\mathbf{x}$: Ce cas de figure, représenté par le lien en pointillé sur la figure 2.1 n'est en fait qu'un cas particulier du cas ou g est mesurable. Puisque g a une influence sur le spectre, on peut espérer construire un modèle de prédiction de g à partir de $\mathbf{x}$, qui nous fournira une estimée $\widehat{g}$ que l'on peut ensuite utiliser dans l'une des méthodes ci-dessus.
- Cas où g est inconnue :

méthode 2.1 - Etalonnage exhaustif : Cette méthode, parfois abusivement dénommée étalonnage robuste, consiste à étalonner le modèle sur une base contenant à la fois des variations de y et de g. Comme nous l'avons vu en 1.3.2, cette précaution est nécessaire pour que la construction des variables latentes tienne compte de l'effet de g. Cette méthode est simple et intuitive ; c'est son principal avantage et la raison pour laquelle elle est très couramment utilisée. Son principal défaut est qu'elle requiert des bases d'étalonnage très volumineuses, surtout quand on s'adresse à plusieurs grandeurs d'influence.

méthode 2.2 - Etalonnage conjoint : L'idée de cette méthode est de construire un modèle qui prédit coinjointement y et g, de manière à contraindre l'indépendance des deux prédictions. Ceci est possible avec l'algorithme original de la PLS (NIPALS), qui gère des réponses multi-colonnes. L'énorme avantage de cette méthode est qu'elle fournit une estimation $\widehat{g}$ de la grandeur d'influence. Par contre, la base d'étalonnage devra être construite comme précédemment et cette méthode présente donc le même inconvénient que l'étalonnage exhaustif.

méthode 2.3 - Sélection de variables orientée : Cette méthode se propose de sélectionner les variables à la fois utiles à la prédiction de y et peu sensibles à la variation de g. Un des moyens les plus simples consiste à opérer une sélection basée sur la minimisation de l'erreur de prédiction sur un ensemble de test contenant des valeurs de g différentes. On pourra employer un des nombreux algorithmes de recherche de minimum, comme les algorithmes génétiques, les procédures pas à pas, etc.

méthode 2.3 - Orthogonalisation par rapport à la grandeur d'influence : J'ai développé une méthode, baptisée EPO, pour External Parameter Orthogonalisation, qui est détaillée dans la section suivante. Elle consiste à enlever de l'espace spectral le sous espace porteur des perturbations engendrées par g.

2.2.2 Matériel et méthodes

Nous avons appliqué cette stratégie à l'effet de la température sur la prédiction du taux de sucres des pommes. Nous avons créé pour cela un jeu de données S^0 de 80 individus, dont les spectres ont été acquis à température constante et deux jeux indépendants S^1 et S^2 constitués de 10 pommes mesurées à 8 températures différentes (de 5 à 40^oC, par pas de 5^oC). Un modèle étalonné sur S^0 et testé sur S^2 nous a servi de test de référence en l'absence de correction (méthode 0). Les différentes méthodes de correction ont été établies sur S^0 et S^1 et testées sur S^2. Les résultats de ce test ont été

exprimés à l'aide : du SEP, du R^2, de la moyenne quadratique des biais constatés aux 8 températures ($RMBias$) et du RPD.

Deux types d'étalonnage multi-varié ont été testés : Une PLS sur la plage 826 - 955 nm, correspondant à $p = 41$ variables ; Une MLR sur un sous ensemble de variables sélectionnées par un processus ad-hoc. La déclinaison des méthodes précédentes sur ces deux types d'étalonnage a produit 14 modèles, codés M.x.x et P.x.x (M signifie MLR et P signifie PLS) :

- Dans M.1.1 et P.1.1, la matrice des spectres $\mathbf{X}$ de S^1 a été augmentée de deux colonnes contenant g et g^2
- Pour M.2.2 et P.2.2, 8 modèles ont été calculés pour les 8 températures présentes dans S^1 ;
- Pour M.1.3 et P.1.3, un premier modèle a été calculé sur S^0, donnant b_0, puis un test sur les 8 températures contenues dans S^1 a permis de calculer le biais en fonction de la température, modélisé par une fonction de g, g^2 et b_0 ;
- M.2.1 et P.2.1 ont simplement été étalonnés sur S^1 ;
- Le modèle P.2.2 a été étalonné sur S^1 (M.2.2 n'existe pas) ;
- Pour M.2.3 et P.2.3, la sélection s'est faite en étalonnant sur S^0 et en testant sur S^1. Pour la MLR, l'algorithme utilisé était celui de la stepwise et pour la PLS, celui des algorithmes génétiques.
- La correction EPO du modèle P.2.4 a été calculée grâce à l'ensemble S^1, puis l'étalonnage a été fait sur S^0 (M.2.4 n'existe pas).

2.2.3 Résultats et discussion

Les tables 2.1 et 2.2 résument les performances des différents modèles, en comparaison des modèles non corrigés, dénotés M0 et P0.

	M0	M.1.1	M.1.2	M.1.3	M.2.1	M.2.3	M.2.5
Nombre de variables	5	7	5	5+3	5	6	4+5
RPD	1.22	2.80	2.96	3.58	2.71	4.37	2.84
SEP^oBrix	1.67	0.73	0.69	0.57	0.75	0.49	0.72
$RMBias^o$Brix	1.61	0.52	0.45	0.28	0.54	0.20	0.51
R^2	0.57	0.95	0.94	0.94	0.95	0.94	0.95

TAB. 2.1 – Résultats du test sur S^2 des modèles basés sur la MLR.

	P0	P.1.1	P.1.2	P.1.3	P.2.1	P.2.2	P.2.3	P.2.4	P.2.5
Nb variables latentes	5	8	6 or 7	5+2	7	8	5	3+4	4+8
RPD	0.72	2.80	2.49	3.10	2.80	3.01	4.28	4.20	2.73
SEP^o Brix	2.82	0.73	0.81	0.66	0.72	0.68	0.50	0.51	0.75
$RMBias^o$Brix	2.77	0.54	0.64	0.31	0.54	0.50	0.30	0.36	0.57
R^2	0.31	0.95	0.94	0.91	0.96	0.96	0.94	0.97	0.95

TAB. 2.2 – Résultats du test sur S^2 des modèles basés sur la PLS.

Une discussion approfondie de ces résultats, accompagnée d'une interprétation des modèles d'un point de vue spectrométrique peuvent être trouvées dans [Article I]. Notons tout de même que toutes les méthodes testées apportent une correction spectaculaire. Les meilleurs résultats sont obtenus pour les méthodes qui cherchent à enlever du spectre les perturbations avant de réaliser l'étalonnage, à savoir la sélection de variables et l'EPO. En d'autres termes ceci indique qu'il vaut mieux débarrasser les spectres de leurs perturbations plutôt que d'essayer de corriger le modèle.

La sélection de variables apparaît donc comme un bon moyen d'améliorer la robustesse d'un modèle, comme nous l'avions déjà constaté dans [Roger and Bellon-Maurel, 2000]. Toutefois, cette méthode travaille sur la base canonique de $\mathbb{R}^p$, où l'on a vu que les variables sont très redondantes.

C'est cette constatation qui a présidé au développement de l'EPO, une autre méthode de réduction de la dimension du modèle qui opère dans un espace plus adapté. Cette méthode fait l'objet de la section suivante.

2.3 EPO : Réduction des effets d'une grandeur d'influence par projection orthogonale

Cette section est dédiée à une contribution originale, publiée dans *Chemometrics and Intelligent Laboratory Systems* ([Article II]) et présentée au congrès *ICNIR'2003*.

Cette méthode a pour but d'améliorer la robustesse d'un modèle vis à vis d'une grandeur d'influence G, en projetant les spectres dans un sous espace approprié de $\mathbb{R}^p$. L'EPO appartient à la classe de méthodes qui ne nécessitent pas la connaissance de g au moment de la prédiction.

2.3.1 Théorie

Soit $\mathbf{X}$ une matrice $(n \times p)$ de n spectres acquis sous l'influence d'une grandeur G. Soient $\mathbf{y}$ les n valeurs de la grandeur Y que l'on veut mesurer et $\mathbf{g}$ celles de la grandeur G. Soit r le rang de $\mathbf{X}$.

Principe

Soit $\vec{S}$ l'espace de dimension p de la mesure spectrale. L'EPO vise à décomposer cet espace en deux sous espaces orthogonaux : $\vec{G}$ contenant majoritairement de l'information reliée à G et $\vec{Y}$ contenant majoritairement celle reliée à Y.

Du point de vue de notre échantillon de données, on cherche à décomposer la matrice $\mathbf{X}$ en une partie *parasitée* $\mathbf{X}^-$ et une partie *utile* $\mathbf{X}^+$, avec $rg(\mathbf{X}^-) = k$ et $rg(\mathbf{X}^+) = r-k$. Si $\mathbf{P}$ est le projecteur sur $Col(\mathbf{X}^-)$, cette décomposition est obtenue par :

$$\begin{aligned} \mathbf{X} &= \mathbf{X}^- + \mathbf{X}^+ \\ \mathbf{X} &= \mathbf{X}\mathbf{P} + \mathbf{X}(\mathbf{I} - \mathbf{P}) \end{aligned}$$

Le prétraitement EPO consiste alors simplement à étalonner le modèle sur $\mathbf{X}^+ = \mathbf{X}(\mathbf{I} - \mathbf{P})$. Comme cette projection est orthogonale à $Col(\mathbf{X}^-)$, aucun prétraitement n'est à réaliser lors de l'utilisation du modèle. En effet, comme le vecteur $\mathbf{b}$ du modèle appartient à $Col(\mathbf{X}^+)$, toute perturbation $\delta\mathbf{x}^-$ portée par $Col(\mathbf{X}^-)$ sera orthogonale à $\mathbf{b}$, donc $\delta\mathbf{x}^{-T}\mathbf{b} = 0$.

Soit $\mathbf{U}$ la matrice $(p \times r)$ contenant une base orthonormée de $Col(\mathbf{X})$. On a :

$$\mathbf{X} = \mathbf{t}_1\mathbf{u}_1^T + \mathbf{t}_2\mathbf{u}_2^T + \cdots + \mathbf{t}_r\mathbf{u}_r^T \tag{2.1}$$

Supposons en outre que $\mathbf{U}$ est construite de telle sorte que $\mathbf{u}_1$ maximise la dépendance entre $\mathbf{g}$ et $\mathbf{t}_1$, $\mathbf{u}_2$ maximise la dépendance entre $\mathbf{g}$ et $\mathbf{t}_2$ sur l'orthogonal à $\mathbf{u}_1$, etc. Alors, on adoptera :

$$\begin{aligned} \mathbf{P} &= \mathbf{U}^-\mathbf{U}^{-T} \\ \text{avec} \quad \mathbf{U}^- &= [\mathbf{u}_1 \cdots \mathbf{u}_k] \end{aligned}$$

Identification du sous espace

Il y a certainement plusieurs façons d'identifier $\mathbf{U}^-$. Nous en proposons une, basée sur une expérimentation simple qui ne nécessite de mesurer ni la valeur g de la grandeur d'influence, ni la valeur y du composé à prédire.

Soit un échantillon de m produits présentant des valeurs y quelconques. Soient $(\mathbf{X}^1, \cdots, \mathbf{X}^q)$ les q matrices des m spectres de cet échantillon acquis à différentes valeurs $(g^1, \cdots, g^q)$ de G.

Soit $\mathbf{M}$ la matrice $(q \times p)$ dont la ligne i contient le spectre moyen de $\mathbf{X}^i$:

$$\mathbf{M} = \begin{bmatrix} \overline{\mathbf{x}^1}^T \\ \vdots \\ \overline{\mathbf{x}^q}^T \end{bmatrix}$$

Soit $\mathbf{D}$ la matrice $(q \times p)$ dont la ligne i contient la différence entre le spectre moyen acquis à $g = g^i$ et celui acquis à $g = g^1$:

$$\mathbf{D} = \begin{bmatrix} 0 \\ \mathbf{m}_2^T - \mathbf{m}_1^T \\ \vdots \\ \mathbf{m}_q^T - \mathbf{m}_1^T \end{bmatrix}$$

Soit s le rang de $\mathbf{D}$ (généralement $s = q - 1$). Une ACP sur $\mathbf{D}$, sans centrage ni réduction, nous fournit la décomposition suivante :

$$\mathbf{D} = \mathbf{t}_1\mathbf{u}_1^T + \mathbf{t}_2\mathbf{u}_2^T + \cdots + \mathbf{t}_s\mathbf{u}_s^T$$

La base $[\mathbf{u}_1 \cdots \mathbf{u}_s]$ est un bon candidat pour l'estimation de $\mathbf{U}^-$, car :
- L'ACP fournit une base orthonormée
- Comme les spectres de $\mathbf{M}$ sont le résultat de la moyenne du même échantillon, la variance portée par $\mathbf{D}$ est essentiellement due aux variations de g
- Le vecteur $\mathbf{u}_1$ est celui qui explique le plus de variance de $\mathbf{D}$, donc d'effet des variations de g. Idem pour le vecteur $\mathbf{u}_2$ sur l'orthogonal à $\mathbf{u}_1$, etc.

En définitive, nous proposons donc d'identifier $\mathbf{U}^-$ aux k premières composantes principales des différences moyennes constatées pour q valeurs de g. En procédant de cette manière, il est possible d'identifier un espace de dimension au plus égale à $q - 1$.

