Année : 2008 N° Ordre :

Université des Sciences et Technologies de Lille

Habilitation à Diriger des Recherches

Discipline : Sciences Physiques

Cyril Ruckebusch
LAboratoire de Spectrochimie Infrarouge et Raman
UMR 8516

Résolution et modélisation chimiométrique en spectroscopie moléculaire

Soutenue le 9 juin 2008 devant la commission d'examen :

- M. Jean-Pierre Fouassier, Professeur, Université de Mulhouse
- M. Romà Tauler, Professeur, CSIC et Université de Barcelone
- M. Pascal Pernot, Chargé de Recherche CNRS, Université de Paris 11
- M. Jean-Louis Bon, Professeur, Université de Lille 1
- M. Guy Buntinx, Directeur de Recherche CNRS, Université de Lille 1
- M. Jean-Pierre Huvenne, Professeur, Université de Lille 1

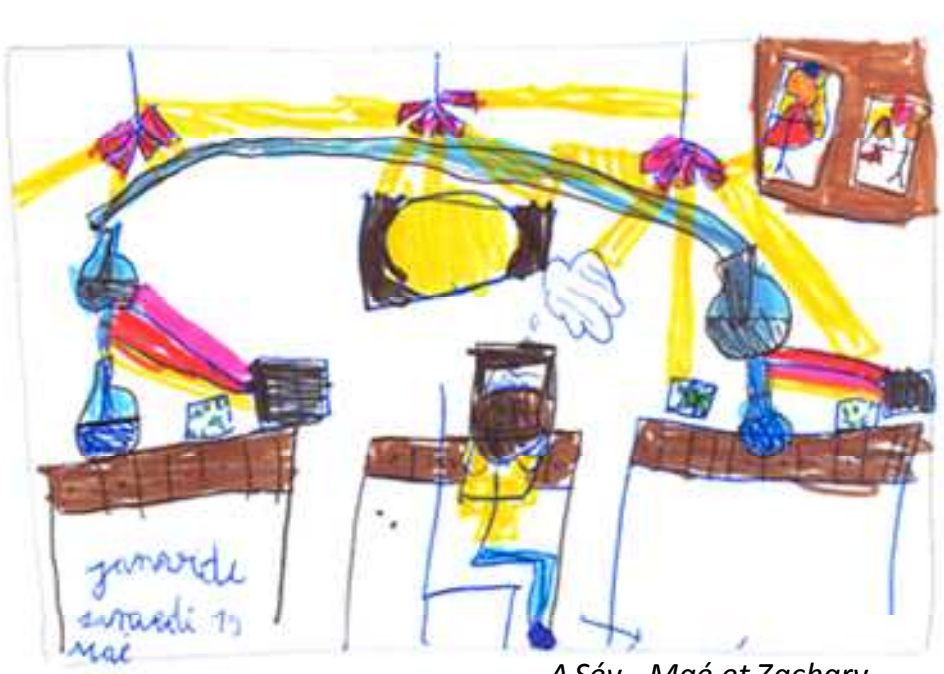
Préface

Ce mémoire comporte trois chapitres. Le premier est un curriculum vitae qui résume mon parcours professionnel, mes activités scientifiques et pédagogiques ; le second chapitre détaille mon travail de recherche et donne l'orientation de mes futurs travaux ; le troisième chapitre regroupe quelques publications sélectionnées pour illustrer la discussion proposée.

Mon travail de recherche consiste à développer des modèles chimiométriques pour représenter et analyser des données spectroscopiques. Ces observations décrivent des réactions chimiques, des systèmes physico-chimiques évolutifs, des mélanges ou des échantillons naturels. Par modèle chimiométrique, on entend à la fois le sens mathématique (descriptif) et le sens statistique (apprentissage) du terme. Cela permet d'englober dans cette définition les méthodes d'analyse factorielle, les méthodes de résolution de courbes et les modèles de classification ou de régression. Au-delà des aspects prédictifs pragmatiques, il y a en chimiométrie des aspects méthodologiques et philosophiques. Ils peuvent être considérés dans le cadre de la modélisation statistique, approche plus générale qui permet l'ouverture à d'autres disciplines.

Au quotidien, j'apprécie le caractère pluridisciplinaire de mon travail à l'interface des mathématiques appliquées, de la spectroscopie, de la physico-chimie et de l'instrumentation. J'apprécie également le dynamisme lié aux différents cadres dans lesquels émargent ces travaux de recherche : activités du laboratoire, collaborations académiques et contrats industriels. Je trouve assez enthousiasmant d'être impliqué dans des projets au-delà du monde universitaire, dans le développement d'instrumentations et de méthodes innovantes. C'est cette vision des choses que j'essaie de transmettre dans l'encadrement des étudiants au LAboratoire de Spectrochimie Infrarouge et Raman et lors de mes enseignements à l'Ecole Polytechnique Universitaire de Lille.

Enfin, ce mémoire est l'occasion d'exprimer de sincères remerciements à tous les chercheurs qui de près ou de loin ont participé à ces travaux. Je réserve une mention particulière à mes collègues et amis de l'équipe Chimiométrie et caractérisation moléculaire. Je suis également très reconnaissant aux membres du jury qui me font l'honneur de juger ce travail.



A Sév., Maé et Zachary

Table des matières

Chapitre 2

Activités et projet de recherche

1 Introduction

Mes travaux de recherche contribuent aux développements chimiométriques en spectroscopie moléculaire pour la description, la résolution ou la modélisation de systèmes physico-chimiques complexes. Au-delà de la synthèse des articles publiés (certains résultats sont repris en encadré dans le texte), l'objectif est de proposer une approche relativement conceptuelle des méthodes de chimiométrie, pour permettre une meilleure ouverture en chimie analytique ou dans d'autres disciplines scientifiques. L'objectif est également de rappeler que les activités académiques, les activités de recherche et les activités liées à la valorisation de la recherche ou à l'innovation forment un ensemble, souvent très pluridisciplinaire, qui est et doit rester la « marque de fabrique » du métier d'enseignant-chercheur.

Le premier axe de recherche que je détaillerai concerne les méthodes multivariées de résolution de courbes[1-4] pour l'analyse des données spectroscopiques des systèmes chimiques évolutifs, tels que les réactions photoinduites. L'idée générale est la description de matrices de données dites *2-voies*, les matrices de données spectro-cinétiques par exemple, par un modèle bilinéaire des contributions des constituants purs du système. L'originalité est que seule la structure algébrique de la décomposition mathématique est imposée. Cela permet de s'affranchir des contraintes rigides des méthodes plus classiques d'analyse factorielle : orthogonalité des axes en analyse en composantes principales ou indépendance statistique en analyse en composantes indépendantes. En effet, même si elles garantissent l'unicité de la décomposition, ces contraintes fournissent des solutions abstraites. En contrepartie, des contraintes flexibles sont donc imposées au cours de la résolution pour limiter l'ambigüité de rotation de la décomposition matricielle et privilégier les solutions en accord avec la nature du problème, celles qui sont non négatives par exemple. En spectroscopie moléculaire, nos travaux contribuent depuis quelques années aux développements méthodologiques avec pour objectifs la recherche de solutions pertinentes et interprétables des phénomènes étudiés. Nous nous focaliserons sur l'analyse multivariée des données spectroscopiques résolues en temps. Différents exemples d'applications couvrant les domaines temporels accessibles instrumentalement au laboratoire seront proposés. Aux temps très courts, nous pointerons les limitations des méthodes uniquement orientées données, notamment lorsqu'il devient nécessaire de considérer *a priori* les caractéristiques instrumentales. Nous reviendrons également sur le positionnement des méthodes de résolution de courbes dans le cadre des mathématiques appliquées au traitement du signal. Enfin, les perspectives de travail ouvriront sur des approches mixtes, associant méthodes de résolution de courbes et approches statistiques de séparation de sources indépendantes ou intégrant l'analyse bayésienne.

Les méthodes supervisées de modélisation font également l'objet de recherches dans le groupe depuis une dizaine d'années. L'idée est la modélisation d'une grandeur analytique obtenue par une méthode chimique de référence sur la base d'observations spectroscopiques des échantillons. L'objectif est, lors d'analyses ultérieures, la prédiction de la grandeur d'intérêt à

partir d'une mesure spectroscopique. Je ferai le bilan des résultats obtenus dans cet axe de recherche sur des systèmes chimiques complexes et des échantillons naturels, issus d'expériences de laboratoire ou de procédés industriels. Nous constaterons que les problèmes chimiométriques explosent très vite en taille et en complexité dès lors qu'on considère des données spectroscopiques. Dans ces conditions, des approches plus flexibles, qui intègrent explicitement la recherche de modèles parcimonieux pour gérer le fléau de la dimension, seront envisagées. Ces méthodes sont issues de la théorie de l'apprentissage statistique[5-7] dont le champ d'application couvre tous les domaines scientifiques où les chercheurs sont confrontés à un volume de données expérimentales important, la chimie, la biologie ou les sciences humaines. La théorie de l'apprentissage statistique fournit un formalisme intéressant pour le développement, notamment par l'implémentation de termes de pénalisation de l'erreur empirique, et l'interprétation des modèles supervisés. Nous détaillerons principalement les approches novatrices des méthodes à noyaux, type séparateur à vaste marge. Les premiers résultats obtenus seront présentés en insistant sur les perspectives de travail pour le traitement des problématiques scientifiques du groupe, notamment celles liées à l'interprétation et à la pérennité des modèles prédictifs.

2 Résolution des spectres des systèmes chimiques évolutifs

L'observation des processus réactionnels en photochimie repose sur la mise en œuvre de techniques performantes de spectroscopie résolue en temps. Au-delà des aspects expérimentaux, une connaissance approfondie des mécanismes réactionnels requiert la résolution des données spectroscopiques, c'est à dire la caractérisation moléculaire des espèces intermédiaires inconnues, de courte durée de vie, qui possèdent des signatures spectrales souvent proches, ainsi que la proposition d'un modèle cinétique pour l'interprétation des phénomènes étudiés.

Pour l'analyse des systèmes chimiques évolutifs, les méthodes multivariées de résolution de courbes, en particulier les méthodes basées sur les moindres carrés alternés, présentent l'avantage d'être résolument orientées données. La connaissance explicite d'un modèle physico-chimique n'est pas nécessaire. Ces méthodes reposent sur un modèle mathématique bilinéaire de structure des données et implémentent des contraintes. Celles-ci sont inspirées par la nature physique ou chimique du problème à résoudre et appliquées indépendamment sur les spectres ou les profils de concentration. Elles sont également applicables à l'analyse simultanée de plusieurs systèmes, permettant de proposer une solution globale. Enfin, il peut être envisagé de coupler la résolution aux paramètres cinétiques d'un modèle pour tirer parti de la robustesse des solutions paramétriques.

Dans ce chapitre, nous montrons comment ces approches sont adaptables aux données spectroscopiques résolues en temps, en infrarouge à transformée de Fourier ou en absorption transitoire UV-visible. Nous présentons également les développements envisagés pour permettre le traitement des spectres enregistrés aux temps ultracourts en régime femtoseconde. Enfin, les méthodes proposées sont replacées dans le cadre plus général du traitement du signal.

2.1 Aspects théoriques et méthodologiques

Les méthodes multivariées de résolution de courbes (MCR, *Multivariate Curve Resolution*) forment aujourd'hui une famille d'une vingtaine d'algorithmes, plus ou moins utilisés en chimiométrie. Historiquement, ces méthodes sont issues de l'analyse factorielle.[8,9] Il faut cependant considérer le fait que l'on exploite des structures évolutives de données, c'est à dire que les lignes des matrices de données sont ordonnées. Le principe de ces méthodes dérive également de la correspondance entre le rang algébrique d'une matrice de données et le nombre de contributions chimiques du système,[10] tout au moins en l'absence de déficience de rang.

L'objectif des méthodes MCR est la décomposition *(2.1)* de la matrice des données spectrales **D**, de dimension (*m*, *n*), contenant l'information non sélective enregistrée sur le système évolutif. Cette matrice *2-voies* (qui regroupe les échantillons dans un tenseur d'ordre 2) est décomposée en deux matrices, **C** de dimension (*m*, *k*) et $\mathbf{S}^{\mathbf{T}}$ de dimension (*k*, *n*). Celles-ci contiennent respectivement les profils de concentration et les spectres d'absorption des *k* composantes individuelles (idéalement, *k* espèces chimiques dites pures). La matrice **E**, de dimension (*m*, *n*), regroupe les résidus de cette modélisation.

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^{\mathbf{T}} + \mathbf{E} \qquad (2.1)$$

D'un point de vue mathématique, les méthodes MCR reposent donc uniquement sur la décomposition bilinéaire de la matrice de données. Ces méthodes ne requièrent pas la connaissance *a priori* d'un modèle physico-chimique (un modèle basé sur les équations cinétiques, par exemple) pour décrire l'évolution du mélange. D'un point de vue numérique, les algorithmes utilisés sont généralement des variantes de la méthode des moindres carrés ordinaires tel que l'algorithme ALS[11] (*Alternating Least Squares*), mais d'autres expressions de la fonction de coût à minimiser peuvent être envisagées.

2.1.1 Optimisation en résolution MCR

Etant donné une matrice **D**, la résolution MCR la plus simple consiste à estimer la matrice **C** (ou la matrice $\mathbf{S}^{\mathbf{T}}$) solution du problème d'optimisation des moindres carrés sans contrainte posé en *(2.2)*. La solution *(2.3)* est une solution analytique obtenue simplement par résolution d'un système d'équations linéaires.[⊥]

$$\underset{\mathbf{S}^{\mathbf{T}}}{Minimiser} \left\|\mathbf{E}\right\|^2 = \underset{\mathbf{S}^{\mathbf{T}}}{Minimiser} \left\|\mathbf{D} - \mathbf{C}\mathbf{S}^{\mathbf{T}}\right\|^2 \qquad (2.2)$$

$$\hat{\mathbf{C}} = \mathbf{D}\,\mathbf{S}\,(\mathbf{S}^{\mathbf{T}}\mathbf{S})^{-1} = \mathbf{D}\,\mathbf{S}^{+} \qquad (2.3)$$

[⊥] La notation $\mathbf{S}^{+}$ (respectivement, $\mathbf{C}^{+}$) représente la matrice pseudo-inverse de **S** (respectivement, **C**)

Si on considère maintenant le cas de l'optimisation MCR sous contrainte, les matrices **C** et $\mathbf{S}^T$ des espèces pures sont estimées comme précédemment mais les connaissances génériques dont on dispose sont utilisées pour restreindre l'ensemble des valeurs prises par leurs éléments. La contrainte de non négativité est un peu le cas d'école en chimie, en particulier lors de l'estimation des concentrations à partir des données spectrales. On recherche les valeurs des éléments des matrices **C** et $\mathbf{S}^T$ qui minimisent le critère d'erreur quadratique *(2.4a)* soumis à des contraintes linéaires du type de celles exprimées en *(2.4b)*.

$$\underset{\mathbf{C},\mathbf{S}^T}{\text{Minimiser}} \left\|\mathbf{E}\right\|^2 = \underset{\mathbf{C},\mathbf{S}^T}{\text{Minimiser}} \left\|\mathbf{D} - \mathbf{C}\mathbf{S}^T\right\|^2 \qquad (2.4a)$$

$$\textit{Soumis à } c_{ij} \geq 0, s_{ij} \geq 0 \qquad (2.4b)$$

Le problème d'optimisation quadratique sous contrainte est détaillé en annexe (annexe 1). D'un point de vue géométrique, chaque contrainte définit un demi-espace et l'intersection de tous les demi-plans autorisés représente l'ensemble des vecteurs acceptables. Le problème se ramène alors à trouver, parmi cet ensemble de solutions, le point qui minimise l'erreur quadratique. Imposer des contraintes revient donc à favoriser un certain type de solution, ce qui peut améliorer la résolution des spectres et des profils de concentration. Néanmoins, puisqu'il existe un ensemble de solutions équivalentes qui supportent les mêmes restrictions (voir paragraphe *2.1.1b*), il n'est pas possible de démontrer leur validité. En pratique, il faut donc s'assurer que les contraintes imposées sont en accord avec la nature physique du problème d'une part et d'autre part qu'elles sont actives (contraintes d'égalité).

2.1.1a Algorithme MCR-ALS

Les étapes de l'algorithme MCR-ALS[2,4] sont reprises ci-dessous dans le cas où l'on dispose d'une estimation initiale de la matrice **C** (dans le cas où l'on démarre de la matrice $\mathbf{S}^T$, il suffit d'intervertir les étapes 3 et 4).

1. Détermination du nombre de contributions chimiques au système étudié, éventuellement par analyse du rang de la matrice de données **D**.
2. Construction d'estimations initiales des variables pures, profils de concentration **C** (ou spectres $\mathbf{S}^T$) à partir de la connaissance chimique du problème ou de méthodes d'analyse factorielle de **D**.
3. Etant données **D** et **C**, calcul de $\hat{\mathbf{S}} = \mathbf{C}^+\mathbf{D}$ par moindres carrés sous contrainte.
4. Etant données **D** et $\mathbf{S}^T$, calcul de $\hat{\mathbf{C}} = \mathbf{D}\mathbf{S}^+$ par moindres carrés sous contrainte.
5. Reproduire **D** à partir de $\hat{\mathbf{C}}$ et $\hat{\mathbf{S}}$, reprendre à l'étape 3 si la convergence n'est pas atteinte.

La résolution MCR nécessite donc tout d'abord de déterminer le sous-espace le plus approprié pour décrire les données. Dans un deuxième temps, il faut effectuer une rotation des vecteurs de base de ce sous-espace pour les amener à satisfaire les contraintes et obtenir des solutions possédant un sens chimique. L'optimisation ALS est la méthode la plus utilisée de part sa robustesse. Elle implique simplement un ensemble d'étapes de régression linéaire pour résoudre des modèles locaux, les facteurs des matrices **C** et $\mathbf{S}^T$ étant fixés alternativement. A chaque itération du cycle d'optimisation, les matrices **C** et $\mathbf{S}^T$ qui répondent au problème de minimisation de l'erreur **E** sont calculées. Les contraintes sont appliquées directement en forçant les solutions obtenues par moindres carrés ou en appliquant des algorithmes spécifiques.

2.1.1b Ambigüité des solutions

■ Les solutions de l'analyse en composantes principales sont parfaitement définies mathématiquement mais ne possèdent pas de sens physique ; les solutions des méthodes multivariées de résolution sont en accord avec la physique et la chimie du système mais ne sont pas uniques.

Les méthodes d'analyse factorielle comme l'analyse en composantes principales (ACP) produisent des solutions mathématiques au problème inverse défini précédemment *(2.1)*. La décomposition matricielle étant effectuée sous contrainte de variance maximale, d'orthogonalité et de normalisation des facteurs successifs, les solutions obtenues sont uniques et sans ambigüité. Néanmoins, ces contraintes très rigides ne sont, sauf exception, pas satisfaites par les grandeurs physico-chimiques du problème. En conséquence, les facteurs de l'ACP sont parfois très éloignés des solutions acceptables et ne permettent pas d'estimer l'allure des contributions chimiques réelles.

A l'opposé, les limitations des méthodes MCR pour résoudre le problème *(2.1)* sont principalement associées à l'absence d'unicité des solutions calculées. Cela signifie qu'il est possible de reproduire mathématiquement les données expérimentales avec la même précision en calculant des matrices **C** et $\mathbf{S}^T$ qui contiennent des combinaisons linéaires des contributions pures.

Pour illustrer cette ambigüité, considérons schématiquement le cas de l'analyse d'un mélange binaire sous contrainte de non-négativité des spectres et des concentrations.[12] Dans la situation schématique proposée (*Fig. 2.1*), les projections des échantillons (les *scores*) dans le plan des deux premières composantes principales d'une ACP génèrent une droite ; autrement dit le système est de rang est 2. Les points les plus éloignés sur cette droite correspondent aux échantillons les plus purs et matérialisent les limites intérieures des profils estimés. Néanmoins, si on extrapole sur cette droite dans les deux directions, les solutions obtenues sont strictement équivalentes tant que les profils de concentration et les spectres ne prennent pas de valeurs négatives. Ces solutions, obtenues dans le même sous-espace de l'ACP, correspondent effectivement à la même erreur quadratique. Les limites extérieures de validité sont atteintes aux points pour lesquelles des valeurs négatives des solutions sont observées. Finalement, pour les

deux composantes du mélange, deux limites encadrent donc un ensemble de solutions pour lesquelles la qualité de l'approximation des données ne change pas.

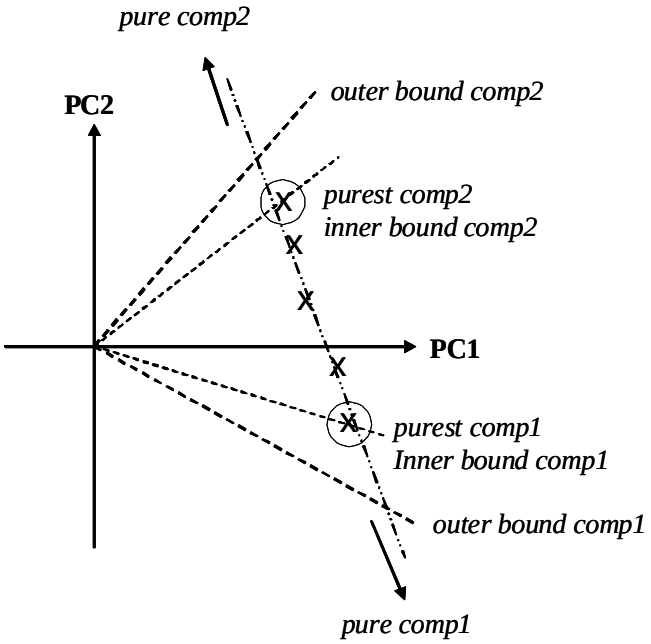


Fig. 2.1 Détermination graphique des bandes de faisabilité.[12]

Mathématiquement, les ambigüités de rotation et d'intensité peuvent s'écrire sous la forme *(2.5)*. Si aucune contrainte n'est imposée, alors il existe un nombre illimité de solutions possibles pour l'équation *(2.5a)* et l'équation *(2.5b)*, quelle que soit la matrice non singulière **U** considérée et avec *k* un nombre réel quelconque. L'objectif des contraintes est donc de restreindre l'ensemble des solutions possibles pour la matrice **U** qui effectue la rotation de **C** vers **C'** et de **S** vers **S'**.

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^{\mathbf{T}} + \mathbf{E} = \mathbf{C}\,(\mathbf{U}\mathbf{U}^{-1})\,\mathbf{S}^{\mathbf{T}} + \mathbf{E} = (\mathbf{C}\mathbf{U})(\mathbf{U}^{-1}\mathbf{S}^{\mathbf{T}}) = \mathbf{C}'\,\mathbf{S}'^{\mathbf{T}} + \mathbf{E} \qquad (2.5a)$$

$$\mathbf{D} = \mathbf{C}\,\mathbf{S}^{\mathbf{T}} + \mathbf{E} = \mathbf{C}\,(\frac{1}{k}k)\,\mathbf{S}^{\mathbf{T}} + \mathbf{E} = (\frac{\mathbf{C}}{k})\,(k\mathbf{S}^{\mathbf{T}}) + \mathbf{E} = \mathbf{C}''\,\mathbf{S}''^{\mathbf{T}} + \mathbf{E} \qquad (2.5b)$$

L'ambigüité d'intensité est en général résolue par normalisation des spectres. Par contre, la résolution sans ambigüité de rotation n'est possible que dans certains cas, très favorables, qui présentent des conditions de sélectivité garantissant l'unicité de la solution. Notons également que le calcul des bandes de faisabilité[13,14] (qui n'est pas abordé ici) peut être effectué par le biais de l'estimation des matrices de rotation spécifiques de chaque espèce pure.

2.1.2 Aspects méthodologiques

En dépit du caractère abstrait des facteurs de l'ACP, cette approche est l'outil principal pour étudier la structure et la complexité mathématique des données, et pour identifier les sources de variances potentiellement associées aux contributions chimiques. En effet, les facteurs de l'ACP sont (au bruit près) des combinaisons linéaires des contributions des composantes chimiques. Ces facteurs permettent donc d'accéder au rang de la matrice des données. D'autre part, comme l'illustre la figure précédente (*Fig. 2.1*), l'ACP définit le sous-espace dans lequel l'optimisation est effectuée. Dans le cas de l'algorithme MCR-ALS, les données expérimentales sont modélisées à partir d'estimations initiales des matrices **C** ou $\mathbf{S}^T$ obtenues par des méthodes d'analyse factorielle,[15] ou par des méthodes plus interactives.[16] Les solutions initiales qui satisfont déjà certaines des contraintes imposées sont favorisées pour conforter l'étape d'optimisation.

2.1.2a Rang d'une matrice de données, rang local

■ Idéalement, le rang d'une matrice de données est en accord avec le nombre de contributions du système étudié.

En pratique, toute technique d'analyse factorielle suppose le nombre de composantes du système (*k*, nombre de contributions chimiques) connu ou estimé au préalable. Si chaque composante d'un mélange se voit associée un spectre et un profil de concentration caractéristiques, alors le rang algébrique de la matrice des données (encore appelé pseudo-rang) est en général en très bon accord avec le nombre de contributions chimiques du système (ou rang chimique). En pratique, cela n'est vérifié qu'à condition que les données soient enregistrées dans des conditions expérimentales permettant d'assurer au mieux une séparation entre les valeurs singulières associées aux contributions chimiques et les autres (artéfacts, bruit, ….).

La notion de rang local est une notion très importante pour l'implémentation de contraintes efficaces. Le rang local peut être défini comme une extension du concept de sélectivité. Dans certaines fenêtres (il peut s'agir d'une fenêtre temporelle, d'une zone du domaine spectral, *etc.*) ou dans certaines régions (en imagerie) de la matrice **D**, il est possible d'affirmer la présence d'une espèce en l'absence des autres (sélectivité) ou de certaines espèces en l'absence des autres (rang local). Le rang local peut être estimé à partir de la connaissance partielle du problème chimique ou par l'intermédiaire de méthodes d'analyse factorielle spécifiques. On retrouve alors les problèmes évoqués précédemment en ce qui concerne le bruit, la détection d'intermédiaires minoritaires ou la délimitation précise de la région sur laquelle la contrainte devra s'appliquer. C'est pour pallier ces difficultés que les contraintes de rang local sont implémentées comme des contraintes d'inégalité, ce qui signifie que la valeur limite pour la détection n'est pas spécifiée strictement à zéro mais à une valeur supérieure.[17]

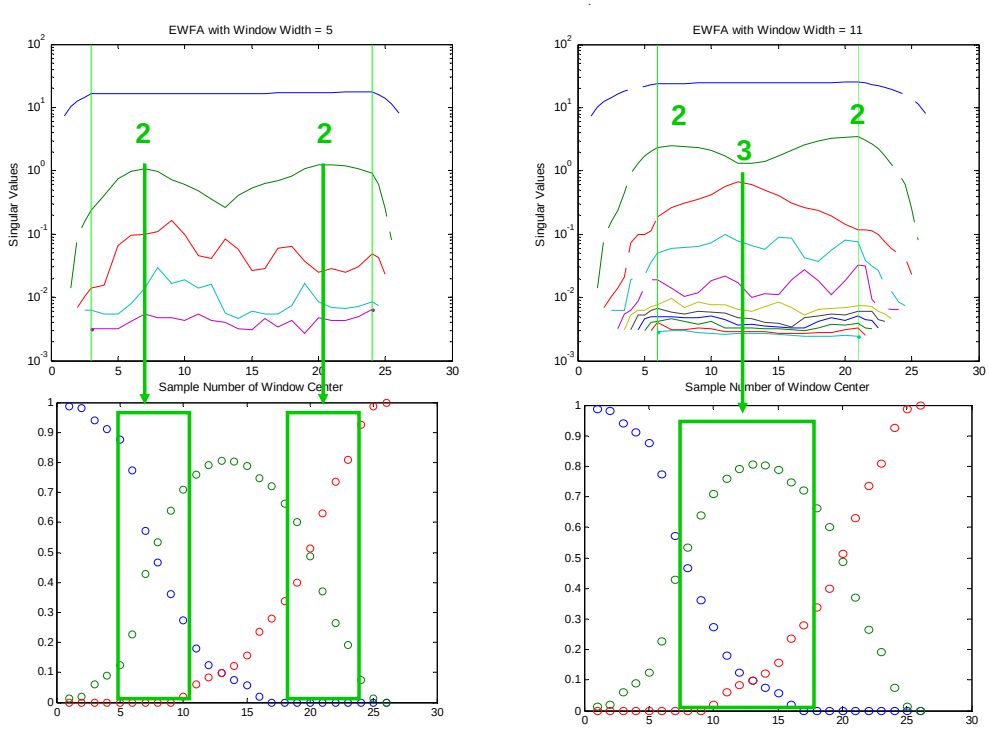


Fig. 2.2 Rang local estimé par Evolving Window Factor Analysis. Profils des valeurs singulières (en haut) pour deux fenêtres de tailles différentes et profils ***C*** *correspondants (en bas), la taille de la fenêtre est matérialisée par un rectangle, la variable évolutive est ici le pH.* [⊥]

La figure précédente (*Fig. 2.2*) propose une illustration du concept de rang local. Par l'analyse des profils de valeurs singulières associées, il est possible d'estimer le nombre d'espèces présentes dans une fenêtre de taille prédéterminée. Il est également intéressant de pouvoir observer l'évolution de cette information lorsque la fenêtre est déplacée dans la direction de la grandeur évolutive.

2.1.2b Déficience de rang

Un système est dit de rang déficient si le pseudo-rang observé est plus faible que le nombre d'espèces absorbantes c'est-à-dire le rang chimique. Cela peut être le cas si les espèces pures présentent des variables, spectres ou profils de concentration, colinéaires ; ou plus ou moins colinéaires pour des rapports signal sur bruit peu favorables.

Mathématiquement, le rang de **D** vérifie la propriété suivante *(2.6)*.

$$Rang\ (\mathbf{D}) \leq min\ (rang\ (\mathbf{C}), rang\ (\mathbf{S}^{\mathbf{T}})) \qquad (2.6)$$

[⊥] Extrait du cours *From factor analysis to multivariate curve resolution* du Master *Advanced Spectroscopy in Chemistry*

Dans le cas de données évolutives, la matrice **C** peut présenter une déficience de rang intrinsèque pour différentes raisons, lorsque certaines espèces chimiques présentent des profils de concentration identiques, si elles sont consommées simultanément *(2.7a)* ou lors du suivi d'un système impliquant des réactions parallèles *(2.7b)*. On peut démontrer que le rang peut s'écrire sous la forme *(2.8)* dans le cas de systèmes fermés[18] avec *R* le nombre de réactions considérées et *k* le nombre d'espèces.

$R_1+R_2 \rightarrow P$, $[R_1]=[R_2]$ *(2.7a)*

$R_1 \rightarrow P_1$, $R_2 \rightarrow P_2$ *(2.7b)*

$$Rang(\mathbf{C}) = min(R+1, k) \qquad (2.8)$$

En pratique, la résolution des profils de concentration et des spectres des espèces pures composant un mélange réactionnel inconnu n'est possible que si les problèmes de déficience de rang sont identifiés et résolus, expérimentalement ou par des stratégies chimiométriques. Ces dernières consistent globalement à réinjecter de l'information chimique, qu'il s'agisse du spectre d'une des espèces (enregistré hors réaction et considéré représentatif de la situation en mélange), de l'analyse simultanée de plusieurs matrices de données expérimentales ou de l'implémentation de modèles cinétiques.

2.1.2c Stratégies de résolution de matrices augmentées

■ L'analyse de systèmes ou de procédés complexes nécessite l'acquisition, dans des conditions expérimentales différentes, de plusieurs lots de données pour décrire la totalité des phénomènes d'intérêt.

L'analyse simultanée de plusieurs matrices de données (formant une matrice dite augmentée) est une stratégie permettant de palier certaines limitations des méthodes MCR, étant entendu que le succès de l'augmentation de matrices est lié au choix des matrices individuelles, qui doivent partager les mêmes contributions. Cela facilite la résolution d'une solution globale commune à l'ensemble des systèmes constituant la matrice augmentée et, dans le même temps, l'obtention d'une information plus robuste, c'est à dire moins soumise aux ambigüités que celle obtenue sur les matrices individuelles.

L'analyse de matrices augmentées peut également permettre de lever la déficience de rang sur un des systèmes constituant l'augmentation, par exemple en analysant conjointement différentes répétitions d'une réaction aux conditions initiales variables. La figure (*Fig. 2.3*) illustre une stratégie d'augmentation de matrices en colonnes (*column-wise matrix augmentation*). Dans cet exemple, trois répétitions d'une expérience réalisées dans des conditions expérimentales différentes ont été analysées simultanément. L'analyse globale de l'ensemble est directe et transparente. Les profils de concentration de trois systèmes peuvent être enregistrés indépendamment mais doivent partager les mêmes espèces (en pratique, au moins certaines d'entre-elles), ce qui signifie que les spectres des espèces pures sont considérés

invariants d'un système à l'autre. Il faut noter que l'augmentation de matrices permet dans ce cas d'assurer un caractère semi-quantitatif, voire quantitatif, aux solutions proposées.