Choix de k

La dimension du sous espace que l'on veut ôter par projection orthogonale, c'est à dire k, doit être correctement choisie. En effet, comme dans la plupart des cas les effets de G et de Y ne sont pas indépendants, il faut trouver un compromis entre la sensibilité à Y et l'indépendance vis à vis de G. Si k est trop petit, la correction sera sous optimale et des effets de G persisteront. S'il est trop grand, on risque de trop éroder les spectres et de perdre de l'information utile à la prédiction de y.

Typiquement, ce genre de compromis peut être trouvé en examinant l'évolution d'une erreur de validation en fonction de k. Mais, ici encore, nous proposons une autre méthode qui n'exige de connaître ni g, ni y.

Considérons les m groupes constitués par les q spectres du même individu mesuré aux q valeurs de g. Soit $\Lambda(k)$ le Lambda de Wilks[1] calculé sur $\mathbf{X}^+$. Ce critère mesure la séparation linéaire des groupes dans l'espace $Col(\mathbf{X}^+)$. A mesure que k augmente, la variabilité due à G diminue et les groupes se

[1] Le lecteur se reportera au chapitre 4 pour plus de détails sur ce critère.

resserrent, c'est à dire que $\Lambda(k)$ augmente. S'il existe une décomposition "nette" entre $\mathbf{X}^+$ et $\mathbf{X}^-$, on doit observer une rapide augmentation de $\Lambda(k)$, jusqu'à atteindre la dimension de l'espace parasite, puis un plateau.

2.3.2 Matériel et méthodes

Cette méthode a été testée sur les données décrites dans la section 2.2.2, à la différence que toute la plage spectrale a été utilisée, sans aucun prétraitement. Les données du jeu S^1 ont servi à calculer $\mathbf{U}^-$. La projection $\mathbf{I} - \mathbf{U}^-\mathbf{U}^{-T}$ a été appliquée à S^0. Un modèle PLS a ensuite été étalonné sur ce jeu de données et testé sur S^2.

2.3.3 Résultats et discussion

L'évolution de $\Lambda(k)$ est reportée sur la figure 2.2. Le plateau attendu apparaît clairement pour $k = 4$, que l'on adopte. La figure 2.3 montre l'effet de la correction : le SEP passe de 4.68 Brix à 0.52 Brix.

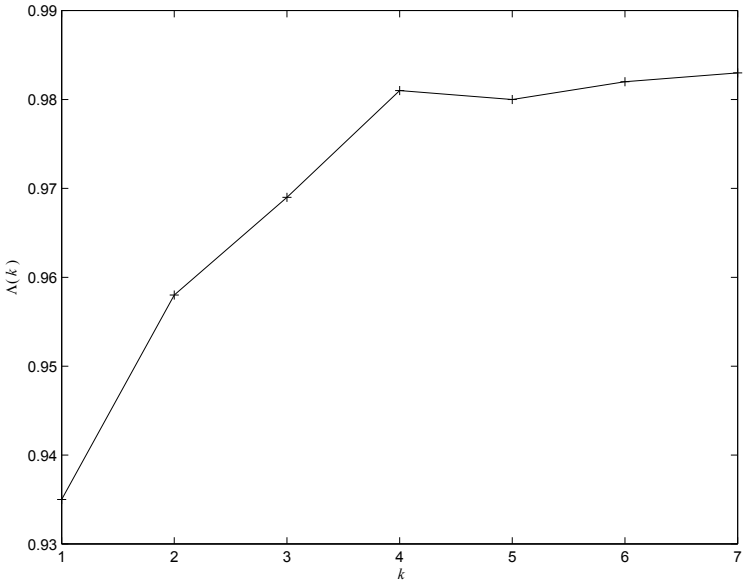

FIG. 2.2 – Evolution de $\Lambda(k)$.

Outre ce résultat spectaculaire, l'EPO offre certains avantages :

- Comme la correction s'effectue dans un espace multidimensionnel, des effets complexes peuvent être corrigés. Ainsi, dans l'application présentée ci-dessus, les perturbations dues à la température sont majoritairement contenues dans un espace de dimension 4. Ceci est d'ailleurs conforme à l'expertise, qui stipule que les effets de la température déforment horizontalement le spectre, à cause des variations des fréquences fondamentales et de la disparition de certaines liaisons faibles. Ces déformations ne sont pas linéaires, mais peuvent être contenues dans un espace de dimension réduite.
- Les vecteurs de la base de $\mathbf{U}^-$ ayant la dimension d'un spectre, ils peuvent être interprétés en comparaison avec des spectres caractéristiques. Ils constituent donc une source d'informations interprétables par le spectroscopiste. Cette boucle de rétro-action de la chimiométrie vers les connaissances fondamentales est très appréciée des utilisateurs. En effet, cela permet "d'ouvrir la boîte noire" et donc : d'une part d'apporter du crédit à la méthode chimiométrique ; d'autre

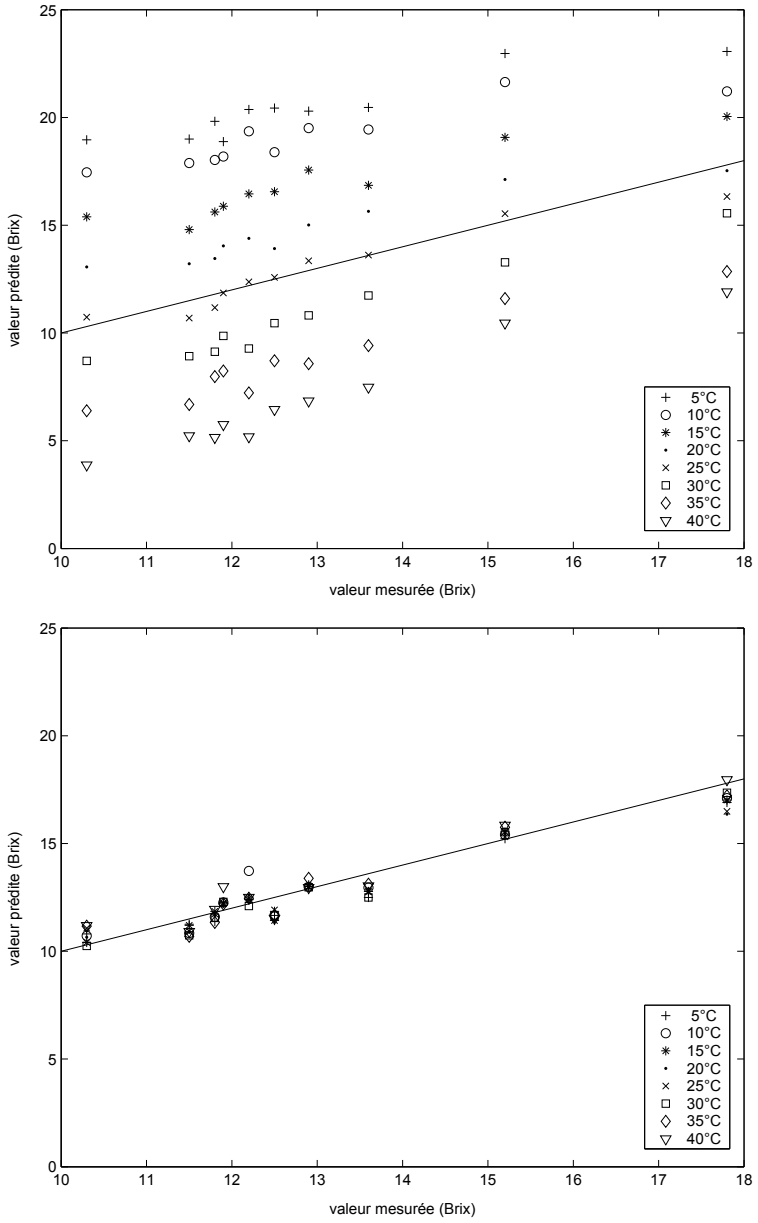

FIG. 2.3 – Prédictions sur S^2 sans prétraitement (en haut) et après correction par EPO (en bas).

part de mieux comprendre l'influence de la grandeur sur les spectres. Dans [Article II], nous avons montré que les influences de la température étaient principalement dues à :
  - Une translation du pic de l'eau à 760 nm ;
  - Une translation du pic des sucres à 915 nm ;
  - Une dilatation du pic d'absorbtion centré sur 838 nm (liaisons O-H).
- Puisque la correction est réalisée par projection orthogonale, le modèle reste valable lorsque il est utilisé en dehors de toute influence de G.

2.3.4 Conclusion

La méthode EPO apporte une réponse aux cas où la grandeur d'influence ne peut être ni contrôlée, ni mesurée. Grâce à une base de données simple à constituer, elle permet de traiter une base d'étalonnage existante pour la rendre indépendante des variations de G. Cette particularité est très intéressante, car elle permet de préserver un capital de données souvent précieux. L'application de l'EPO au cas de la température du produit a montré des résultats spectaculaires, bien que les influences de cette grandeur soient connues pour être complexes.

Cette méthode présente un potentiel très important pour nos applications. De plus, elle est parfaitement applicable au cas des grandeurs d'influence discrètes, comme l'origine du produit. Elle a d'ailleurs été appliquée avec succès au problème du transfert d'étalonnage entre spectromètres par une équipe du Département de Statistiques du *University College of London* ([Andrew and Fearn, 2004]).

Des travaux doivent être poursuivis pour étudier la prise en compte de plusieurs grandeurs d'influence.

2.4 Perspectives de recherches

La prise en compte des grandeurs d'influence est une condition *sine qua non* de la faisabilité des prototypes que nous développons. C'est pourquoi nous avons adopté une démarche très pragmatique, comme exposé dans la section 2.2. Certaines grandeurs d'influence étant ni contrôlables, ni mesurables, j'ai été amené à effectuer des recherches spécifiques à ce cas et à développer l'EPO, comme exposé dans la section 2.3.

Comme je l'ai déjà abordé dans le chapitre 1, un spectre est le résultat d'une combinaison d'effets dûs à plusieurs grandeurs, dont seulement certaines nous intéressent et qui s'entremêlent dans les variables mesurées. C'est pourquoi l'analyse des spectres doit se faire en terme de structures vectorielles, c'est à dire en se restreignant à un sous espace de l'espace de mesure.

La PLS identifie le sous espace dans lequel les variations spectrales sont les plus reliées aux variations de y. L'EPO se charge d'identifier celui qui est porteur des effets spectraux dûs aux variations d'une grandeur d'influence. On peut sans peine imaginer d'autres techniques qui mettront à jour d'autres espaces.

D'un autre côté, les spectres purs et les spectres d'interaction, présentés en 1.3, sont eux aussi représentatifs d'un sous espace. De même pour les phénomènes physiques, tels que la diffusion.

Il me semble donc très intéressant d'étudier, dans un avenir proche, comment toutes ces structures vectorielles, qu'elles soient issues de décompositions algébriques ou de connaissances théoriques, peuvent faire progresser notre connaissance de la mesure spectrométrique et améliorer son exploitation. Ainsi, par exemple, je propose de rechercher un moyen de combiner l'approche de l'étalonnage basé sur les données (comme la PLS) avec celle de l'étalonnage basé sur les connaissances fondamentales (comme les modèles de mélanges).

Chapitre 3

La maintenance de la robustesse des modèles utilisés en ligne

Ce chapitre est dédié à la troisième voie de recherche identifiée en 1.4, c'est à dire à la maintenance de la robustesse des modèles utilisés en ligne. Après une introduction posant les notations et les hypothèses, une première partie expose brièvement le principe qui a servi de base à une comparaison de plusieurs méthodes classiques, en référence à l'article [Article III]. La seconde partie expose une contribution originale, publiée dans [Article IV]. Enfin, la troisième partie propose des perspectives de recherche.

3.1 Notations - Hypothèses

On note Y la grandeur que l'on cherche à mesurer, y sa valeur et t le temps. La base d'étalonnage, contenant n_0 individus, est notée $(\mathbf{X}_0, \mathbf{y}_0)$. Les spectres acquis au cours du procédé sont notés $\mathbf{X}_t$. Les valeurs de Y correspondantes (et inconnues) sont notées $\mathbf{y}_t$.

Nous supposerons que :

- Pour les mesures spectrométriques, les problèmes de robustesse aux grandeurs d'influence présentent une certaine pérennité, c'est à dire que la constante de temps d'évolution de G est très nettement supérieure à la période de mesure du capteur spectrométrique.
- On dispose de n_τ mesures de références de Y acquises au cours du procédé, peu nombreuses, permettant de recaler le capteur. Ces valeurs, accompagnées des spectres acquis simultanément par le capteur, forment une *base de recalage*, notée $(\mathbf{X}_\tau, \mathbf{y}_\tau)$.

Nous caractériserons la robustesse par le rapport $R_c = SEP_0/SEP$, où SEP_0 est l'erreur standard de test constatée au laboratoire (conditions standard) et SEP est l'erreur standard constatée en présence du problème de robustesse. Plus ce critère est grand, meilleure est la robustesse du modèle. Ce critère pourra être calculé en fonction du temps en calculant le SEP sur une fenêtre temporelle glissante.