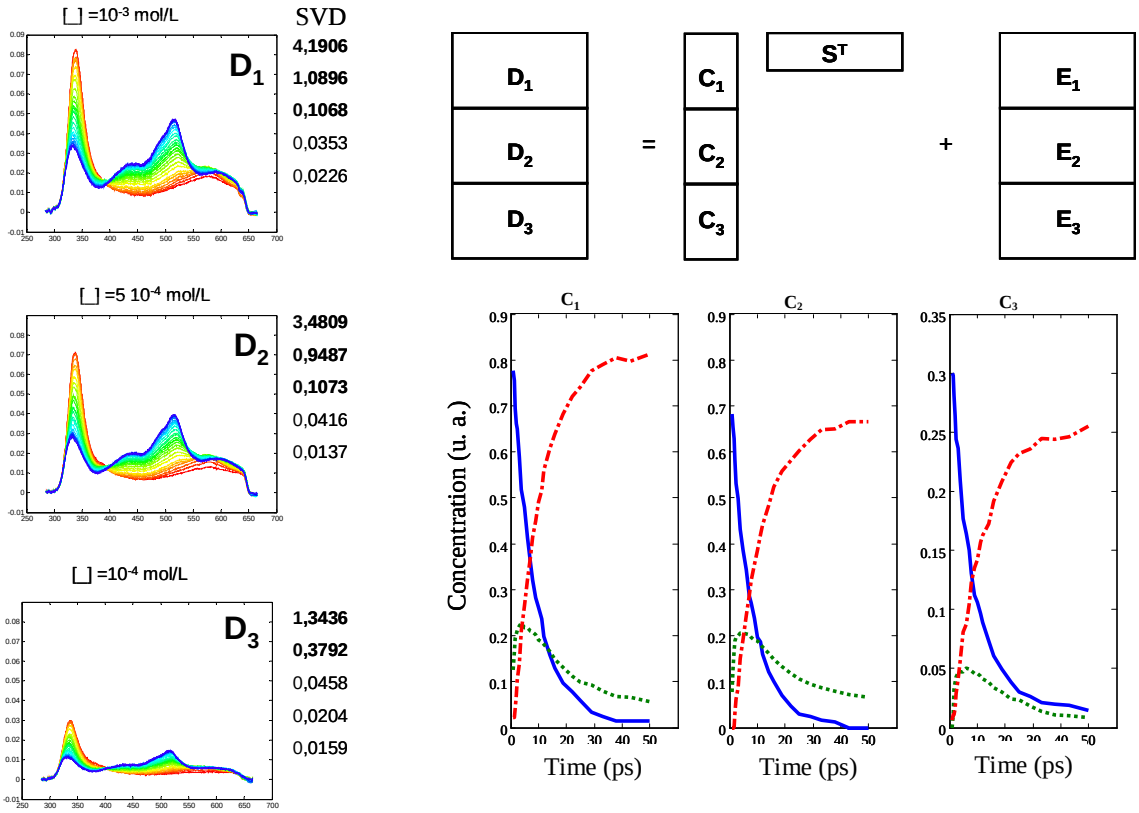


Fig. 2.3 Analyse de la matrice augmentée en colonnes $\boldsymbol{D} = [\boldsymbol{D_1}; \boldsymbol{D_2}; \boldsymbol{D_3}]$. *Les profils cinétiques individuels* $[\boldsymbol{C_1}; \boldsymbol{C_2}; \boldsymbol{C_3}]$ *sont proposés. Les spectres purs* $\boldsymbol{S^T}$ *sont communs aux 3 lots de données.*[19]

Plus généralement, de nombreuses alternatives sont possibles pour la construction de matrices de données, en lien avec les stratégies expérimentales propres à chaque domaine ou à chaque technique analytique. La figure ci-dessous (*Fig. 2.4*) montre deux alternatives différentes pour la fusion de données spectroscopiques résolues en temps.[20]

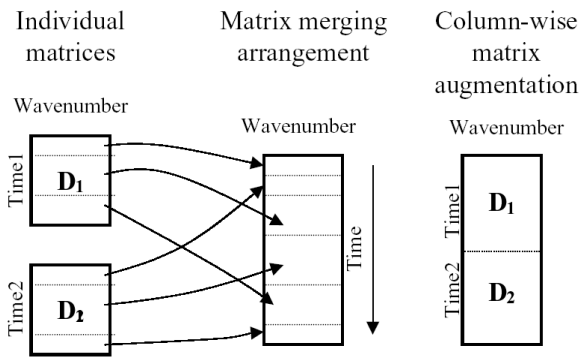


Fig. 2.4 Deux approches différentes pour la fusion de matrices de données spectroscopiques en IRTF step-scan.[20]

2.1.3 Développements

Ce paragraphe a pour premier objectif, de replacer les méthodes de résolution MCR sous contrainte dans un contexte plus académique, celui de l'étude des cinétiques chimiques. Bien que ces méthodes ne nécessitent pas la connaissance explicite d'un modèle physico-chimique, l'optimisation des paramètres descriptifs des modèles réactionnels et le calcul des constantes cinétiques, intégrés dans la procédure ALS, permettent de réduire significativement l'ambigüité des solutions sans se défaire des avantages évoqués jusqu'ici. Ces développements sont effectués dans le cadre d'une collaboration étroite avec le groupe de chimiométrie des équilibres en solution de l'Université de Barcelone.

D'autre part, si on revient à la définition même du problème d'optimisation MCR, des approches conceptuellement différentes peuvent être développées pour la résolution. Elles consistent à implémenter des critères basés sur la pénalisation des moindres carrés. Le problème se ramène alors à un problème d'optimisation sans contrainte. Dans certaines conditions, notamment si les paramètres de régularisation sont fixés, cela peut permettre d'assurer l'unicité de la solution.

2.1.3a Algorithmes hybrides

■ Au-delà des aspects descriptifs, les méthodes de résolution de courbes autorisent l'introduction de contraintes associées aux équations d'un modèle chimique dans la boucle de l'algorithme ALS.

Il y a fondamentalement deux façons différentes d'extraire de l'information d'un ensemble de spectres mesurés : les analyses orientées données type MCR (*model-free* ou *soft-modeling*, le vocabulaire dépendant du domaine d'application) et celles basées sur un modèle physico-chimique quantitatif des données (*model-based* ou *hard-modeling*). Dans le premier cas, l'objectif est essentiellement descriptif, les résultats peuvent être considérés en tant que tel ou servir de base à des analyses ultérieures, pour suggérer un modèle descriptif. Dans le deuxième cas, on cherche à modéliser les données à partir de la connaissance explicite du système d'équations associé au modèle chimique. Les paramètres calculés, les constantes réactionnelles, ont un sens analytique explicite. Ils permettent la prédiction *a priori* du comportement chimique, la comparaison à d'autres systèmes, le calcul des spectres d'absorption, *etc*.

Néanmoins, dans de nombreuses situations, particulièrement en chimie analytique ou en suivi de procédés, on ne dispose pas d'une connaissance exhaustive du système chimique. Il n'est donc pas envisageable d'expliciter *a priori* le système complet par un modèle. C'est le cas lorsque le système étudié est soumis à des grandeurs d'influence difficilement contrôlables ou lorsque les espèces cibles sont analysées en présence d'espèces interférentes par exemple. C'est dans ce contexte, lorsque le système ne peut être décrit que partiellement par des équations

cinétiques, que s'entend l'intérêt d'approches de résolution mixtes, encore appelées approches hybrides (*hard-soft modeling* ou *grey-modeling*).[21,22]

L'idée est d'implémenter, dans la boucle ALS, une contrainte cinétique sur les profils de concentration de la matrice **C**, tout au moins sur certaines colonnes de cette matrice, pour forcer ces profils de concentration à suivre un modèle cinétique particulier. En pratique, les paramètres non linéaires à optimiser pour la minimisation des résidus sont uniquement les constantes cinétiques, le terme d'erreur quadratique *(2.9)* pouvant s'écrire comme fonction linéaire de ces paramètres non-linéaires par les biais des équations du système.[23]

$$\underset{k1,k2,\ldots}{Minimiser} \left\|\mathbf{E}\right\|^2 = \underset{k1,k2,\ldots}{Minimiser} \left\|\mathbf{D}(\mathbf{I} - \mathbf{C}\mathbf{C}^{+})\right\|^2 \qquad (2.9)$$

Le problème d'optimisation s'exprime ainsi comme un problème de moindres carrés non linéaires séparables.[24] Le terme exprimé dans *(2.9)* est appelé la variable projetée de **D**. Les algorithmes classiques de régression non linéaire de type Newton-Gauss peuvent donc être utilisés, avec l'avantage de travailler dans un sous-espace réduit (l'espace colonnes **C**). L'estimation des paramètres linéaires pour le calcul de $\mathbf{S}^{\mathbf{T}}$ est ensuite effectuée à chaque itération en considérant la valeur optimale obtenue précédemment. On notera que l'utilisation d'algorithmes de projection de variables permet de s'assurer que l'on résout un problème numérique bien conditionné (la stabilité numérique est la principale limitation de l'optimisation dans le sous-espace direct pour les problèmes complexes). Seuls les profils de concentration associés au modèle cinétique sont intégrés dans le modèle paramétrique, les autres peuvent être exclus et soumis uniquement aux contraintes implémentées dans la boucle ALS classique, comme l'illustre schématiquement la figure ci-dessous (*Fig. 2.5*).

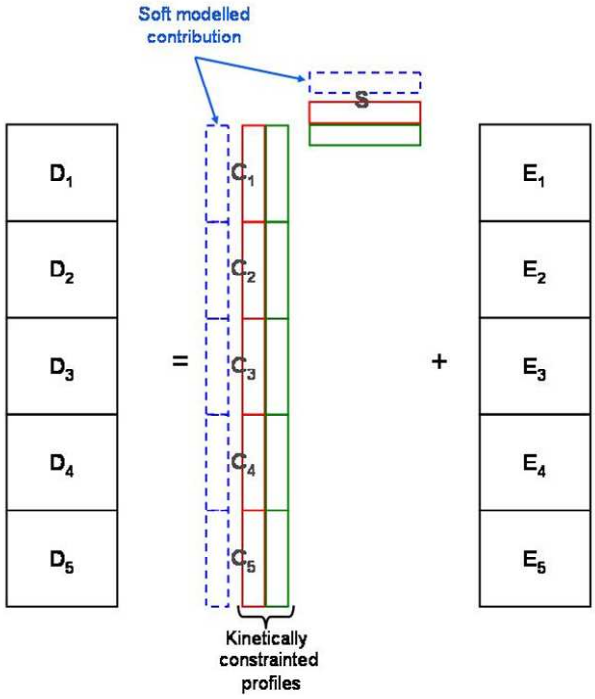


Fig. 2.5 Approche hybride pour la modélisation cinétique dans MCR-ALS.[26]

L'intérêt est donc de conserver les avantages liés à la flexibilité des approches *soft-modeling* tout en rationalisant une partie des résultats et en réduisant de manière significative, voire complètement,[25] les ambigüités d'intensité et de rotation des solutions proposées.

2.1.3b Approches alternatives : un problème de factorisation de matrices non négatives

Le problème d'optimisation MCR peut également être exprimé comme une variante d'un problème de factorisation en matrices non négatives.[27] Il s'agit d'un problème d'optimisation non linéaire pour lequel on définit une fonction de coût $L(\mathbf{C},\mathbf{S}^T)$ qui, d'une certaine manière, intègre les contraintes. La notion de fonction de coût est un concept important qui sera repris au paragraphe 3 concernant les modèles d'apprentissage supervisés. Pour un problème donné, l'idée générale est d'exprimer la distance qui sépare une solution particulière de la solution optimale. Si la fonction de coût s'écrit uniquement comme l'erreur quadratique des résidus, le problème de minimisation se ramène à celui défini précédemment. La plupart des développements consistent donc à affiner cette fonction de coût.

Les différentes approches proposées peuvent être regroupées et la fonction $L(\mathbf{C},\mathbf{S}^T)$ s'écrit alors comme une somme de contributions individuelles du type de l'équation *(2.10)*. A la contribution de l'erreur quadratique peuvent être ajoutés des termes de pénalisation des éléments négatifs des matrices calculées, ainsi que des termes de régularisation permettant de réduire l'ambigüité de rotation des facteurs calculés.[28,29]

$$L(\mathbf{C},\mathbf{S}^T) = \sum_{i,j=1}^{m,n} (e_{ij}/\sigma_{ij})^2 - \alpha \sum_{i,k=1}^{m,p} log\, c_{ik} - \beta \sum_{i,k=1}^{m,p} log\, s_{ik} + \gamma \sum_{i,k=1}^{m,p} a_i c_{ik}^2 + \delta \sum_{k,j=1}^{p,n} b_j s_{kj}^2 \qquad (2.10)$$

Posé de cette manière, le problème d'optimisation linéaire est un problème sans contrainte que l'on peut résoudre par des méthodes classiques. Le premier terme revient à effectuer un calcul de moindres carrés pondérés. Les termes logarithmiques préviennent l'émergence de solutions négatives pour les valeurs de $\mathbf{C}$ et $\mathbf{S}^T$ (les estimations initiales vérifiant les contraintes de non-négativité). Les coefficients α et β permettent de contrôler la force de la pénalisation sur chacun des termes. En pratique, ces termes sont approchés localement par des fonctions quadratiques,[28] ce qui permet l'application de méthodes de type Newton-Gauss. Rien n'empêche donc l'ajout de contraintes supplémentaires dans l'expression de cette fonction de coût, tant que celle-ci prend la forme d'une somme de termes quadratiques.

Les termes de régularisation auxquels sont associés les coefficients γ et δ pénalisent les valeurs trop importantes des éléments des matrices $\mathbf{C}$ et $\mathbf{S}^T$. De manière moins formelle, on dit que les termes de régularisation doivent ajouter de la douceur, c'est à dire sélectionner la plus petite valeur parmi toutes les solutions respectant les autres contraintes. Cela revient à limiter l'ambigüité de rotation des facteurs calculés. Nous proposons d'illustrer ce compromis sur la représentation proposée ci-dessous (*Fig. 2.6*).

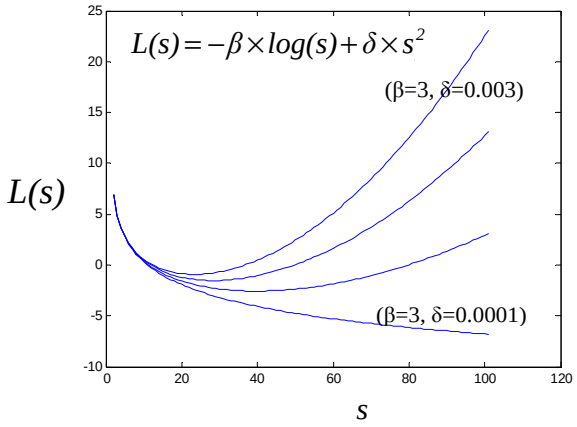


Fig. 2.6 Visualisation de l'effet combiné d'un terme de pénalisation et d'un terme de régularisation.

Enfin, les coefficients a_i et b_i permettent d'éliminer les facteurs d'échelle entre les colonnes de **C** et les lignes de $\mathbf{S}^{\mathbf{T}}$. L'existence de ces coefficients est liée au fait que la méthode de factorisation de matrices positives a été à l'origine développée pour le traitement de données environnementales, données types pour lesquelles les valeurs des variables ou des individus peuvent être mesurées sur des échelles correspondant à des ordres de grandeur très différents. Ces coefficients peuvent être omis dans le cas de l'analyse de données spectroscopiques. Le problème se ramène donc à un problème d'optimisation sans contrainte qui produit une solution unique pour des coefficients de pénalisation et de régularisation fixés (le problème est en quelque sorte déplacé sur le choix des coefficients).

Tauler[30] a proposé très récemment une approche MCR basée sur la minimisation d'une fonction de coût du type de celle proposée équation *(2.11)*. Celle-ci représente une somme de fonctions scalaires associées aux différentes contraintes appliquées (*norm* pour la normalisation, *nonneg* pour non-négativité, *unimod* pour unimodalité), plus exactement aux déviations de la solution calculée par rapport aux valeurs imposées par ces contraintes. Les différentes fonctions scalaires implémentées prennent des valeurs importantes lorsque les contraintes ne sont pas respectées, et inversement. La fonction f_{nonneg} correspondant à la non-négativité des spectres et des profils de concentration est reprise en *(2.12)* à titre d'exemple. On peut remarquer que cette fonction $f(\mathbf{U})$, où **U** est la matrice de rotation exprimée précédemment équation *(2.5)*, n'est pas une fonction de coût quadratique de l'erreur puisque la rotation est par définition effectuée dans le sous-espace des solutions.

$$f(\mathbf{U}) = f_{norm}(\mathbf{U}) + f_{nonneg}(\mathbf{U}) + f_{unimod}(\mathbf{U}) + \ldots \qquad (2.11)$$

$$f_{nonneg}(\mathbf{U}) = \sum_i (c_i < 0)^2 + \sum_j (s_i < 0)^2 \qquad (2.12)$$

L'objectif est en fait de trouver la matrice de rotation **U** qui minimise $f(\mathbf{U})$ sans modifier l'erreur d'optimisation, c'est-à-dire en considérant que l'on reste dans le sous-espace de l'ACP. Toutes les solutions obtenues modélisent de manière équivalente les données puisque leurs

rotations ne modifient pas l'erreur quadratique, $f(\mathbf{U})$ traduisant uniquement le respect des contraintes. Plusieurs questions restent posées au regard de l'interprétation faites de la fonction de coût *(2.11)*, comme la possibilité d'affecter les différents termes de cette fonction de coefficients de régularisation permettant de favoriser un type de contraintes.