3.2 Amélioration intrinsèque de la robustesse de l'étalonnage

Cette courte section est reliée à un article publié dans *Trends in Analytical Chemistry* en 2004 ([Article III]). Avant de s'intéresser à la maintenance de la robustesse d'un modèle, nous avons exploré les différentes méthodes qui, appliquées au moment de l'étalonnage, sont susceptibles d'en améliorer la robustesse intrinsèque. Nous avons donc établi une revue des méthodes les plus classiques et comparé leur apport à la robustesse du modèle, de manière théorique, grâce à l'équation 1.10 (page 21),

reportée ci-après :

$$|\delta\widehat{y}| = \|\delta\mathbf{x}\| \cdot \|\mathbf{b}\| \cdot |\cos(\delta\mathbf{x}, \mathbf{b})|$$

D'après cette équation géométrique simple, trois voies peuvent être utilisées pour réduire l'erreur : réduire $\|\delta\mathbf{x}\|$; réduire $\|\mathbf{b}\|$; réduire $|\cos(\delta\mathbf{x}, \mathbf{b})|$. Ces trois voies correspondent à trois grands groupes de méthodes : les prétraitements géométriques des spectres (normalisation, filtrage, dérivation), la sélection de variables et les projections orthogonales, telles que l'EPO (Cf. 2.3).

3.3 Maintenance en ligne de la robustesse de l'étalonnage

Cette section est reliée à un article soumis pour publication à *Chemometrics and Intelligent Laboratory Systems* ([Article IV]). La méthode qui y est développée, baptisée DOP pour Dynamic Orthogonal Projection, a pour but de maintenir la robustesse d'un étalonnage IR, utilisé pour la mesure en ligne sur un procédé, moyennant les hypothèses formulées en 3.1.

Plaçons nous au temps t, au cours du suivi d'un procédé. Un certain nombre de mesures de références $\mathbf{y}_\tau$, prises dans le passé, nous apportent une information sur l'erreur commise par le modèle d'étalonnage. Le problème consiste alors à utiliser au mieux cette information pour améliorer la robustesse du modèle pour les mesures à venir.

Les méthodes classiquement employées pour répondre à ce problème sont de deux ordres :

- La correction *post modèle* consiste simplement à identifier une loi de correction entre $\widehat{\mathbf{y}}_\tau$, les valeurs prédites par le modèle et $\mathbf{y}_\tau$, les valeurs que le modèle aurait dû prédire, puis à appliquer cette correction sur les futures prédictions. Différentes fonctions de correction peuvent être élaborées, de la simple correction d'un biais jusqu'à des corrections polynômiales. La simplicité de cette méthode est son principal avantage. Par contre, si la perturbation disparaît, la correction agit dans le mauvais sens et introduit une erreur. De plus, cette méthode apporte très peu d'information sur la perturbation à l'origine du problème.
- Le ré-étalonnage du modèle consiste à modifier le modèle pour qu'il tienne compte des nouvelles données apportées par $(\mathbf{X}_\tau, \mathbf{y}_\tau)$. Selon le nombre de mesures de recalage disponibles, on peut envisager de ré-étalonner un modèle sur $(\mathbf{X}_\tau, \mathbf{y}_\tau)$, ou sur $(\mathbf{X}_0, \mathbf{y}_0) \bigcup (\mathbf{X}_\tau, \mathbf{y}_\tau)$. Ici encore, la méthode est simple, mais elle est peu informative et peut nécessiter un grand nombre de points de recalage.

La méthode DOP propose une autre voie. Elle utilise l'information fournie par les points de recalage pour identifier un espace vectoriel contenant les perturbations spectrales survenues, projette les spectres de la base d'étalonnage orthogonalement à cet espace et enfin ré-étalonne le modèle sur cette base d'étalonnage modifiée.

3.3.1 Théorie

Le principe de base de DOP repose sur la notion de *standards virtuels*. Classiquement, un standard est un échantillon mesuré à la fois dans les conditions de l'étalonnage et dans les conditions perturbées. Les standards virtuels de DOP sont constitués du couple $(\widehat{\mathbf{X}}_\tau, \mathbf{X}_\tau)$, où $\widehat{\mathbf{X}}_\tau$ sont les spectres que l'on aurait dû mesurer en l'absence de perturbation et $\mathbf{X}_\tau$ les spectres effectivement mesurés, simultanément aux mesures de recalage $\mathbf{y}_\tau$. Une fois ces standards disponibles, la matrice des différences $\mathbf{D} = \widehat{\mathbf{X}}_\tau - \mathbf{X}_\tau$ est calculée. Cette matrice constitue un nuage de n_τ points de $\mathbb{R}^p$, représentatifs des influences spectrales à corriger. Il suffit alors, pour identifier le sous espace vectoriel recherché, d'opérer de manière similaire à l'EPO (Cf. 2.3), c'est à dire par les composantes d'une ACP calculée sur $\mathbf{D}$.

Pour calculer les standards virtuels, c'est à dire l'estimation de $\widehat{\mathbf{X}}_\tau$, nous proposons de réaliser une combinaison linéaire des spectres $\mathbf{X}_0$. Les coefficients de cette combinaison linéaire sont donnés

par une fonction de noyau centrée sur $\mathbf{y}_\tau$ et appliquée sur $\mathbf{y}_0$. En d'autres termes, on calcule une combinaison des $\mathbf{y}_0$ qui approche $\mathbf{y}_\tau$ et on l'applique à $\mathbf{X}_0$.

Le processus complet est donc le suivant :

1. L'application de n_τ fonctions de noyau centrées sur les éléments de $\mathbf{y}_\tau$ fournit une matrice $\mathbf{A}_{(n_\tau \times n_0)}$:

$$a_{ij} \quad = \quad F_{y_{\tau i}}(y_{0j}) \quad \text{où } F_{y_{\tau i}} \text{ est une fonction de noyau centrée sur } y_{\tau i}$$

2. Les spectres $\widehat{\mathbf{X}}_\tau$ sont ensuite estimés par :

$$\widehat{\mathbf{X}}_\tau \quad = \quad \mathbf{A}\mathbf{X}_0$$

3. Les spectres de différence $\mathbf{D}$ entre $\mathbf{X}_\tau$ et $\widehat{\mathbf{X}}_\tau$ sont calculés :

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

4. Une base orthonormée $\mathbf{P}$ de l'espace généré par $\mathbf{D}$ est estimée par ACP :

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

où $\mathbf{P}$ contient les facteurs de $\mathbf{D}$ et $\mathbf{T}$ les coordonnées factorielles.

5. La base d'étalonnage est corrigée par projection orthogonale :

$$\mathbf{X}_0^* \quad = \quad \mathbf{X}_0(\mathbf{I} - \mathbf{P}\mathbf{P}^T)$$

6. Un nouvel étalonnage est calculé sur $(\mathbf{X}_0^*, \mathbf{y}_0)$. Puisque la base est corrigée par projection orthogonale, la correction est embarquée dans le modèle. Par conséquent, aucun prétraitement n'est à réaliser ensuite sur les spectres lors de l'utilisation du modèle.

3.3.2 Matériel et méthode

La méthode DOP a été testée sur le suivi des fermentations alcooliques, à l'échelle du pilote (100 l), pour maintenir la robustesse d'un étalonnage de mesure de l'éthanol par transmission NIR en présence de variations de température du produit. En parallèle de la mesure NIR, l'éthanol était mesuré par intégration du débit de CO_2. Une première fermentation a été suivie dans des conditions isothermes (25^{o}C). La base d'étalonnage $(\mathbf{X}_0, \mathbf{y}_0)$, constituée de $n_0 = 300$ individus a été extraite de cette première expérience. Ensuite, une deuxième fermentation a été réalisée dans les mêmes conditions, avec le même moût, mais en faisant varier la température, comme indiqué par la figure 3.1. Sur cette même figure sont reportés les 5 instants de recalage, qui ont été choisis comme suit : 2 points ont été placés au début et à la fin de la première phase, qui se déroulait à la même température que l'étalonnage ; 3 autres points ont été équi-répartis pendant la phase de montée en température.

Les fonctions de noyau utilisées dans DOP ont été choisies gaussiennes, avec une largeur égale à $\varepsilon\sigma(\mathbf{y}_0)$. Le nombre k de composantes principales de l'ACP, c'est à dire la dimension du sous espace corrigé, a été choisi sur la base de la part de variance de $\mathbf{D}$ capturée, v. Plusieurs valeurs de ε et de v ont été testées.

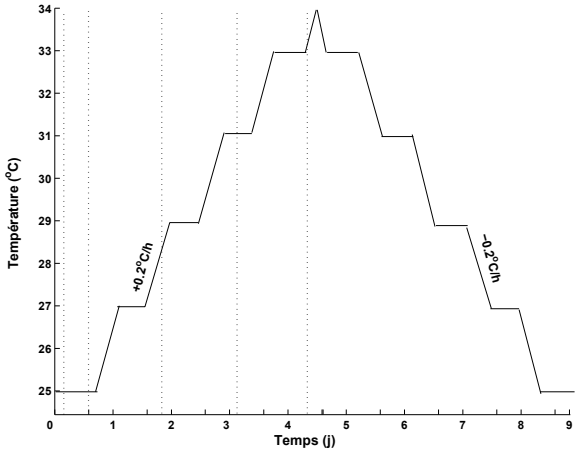

FIG. 3.1 – Profil d'évolution de la température durant la fermentation de test. Les lignes verticales indiquent les instants τ de recalage.

3.3.3 Résultats et discussion

La méthode DOP possède deux paramètres de réglage : la largeur du noyau, ici donnée par ε ; et la dimension du sous espace ôté par projection orthogonale, i.e. le nombre de composantes principales de l'ACP, ici choisi par la part de variance v retenue.

Supposons que $\mathbf{y}_0$ suit distribution uniforme, sur l'intervalle $[a, b]$. Soit Δ l'intervalle moyen entre deux valeurs de $\mathbf{y}_0$: $\Delta = (b-a)/(n_0-1)$. Il parait raisonnable de choisir un noyau dont la largeur soit du même ordre que cet écart, c'est à dire : $4\varepsilon\sigma(\mathbf{y}_0) \simeq \Delta$. Or, $\sigma(\mathbf{y}_0) = (b-a)/2\sqrt{3}$. Finalement, on obtient pour ε l'ordre de grandeur suivant : $\varepsilon \simeq 1/n_0$.

La figure 3.2 montre l'évolution du critère global de robustesse R_c calculé sur la fermentation anisotherme, en fonction de ε, pour différentes valeurs de v. Il apparaît clairement une zone de réglage optimal, très large et quasiment indépendante de v. La droite verticale indique la valeur théorique de $\varepsilon = 1/n_0$, qui correspond bien à la zone du maximum de robustesse. Nous avons donc choisi d'adopter $\varepsilon = 3.3\ 10^{-3}$ et $v = 99\%$.

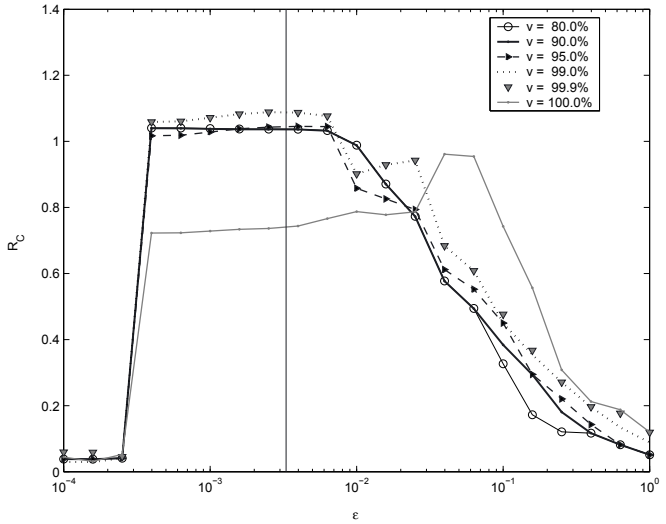

FIG. 3.2 – Évolution du critère de robustesse $R_C = SEP_0/SEP$, en fonction de la largeur relative du noyau (ε) et de la part de variance de $\mathbf{D}$ capturée par $\mathbf{P}$, v.

La figure 3.3 permet de comparer le modèle brut et le modèle corrigé par DOP. On y voit que le modèle brut présente deux problèmes de robustesse :

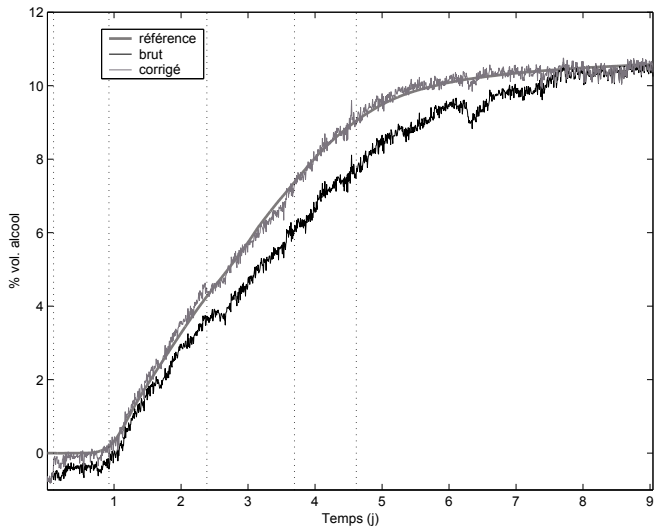

FIG. 3.3 – Tests du modèle brut et du modèle corrigé par DOP sur la fermentation anisotherme.

- En début de process, un biais négatif, d'environ -0.5% vol. est présent (Cf. figure 3.4). Les causes précises de ce problème n'ont pas pu être expliquées : les conditions opératoires étaient strictement identiques à celles de la fermentation ayant servi à l'étalonnage. Seuls des paramètres physiques du moût, comme la turbidité, peuvent être mis en cause. Notons également que ce biais n'est plus présent en fin de fermentation, lorsque la température redevient identique à celle de l'étalonnage.
- Pendant la fermentation, alors que la température varie, on constate une dérive du modèle qui atteint environ -1% vol. aux environs du jour 5, lorsque la température est la plus élevée.

Ces deux problèmes de robustesse sont visibles sur la figure 3.5, montrant l'évolution du critère de robustesse en fonction du temps. En début de procédé, le biais de prédiction cause un R_c voisin de 0.3, évoluant lentement jusqu'à 0.6 à la moitié du deuxième jour. Il est fort probable que le problème responsable du biais s'estompait lentement. À partir du deuxième jour, l'effet de la température se voit clairement : le R_c décroît au fur et à mesure que la température croît (jusqu'au jour 5), puis remonte pour atteindre finalement une valeur proche de 1, en fin de procédé.

L'effet de DOP est très nettement visible sur les figures 3.3 et 3.4 :

- Le biais du début de procédé est corrigé dès le premier point de recalage et confirmé au deuxième.
- On constate une légère dérive pendant le deuxième jour, complètement corrigé par le troisième point de recalage. Le même phénomène peut être observé pour le point de recalage suivant. Ensuite, le modèle reste robuste, jusqu'à la fin du procédé.

Le gain apporté par DOP est bien illustré par le critère de robustesse reporté sur la figure 3.5. On y constate que, dès le premier point de recalage (au retard dû au lissage près), le critère de robustesse passe au dessus de 1 et y reste quasiment en permanence jusqu'à la fin du procédé.

Le biais du début de procédé a été corrigé en utilisant un seul point de recalage, donc un espace de dimension 1. Ceci indique que la perturbation était de la forme $\alpha \mathbf{k}_0$, où $\mathbf{k}_0$ est un spectre constant. Ceci renforce l'hypothèse d'un phénomène de diffusion, éventuellement dû à une différence de turbidité entre l'étalonnage et le test.

À partir du troisième point de recalage, la correction effectuée par DOP tient compte à la fois du problème de décalage initial et des perturbations dues à la température. Ceci montre la capacité de

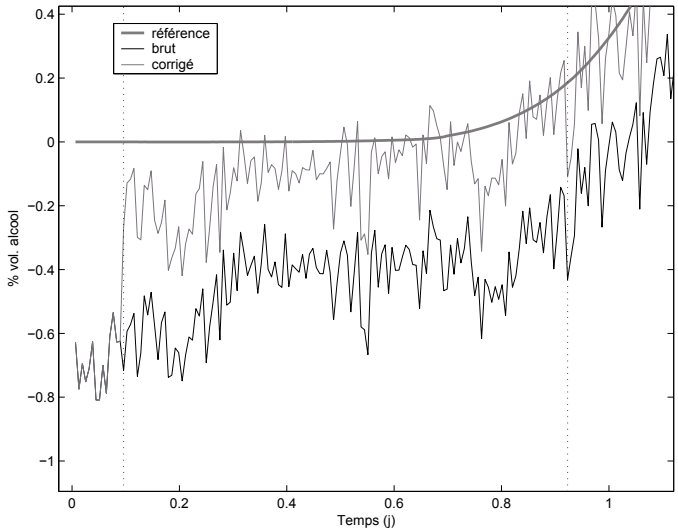

FIG. 3.4 – Tests du modèle brut et du modèle corrigé par DOP sur la fermentation anisotherme. Zoom sur la première journée

la méthode à gérer plusieurs problèmes de robustesse simultanément. Au dernier point de recalage, 2 dimensions ont été utilisées par DOP pour corriger l'espace. La figure 3.6 montre l'effet de la correction DOP au point 5. La variation de température a pour effet de distordre les spectres, notamment en déplaçant horizontalement le minimum. L'effet de la correction amenée par DOP est spectaculaire, comme en témoignent les spectres corrigés [1].

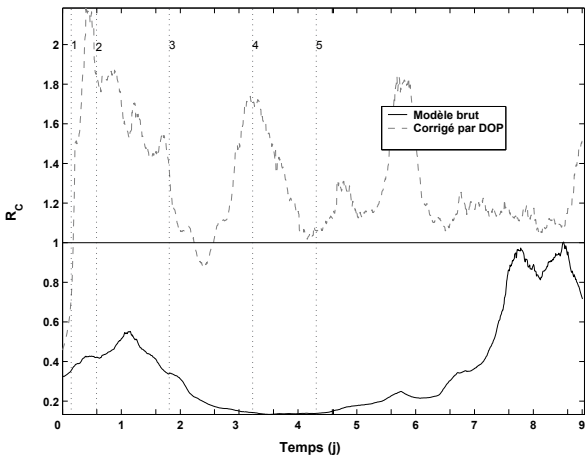

FIG. 3.5 – Évolution du critère de robustesse du modèle brut et du modèle corrigé par DOP pendant la fermentation anisotherme. Calculs effectués sur une fenêtre glissante de 15 h.

3.3.4 Conclusion

La méthode DOP apporte une solution originale au problème de maintien de la robustesse des étalonnages multi-variés. Son originalité repose essentiellement sur la notion de standard virtuel, dont le calcul passe par les seules valeurs de référence $\mathbf{y}_\tau$. Cette étape peut paraître osée, dans la mesure

[1] À remarquer que seuls les spectres d'étalonange sont effectivement corrigés de la sorte ; lors du test, la correction est *automatiquement* réalisée par les coefficients du modèle.

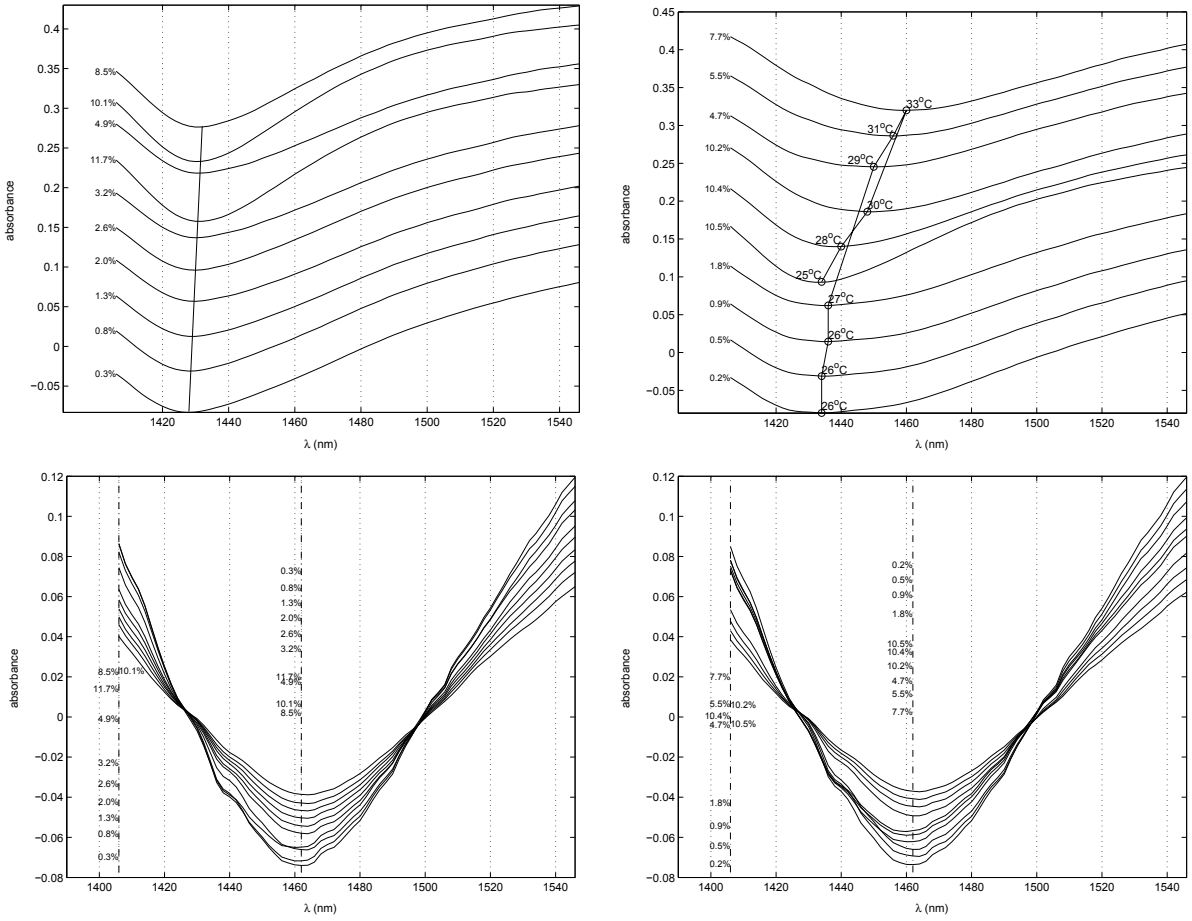

FIG. 3.6 – Effet de la correction DOP au point 5 : À gauche ; spectres d'étalonnage (isotherme) - À droite ; spectres de test (anisotherme) - En haut ; Spectres bruts - En bas ; spectres corrigés.

où une même valeur de référence peut correspondre à plusieurs spectres. Cependant, l'estimation du standard virtuel, en opérant par moyenne (fonction de noyau), s'affranchit de cet écueil, à condition que la base d'étalonnage soit suffisamment riche. Ainsi, si cette base contient un grand nombre d'exemples de spectres, la différence spectrale qui n'est pas due à une différence de y est elle vue par DOP comme du bruit à éliminer. En poussant ce raisonnement à l'extrême, si on applique DOP à chacun des individus de la base d'étalonnage, on obtient finalement un prétraitement de la base qui consiste à éliminer par projection orthogonale tout ce qui, dans les spectres, n'est pas relié à y, c'est à dire à réaliser une orthogonalisation du type O-PLS, comme décrit dans [Trygg, 2001].

La méthode DOP possède de nombreux avantages :

- La base d'étalonnage $(\mathbf{X}_0, \mathbf{y}_0)$ est utilisée, ce qui valorise d'une part, les données expérimentales qui ont servi à son élaboration et d'autre part, toutes les optimisations qui y ont été appliquées : détection d'outliers, sélection optimale des échantillons, etc. Cette qualité distingue DOP des méthodes de ré-étalonnage local : en effet, le modèle qui est reconstruit à chaque correction de DOP intègre bien toute la variabilité représentative du processus, et pas seulement celle observée dans $\mathbf{y}_\tau$.
- La correction est embarquée dans le modèle, c'est à dire qu'aucun prétraitement des spectres n'est à réaliser on-line. Ceci est très important dans le cas de l'utilisation de spectromètres incluant un logiciel d'étalonnage *fermé*. L'application de DOP nécessite simplement de pouvoir accéder aux spectres enregistrés et à la base d'étalonnage. La correction peut alors être implémentée comme un module autonome et indépendant, auquel on fournit $\mathbf{X}_0, \mathbf{y}_0, \mathbf{X}_\tau, \mathbf{y}_\tau$ et qui fournit les spectres corrigés $\mathbf{X}_0^*$.
- La correction DOP rend le modèle indépendant des perturbations corrigées. De la sorte, lorsque une ou plusieurs perturbations changent en intensité, voire disparaissent, il n'y a aucune conséquence néfaste sur le modèle. Cette caractéristique confère à DOP une supériorité très nette par rapport aux méthodes de correction classiques, comme l'ajustement biais pente des prédictions.
- Seulement deux paramètres sont à régler. De plus, il semble que la méthode soit peu sensible à la valeur de ces paramètres. La largeur du noyau, donnée par ε, peut être ajustée théoriquement en fonction de la distribution de la base d'étalonnage. La dimension du sous espace peut être déterminée en fonction de la part de variance capturée.
- Lors de la correction, DOP définit un projecteur $(\mathbf{P}\mathbf{P}^T)$, chargé d'ôter la partie polluée des spectres. L'application de ce projecteur aux spectres $\mathbf{X}_t$ permet de visualiser l'évolution de cette pollution, dans un espace compréhensible par le spectroscopiste. Cette propriété n'a pas été clairement illustrée dans la présente application, où l'effet température n'est pas très parlant. Elle est par contre très nettement montrée dans la thèse de M. Zeaiter ([Zeaiter, 2004]), sur une application où le facteur d'influence est de nature chimique.
- Plusieurs effets perturbateurs peuvent être gérés simultanément.

3.4 Perspectives de recherches

Les résultats apportés par DOP ouvrent de nombreuses perspectives, outre celles déjà citées pour l'EPO (Cf. 2.4) :

- En introduisant le concept de standards virtuels, elle ouvre la voie à de nouvelles méthodes de transfert d'étalonnage. Ce problème de transfert est généralement rencontré lors du changement de spectromètre ou d'une opération de maintenance, comme un changement de lampe. La méthode généralement employée, pour ne pas perdre les bases d'étalonnage existantes, consiste à mesurer des échantillons standard sur l'instrument de laboratoire (maître) et sur le nouvel instrument (esclave). Cette opération peut cependant poser des problèmes. Par exemple, quels standards utiliser quand les produits à analyser sont des objets biologiques périssables ? Il me semble donc particulièrement intéressant d'étudier le potentiel de DOP pour résoudre le pro-

blème du transfert d'étalonnage, sans utilisation d'échantillons standard.