2.2 Applications en spectroscopie IRTF rapide

La spectroscopie vibrationnelle permet d'obtenir des informations concernant la structure des molécules complexes, typiquement les protéines ou les systèmes biologiques. En particulier, la spectroscopie différentielle est utilisée pour suivre de faibles variations d'absorbance de l'échantillon en présence de larges fonds d'absorption dus à l'environnement. Lorsque les spectres sont échantillonnés au cours du temps, la spectroscopie infrarouge à transformée de Fourier (IRTF) autorise le suivi de systèmes chimiques complexes évoluant rapidement. L'analyse multivariée de ces données spectroscopiques résolues en temps, assemblées en matrices *2-voies*, permet alors de déduire des informations sur le nombre d'espèces intermédiaires observables ainsi que sur leurs caractéristiques spectrales et cinétiques.

Les premiers travaux effectués au laboratoire ont concerné l'étude du photocycle de la bactériorhodopsine comme système modèle pour la validation d'un couplage entre excitation laser et spectromètre IRTF *step-scan*. Ces travaux ont révélé les problèmes liés à l'analyse des spectres différentiels par les approches chimiométriques et à l'interprétation des profils cinétiques et des spectres résolus. Des solutions méthodologiques et algorithmiques ont été proposées pour pallier la déficience de rang des données spectroscopiques différentielles. Ces méthodes ont été adaptées et appliquées ensuite à l'analyse d'un centre réactionnel photosynthétique en spectroscopie IRTF *rapid-scan*.

2.2.1 Validation d'une instrumentation IRTF *step-scan*

La spectroscopie IRTF en mode *step-scan*[31] est une technique spectroscopique résolue en temps (~10 ns) qui combine le déclenchement d'une réaction (photo)chimique avec la détection différentielle rapide dans l'infrarouge. Le principe de la mesure repose sur l'enregistrement pas à pas, c'est-à-dire pour chaque différence de marche, des signaux de différence associés aux retards successifs par rapport au déclenchement de la réaction.[⊥] Les interférogrammes complets ne sont reconstitués qu'une fois tous les pas de mesures échantillonnés. La résolution temporelle n'est donc pas limitée par la vitesse de déplacement du miroir mobile, comme en mode de fonctionnement *rapid-scan*, mais par la vitesse de déclenchement de la réaction chimique et par le détecteur IR et son électronique. Néanmoins, ce type de mesure séquentielle restreint le domaine d'application à l'analyse de systèmes réversibles, sauf lorsque l'échantillon peut être renouvelé par circulation. Parmi les systèmes photochimiques utilisables pour la validation du couplage de l'excitation laser et du spectromètre IRTF *step-scan*, notre choix s'est porté sur la bactériorhodopsine.

[⊥] Le déclenchement rapide et contrôlé de la réaction peut se faire par voie optique, synchronisant un laser Nd :YAG pulsé à l'interféromètre du spectromètre IRTF.

1-

Le photocycle de la bactériorhodopsine a été très largement étudié. Ce processus modèle de pompes à protons photoinduite implique plusieurs intermédiaires métastables successifs possédant des durées de vie s'étendant de la picoseconde à la milliseconde. Les mécanismes descriptifs sont connus mais ne sont pas encore complètement élucidés au niveau moléculaire, en particulier en ce qui concerne la transition entre les intermédiaires L et M.

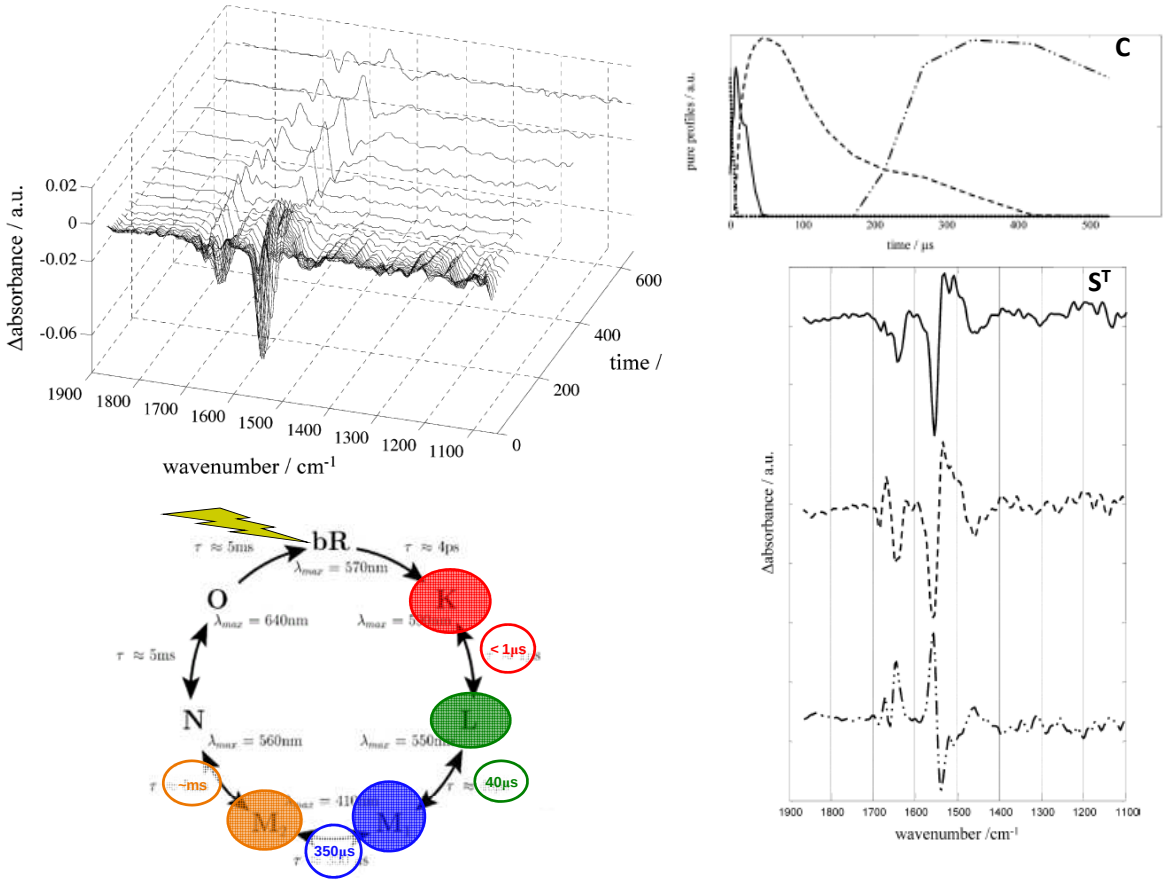


Fig. 2.7 Présentation schématique des spectres en résolution temporelle et des résultats MCR-ALS obtenus sur le photocycle de la bactériorhodopsine en IRTF step-scan.[32]

Les résultats obtenus sont repris globalement ci-dessus (*Fig. 2.7*). Les spectres observés dans la gamme temporelle 1-500 µs sont représentés. Les contributions observées dans les régions Amide I et Amide II sont associées majoritairement à des changements structuraux de la protéine et de son environnement. La bande de vibration d'élongation C=C observée à ~1550 cm^{-1} ($\Delta A<0$) est généralement choisie pour suivre l'évolution du photocycle. Les spectres $\mathbf{S^T}$ et profils cinétiques **C** correspondant aux photo- intermédiaires L, M_1 et M_2 ont été obtenus par analyse MCR, sans *a priori* sur le mécanisme réactionnel. Ces résultats ont pu être confirmés en analyse sur des matrices augmentées sur différentes gammes temporelles et ont été partiellement interprétés, au vu de la connaissance du cycle de la protéine.[33]

2.2.2 Chimiométrie des spectres différentiels

■ La structure des matrices de données différentielles requiert une approche spécifique pour le traitement et l'interprétation des données.

D'un point de vue chimiométrique, les matrices de spectres différentiels présentent une déficience de rang de premier ordre. En effet, chaque spectre de différence est combinaison linéaire d'au moins deux contributions, celle enregistrée au temps considéré et celle qui correspondrait au signal au temps initial. La nature intrinsèque de cette déficience de rang ne permet pas de profiter directement d'approches de type augmentation de matrices, pourtant performantes sur les données spectroscopiques classiques. De plus, les problèmes de déficience de rang rencontrés plus couramment lors du suivi de réaction ou de procédés (voir paragraphe *2.1.2b*) doivent également être considérés.

Une partie du travail de la thèse de Lionel Blanchet, effectuée en cotutelle avec l'Université de Barcelone, a consisté à démontrer l'existence d'une déficience de rang systématique pour les données de différence, que le système soit à l'origine de rang plein ou non. Cela confirme l'intuition que les déficiences de rang s'ajoutent. Nous nous sommes ensuite focalisés sur l'adaptation des approches hybrides intégrant des contraintes cinétiques dans la résolution ALS pour contourner la déficience de rang et permettre la résolution des matrices de données différentielles. L'attention à également été portée sur l'interprétation des données différentielles.

2-

L'équation *(2.8)* pour le calcul du rang peut être généralisée pour permettre une estimation rigoureuse du nombre de contributions modélisables lors du suivi de réactions en spectroscopie différentielle. Dans ce cas, le caractère systématique de la déficience de rang a pu être démontré.[34] Le schéma proposé ci-dessous (*Fig. 2.8*) reprend les étapes principales des calculs. Nous avons modélisé deux situations initiales, partant de matrices de données simulées de rang plein (système A→B→C) ou de matrices de données présentant initialement une déficience de rang d'ordre 1 (système $A_1 \rightarrow A_2$ et $B_1 \rightarrow B_2$). Les systèmes sont considérés comme des systèmes fermés ce que traduit mathématiquement la propriété de conservation exprimée sur les concentrations. L'écriture matricielle **ΔD** du système en différence est ensuite proposée, pour les deux cas envisagés ici.

Le premier résultat est la confirmation de la perte de rang puisque la matrice **ΔD** s'écrit sous la forme d'une somme de deux contributions uniquement. Le deuxième résultat démontre qu'en spectroscopie différentielle, les lois cinétiques sont les mêmes qu'en spectroscopie directe puisque les profils de concentration prennent la même forme. De plus, dans le cas de mécanismes réactionnels simultanés, la description des différentes cinétiques individuelles peut se faire sur des colonnes différentes de la matrice **C**, permettant l'implémentation de contraintes cinétiques indépendantes ce qui s'avère très important pour les applications.

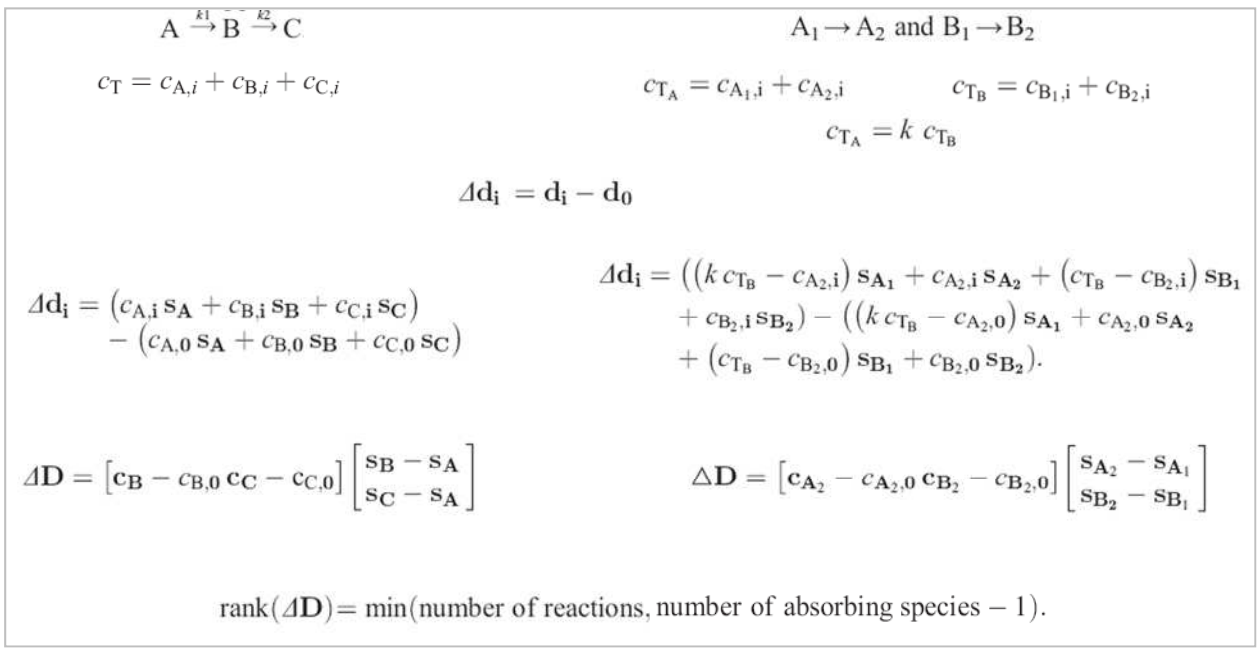


Fig. 2.8 Principales étapes de la démonstration de l'équation de calcul du rang pour des données spectroscopiques de différence, pour un système de rang plein (colonne de gauche) et pour un système en déficience de rang (colonne de droite).[34]

2.2.3 Etude d'un centre réactionnel photosynthétique par spectroscopie IRTF *rapid-scan*

Les centres réactionnels photosynthétiques présentent un comportement modèle de certaines réactions bioénergétiques déclenchées et contrôlées par des impulsions lumineuses. Ils sont donc particulièrement étudiés, notamment en spectroscopie IRTF *rapid-scan* pour l'étude des mécanismes réactionnels. Même si la résolution temporelle accessible en *rapid-scan* (~100 ms) ne permet pas en général de couvrir l'ensemble des phénomènes d'intérêt, cette technique conserve l'avantage d'être applicable sur des systèmes non réversibles.

La spectroscopie IRTF *rapid-scan* a été appliquée au centre réactionnel *Rhodobacter sphaeroides* pour suivre au cours du temps les variations d'absorbance induites par l'éclairement.[35] En particulier, il est possible d'observer la formation de molécules d'ubiquinol qui est l'étape limitante de la réaction globale. Notre contribution a consisté à développer une approche de résolution multivariée des spectres IRTF de différence pour la caractérisation moléculaire des espèces impliquées dans le processus biochimiques et l'établissement d'un modèle cinétique. L'originalité est que le modèle semi-paramétrique développé permet de gérer les changements de conformation protéiques du milieu biologique environnant. L'approche proposée repose également sur l'analyse simultanée de différents lots de données, couvrant la même gamme temporelle mais acquis dans des conditions expérimentales différentes. En

particulier, l'influence des conditions d'éclairement sur les processus cinétiques a pu être évaluée.

3-

Nous proposons la résolution MCR de spectres IRTF différentiels enregistrés en mode *rapid-scan* pour la caractérisation des processus chimiques photoinduits dans les membranes photosynthétiques de la bactérie pourpre *Rhodobacter Sphaeroides*. Les réactions induites sous éclairement sont reprises schématiquement ci-dessous (*Fig. 2.9*). L'objectif est plus précisément l'étude du mécanisme de la réduction de l'ubiquinone (Q) et la formation de l'ubiquinol (QH_2).

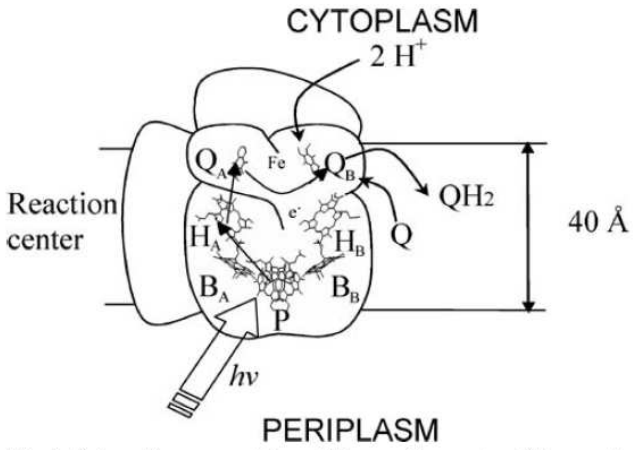


Fig. 1 Schematic representation of the reaction centre of the purple bacterium *Rhodobacter sphaeroides* (*H*, bacteriopheophytin; *P*, primary donor; *B*, bacteriochlorophyll; *Q*, ubiquinone; QH_2, ubiquinol).

Fig. 2.9 Représentation du centre réactionnel Rhodobacter sphaeroides.[36]

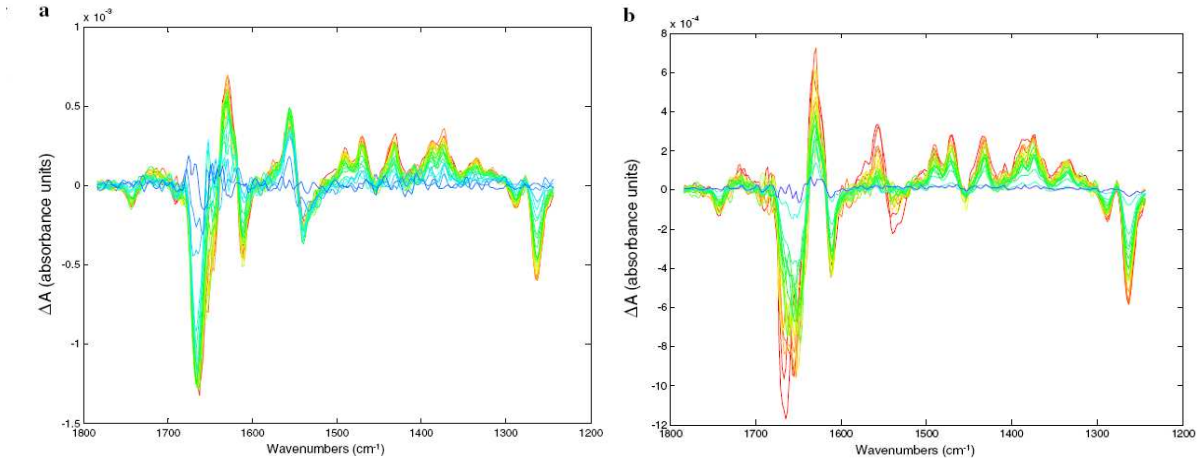


Fig. 2.10 Spectres IRTF rapid-scan du centre réactionnel enregistrés sous éclairement durant 4,3 s (a) puis dans l'obscurité pendant 55 s (b).[36]

La réaction photochimique est suivie sous éclairement continu puis la relaxation de l'échantillon est enregistrée dans l'obscurité. Différentes expériences ont été répétées en faisant varier principalement l'intensité lumineuse mais également les conditions d'acquisition des spectres. Les spectres enregistrés sont repris sur la figure *Fig. 2.10*. Les évolutions spectrales concernent les bandes amides des protéines (Amide I, $\Delta A<0$ à 1666 cm^{-1} et Amide II, $\Delta A>0$ à 1556 cm^{-1}) et la formation de l'ubiquinol QH_2 ($\Delta A>0$ dans l'intervalle 1500-1300 cm^{-1}). Une bande très caractéristique du problème est également observée à 1264 cm^{-1} ($\Delta A<0$), souvent considérée comme indicatrice de la consommation d'ubiquinone (Q) lors de la réaction.

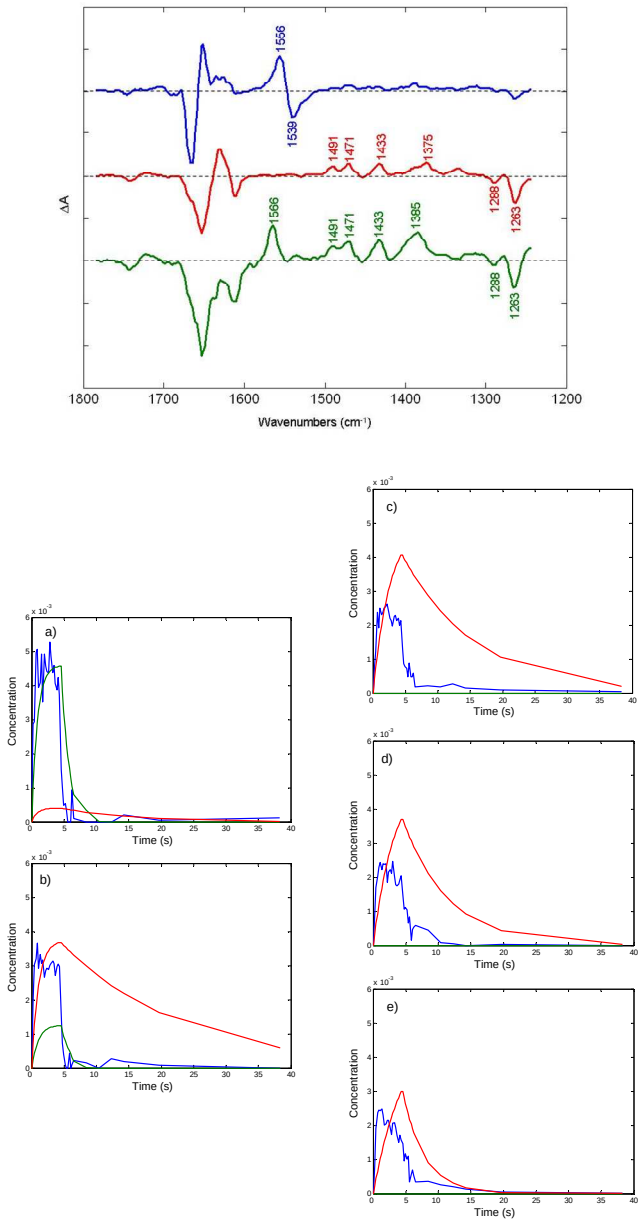


*Fig. 2.11 Résolution MCR-ALS sur matrice augmentée en colonnes [**D_1**; **D_2** ; **D_3**; **D_4**; **D_5**]. Spectres de différence **S^T** extraits et profils cinétiques [**C_1** - **C_5**] correspondants (a-e). Contributions de l'environnement protéique (en bleu), de la photo-réduction de l'ubiquinone (en rouge) et de la photo-réduction de l'ubiquinone sous éclairement intense (en vert).*[26]

La figure proposée (*Fig 2.11*) regroupe les résultats obtenus par l'analyse MCR-ALS hybride[62] en augmentation de matrices, sur 5 lots de données, avec modélisation cinétique de la photo-réduction de l'ubiquinone (l'organisation des données en matrice augmentée en colonnes est celle présentée précédemment, *Fig 2.5*). La première contribution correspond globalement à des modifications structurales des protéines. La deuxième contribution spectrale est caractéristique de la photo-réduction de Q ($\Delta A<0$ à 1263 cm^{-1}) et de la formation de QH_2 ($\Delta A >0$ à 1491 cm^{-1}, 1471 cm^{-1}, …). Cette composante à fait l'objet de l'implémentation d'une contrainte cinétique de premier ordre pour la résolution ($Q \rightarrow QH_2$ sous éclairement ; $QH_2 \rightarrow Q$ lors de la relaxation). Par ailleurs, l'étude a permis de mettre en évidence une perte d'efficacité du processus photochimique de production de QH_2. Cet« effet de fatigue » est observé lorsque l'expérience est répétée sur un même échantillon dans des conditions expérimentales parfaitement identiques (matrices $\mathbf{D_3}$-$\mathbf{D_5}$ et profils *Fig. 2.11c-e*).

La troisième contribution n'a été observée que dans des conditions d'éclairement intense (matrices $\mathbf{D_1}$ et $\mathbf{D_2}$ et profils *Fig 2.11a-b*). Cette espèce a pu être modélisée par la même contrainte cinétique et semblerait traduire un processus alternatif de formation de l'ubiquinol,[26] observable uniquement sous éclairement fort (une alternative serait l'oxydation d'un médiateur redox utilisé[36]). L'interprétation des spectres extraits semble confirmer cette hypothèse. En effet, les bandes caractéristiques de l'ubiquinol peuvent être observées (avec des intensités relatives inversées) et, contrairement à ce qui a été observé pour la première contribution, ces bandes sont associées à des changements de conformation de l'environnement protéique ($\Delta A >0$ à 1566 cm^{-1}). Cette interprétation signifierait que l'échange de quinone se fait préférentiellement avec le milieu sous éclairement important. Cela reste à confirmer notamment par l'étude de centres réactionnels isolés.

2.3 Application en spectroscopie d'absorption résolue en temps

La spectroscopie d'absorption transitoire en régime femtoseconde est une technique expérimentale de pointe pour l'observation des états électroniques excités et des intermédiaires à très courte durée de vie mis en jeu dans les réactions photochimiques.[⊥] En effet, l'utilisation de lasers pulsés permet de déclencher les réactions et de suivre leurs dynamiques à des échelles de temps subpicoseconde. L'observation des processus réactionnels photoinduits repose donc en premier lieu sur des instrumentations performantes. Mais pour tendre vers une connaissance plus approfondie des mécanismes réactionnels de systèmes complexes, la résolution ou la modélisation des matrices de spectres reste souvent nécessaire.

Nous ferons dans ce chapitre la distinction entre les traitements de données applicables aux spectres enregistrés aux temps courts (de l'ordre de la picoseconde et au-delà), où se situe notre travail à ce jour, et ceux envisageables aux temps ultracourts (inférieurs à la picoseconde) qui concernent nos perspectives de travail. Se limitant aux temps courts, les matrices de données obtenues en spectroscopie d'absorption transitoire peuvent s'exprimer directement comme produits de matrices cinétiques et de matrices de spectres. Nous reviendrons sur les résultats obtenus lors de l'étude de la photophysique de la benzophénone. Pour cette molécule modèle, une controverse subsiste concernant la description exacte des mécanismes impliqués dans la relaxation ultrarapide de l'état singulet photoexcité, $S_1(n,\pi^*)$, vers l'état triplet, $T_1(n,\pi^*)$.

2.3.1 Chimiométrie des données spectroscopiques aux temps courts

■ Aux temps courts, la structure des données ne s'écarte pas de la décomposition bilinéaire directe, au moins en première approximation.

Les mesures résolues en temps peuvent être effectuées dans différentes configurations expérimentales. Les données peuvent être obtenues simultanément à toutes les longueurs d'onde pour un retard donné par rapport à l'impulsion d'excitation (mode multicanal). Elles peuvent également être acquises en fonction du retard par rapport à l'impulsion à une longueur d'onde déterminée (mode monocanal). Les difficultés liées aux rapports signal sur bruit, à la répétabilité des mesures ou à la nature stochastique des données sont propres à chaque mode de mesure et à chaque type d'expérience spectro-cinétique.

Les traitements des données spectroscopiques résolues en temps sont assez peu décrits dans la littérature.[37-39] La plupart des approches proposées reposent sur l'implémentation d'un modèle physico-chimique paramétrique (appelé modèle cible, *target analysis*). Il peut s'agir d'un modèle construit sur la base des équations cinétiques ou d'un modèle spectral pour lequel on

[⊥] L'ensemble des processus physico-chimiques produits par action de la lumière UV-visible

explicite les paramètres propres des contributions individuelles permettant de reconstruire les spectres, en fluorescence résolue en temps par exemple. Comme nous l'avons exprimé précédemment, l'approche chimiométrique se focalise plutôt sur la structure algébrique intrinsèque des données spectro-cinétiques, sans nécessairement formuler de modèle physique ou photochimique des données. L'intérêt supplémentaire des méthodes hybrides que nous proposons réside dans la possibilité de coupler la résolution sans *a priori* à l'ajustement des paramètres d'un modèle cinétique.

Aux temps courts, nous considérons que les matrices de données *2-voies* enregistrées en spectroscopie d'absorption en régime femtoseconde ne s'écartent pas, au moins en première approximation, de la décomposition bilinéaire directe proposée dans l'équation *(2.1).* Les hypothèses faites concernent alors d'une part la superposition des propriétés individuelles des différentes composantes du système étudié et d'autre part la séparation des propriétés cinétiques C(t) et spectroscopiques S(λ), au bruit près.

2.3.2 Etude de la photophysique de la benzophénone

La benzophénone est une molécule modèle pour l'étude des processus de relaxation des états excités $S_1(n,\pi^*)$ de cétones aromatiques. La spectroscopie d'absorption transitoire a été utilisée pour suivre précisément la désexcitation de l'état singulet $S_1(n,\pi^*)$ de la benzophénone à l'échelle picoseconde. Notre contribution a consisté à développer une approche de résolution MCR des matrices de données spectro-temporelles.

Les principales difficultés rencontrées sont liées au domaine spectral considéré. D'une part, les spectres UV-visible des différentes espèces, identifiées ou non, sont des spectres possédant de larges bandes d'absorption, peu spécifiques et par nature très corrélés. D'autre part, les facteurs physico-chimiques tels que la polarité du solvant ou la longueur d'onde d'excitation affectent à la fois les profils cinétiques et les spectres des espèces pures, ce qui offre peu de possibilités d'analyses MCR en augmentation de matrices. La question de la robustesse des solutions proposées vis-à-vis des ambigüités de rotation doit donc être traitée, notamment par l'application de méthodes de résolution hybrides avec contraintes cinétiques.

4-

De nombreux travaux se sont intéressés au processus de croisement intersystème (ISC) de la benzophénone. La plupart de ces travaux rapportent un phénomène très rapide pour la transition $S_1(n,\pi^*) \rightarrow T_1(n,\pi^*)$. Cette observation est néanmoins en contradiction avec les règles dites d'El-Sayed qui suggèrent l'implication d'intermédiaires de configuration (π,π^*).

Les spectres d'absorption transitoire de la benzophénone dans l'acétonitrile sont présentés (*Fig. 2.12*) dans la gamme temporelle 0,8-50 ps. Les spectres enregistrés pour des retards compris entre 0,8 et1,7 ps (voir encart *Fig. 2.12*) montrent la formation ultrarapide de l'état singulet $S_1(n,\pi^*)$ suite à la disparition de l'état singulet $S_2(n,\pi^*)$ peuplé initialement. Pour

l'étude picoseconde présentée ici, seuls les spectres enregistrés au delà de 2 ps ont été considérés lors de l'analyse. Le spectre initial est donc caractéristique de l'état $S_1(n,\pi^*)$.

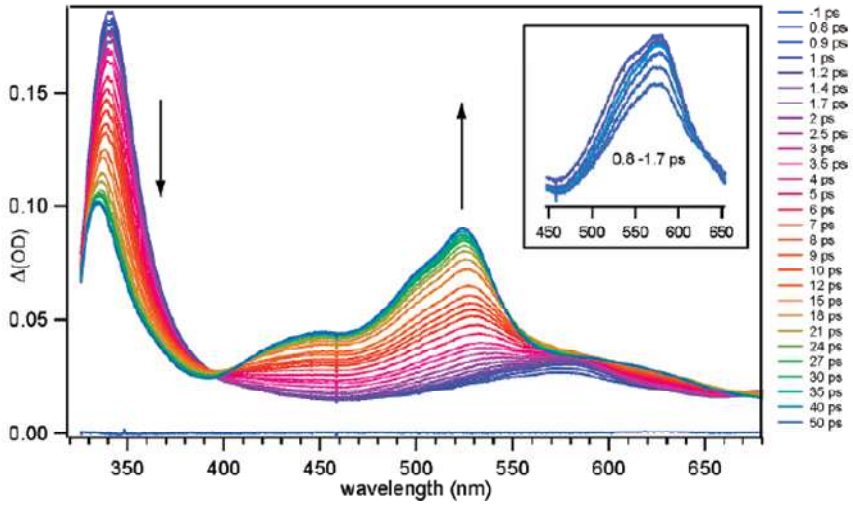


Fig. 2.12 Spectres d'absorption transitoire femtoseconde de la benzophénone dans l'acétonitrile à la longueur d'onde d'excitation 267 nm dans la gamme temporelle 0,8-50 ps.[40]

Une analyse *soft-modeling* des données spectroscopiques résolues en temps dans l'intervalle 2-50 ps a d'abord été effectuée. Les estimations du rang de la matrice des données, du rang local dans la dimension temporelle, et l'analyse des résidus[41] obtenus lors des différentes modélisations MCR (*Fig. 2.13*) ont permis de conclure à l'implication de trois contributions significatives pour la description des données expérimentales.