– En même temps qu'elle corrige la base d'étalonnage et donc le modèle, DOP fournit des infos de diagnostic dans l'espace de la mesure spectrale, sous la forme d'un projecteur de $\mathbb{R}^p$. Nous avons vu, sur un exemple de fermentation, que l'on pouvait gérer l'ensemble des perturbations survenant sur le procédé, en cumulant leurs effet. Cependant, sur un processus plus long, se pose la question de la gestion d'une *mémoire* des perturbations. On peut en effet craindre que, si toutes les perturbations sont cumulées, le signal utile s'érode au point de ne plus fournir de modèle. De manière symétrique, les corrections calculées par DOP sur un procédé à un moment donné doivent pouvoir être utiles à la gestion d'un autre procédé, à un autre moment. Tout cela appelle la gestion d'une mémoire des perturbations spectrales rencontrées sur les procédés. On peut imaginer une bibliothèque de projecteurs, auxquels l'expert du procédé aura attaché une signification et qui, utilisés en continu, fournissent des scores renseignant les opérateurs sur l'état du procédé. Cette perspective de recherche soulève des problèmes relevant de la productique et de l'automatique.

Chapitre 4

La discrimination à partir des spectres

Ce chapitre est dédié à la quatrième voie de recherche identifiée en 1.4, c'est à dire à la discrimination à partir des spectres. Après une introduction posant les notations et rappelant les notions générales de la discrimination, une première partie rapporte une comparaison de plusieurs méthodes classiques, en s'appuyant sur l'article [Article V]. La seconde partie expose une contribution originale, publiée dans [Article VI], qui propose une méthode de discrimination directe, particulièrement adaptée aux spectres. Enfin, la troisième partie propose des perspectives de recherche.

4.1 Introduction - Notations

Le problème de la discrimination consiste à attribuer un label à un individu $\mathbf{x}$. Ce label peut prendre sa valeur dans un ensemble fini de symboles, connus *a priori*. Ce qui différencie la discrimination de la régression, c'est que l'ensemble des labels possibles ne possède ni relation d'ordre, ni structure algébrique. Les critères de covariance, de corrélation, de moindres carrés entre les variables prédictives et la réponse ne peuvent donc plus être calculés directement. La discrimination linéaire suppose que les individus d'un même label se situent dans une même région de l'espace, c'est à dire un *groupe* ou une *classe*. Enfin, je considérerai que le terme de discrimination est réservé aux méthodes qui opèrent par apprentissage supervisé, contrairement aux techniques de *classification*, que je n'aborderai pas.

En supplément aux notations générales (page 9), j'adopte pour ce chapitre les hypothèses et notations suivantes :

La matrice $\mathbf{X}$ est supposée centrée. Soient $\{1, \cdots, c\}$ c classes d'effectifs $\{n_1, \cdots, n_c\}$. On supposera $n_i \neq 0$, pour tout i. Soit $\mathbf{Y}_{(n\times c)}$ la matrice de codage disjonctif complet de l'appartenance des individus de $\mathbf{X}$ aux classes, i.e. $y_{ij} = 1$ si l'individu i appartient à la classe j, et 0 sinon.

Soit $\mathbf{T}$ la matrice de variance-covariance totale :

$$\mathbf{T} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$$

Soit $\mathbf{B}$ la matrice de variance-covariance inter-classes :

$$\mathbf{B} = \frac{1}{n-1}\mathbf{X}^T\mathbf{Y}\left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T\mathbf{X}$$

Soit $\mathbf{W}$ la matrice de variance-covariance intra-classes :

$$\mathbf{W} = \mathbf{T} - \mathbf{B}$$

Les performance d'un modèle de discrimination, produisant une estimation $\widehat{\mathbf{Y}}$ de $\mathbf{Y}$ sont jugées au travers de la matrice de confusion : $\mathbf{C} = \widehat{\mathbf{Y}}^T\mathbf{Y}$; c_{ij} = nombre d'individus de la classe j attribués à

la classe i. A partir de cette matrice, l'erreur de prédiction globale est calculée par $E(\%) = 100(1 - \text{tr}(\mathbf{C})/n)$. L'erreur de prédiction relative à la classe i est donnée par $E_i(\%) = 100(1 - c_{ii}/n_i)$.

Nous supposerons en outre que : La matrice $\mathbf{X}$ est de rang maximal, i.e. $r = \text{rank}(\mathbf{X}) = \min(n - 1, p)$; La dimension de l'espace engendré par $\mathbf{X}$ est suffisante pour construire un espace discriminant de dimension maximale, i.e. $r \geq c - 1$; La matrice $\mathbf{B}$ est de rang maximal, i.e. $\text{rank}(\mathbf{B}) = c - 1$ (il n'existe pas deux classes dont les centres de gravité coïncident).

Le problème de l'analyse discriminante linéaire (DA) consiste à déterminer une matrice $\mathbf{U}(p \times k)$ telle que $\mathbf{Z} = \mathbf{XU}$ présente une séparation des classes optimale. La détermination de $\mathbf{U}$ est basée sur la maximisation de la variance inter-classes ($\mathbf{U}^T\mathbf{BU}$) par rapport à la variance totale ($\mathbf{U}^T\mathbf{TU}$) ou à la variance intra-classes ($\mathbf{U}^T\mathbf{WU}$). L'utilisation de l'un ou l'autre de ces critères est équivalente. L'école anglo-saxonne utilise plutôt le rapport à la variance intra-classes, donnant un nombre variant entre 0 et $+\infty$. L'école française utilise le rapport à la variance totale, donnant un nombre variant entre 0 et 1. Nous adopterons ce dernier point de vue.

Le rapport des variances inter-classes et totale est mesuré par le Lambda de Wilks, classiquement défini par :

$$\Lambda(\mathbf{XU}) = \frac{\det(\mathbf{U}^T\mathbf{BU})}{\det(\mathbf{U}^T\mathbf{TU})}$$

Si $\mathbf{XU}$ n'est pas de plein rang (par exemple si $k > n$), la définition ci-dessus n'est plus valable et peut être remplacée par :

$$\Lambda(\mathbf{XU}) = \frac{\text{tr}(\mathbf{U}^T\mathbf{BU})}{\text{tr}(\mathbf{U}^T\mathbf{TU})}$$

Lorsqu'aucune confusion n'est possible, nous noterons $\Lambda(\mathbf{U}) = \Lambda(\mathbf{XU})$ et $\Lambda(\mathbf{u}) = \Lambda(\mathbf{Xu})$, dans le cas où $\mathbf{U}$ est réduite à un vecteur discriminant $\mathbf{u}$.

Soit donc $\mathbf{u}$ un vecteur unitaire. On cherche à maximiser $\Lambda(\mathbf{u})$, forme linéaire de $\mathbb{R}^p$. En exprimant la condition de nullité de son gradient, il vient :

$$(\mathbf{B} - \Lambda(\mathbf{u})\mathbf{T})\,\mathbf{u} \;=\; \mathbf{0}$$

Dans le cas des problèmes bien dimensionnés (où $\mathbf{T}$ est inversible), cette relation donne naissance à l'analyse factorielle discriminante (AFD). La matrice $\mathbf{U}$ est donnée par les $c - 1$ vecteurs propres associés aux valeurs propres non nulles de $\mathbf{T}^{-1}\mathbf{B}$. Chacune de ces valeurs propres est d'ailleurs le Lambda de Wilks du vecteur propre associé.

Depuis la première contribution à ce sujet, due à Fisher en 1936 ([Fisher, 1936]), plusieurs variantes ont été proposées, dont une bonne revue peut être trouvée dans [Indahl et al., 1999]. Dans le domaine de la reconnaissance de formes, l'analyse discriminante est aussi très largement étudiée. Un nombre important de méthodes y ont été produites, dont un bon échantillon pourra être trouvé dans [Foley and Sammon, 1975], [Liu et al., 1992] et [Xiao-Jun et al., 2004], mais aucune ne s'intéresse aux problèmes mal conditionnés, tels que celui qui nous intéresse. De manière analogue au cas de la régression, les problèmes de dimensionnenent et de conditionnement sont généralement abordés par la réduction de dimension. Cette opération s'opère : soit par une analyse en composantes principales (PCA-DA) ; soit par une PLS entre les spectres et les degrés d'appartenance aux classes (PLS-DA) ; soit par une sélection de variables suivie d'une analyse discriminante classique (par exemple une Stepwise DA). A noter que deux types de PLS-DA existent :

- Le premier consiste à calculer une PLS-2 (dédiée au cas de plusieurs réponses) entre $\mathbf{X}$ et $\mathbf{Y}$, puis à réaliser une analyse discriminante (p.ex. une AFD) sur l'espace des variables latentes. C'est cette méthode que nous appellerons PLS-DA.
- Le deuxième étalonne directement un modèle de prédiction entre $\mathbf{X}$ et $\mathbf{Y}$. Lors de l'utilisation du modèle, l'affectation à une classe se décide sur le maximum des degrés prédits. Nous appellerons cette méthode la DPLS.

4.2 Comparaison de méthodes de discrimination appliquées à un problème mal dimensionné et mal conditionné

Cette section est relative à un article paru dans *Biotechnology and Bioengineering*, en 2002 ([Article V]). Le but en était de passer en revue et de comparer différentes méthodes de discrimination traitant des courbes. L'application concernait la reconnaissance du cépage d'un moût à partir de sa courbe de cinétique fermentaire. Le problème de dimensionnement et de conditionnement est similaire à celui rencontré avec les spectres.

4.2.1 Matériel et méthodes

La base de données utilisée dans cette étude était composée de 42 individus, issus de 5 cépages : Grenache (gn), Carignan (ca), Cinsault (ci), Mourvèdre (mo) et Syrah (sy). Les échantillons provenaient de la même vigne, vendangée à différentes années, fournissant 8 ou 9 individus par classe, comme indiqué par la table 4.1.

Cépage	Grenache	Carignan	Cinsault	Mourvèdre	Syrah
Code	gn	ca	ci	mo	sy
Effectif	9	8	8	8	9

TAB. 4.1 – Constitution de la base de données.

Chacun des $n = 42$ échantillons a été vinifié dans les mêmes conditions ; la courbe de vitesse de production de CO_2 a été enregistrée et digitalisée sur $p_1 = 80$ points. Un premier jeu de données (42×80), constituait donc un problème mal dimensionné. A partir de ces données, $p_2 = 11$ variables ont été synthétisées, pour constituer un problème bien dimensionné. Ces variables étaient issues à la fois de l'expertise de l'œnologue et d'une ACP réalisée sur les courbes, comme indiqué par la table 4.2.

Numéro de variable	Code	Signification
1	V_{max}	Vitesse maximale
2	V_{40}	Vitesse à 40 g/l de CO_2 produit
3	V_{60}	Vitesse à 60 g/l de CO_2 produit
4	V_{20-60}	Vitesse moyenne entre 20 g/l et 40 g/l de CO_2 produit
5	$V_{80\%}$	CO_2 produit pendant que $V > 0.8V_{max}$
6	CO_{2max}	Quantité maximale de CO_2 produit
7	F_1	
8	F_2	Coordonnées factorielles sur les 5 premières
9	F_3	composantes d'une ACP calculée sur les
10	F_4	30 premiers points des courbes
11	F_5	

TAB. 4.2 – Signification des variables du deuxième jeu de données.

Trois méthodes de discrimination ont été testées : l'AFD, la DPLS et les réseaux de neurones artificiels (RNA). Ces méthodes ont été combinées avec deux types de sélection de variables : procédure stepwise (Step) et algorithmes génétiques (AG). Au total, 8 modèles ont été testés, comme reporté dans la figure 4.1. Compte tenu du petit nombre d'individus et comme le but était de comparer des méthodes, aucun test réel des modèles n'a été effectué ; toutes les performances ont été évaluées par validation croisée leave-one-out.

L'algorithme stepwise opérait sur la maximisation du Lambda de Wilks, avec retour arrière possible. Le nombre maximal de variables étaient ainsi triées par ordre décroissant d'intérêt, puis réintroduites une à une dans une validation croisée. Le minimum de l'erreur a donné la meilleure sélection.

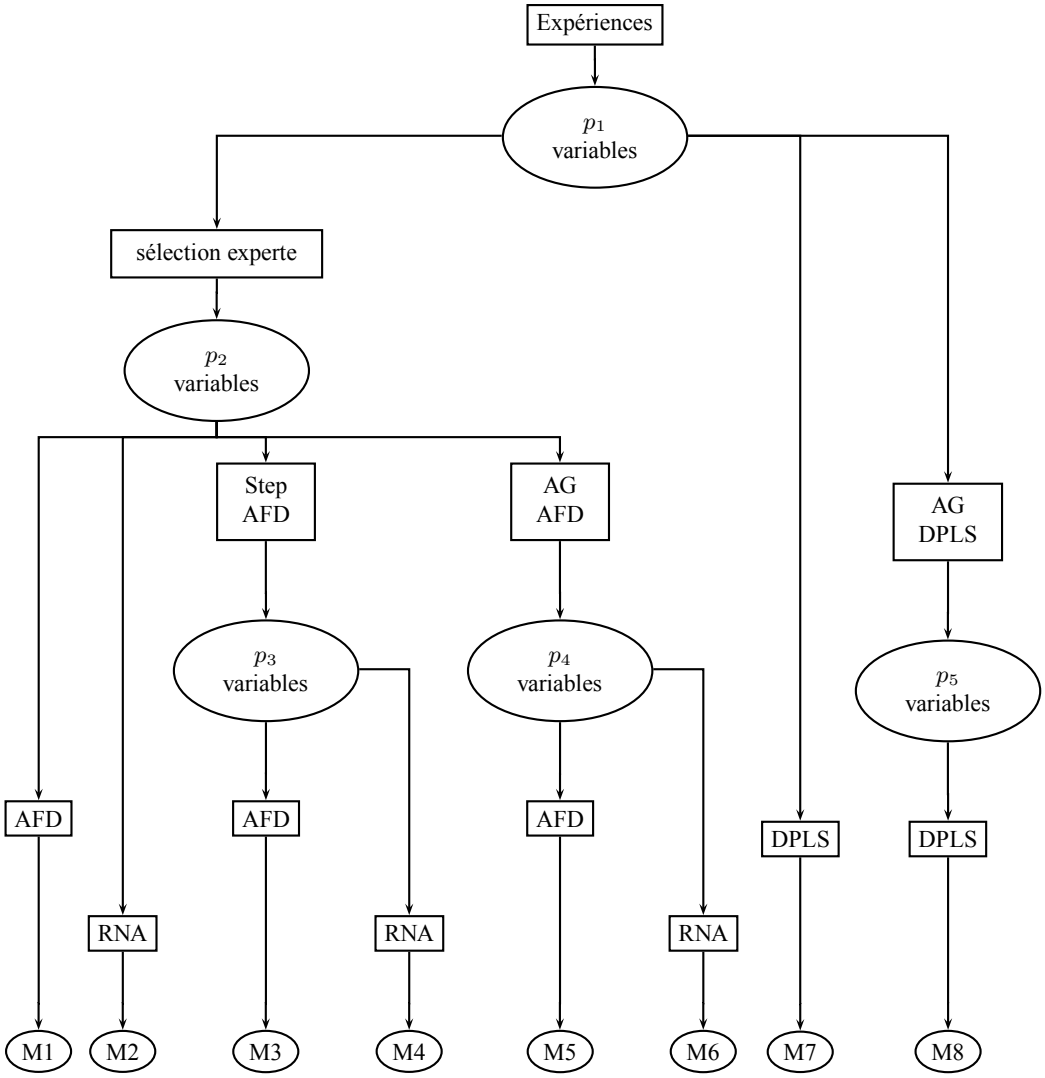

FIG. 4.1 – Récapitulatif des modèles testés.