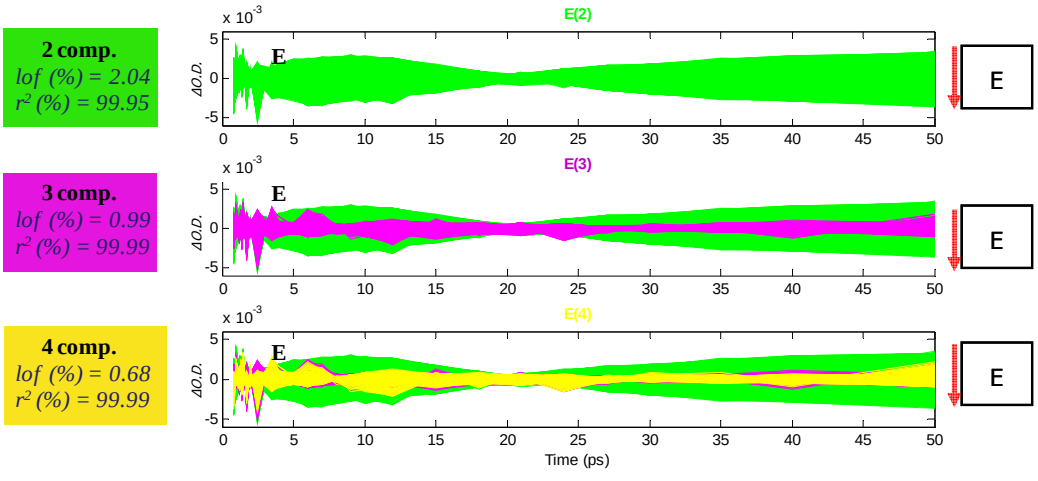


Fig. 2.13 Résidus cumulés en fonction du temps pour des résolutions MCR à 2, 3 et 4 contributions.[19]

Les résultats obtenus finalement sont présentés ci-dessous (*Fig. 2.14*). Pour tenter d'interpréter les solutions du point de vue de la photophysique, une résolution MCR hybride

intégrant les contraintes descriptives d'un modèle cinétique du premier ordre $S_1(n,\pi^*) \rightarrow IS \rightarrow T_1(n,\pi^*)$ considérant un état triplet intermédiaire IS a été proposée. Ce modèle est le plus robuste (le plus simple) au sens où il intègre le nombre minimum de contributions pour décrire complètement l'évolution des données expérimentales. Les résultats se sont avérés en bon accord avec ceux obtenus lors de l'analyse effectuée sans imposer de contrainte cinétique. D'un point de vue plus fondamental, la nature exacte de l'interaction de cet état triplet intermédiaire IS avec l'état singulet $S_1(n,\pi^*)$ et l'état triplet $T_1(n,\pi^*)$ reste néanmoins à déterminer. Dans cet objectif, les résultats ont également été interprétés par analogie avec ceux obtenus pour la 4-méthoxybenzophénone, molécule pour laquelle la possibilité d'une inversion des états excités triplets $T_1(n,\pi^*)$-$T_2(\pi,\pi^*)$ en fonction de la polarité du solvant est avérée.[40]

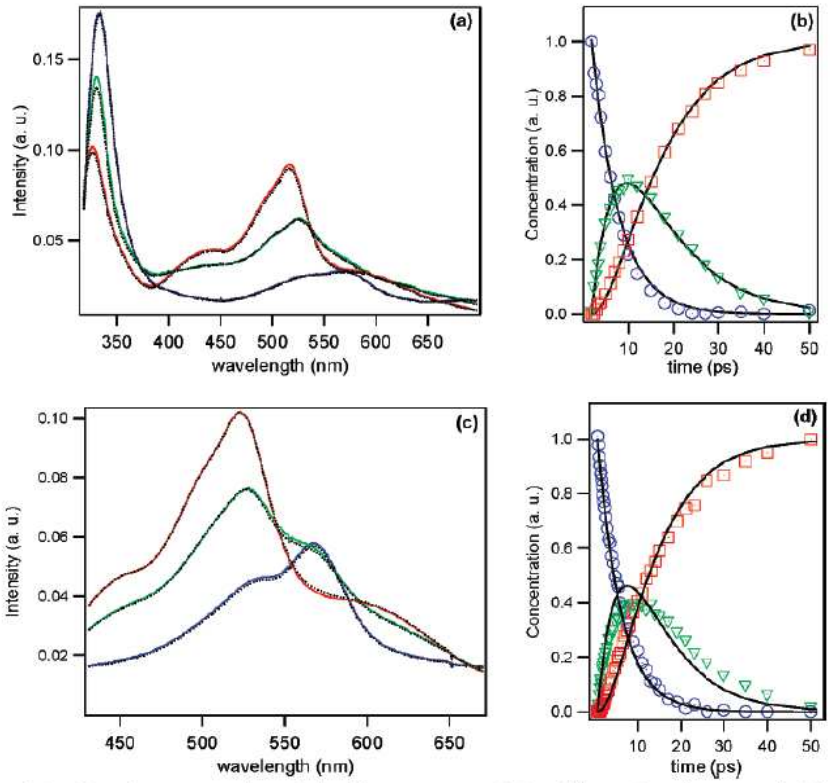


Figure 4. Spectra and time-dependent concentrations obtained from MCR-ALS (colored lines and markers, respectively) and from HS-MCR (black lines in both cases) analysis of subpicosecond time-resolved absorption data of BP excited at 267 nm, (a) and (b); and 383 nm, (c) and (d) in ACN. The color code is the following: blue for S_1; green for IS; red for T_1.

Fig. 2.14 Spectres et profils de concentration résolus par MCR-ALS et approche MCR-ALS hybride (k_1~0,15 ps^{-1} et k_2~0,09ps^{-1})[40]

La résolution chimiométrique de ces données a donc permis de révéler l'implication d'un intermédiaire dans la transition $S_1(n,\pi^*) \rightarrow T_1(n,\pi^*)$ attribué à l'état triplet $T_2(\pi,\pi^*)$ et de proposer pour la première fois un processus de relaxation en deux étapes en accord avec la prédiction des règles d'El-Sayed.

2.4 Perspectives aux temps ultracourts

Aux temps inférieurs à la picoseconde, la structure des matrices de données *2-voies* générées par les techniques spectroscopiques en régime femtoseconde peut s'écarter de la décomposition bilinéaire. En effet, la résolution temporelle des mesures devient limitée par des caractéristiques instrumentales, notamment par la largeur des impulsions laser. Les phénomènes perturbateurs pour les montages de type pompe-sonde peuvent également être liés à la dispersion des différentes composantes spectrales ou à l'existence de structures propres de corrélations de l'erreur de mesure.

Dans ce paragraphe, nous pointerons les adaptations et les développements à envisager pour l'analyse des données spectro-cinétiques aux temps ultracourts. L'introduction de la fonction de réponse instrumentale dans les modèles cinétiques hybrides, ou semi-paramétriques, et la gestion des structures de corrélation seront abordées. D'autres approches s'appuyant sur des collaborations à développer sont également dans nos perspectives de travail.

2.4.1 Structure des données spectroscopiques aux temps ultracourts

■ Aux temps ultracourts, la question de la structure des données spectro-cinétiques doit être posée.

La fonction de réponse instrumentale (IRF, *instrumental response function*) limite la résolution temporelle des mesures. Plus précisément, la largeur temporelle de l'IRF détermine le plus court délai observable dans des conditions expérimentales déterminées. Idéalement, cette fonction est connue et peut être prise en compte pour l'analyse des phénomènes observés aux délais sub-picoseconde. L'IRF est obtenue en effectuant la convolution des fonctions décrivant la forme de l'impulsion d'excitation et de l'impulsion d'analyse. En pratique, cette fonction est donc généralement estimée par une fonction gaussienne du temps (typiquement quelques centaines de femtosecondes) dont on ajuste les paramètres liés à la position et à la largeur à mi-hauteur. Aux temps ultracourts, une complication supplémentaire des mesures en spectroscopie dispersive concerne la dépendance de la réponse instrumentale avec la longueur d'onde d'analyse du fait de phénomènes de dispersion de vitesse de groupe au cours de la traversée des différents milieux optiques (GVD, *group velocity dispersion*).[⊥]

La fonction de réponse instrumentale est donc en toute rigueur fonction du retard et de la longueur d'onde considérés. La question de la séparabilité des matrices de données en deux matrices indépendantes, $\mathbf{C}$ d'une part et $\mathbf{S}^{\mathbf{T}}$ d'autre part, doit donc être posée. En première approche, les corrections peuvent être considérées indépendamment l'une de l'autre. La

[⊥] Lors de la traversée d'un milieu optique (à dispersion positive), la vitesse de propagation de la lumière est plus élevée aux courtes longueurs d'onde

dispersion liée à la GVD est souvent estimée sur une mesure de référence et une correction est ensuite appliquée. La contribution temporelle peut, quant à elle, être prise en compte de manière plus explicite, directement dans le modèle cinétique. Le signal temporel observé à une longueur d'onde donnée est alors écrit comme le produit de convolution du signal vrai, C(t), par la fonction d'appareil, IRF(t).

Jusqu'ici, la largeur temporelle de la fonction d'appareil a été négligée dans les modèles cinétiques considérés (cela revient à exprimer un produit de convolution des profils de concentration par des pics de Dirac). De notre point de vue, la structure des modèles MCR hybrides avec contraintes cinétiques doit permettre l'implémentation de ce type de modification, le produit de convolution étant effectué sur les colonnes de la matrice **C**. Une autre possibilité consiste à remettre en cause la bilinéarité de la décomposition par l'implémentation de modèles paramétriques dits spectro-cinétiques, impliquant notamment la considération d'un paramètre dépendant de la longueur d'onde dans l'écriture des fonctions temporelles.[39] L'analyse bayésienne des signaux que nous aborderons au paragraphe 2.4.3 est également une alternative. Dans les deux cas, la description mathématique complète des phénomènes est requise.

5-

L'effet de la correction de la dispersion de vitesse de groupe sur les données spectroscopiques aux temps ultracourts est représenté ci-dessous (*Fig. 2.15*). Les phénomènes de dispersion de l'impulsion peuvent être approchés par une fonction polynomiale de la longueur d'onde et pris en considération par le biais d'un prétraitement mathématique des données.

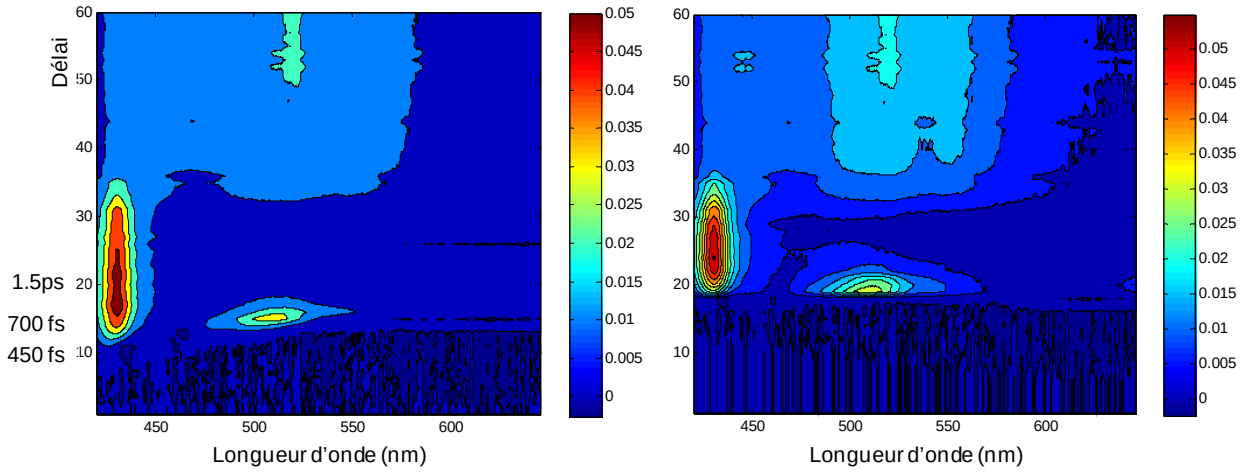


Fig. 2.15 Spectres d'absorption transitoire en régime pompe-sonde (absorbance en fonction du délai pompe-sonde et de la longueur d'onde) : spectres bruts (à gauche) et spectres corrigés (à droite). Sur ces représentations, le temps t= 0 est positionné arbitrairement.

2.4.2 Structure de corrélation des incertitudes de mesure

■ Les structures de corrélation des erreurs de mesure peuvent être estimées et éventuellement prises en compte dans l'analyse des données.

En spectroscopie optique, les erreurs de mesure sont reliées aux propriétés des photons. Pour les mesures spectroscopiques résolues en temps, l'origine de ces incertitudes est encore mal définie, hormis peut être pour les mesures de spectroscopie de fluorescence. Par défaut, on considère la plupart du temps que les observations sont indépendantes et affectées d'un bruit gaussien additif. L'hypothèse d'indépendance des observations est justifiée lorsque les mesures sont effectuées séquentiellement (mode monocanal) mais peut être remise en cause dans le cas de mesures en régime pompe-sonde (mode multicanal). Puisqu'un spectre complet est observé simultanément à un délai fixé, il peut exister des incertitudes communes aux mesures effectuées à différentes longueurs d'onde. Le cas échéant, on observe alors une dépendance statistique dans la réponse à chaque longueur d'onde (on parle également d'erreur systématique, la corrélation résultant du fait que les mesures ont un terme de biais commun).

La structure de la matrice de variance-covariance dépend du mode d'acquisition des mesures et donc de la structure des données. Si les mesures sont enregistrées point par point, il n'y a idéalement pas de corrélation (puisqu'on mesure un biais différent pour chaque point) et la matrice de variance-covariance n'a d'éléments non nuls que sur la diagonale. Si les mesures sont obtenues ligne par ligne, il y a une valeur de biais pour chaque ligne ce qui induit des structures observables localement hors diagonale. On notera que lorsque l'acquisition concerne simultanément tous les points, la matrice de variance-covariance est potentiellement complète.

Le cube formé par l'ensemble des matrices de variance-covariance calculées peut être analysé, notamment par des méthodes *3-voies*, pour proposer l'indentification des différents facteurs à l'origine de cette structure de l'erreur de mesure.[42] L'étape suivante est finalement d'incorporer dans les traitements de données, notamment lors de résolutions MCR, l'information concernant la structure de l'erreur comme une connaissance *a priori*. C'est l'idée des moindres carrés pondérés, des méthodes basées sur l'estimation du maximum de vraisemblance ou encore de l'analyse bayésienne ; le choix d'un *prior* adapté permet de s'assurer que l'on ne détruit pas la structure de corrélation des incertitudes de mesure lors de la modélisation des données.

6-

Une partie des travaux en cours concerne l'analyse de la structure des erreurs de mesure en régime pompe-sonde. Les réponses enregistrées à chaque longueur d'onde et les incertitudes associées sont potentiellement statistiquement dépendantes, mais de matrice de covariance inconnue. Le moyen le plus direct pour estimer les matrices de variance-covariance de l'erreur de mesure est l'analyse de répétitions de la mesure d'une matrice *2-voies* (on notera que le concept est le même que pour l'estimation de la variance pour une mesure scalaire)

La figure (*Fig. 2.16*) illustre l'analyse du cube de données, construit par répétition d'une expérience de spectroscopie pompe-sonde d'absorption transitoire.

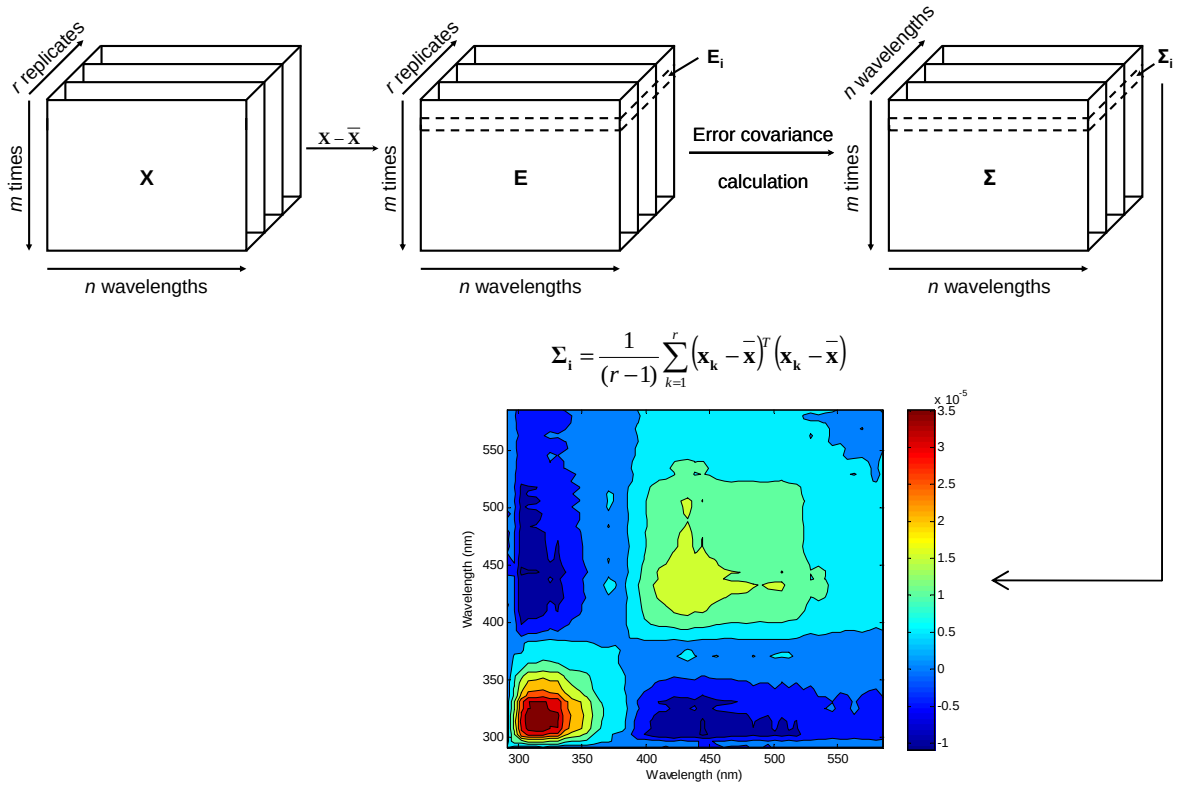


Fig. 2.16 Calcul et représentation de la matrice de variance-covariance de l'erreur dans la direction spectrale.

2.4.3 Approches bayésiennes et approches mixtes

■ La résolution exacte d'un système dépend des propriétés intrinsèques des données en termes de complexité et de sélectivité mais également, au-delà des aspects mathématiques, de la transcription de la connaissance chimique du problème sous forme de contraintes ou d'informations *a priori*.

L'analyse bayésienne des données n'est encore que peu appliquée aux problèmes de chimiométrie.[43,44] Elle requiert une information *a priori* sur les paramètres et les incertitudes de mesure, information qui peut être directement utilisée dans l'analyse pour améliorer la qualité des modèles. Dans le formalisme bayésien,[⊥] les points expérimentaux sont considérés comme le

[⊥] Dans le formalisme probabiliste (par opposition au formalisme fréquentiste), la probabilité traduit l'incertitude sur la valeur d'un paramètre

résultat d'un tirage aléatoire de l'échantillon selon une loi de probabilité conditionnelle, appelée densité *a posteriori*. La difficulté est d'échantillonner cette distribution pour laquelle il n'existe pas d'expression analytique. Au final, la densité *a posteriori* regroupe toute l'information dont on dispose concernant les paramètres d'un modèle optimisé. Elle s'écrit, par la formule de Bayes, en fonction de la densité *a priori* des paramètres et de la fonction de *vraissemblance* qui traduit la probabilité d'obtenir les valeurs observées, étant donné les valeurs des paramètres du modèle.

L'analyse bayésienne paramétrique peut être appliquée à la modélisation cinétique des données spectroscopiques résolues en temps.[45] Elle permet d'évaluer la pertinence des résultats pour l'identification des modèles, par l'interprétation des valeurs des paramètres et l'estimation des intervalles de confiance associés. Les modèles sont développés sur la base de l'écriture explicite de la réaction cinétique voire des formes paramétriques des spectres.[46,47] Les résultats obtenus sont convaincants lorsqu'on s'intéresse à des spectres larges, modélisés par des fonctions unimodales. Néanmoins, dans le cas de spectres plus complexes, en spectroscopie IRTF par exemple, les modèles semblent difficilement applicables.

Le problème inverse de résolution des données spectroscopiques peut également être traité comme un problème de séparation de sources en traitement du signal puisque les modèles matriciels sont de même forme (seules les contraintes appliquées varient). Si l'on se focalise sur la méthode d'analyse en composantes indépendantes (ICA, *independant component analysis*), la décomposition repose sur la notion d'indépendance statistique, c'est-à-dire que les variables latentes (les sources, les composantes pures d'un mélange, *etc.*) sont considérées non gaussiennes et mutuellement indépendantes.[48,49] Ces critères présentent l'avantage d'assurer l'unicité de la solution en dehors de toute contrainte d'orthogonalité. Néanmoins, si l'implémentation d'un critère de positivité est requise, pour l'analyse de données spectroscopiques par exemple, le critère d'orthogonalité des sources doit être rétabli pour garantir l'unicité de la décomposition. Les applications de l'ICA sont donc restées relativement limitées dans ce domaine.

Très récemment, une approche bayésienne non paramétrique, basée sur l'attribution de fonctions gamma pour décrire la connaissance *a priori* sur $\mathbf{C}$ et sur $\mathbf{S}^T$, a été proposée (BPSS,[50] *Bayesian Positive Source Separation*). Ce modèle permet, sous condition d'indépendance statistique des sources, d'encoder la non-négativité des distributions des spectres purs et des profils de concentration et, plus spécifiquement, le fait que la plupart des valeurs observées soient nulles ou proches de zéro (on parle de matrices creuses, *sparses* en anglais). La méthode s'apparente à un calcul de moindres carrés sous contrainte, analogue à celui codé dans la fonction de coût *(2.10)* pour exprimer le problème de factorisation en matrices positives. Dans le cas de l'approche BPSS, les paramètres optimaux de régularisation sont estimés par l'analyse bayésienne, pour chaque composante du mélange et sans indétermination. L'ambigüité rotationnelle est donc complètement levée mais l'efficacité de la méthode repose sur le choix de fonctions appropriées, ce qui restreint le domaine d'application. Ainsi, même si les fonctions gamma proposées s'avèrent très efficaces pour encoder l'information portée par des données très sélectives, elles s'avèrent moins performantes pour les spectres larges couvrant tout le domaine

spectral étudié ; les exemples proposés jusqu'ici illustrent d'ailleurs des problèmes de spectroscopie Raman.[50,51]

Quelle que soit la méthode considérée, *hard-modeling* ou *soft-modeling*, paramétrique ou non-paramétrique, la modélisation (au sens large) repose sur les résultats des analyses factorielles, au moins pour les estimations du rang ou de la complexité du signal. Partant de ce constat, le développement de méthodes bayésiennes semi-paramétriques semble assez naturel et doit pouvoir être envisagé. L'objectif serait d'étendre la modélisation paramétrique à des spectres multimodaux, plus spécifiques, mais dont la complexité rend la description paramétrique très délicate. Pour ce faire, des développements implémentant un modèle bayésien paramétrique sur les profils de concentration dans des algorithmes hybrides sont envisagés en collaboration avec le Laboratoire de Chimie Physique de l'Université Paris-Sud (UMR 8000). D'autre part, concernant les approches bayésiennes non-paramétriques de résolution du problème dit de séparation de sources, développées au Centre de Recherche en Automatique de l'Université de Nancy (UMR 7039), les limites évoquées doivent pouvoir être au moins partiellement levées par l'implémentation de contraintes dans la direction temporelle.

3 Modèles prédictifs en spectroscopie

Le développement de modèles multivariés pour la prédiction de grandeurs analytiques de référence à partir de données expérimentales représente une part importante des activités de recherche en chimiométrie. Néanmoins, la modélisation peut s'avérer très critique dès lors que les variables prédictives échantillonnent des données spectroscopiques, notamment en spectroscopie vibrationnelle. L'optimisation concerne alors la gestion de la dimension des données, la sélection de l'information pertinente, mais également la détermination de la structure et des paramètres des modèles pour la généralisation.

Les modèles chimiométriques peuvent s'inscrire dans le cadre de la théorie de l'apprentissage statistique qui intègre explicitement l'optimisation du compromis entre biais et variance. Nous insistons sur les développements récents en classification supervisée à l'origine des méthodes de séparateurs à vaste marge, également appelées machines à vecteurs de support (SVM, *support vector machines*). L'algorithme d'optimisation est par nature adapté aux données spectroscopiques, la formulation du problème reposant d'une part sur l'écriture d'un nombre relativement limité de contraintes dans un espace de grande dimension et d'autre part sur la notion de produit scalaire.

Dans la lignée des travaux effectués sur les réseaux de neurones artificiels, nous présentons les résultats obtenus récemment pour la prédiction de propriétés physico-chimiques par la mesure du spectre proche infrarouge. Les performances des modèles SVM de classification sont discutées, l'accent étant mis sur l'optimisation de l'étape d'apprentissage et l'interprétation des résultats.

3.1 Applications en spectroscopie vibrationnelle

Les techniques de spectroscopie vibrationnelle possèdent les caractéristiques métrologiques et les avantages des méthodes optiques, combinés aux atouts des approches moléculaires pour la caractérisation. Ces techniques permettent donc l'analyse rapide et non destructive d'échantillons aux propriétés optiques très variables et autorisent des applications analytiques performantes.

Lors de l'analyse d'échantillons complexes, inconnus et multi-composants, le spectre observé peut représenter une information assez peu spécifique ; il faut également considérer l'effet de nombreuses grandeurs d'influence potentielles, facteurs physico-chimiques ou environnementaux. Mais le spectre représente une information très reproductible, notamment dans le domaine proche infrarouge, ce qui permet des approches chimiométriques inductives lorsqu'on dispose de mesures chimiques de référence. L'idée est donc la modélisation des données expérimentales pour à terme prédire les grandeurs analytiques d'intérêt.

3.1.1 Bilan des travaux

■ Par leurs aspects applicatifs ou contractuels, les approches inductives constituent une part importante des activités de recherche en chimiométrie.

Sur la période 1997-2002, les activités de recherche ont concerné l'acquisition,[52,53] l'optimisation de modèles supervisés pour le suivi de procédés,[54-56] et la prédiction de la nature[57,58] ou de la composition[53,59,60] de matières premières et de produits manufacturés. Dans ces travaux, les résultats ont toujours été discutés en fonction de la technique d'échantillonnage, des prétraitements mathématiques des spectres et de la structure des données. Une attention particulière a été portée aux modèles de réseaux de neurones artificiels.[53,54,56,61] Certains résultats originaux sont détaillés dans l'exemple repris par la suite.

Dans le même temps, nos recherches ont concerné plus spécifiquement l'optimisation de l'architecture des réseaux de neurones pour la gestion de la dimension des données spectroscopiques pour des bases de données contenant peu d'échantillons,[57,62] montrant l'intérêt de favoriser des modèles hiérarchiques.[63] On retiendra que la complexité mathématique d'un modèle et la dimension physico-chimique du problème, c'est-à-dire le nombre de contributions physico-chimiques identifiables, doivent être en adéquation. Les modèles non-paramétriques ont également été appliqués à la problématique du transfert d'étalonnage.[64,65] Dans l'ensemble, ces travaux ont participé aux développements méthodologiques pour les données spectroscopiques dans un contexte de recherche pluridisciplinaire même si, avec le recul, certains travaux peuvent paraître aujourd'hui relativement datés du fait notamment de l'évolution des capacités de calcul.

L'atout principal des méthodes de réseaux de neurones reste que la structure du modèle n'est pas spécifiée *a priori* mais, au contraire, déterminée à partir des données d'apprentissage (c'est le sens du terme non-paramétrique : le nombre et la nature des paramètres ne sont pas fixés à l'avance). Les inconvénients des algorithmes d'apprentissage des réseaux de neurones sont liés à l'initialisation des poids, dont dépend la solution obtenue, ainsi qu'à l'optimisation des méta-paramètres.

7-

Nous revenons tout d'abord sur les travaux concernant le suivi par spectrométrie de vibration de la réaction *batch* d'hydrolyse de l'hémoglobine bovine par la pepsine (procédé développé par l'équipe du Pr Guillochon, Laboratoire Probiogem, EA1026 Lille). L'objectif est la production de peptides aux propriétés organoleptiques et fonctionnelles spécifiques par voie naturelle et chimiquement douce de transformation de l'hémoglobine. Nous nous sommes focalisés sur l'interprétation numérique et physico-chimique des réseaux de neurones, modèles non-paramétriques qui nécessitent un contrôle rigoureux de l'apprentissage par l'optimisation du compromis biais-variance. Nous avons proposé une analyse des traitements des différentes unités composant l'architecture du réseau.

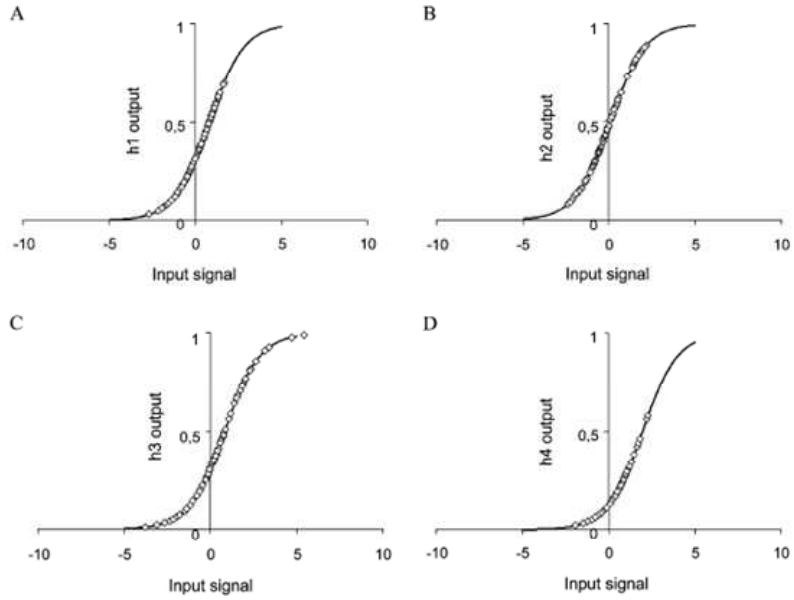


Fig. 3.1 Activation des fonctions de transfert des neurones (h_1-h_4) de la couche intermédiaire.[19]

Les figures reprises ici (*Fig. 3.1* et *Fig. 3.2*) illustrent les fonctions élémentaires des unités de traitement de la couche intermédiaire. La visualisation des activations de ces neurones permet d'analyser individuellement le traitement de chaque unité, en particulier d'évaluer le niveau de non-linéarité des traitements de l'information (ici les absorbances pondérées venant de la couche d'entrée). Ainsi, nous avons pu démontrer que certaines unités sont plutôt affectées aux traitements linéaires (le neurone h2 gère la présence d'une structure en *clusters* des données

batch, appliquant en quelque sorte un biais à chaque *batch*) tandis que d'autres (le neurone h3 en particulier) gèrent les aspects non linéaires. On notera également la similitude des traitements des unités intermédiaires h1 et h4 ce qui peut amener à une prise de décision pour l'optimisation de l'architecture, au-delà des heuristiques classiques uniquement basés sur l'estimation de l'erreur de prédiction.

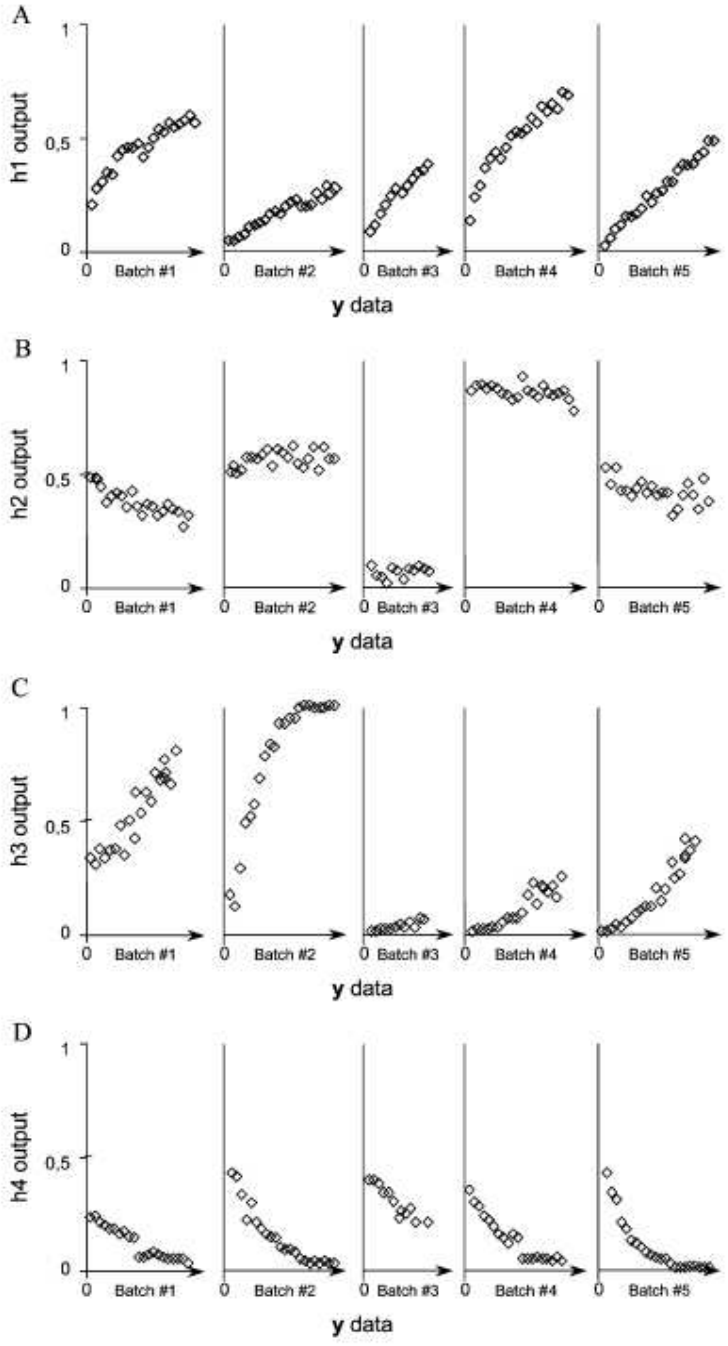


Fig. 3.2 Valeurs produites par les neurones (h_1-h_4) de la couche intermédiaire en fonctions de la valeur à prédire y, ici le degré d'avancement de l'hydrolyse.[61]

3.1.2 Sélection de l'information pour l'optimisation des modèles

■ A l'heure des méthodes expérimentales dites haut-débit, les problèmes de gestion de l'information posés par la dimension des données spectroscopiques sont généralisables à de nombreux domaines de la chimie.