L'algorithme génétique utilisé était issu d'une conception propre, comme publié dans [Roger and Bellon-Maurel, 2000]. Il a opéré de manière classique : Une population de 16 *chromosomes* de p bits (codant la sélection des variables) a tout d'abord été générée aléatoirement. A chaque génération, toutes les sélections codées par les chromosomes sont testées en validation croisée, affectant à chaque chromosome une *fitness*. La meilleure moitié de la population est gardée, puis utilisée pour générer autant de chromosomes fils par *crossing-over* et *mutation*. L'algorithme stoppe lorsque 60% des chromosomes sont identiques. Le processus a été répété 100 fois. Lorsqu'il a été appliqué aux p_1 variables (modèle 8), ont été retenues les variables sélectionnées plus de 50 fois. Lorsqu'il a été appliqué aux p_2 variables (modèles 5 et 6), c'est la meilleure sélection qui a été retenue.

Le réseau de neurones artificiel était constitué de 3 couches, c'est à dire qu'il avait 1 seule couche cachée. Les fonctions d'activation de cette couche ont été choisies sigmoïdales, et celles de la couche de sortie, linéaires. Le nombre de neurones de la couche cachée a été testé entre 1 et 15 par validation croisée, afin de trouver la meilleure configuration.

Enfin, le réglage du modèle par DPLS s'est fait en observant l'évolution de l'erreur de validation croisée en fonction du nombre de variables latentes utilisées.

4.2.2 Résultats et discussion

Les résultats avec 5 classes ont montré que la classe sy été toujours confondue avec les autres. Elle a donc été ôtée de l'analyse. Le tableau 4.3 montre l'ensemble des résultats pour les 8 modèles.

	Erreur de validation croisée		
Variables	AFD	DPLS	RNA
11 variables synthétiques	27%		30%
Variables choisies par Stepwise $V_{20-60}, CO_{2max}, V_{max}, F_3$	15%		27%
Variables choisies par Algorithmes Génétiques $V_{max}, V_{60}, CO_{2max}, F_2$	9%		12%
Courbe entière (80 variables)		12%	
Sélection Algorithmes Génétiques		6%	

TAB. 4.3 – Résultats des 8 modèles de discrimination.

Ces résultats illustrent bien les capacités d'apprentissage des différentes méthodes :

- La puissance de la PLS est vérifiée, puisque, sans aucun prétraitement ni sélection, elle réalise un score de 12% d'erreur. Elle apparaît bien comme un bon outil "boîte noire", qui est capable d'exprimer les relations linéaires existantes, sans requérir de connaissances expertes. Elle offre cependant la possibilité d'expliquer ces relations, en analysant les b-coefficients du modèle. Ainsi, on a pu constater que les régions présentant des coefficients élevés correspondaient aux zones considérées comme importantes par l'expert.
- L'amélioration apportée par la sélection de variables est très nette. Dans le cas de la DPLS associée aux algorithmes génétiques, cela conduit au meilleur modèle (6% d'erreur seulement). Les zones sélectionnées sont conformes aux connaissances de l'expert, à savoir : autour du

maximum, lors du ralentissement et en fin de fermentation. Dans les cas de l'AFD et des RNA, la sélection par algorithmes génétiques est meilleure que celle par stepwise. Ceci était prévisible, car l'utilisation que nous avons faite ici des algorithmes génétiques revenait certainement à parcourir la totalité de l'espace des solutions et à sélectionner la meilleure.
- Quel que soit l'ensemble de variables utilisé, l'AFD se montre meilleure que les RNA. Ceci est certainement dû au faible nombre d'exemples disponibles pour réaliser l'apprentissage.
- Enfin, il faut noter que certaines variables extraites des courbes se montrent assez performantes, mais seulement après sélection. En effet, une AFD sur les 4 variables sélectionnées par algorithmes génétiques réalise le deuxième meilleur score (9% d'erreur). A noter la complémentarité des 3 premières variables sélectionnées ($V_{max}, V_{60}, CO_{2max}$), qui sont des variables expertes, avec la quatrième (F_2), issue d'une ACP. Ceci peut s'interpréter par le fait que les grandes tendances sont bien connues de l'expert, mais que certains détails, extraits par une méthode statistique, sont nécessaires pour réaliser un modèle performant.

Cet article n'avait pas pour but de construire un modèle réel de discrimination. Il avait plutôt pour vocation de présenter à la communauté du Génie des Procédés Biologiques une panoplie d'outils de discrimination de courbes, de montrer leur paramétrage et de discuter de leurs avantages et inconvénients. L'intérêt de la sélection de variables et de la complémentarité entre connaissances expertes et connaissances expérimentales y est aussi clairement montré.

4.3 Discrimination par parcours des Fonctions Propres Focales

Cette section est dédiée à une contribution originale, soumise pour publication à *Chemometrics and Intelligent Laboratory Systems* ([Article VI]).

Cette méthode cherche des vecteurs discriminants optimaux, sans passer par l'intermédiaire d'une réduction de dimension. Pour ce faire, elle définit des fonctions qu'il suffit de parcourir pour choisir l'espace discriminant optimal.

Nous ne nous intéresserons ici qu'aux problèmes mal conditionnés et mal dimensionnés, tels que rencontrés en spectrométrie. Le lecteur trouvera dans l'article référencé une discussion montrant que ces deux notions sont en fait assez similaires et peuvent être traitées de la même manière. En outre, la méthode proposée ici, lorsqu'elle est appliquée aux cas bien conditionnés, donne les mêmes résultats que l'AFD.

4.3.1 Définition des fonctions propres focales

Pour tout $z \in [0,1]$, soient $F_1(z), \cdots, F_{c-1}(z)$ les $c-1$ plus grandes valeurs propres de $\mathbf{B} - z\mathbf{T}$ en valeur absolue, telles que $F_1(z) \geq F_2(z) \geq \cdots \geq F_{c-1}(z)$. Ceci définit une famille de r fonctions de $[0,1]$ dans $\mathbb{R}$, que nous appelons Fonctions Propres Focales (FPF). A chaque FPF est associée une fonction vectorielle, de $[0,1]$ dans $\mathbb{R}^p$, $\mathbf{u}_i(z)$, telle que $\mathbf{u}_i(z)$ est vecteur propre associé à $F_i(z)$.

4.3.2 Propriétés des fonctions propres focales

Dans [Article VI], nous avons montré que les fonctions propres focales ont les propriétés suivantes :

Propriété 1 *Les FPF sont dérivables et strictement décroissantes.*

Propriété 2 *Dans le cas mal dimensionné ($n \leq p$), les fonctions propres focales sont strictement positives sur* $[0,1[$ *et nulles en* 1.

Propriété 3 *Les FPF présentent une courbure positive.*

Propriété 4 *Le pouvoir discriminant des fonctions vectorielles u_i est croissant, i.e. que la fonction $L_i(z) = \Lambda(\mathbf{u}_i(z))$ est croissante.*

4.3.3 Illustration

Prenons les célèbres données des iris de Fisher [1]. Elles sont constituées de 150 individus, décrits par 4 variables et appartenant à 3 groupes de mêmes effectifs. Le graphe de la figure 4.2 (traits pleins) montre l'évolution des FPF F_1 et F_2 calculées sur ces données, avec une ordonnée logarithmique, afin de magnifier les faibles valeurs. Elles s'annulent respectivement en 0.97 et 0.22, valeurs propres de $\mathbf{T}^{-1}\mathbf{B}$ et solutions de la FDA.

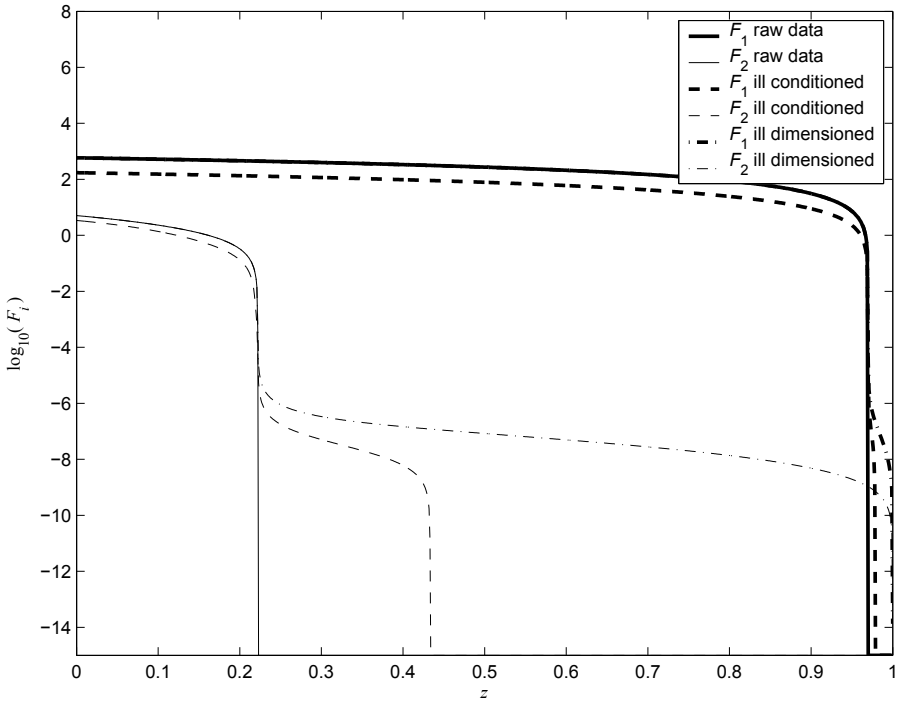

FIG. 4.2 – Fonctions F_i pour les iris de Fisher (en trait continu) ; pour les mêmes données rendues artificiellement mal conditionnées (en trait interrompu) et mal dimensionnées (trait mixte).

Transformons maintenant ces données pour produire un problème mal dimensionné et mal conditionné : multiplions la matrice des données par les 4 premiers facteurs d'une PCA réalisée sur un jeu de spectres de pommes décrits par 176 variables. Nos nouvelles données sont donc maintenant décrites par 176 variables, très corrélées entre elles, puisque le rang de la matrice ainsi obtenue est 4. Ajoutons alors à chacune de ces variables un bruit aléatoire uniforme d'une amplitude égale à 10^{-4} fois son écart-type, pour obtenir une matrice de rang 150 (149 une fois centrée). Ces données illustrent bien le cas de la spectrométrie NIR où l'on sait que les spectres, bien qu'ils soient décrits par un nombre important de variables, sont en réalité formés par une combinaison (plus ou moins) linéaire de quelques spectres purs additionnée de bruits. Sur la figure 4.2 (traits mixtes) sont reportées les FPF calculées pour ces nouvelles données. On y voit que le zéro de ces fonctions a été reporté à $z_1 = z_2 = 1$, mais qu'une inflexion a subsisté à l'endroit des zéros initiaux.

Créons maintenant un problème bien dimensionné mais mal conditionné, en ne retenant que les 40 premières variables du jeu de données précédent. Les FPF correspondant à ces données sont reportées sur la figure 4.2 (traits interrompus). Le comportement de ces fonctions est intermédiaire entre celui du

[1] http ://lib.stat.cmu.edu/DASL/Datafiles/Fisher'sIris.html

cas bien conditionné et celui du cas mal dimensionné. Elles s'annulent pour une valeur de z différente de 1, mais supérieure à la solution de la FDA réalisée sur les données brutes. Si on réalisait une FDA sur ces données, les deux axes discriminants présenteraient un Lambda de Wilks respectivement égal à 0.98 et 0.43, au lieu de 0.97 et 0.22, ce qui est évidemment un résultat illusoire.

4.3.4 Analyse discriminante par le parcours des Fonctions Propres Focales (FPF-AD)

Grâce à la propriété 4, nous savons que, si nous parcourons les FPF dans le sens croissant, l'espace engendré par les vecteurs propres associés est de plus en plus discriminant. Nous savons en outre que, dans les cas mal dimensionnés et/ou mal conditionnés, le zéro des FPF fournit une solution sur-ajustée (solution de l'AFD). L'idée de la FPF-AD est de générer une suite d'espaces discriminants de plus en plus ajustés, en parcourant les FPF dans le sens croissant avec des suites convergeant vers le zéro des FPF. Cette suite est ensuite testée par un critère de sur-apprentissage, comme par exemple une erreur de validation, afin de déterminer l'espace optimal.

Les autres propriétés permettent de définir des suites croissantes et convergeant vers le zéro des FPF. Trois méthodes sont ainsi données à titre d'exemple dans l'article associé :

Le parcours vertical utilise la propriété de bijectivité (héritée de la continuité et de la monotonie) des FPF, pour réaliser un parcours en "tranches" verticales. Toutes les FPF sont scannées de manière synchrone.

Le parcours asynchrone se sert des propriétés géométriques des FPF. Il utilise une suite de Newton Raphson, dont on est sûr qu'elle converge vers le zéro des FPF, à cause de leur décroissance et de leur courbure. Cette méthode scanne les FPF de manière indépendante (asynchrone) et adaptée à la forme de chacune d'elle.

Le parcours orthogonal consiste à implémenter un des deux parcours précédent, mais en orthogonalisant l'espace après le calcul de chaque vecteur. Ce parcours est implémenté par l'algorithme récursif suivant, où s est une suite de parcours vertical ou asynchrone :

1. $i = 1$
2. Pour chaque $s(k)$, calcul de $\mathbf{v} = \mathbf{u}_1(s(k))$
3. Si $i < c - 1$:
   - Orthogonalisation de $\mathbf{X}$ par rapport à $\mathbf{v}$: $\mathbf{X} = \mathbf{X}(\mathbf{I} - \mathbf{v}\mathbf{v}^T)$
   - $i = i + 1$
   - Aller à l'étape 2
4. Sinon, sortie de l'algorithme.

Le résultat est donc un arbre de profondeur $c - 1$. À un nœud de profondeur k est attachée une base orthonormée d'un espace discriminant de dimension k. Cette méthode fournit donc un nombre important de solutions. Elle présente l'avantage de produire des vecteurs discriminants orthogonaux. Par contre, elle requiert un nombre important de calculs, croissant exponentiellement avec le nombre de classes. Il est donc nécessaire d'optimiser le parcours à chaque profondeur de l'algorithme.

4.3.5 Matériel et méthodes

La FPF-AD, implémentée avec les trois types de parcours exposés précédemment, plus la PLS-DA, ont été testées sur un problème de discrimination de variétés de raisins à partir d'une mesure en spectrométrie visible et proche infrarouge. L'expérimentation a porté sur 3 variétés : *carignan* (crg), *grenache blanc* (grb) et *grenache noir*(grn). Pour les variétés crg et grb, les ensembles d'apprentissage et de test sont des lots de 50 individus différents, alors que pour la variété grn, un lot de 50 spectres

a été coupé aléatoirement en deux parties égales. Ainsi, les ensembles d'étalonnage et de test étaient constitués de $n = 125$ individus décrits par $p = 256$ variables.

Quelle que soit la méthode utilisée, une validation croisée leave-one-out a permis de calculer une erreur de validation croisée (CVE), exprimée en pourcentage de mauvaise classification, en regard des paramètres suivants :

- Pour la PLS-DA, le nombre de variables latentes : n_{LV}
- Pour le parcours vertical et le parcours asynchrone : $\beta(k)$
- Pour le parcours orthogonal : parcours de chaque niveau avec la méthode asynchrone paramétrée avec une suite $\beta(k)$
- Pour toutes les méthodes, le nombre de vecteurs discriminants : n_{DV}

4.3.6 Résultats et discussion

La figure 4.3 montre l'évolution de l'erreur de validation croisée pour les quatre modèles. A l'examen de ces courbes, les valeurs suivantes ont été retenues :

- Pour la PLS-DA, $n_{LV} = 10$ et $n_{DV} = 2$
- Pour le parcours vertical : $\beta = 10^{-3.2}$ et $n_{DV} = 2$
- Pour le parcours asynchrone : $\beta = 1, 1.5, 2, \cdots, 10.5, 11$ et $n_{DV} = 2$
- Pour le parcours orthogonal : $\beta^1 = 1, 2, \cdots, 5, 5.1, 5.2, 5.3$ et $\beta^2 = 1, 2, \cdots, 5, 5.5, 5.6, \cdots, 6$

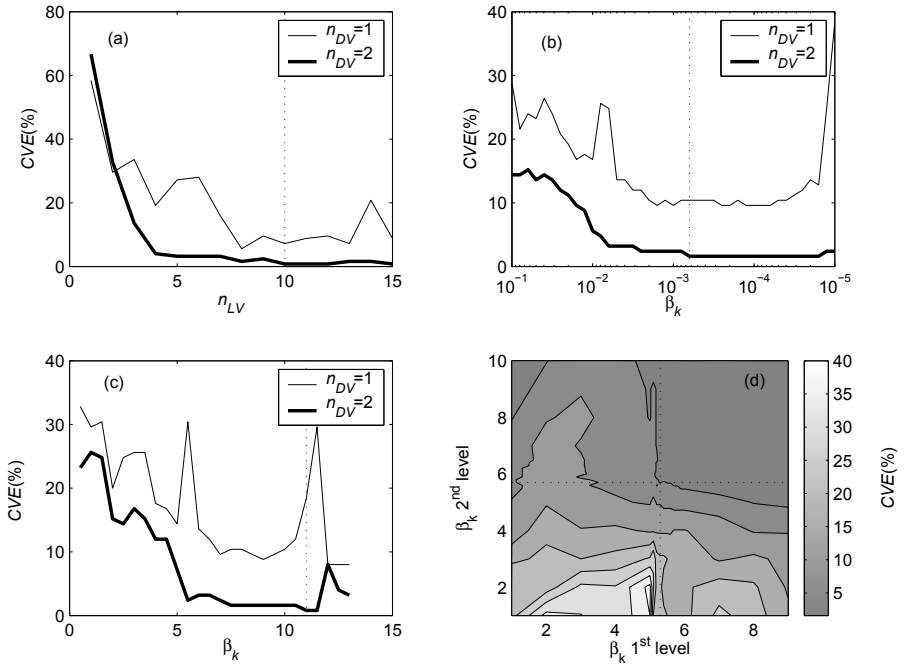

FIG. 4.3 – Evolution de l'erreur de validation croisée (CVE) pour : (a) la PLS-DA, (b) le parcours vertical, (c) le parcours asynchrone et (d) le parcours orthogonal.

Les résultats du test des quatre modèles paramétrés comme décrit ci-dessus sont reportés dans la table 4.4.

La PLS-DA semble moins efficace que l'AD-FPF, surtout si cette dernière est implémentée avec un parcours orthogonal. Cette première constatation doit néanmoins être tempérée par le fait que le modèle PLS-DA, choisi au vu de la cross-validation, est peut être trop ajusté aux données. Ainsi, au vu de la courbe de CVE, on aurait pu choisir 4 ou 8 variables latentes. La table 4.5 montre les résultats du test du modèle PLS-DA avec ces nouvelles valeurs. Le modèle à 4 variables latentes est clairement

PLS-DA				Parcours vertical			
$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn	$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn
crg	44	-	-	crg	44	-	-
grb	6	46	-	grb	6	49	-
grn	-	4	25	grn	-	1	25
PE = 8.0 %				PE = 5.6 %			
Parcours asynchrone				Parcours orthogonal			
$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn	$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn
crg	47	-	-	crg	49	-	-
grb	2	50	-	grb	-	50	-
grn	1	-	25	grn	1	-	25
PE = 2.4 %				PE = 0.8 %			

TAB. 4.4 – Résultat du test des quatre méthodes de discrimination.

sous ajusté ; il n'est pas capable de différencier la classe crg. Le modèle à 8 variables latentes est certes un peu meilleur que celui à 10 variables latentes, mais toujours bien moins bon que le parcours orthogonal.

$n_{LV} = 4$				$n_{LV} = 8$			
$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn	$\widehat{\mathbf{Y}}^T\mathbf{Y}$	crg	grb	grn
crg	37	-	-	crg	42	-	-
grb	12	50	-	grb	8	49	-
grn	1	0	25	grn	-	1	25
PE = 10.4 %				PE = 7.2 %			

TAB. 4.5 – Résultat du test sur le jeu de données du modèle PLS-DA avec 4 et 8 variables latentes.

La figure 4.4 montre l'ensemble de test projeté dans l'espace discriminant, pour le modèle à parcours othogonal, ainsi que les vecteurs discriminants qui forment une base orthonormée de cet espace. La forme de ces deux vecteurs est tout à fait interprétable, en terme de colorimétrie (relation entre les pics dans le visible et la couleur des baies) et de spectrométrie NIR (pic d'absorption de l'eau à 960 nm). Une interprétation détaillée peut être trouvée dans [Article VI].

Méthode	PLS-DA	FPF-AD vert.	FPF-AD async.	FPF-AF orth.
$\mathbf{u}_1^T\mathbf{u}_2$	-0.6153	0.0084	0.0046	0.0000

TAB. 4.6 – Cosinus entre les deux vecteurs discriminants, pour les quatre modèles.

4.3.7 Conclusion

En matière de discrimination sur des données mal conditionnées, la FPF-DA constitue bien une alternative originale. Par rapport à la PLS-DA, elle présente les caractéristiques suivantes :

– Toute l'information est prise en compte, directement, dans le calcul des vecteurs discriminants. Lorsque l'on utilise une PLS-DA, au moment du choix du nombre de variables latentes, on se prive d'une certaine quantité d'information. La décomposition factorielle réalisée par la PLS nous assure uniquement que les dimensions ignorées par le modèle sont telles que la covariance entre les scores et les degrés d'appartenance est plus faible que dans celles retenues. Dans

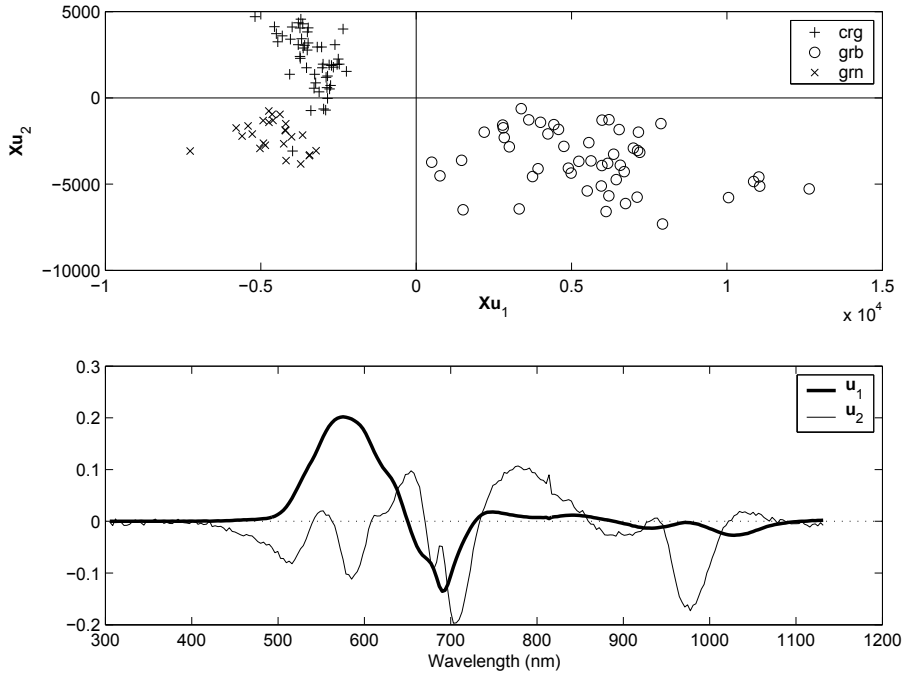

FIG. 4.4 – En haut : Carte factorielle de l'ensemble de test après application du modèle basé sur le parcours orthogonal. En bas : Vecteurs discriminants du modèle.

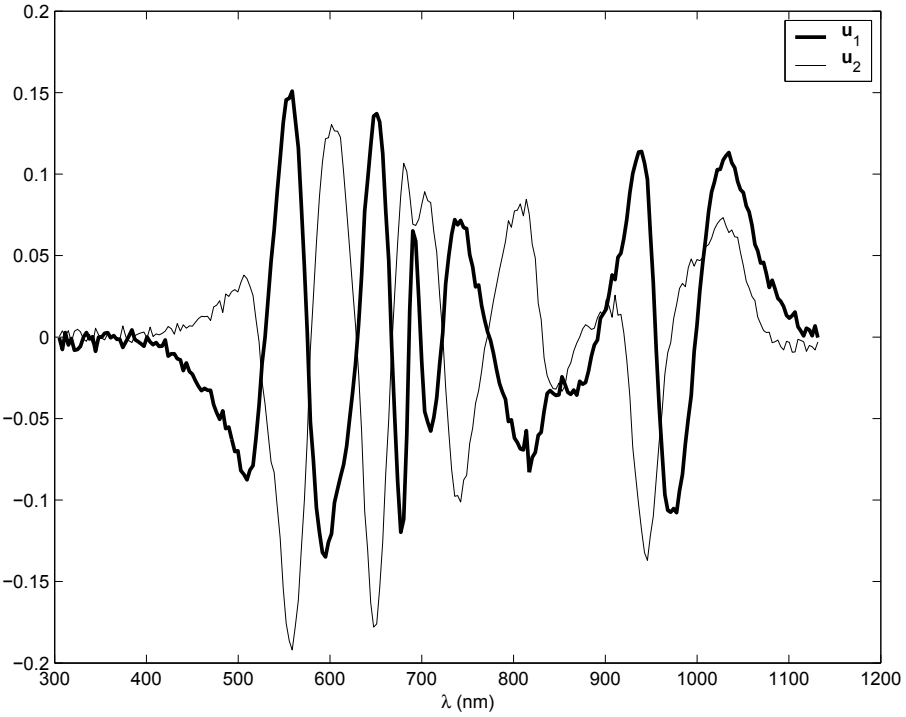

FIG. 4.5 – Vecteurs discriminants du modèle PLS-DA avec 10 variables latentes.

[Ripley, 1996], il est montré que la discrimination par régression sur les degrés d'appartenance n'est équivalente à une véritable analyse discriminante qu'à la condition que les densités de probabilité des classes soient identiques. Le critère de séparabilité utilisé par la FPF-DA se base sur des hypothèses beaucoup moins fortes. Il semble donc que la FPF-AD permette de réaliser des modèles plus justes qu'avec une PLS-DA.

- Les paramètres de réglage des parcours sont continus ; ils peuvent donc être réglés très finement. Pour la PLS-DA, par contre, le paramètre de réglage est discret. Chaque raffinement de l'algorithme consiste à ajouter une dimension au modèle, ce qui peut se révéler brutal.
- Les vecteurs produits sont très peu colinéaires, voire orthogonaux (dans le cas du parcours orthogonal), comme en témoigne la table 4.6, qui donne les cosinus entre les deux vecteurs discriminants pour les quatre méthodes, sur l'exemple ci-dessus. Le premier avantage de cette propriété est d'ordre algébrique : plus les vecteurs discriminants sont indépendants, meilleure est la définition de l'espace discriminant. Le deuxième a trait à l'interprétation des vecteurs discriminants : s'ils sont indépendants, leur interprétation spectroscopique sera facilitée. A titre d'illustration, il suffit de comparer les vecteurs discriminants de la PLS-DA reportés en figure 4.5 avec ceux de la FPF-AD orthogonale, reportés en figure 4.4.

Cependant, la FPF-AD est une méthode gourmande en puissance de calcul. De plus, le parcours orthogonal paraît compliqué à mettre en œuvre pour un problème à plus de 3 classes. En effet, le nombre de paramètres en fonction desquels la variation du $SECV$ doit être examinée devient trop important pour autoriser une visualisation graphique simple. Des optimisations doivent donc être trouvées pour rendre cette méthode plus facile à utiliser.

4.4 Perspectives de recherches

La discrimination avait été identifiée, dans le chapitre 1, comme une voie de recherche à la limite de la problématique de la robustesse. Dans ce cadre, elle permet d'effectuer des diagnostics qui peuvent modifier l'étalonnage multi-varié. Par exemple, dans le cas de la mesure de la qualité des fruits, on peut imaginer détecter la variété pour sélectionner le modèle le plus adapté. Elle peut aussi permettre de détecter en temps réel les spectres aberrants.

Si on élargit un peu le point de vue, je pense que la discrimination deviendra en fait très importante, voire capitale, dans un avenir assez proche. En effet, la spectrométrie est utilisée (depuis relativement peu de temps) comme une technique palliative des mesures destructives classiques. Son domaine de prédilection était donc le laboratoire de chimie analytique et les grandeurs mesurées étaient des concentrations. La chimiométrie s'est donc tout naturellement intéressée à l'étalonnage de la mesure quantitative. Dorénavant, on commence à envisager d'utiliser la spectrométrie comme un moyen invasif mais non destructif, capable de fournir des informations générales sur le produit examiné, à relier à des caractéristiques qui peuvent être de nature qualitative. Notre cadre applicatif en recelle de nombreux exemples :

- En premier exemple : le caractère sucré d'un fruit. Il s'agit d'une variable "faussement" qualitative, puisqu'elle est ordonnée, mais qui peut être traitée par discrimination. Dans les applications de tri des fruits, le taux de sucre exact n'est pas demandé. Si le taux est mesuré, pour réaliser l'opération de tri, des seuils de décision vont être appliqués pour réaliser des classes (par exemple : peu sucré, moyennement sucré et très sucré). Dans ces conditions, ne gagnerait on pas à réaliser directement la prédiction de la classe ?
- Deuxième exemple : la qualité d'un fruit. Aujourd'hui, la qualité d'un fruit est estimée, approximée, par un ensemble de critères objectifs et mesurables : taux de sucre, couleur, fermeté, farinosité, etc. Or, cette qualité est par essence même une variable non quantitative. La véritable mesure de la qualité ne peut se faire que par des analyses sensorielles, à partir de jugements d'experts. La discrimination serait donc tout à fait indiquée pour relier les mesures

physiques telles que la spectrométrie, éventuellement associée à d'autres mesures disponibles, à des classes définies par des experts.

- Troisième exemple : le traitement des images hyperspectrales. De récents progrès technologiques mettent à notre disposition des systèmes de prise d'images hyperspectrales, i.e. pour lesquelles chaque pixel est un spectre. Ce matériel est encore onéreux et lent, mais nous pouvons gager qu'un avenir proche nous apportera des solutions temps réel d'un prix compatible avec nos applications. Or, le traitement des images numériques passe essentiellement par une opération de segmentation, qui a pour but de reconnaître des objets ou des régions. Cette opération n'est rien d'autre qu'une classification qui, si elle est apprise par un algorithme supervisé, devient de la discrimination.

A la lumière des développements exposés dans ce chapitre et des perspectives d'application illustrées par les exemples ci-dessus, je pense qu'en matière de discrimination à partir des spectres, les perspectives de recherche suivantes sont ouvertes :

- Caractérisation qualitative directe d'un objet biologique, en association avec d'autres systèmes et d'autres informations. En d'autres termes, comment utiliser une mesure spectrométrique dans un système d'aide à la décision ou de diagnostic ?
- Dans le cas de variables quantitatives, quel lien peut on réaliser entre une discrimination sur la variable discrétisée et une régression classique ? Est ce que la discrimination est plus robuste qu'une régression suivie d'un seuillage ? Est-ce qu'une discrimination suivie d'une interpolation peut s'avérer meilleure qu'une régression, notamment dans les cas de non linéarité ?
- Dans le cadre du traitement d'images hyperspectrales, comment réaliser une discrimination qui tienne compte conjointement de la dimension spectrale et des dimensions spatiales ?

Conclusion générale

Ce mémoire a présenté la démarche de recherche que j'ai employée pour tenter de résoudre le problème de la robustesse des étalonnages multivariés embarqués dans les capteurs à base de spectrométrie infrarouge. Une analyse de la problématique m'a amené à dégager trois voies de recherche : La prise en compte d'une grandeur d'influence ; La maintenance de la robustesse du modèle ; La discrimination à partir de spectres. À chacune de ces voies était dédié un chapitre, bâti sur deux articles auxquels j'ai fortement contribué : le premier a été utilisé pour dresser un panorama de l'état de l'art, le deuxième pour exposer une contribution originale.

Ces recherches m'ont permis d'améliorer la robustesse des capteurs développés au Cemagref, dans le cadre de projets de recherche. En outre, elles ouvrent nombre de perspectives, pour résoudre les problèmes qui subsistent et pour investir de nouvelles technologies :

– Avec les procédés d'orthogonalisation, tels que décrits dans 2.3 et 3.3, un nouveau point de vue apparaît : Les données constitutives de l'ensemble d'apprentissage forment un nuage de points, dans l'espace de mesure spectrale. Ce nuage définit un sous espace, représenté par des structures latentes, i.e. par une base de spectres. Lorsque l'on utilise une technique d'étalonnage classique, comme la PLSR, on identifie le sous espace dans lequel les variations spectrales sont les plus reliées à la réponse, en considérant le reste de l'espace comme du bruit. En procédant à des corrections par projection orthogonale, on réalise exactement l'inverse. On identifie les bruits structurés, i.e. les spectres *parasites*, pour les enlever de l'espace de mesure. D'un autre côté, nombre de spectres, donc de structures vectorielles, sont disponibles, en provenance de l'expertise ou de l'expérience. Il me semble donc très intéressant d'étudier, dans un avenir proche, comment toutes ces structures vectorielles, qu'elles soient issues de décompositions algébriques ou de connaissances théoriques, peuvent faire progresser notre connaissance de la mesure spectrométrique et améliorer son exploitation. Ainsi, par exemple, je propose de rechercher un moyen permettant d'intégrer les connaissances théoriques sur la diffusion de la lumière dans les méthodes d'étalonnage classiques.

– La méthode DOP (Cf. 3.3), en introduisant la notion de standard virtuel, ouvre la porte à un grand nombre d'applications potentielles, comme le transfert d'étalonnage entre instruments, entre variétés, etc. Cette perspective, très large, devra être abordée méthodiquement. Il sera notamment nécessaire, rapidement, d'étudier finement le comportement de DOP dans le cadre des réponses multiples. D'autre part, la gestion des informations fournies par DOP et leur utilisation en temps réel, conjointement à un système de diagnostic, semble constituer un enjeu intéressant, dans le cadre de l'automatique et de la productique.

– C'est peut être en matière de discrimination que les perspectives de recherches sont les plus novatrices. En effet, la spectrométrie, couplée à la chimiométrie s'est jusqu'à présent tout naturellement tournée vers des applications de quantification, propres à la chimie analytique (d'où le nom de chimiométrie). Or, surtout dans notre cadre applicatif, la connaissance que l'on cherche à produire est bien souvent de nature qualitative. Elle relève bien souvent d'un raisonnement de type diagnostique, c'est à dire d'une discrimination entre des classes. Une telle connaissance est

alors *compatible* avec les systèmes d'aide à la décision, voire avec le paradigme de l'opérateur humain. La question devient alors : comment traiter au mieux l'information contenue dans un spectre, ou dans une image hyperspectrale, pour alimenter un système basé sur la connaissance ?

Bibliographie

[Andrew and Fearn, 2004] Andrew, A. and Fearn, T. (2004). Transfer by orthogonal projection : making near-infrared calibrations robust to between-instrument variation. *Chemometrics and Intelligent Laboratory Systems*, 72(1) :51–56.

[Chauchard et al., 2004] Chauchard, F., Cogdill, R., Roussel, S., Roger, J. M., and Bellon-Maurel, V. (2004). Application of ls-svm to non-linear phenomena in nir spectroscopy : development of a robust and portable sensor for acidity prediction in grapes. *Chemometrics and Intelligent Laboratory Systems*, 71(2) :141–150.

[Fisher, 1936] Fisher, R. (1936). The use of multiple measurements in taxinomic problems. *Ann. Eugenics*, 7 :179–188.

[Foley and Sammon, 1975] Foley, D. and Sammon, J. (1975). An optimal set of discriminant vectors. *IEEE Trans. Comput.*, 24(3) :281–289.

[Indahl et al., 1999] Indahl, U., Sahni, N., Kirkhus, B., and Næs, T. (1999). Multivariate strategies for classification based on nir-spectra–with application to mayonnaise. *Chemometrics and Intelligent Laboratory Systems*, 49 :19–31.

[Lachenal, 2000] Lachenal, G. (2000). Introduction à la spectroscopie infrarouge. In TecDoc, editor, *La spectroscopie infrarouge et ses applications analytiques*, pages 31–75. Lavoisier, 11 rue Lavoisier F75384 Paris, 1ère edition.

[Liu et al., 1992] Liu, K., Cheng, Y., and Yang, J. (1992). An generalized optimal set of discriminant vectors. *Pattern Recognition*, 25(7) :731–739.

[Martens and Naes, 1989] Martens, H. and Naes, T. (1989). *Multivariate Calibration*. Wiley, New York.

[Otto, 1999] Otto, M. (1999). *Chemometrics - Statistics and Computer Aplication in Analytical Chemistry*. Wiley-VCH, D-69469 Weinheim, 1st edition.

[Ripley, 1996] Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge University Press, Cambridge.

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

[Sánchez et al., 2003] Sánchez, N. H., Lurol, S., Roger, J. M., and Bellon-Maurel, V. (2003). Robustness of models based on nir spectra for sugar content prediction in apples. *J. Near Infrared Spectrosc.*, 11 :97–102.

[Seasholtz and Kowalski, 1993] Seasholtz, M. and Kowalski, B. (1993). The parsimony principle applied to multivariate calibration. *Anal. Chim. Acta*, 277 :165–177.

[Trygg, 2001] Trygg, J. (2001). *Parcimonious Multivariate Models*. PhD thesis, UmeåUniversity.

[Wold, 1978] Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, 20 :397–405.

[Xiao-Jun et al., 2004] Xiao-Jun, W., Kittler, J., Jing-Yu, Y., and Shi-Tong, W. (2004). An analytical algorithm for determining the generalized optimal set of discriminant vectors. *Pattern Recognition*, In press.

[Zeaiter, 2004] Zeaiter, M. (2004). *Mesure robuste en ligne des solutés organiques*. PhD thesis, SPBI - Montpellier University.

[Zeaiter et al., 2004] Zeaiter, M., Roger, J. M., Bellon-Maurel, V., and Rutledge, D. N. (2004). Robustness of models developed by multivariate calibration. part i : The assessment of robustness. *TrAC Trends in Analytical Chemistry*, 23, 2 :157–170.