Considérer un grand nombre de variables est incontournable dès lors que l'on s'intéresse à des mesures spectroscopiques d'échantillons complexes. Parallèlement, décrire l'ensemble des sources de variabilité requiert un grand nombre d'échantillons. Néanmoins, plus les mesures sont bruitées et redondantes, plus elles présentent de quasi-colinéarités, et plus la complexité du modèle d'apprentissage augmente. Cela se traduit par une dégradation de la robustesse des modèles, c'est-à-dire de leur applicabilité pour une utilisation future dans des conditions opératoires relativement similaires.

En restreignant significativement le nombre de variables prédictives, la sélection de l'information permet le développement de modèles parcimonieux et peut faciliter l'interprétation. L'évaluation de la pertinence analytique de l'information contenue dans les bases de spectres et la gestion de la quantité d'information lors du développement des modèles prédictifs sont des objectifs de recherche fixés dans le projet de laboratoire 2006-2009. Compte tenu de la difficulté de sélectionner *a priori* les variables d'intérêt (à plus forte raison en spectroscopie proche infrarouge) et de la dimension des données spectrales, la sélection de variables doit être gérée par des heuristiques. Cela signifie que toutes les approches de sélection envisageables sont forcément sub-optimales. Parmi les nombreuses manières de modéliser la relation considérées, aucune n'est universellement la meilleure. Mais l'intérêt des heuristiques reste néanmoins de proposer une solution acceptable dans des cas où les algorithmes basés sur une recherche exhaustive des solutions fournissant des résultats exacts ne sont pas applicables, du fait de la nature du problème.

3.1.2a Critères de pertinence et algorithmes d'optimisation en sélection de variables

D'un point de vue conceptuel, une procédure de sélection de variables nécessite de définir à la fois un critère de pertinence, pour mesurer l'influence des variables X pour la prédiction de Y, et un algorithme d'optimisation. Le critère de pertinence peut être évalué indépendamment du choix d'une structure de modèle, comme un prétraitement. C'est le cas lorsque l'on estime, par exemple, le coefficient de corrélation ou l'information mutuelle entre les variables X et Y. L'autre possibilité est d'évaluer la pertinence de la sélection de variables par le biais du modèle lui-même. C'est l'idée des algorithmes génétiques qui sont des approches stochastiques applicables à l'optimisation des modèles de régression.

L'information portée par une variable X peut s'exprimer comme une entropie de Shannon.[66] Celle-ci exprime quantitativement la pertinence de l'information analytique c'est à

dire la précision associée au résultat d'une mesure (c'est-à-dire l'incertitude d'un résultat non biaisé). L'information mutuelle portée par une (ou plusieurs) variable(s) *X* sur une variable *Y* estime, quant à elle, la quantité d'information contenue dans la (les) variable(s) dépendante(s) qui peut être utile pour la prédiction de la variable indépendante, sans aucune hypothèse structurelle. L'information mutuelle est une mesure de la corrélation entre deux variables aléatoires à l'instar du coefficient de corrélation. L'intérêt est que l'information mutuelle est une mesure plus générale de la corrélation, qui reste appropriée même dans le cas de dépendances non-linéaires. Le calcul requiert néanmoins l'estimation des fonctions densité de probabilité, ce qui peut s'avérer délicat pour des données de grande dimension. Nous avons repris dans nos travaux, les estimateurs proposés par Kraskov et codés par Rossi.[67]

Les algorithmes génétiques[68] sont des outils d'optimisation stochastiques basés sur des heuristiques inspirés du *néo-darwnisme* tels que les opérateurs *mutation* ou *points de croisement*. Le problème d'explosion combinatoire évoqué précédemment est contourné en démarrant d'une distribution d'individus correspondant à des sélections aléatoires de variables. Des modèles de régression sur facteurs sont ensuite construits en parallèle pour chacune d'entre-elles, puis la population évolue pour l'optimisation d'une fonction de coût appropriée. Au final, les individus portent tous la même information génétique c'est-à-dire qu'ils sont constitués des mêmes variables. Dans la plupart des applications, la fonction de coût est simplement une fonction inverse de l'erreur de validation croisée calculée pour un modèle de régression donné, dans des hypothèses structurelles fortes contrairement au cas précédent.

3.1.2b Modèles prédictifs de la composition de textiles

Les travaux récents,[67-69] concernent l'évaluation rapide de la composition ou de propriétés physico-chimiques d'intérêt des textiles par la mesure du spectre proche infrarouge. Ils ont été développés dans le cadre de la thèse d'Alexandra Durand, en marge d'un programme de recherche du Fonds de la Recherche Technologique (voir chapitre 1). Ces problèmes sont des problèmes importants, posés par les applications (répression de fraudes, tri sélectif des textiles ou développement de capteurs spécifiques) pour pallier les contraintes des méthodes analytiques classiques, approches chimiques (dilutions) ou physiques (microscopie). Du point de vue de la spectroscopie, les difficultés sont liées à la variabilité inter-échantillons des spectres enregistrés en réflexion diffuse, même pour des échantillons de composition chimique semblable.

8-

Revenons sur l'étude de la détermination de la composition de produits textiles coton-viscose (deux fibres cellulosiques très proches) sur la base du spectre proche infrarouge. Les résultats[68] obtenus par l'application d'un modèle de régression PLS (*Partial Least Squares*) sur spectres proche infrarouge complets sont de l'ordre de 3,74% en RMSEP (*Root Mean Squarred Prediction Error)* dans la gamme analytique 0-100% pour la teneur en coton. Le modèle développé requiert 11 variables latentes, ce qui correspond à un niveau de complexité élevé. Cela traduit la difficulté de séparer la variance associée aux facteurs chimiques de celle associée aux

autres sources : échantillonnage, variabilité externe, variabilité naturelle des échantillons textiles, *etc.* Du fait de la sensibilité de la spectroscopie proche infrarouge et du manque d'information pour décrire explicitement ces sources, l'analyse est difficile notamment lorsqu'elle est réalisée sur spectres complets.

La figure ci-dessous (*Fig. 3.3*) reprend schématiquement la démarche correspondant au développement d'un modèle de réseaux de neurones artificiels sur la base des variables retenues par une procédure de sélection par information mutuelle.

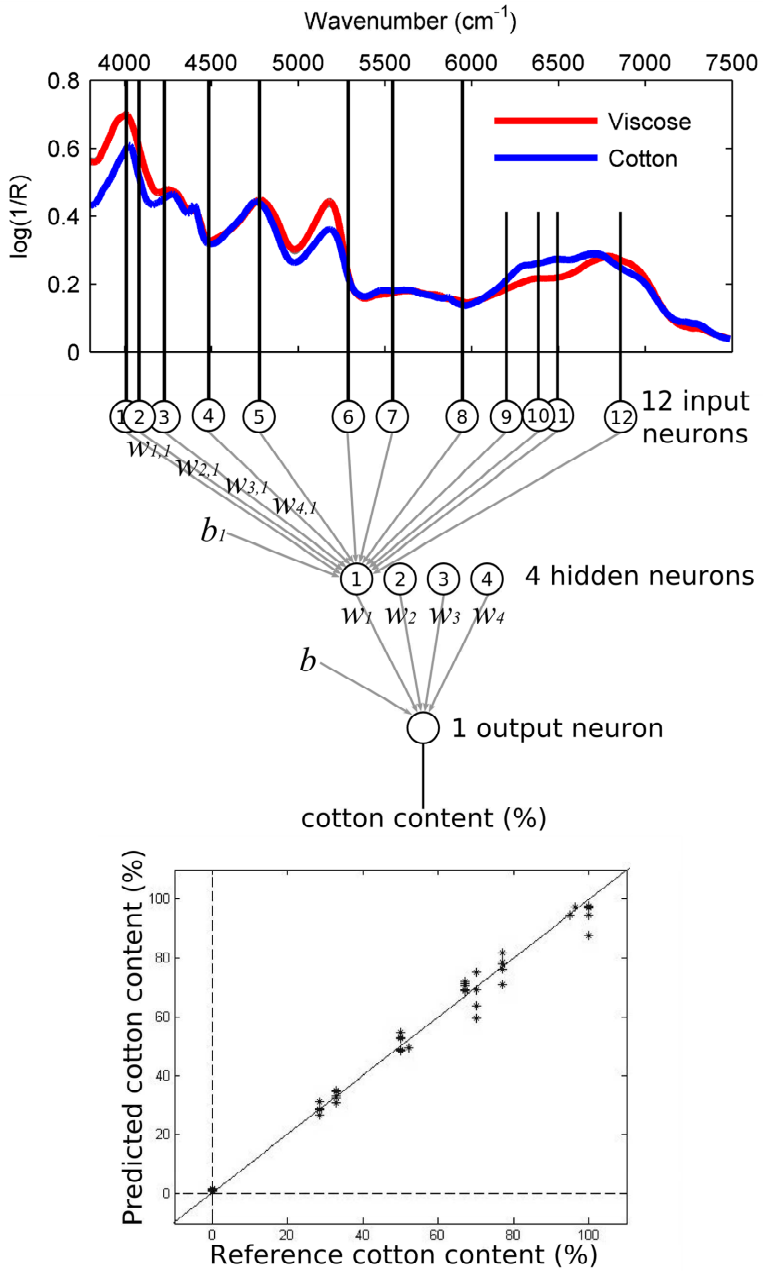


Fig. 3.3 Schéma de principe de l'application d'un modèle de réseaux de neurones pour la prédiction de la composition de textiles à partir de spectres proche infrarouge réduits à 12 points par sélection de variables par information mutuelle

Les résultats sont assez représentatifs de ceux obtenus en général sur des bases de données de spectres proche infrarouge d'échantillons complexes. Douze variables spectroscopiques, distribuées sur l'ensemble du domaine spectral, sont identifiées. Le modèle d'étalonnage multivarié est construit sur la base de ces spectres discrets. L'erreur de prédiction obtenue est 3,43%, en RMSEP, pour un modèle de réseau de neurones artificiels optimisé sur une architecture *12×4×1*. Ces résultats ont été discutés parallèlement à ceux obtenus par une procédure d'algorithmes génétiques intégrant un modèle de régression PLS.[67] On retiendra que les deux méthodes de sélection de variables produisent des résultats équivalents en terme de prédiction. Néanmoins, puisqu'elles reposent sur des heuristiques très différentes, la comparaison directe des deux approches (sur la base de la complexité des modèles, du nombre et de la nature des variables sélectionnées) reste limitée.

3.2 Cadre théorique de la modélisation des données

L'apprentissage supervisé a pour objectif la modélisation et la prédiction d'une (ou de plusieurs) variable(s) qualitative(s) ou quantitative(s) y sur la base de variables prédictives x données. Puisqu'on ne connaît pas la loi qui régit la distribution des données, le modèle est basé sur un ensemble d'apprentissage (x_i, y_i), échantillon de taille n composé d'observations des entrées et des sorties. Ce modèle devra être une généralisation des exemples qui permette de prédire y connaissant x, pour tout (x, y) issu de la distribution des données. Techniquement, la conception et la construction du modèle se ramènent là encore à la minimisation d'une fonction de coût qui pénalise les erreurs en prédiction. L'erreur quadratique est la fonction de coût la plus simple d'un point de vue analytique mais des fonctions plus robustes, incluant des termes de pénalisation, peuvent être implémentées.

3.2.1 Modélisation statistique en apprentissage supervisé

Soit un couple (X,Y) de variables aléatoires observées conjointement et $P(X,Y)$ la loi de probabilité associée. La variable X est un vecteur de $\Re^P$ dont les composantes sont des variables aléatoires et Y est pris dans $\Re$ (on considère dans un premier temps la prédiction d'une variable quantitative). On cherche une fonction $f(X)$ permettant de prédire Y ayant observé X.

Le critère pour choisir la fonction f est le risque $R(f)$ (appelé contraste en mathématiques), c'est à dire l'espérance de l'erreur de prédiction *(3.1)*. La fonction $L(y, f(x))$ correspond à l'expression tout à fait générale d'une fonction de coût dans le cadre d'un problème d'apprentissage supervisé, la fonction de coût standard étant l'erreur quadratique *(3.2)*. Le risque $R(f)$ définit le risque réel, c'est à dire l'erreur qui serait mesurée sur tous les couples (x, y) en tenant compte de la probabilité d'observer chacun d'entre-eux.

$$R(f) = \int L(y, f(x))\ dP(x,y) = E[\ L(Y, f(X))\ ] \qquad (3.1)$$

$$L(y, f(x)) = (y - f(x))^2 \qquad (3.2)$$

En pratique, le problème est de minimiser ce risque dans les cas où la distribution des données, $P(X,Y)$, est inconnue. On ne pourra donc calculer qu'une estimation du risque réel, appelée risque empirique $R_{emp}(f)$, cette estimation étant obtenue à partir d'un ensemble d'échantillons indépendants et répartis au hasard[⊥] sur $P(X,Y)$. Ce risque *(3.3)* correspond à la moyenne observées sur la base de données d'apprentissage (x_i, y_i) de taille n.

[⊥] *iid, independent and identically distributed.*

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \qquad (3.3)$$

Pour le moment restons dans le cas d'un lot d'apprentissage suffisamment grand ; au sens statistique cela signifie que l'on dispose d'au moins une observation pour chaque point de l'espace des données d'entrée. Le calcul *(3.2)* se ramène alors à celui d'un calcul d'espérance par conditionnement en chaque point $X = x$ dont la solution est la fonction de régression *(3.4)*.[7]

$$f(x) = E(Y|X=x) \qquad (3.4)$$

Sans aller plus loin, il est intéressant de constater que de nombreuses méthodes utilisées en chimiométrie peuvent être décrites dans ce formalisme, notamment la régression linéaire et les méthodes de plus proches voisins type *k*-NN (*k-Nearest Neighbours*). Elles consistent à calculer *f(x)* avec des hypothèses différentes pour le modèle. Les moindres carrés supposent que la fonction *f* est correctement estimée par une fonction linéaire globale, ce qui représente une contrainte de structure très rigide. Les prédictions seront donc précises mais potentiellement non justes (biaisées). A l'opposé, les méthodes *k*-NN n'imposent quasiment pas de structure, la fonction *f* étant estimée par une fonction locale constante. Cette flexibilité assure des prédictions justes mais qui pourront s'avérer imprécises (instables, variance élevée) et dépendront des valeurs particulières de quelques données d'entrée lorsque la distribution des échantillons n'est pas suffisamment dense (lorsqu'on s'écarte de l'hypothèse précédente concernant la taille du lot d'apprentissage, c'est-à-dire systématiquement en chimiométrie).

La plupart des méthodes chimiométriques se positionnent donc quelque part entre ces deux situations extrêmes du compromis entre biais et variance. C'est le cas des régressions locales qui relâchent l'hypothèse globale de linéarité ou des méthodes à noyaux (*kernels*) qui pondèrent le voisinage entourant le point considéré.

3.2.1a Régression linéaire

La régression linéaire rentre dans le cadre de l'apprentissage statistique, il s'agit d'une approche basée sur un modèle linéaire *(3.5)* de la fonction *f* avec *X* et *Y* les matrices construites sur la base de données d'apprentissage (x_i, y_i) de taille *n* et β estimé à partir des observations.

$$f(x) \approx x^T\beta \qquad (3.5)$$

$$\beta = [E(XX^T)]^{-1}E(X^TY) \qquad (3.6)$$

$$\hat{f}(x) \approx x^T\hat{\beta} \text{ et } \hat{\beta} = (XX^T)^{-1}X^TY \qquad (3.7)$$

En injectant ce modèle dans l'expression de la fonction de coût *(3.2)*, on obtient l'équation *(3.6)* par différenciation. La solution des moindres carrés[⊥] *(3.7)* est obtenue en remplaçant l'espérance par les moyennes calculées sur les données d'apprentissage.

3.2.1b Plus proches voisins

Dans le cas de la méthode de plus proches voisins, la fonction de régression *(3.4)* est quasiment implémentée directement. L'espérance est estimée par une moyenne sur les données d'apprentissage, la contrainte de conditionnement à l'ensemble des x est relâchée et ne s'applique que sur un voisinage $N_k(x)$ de x avec k proches voisins.

$$\hat{f}(x) = Moy\left(y_i \middle| x_i \in N_k(x)\right) \qquad (3.8)$$

En chaque point x, on calcule la moyenne des réponses y_i correspondant aux $x_i = x$. Ainsi, dans le contexte d'un lot de d'apprentissage arbitrairement large, les méthodes de plus proches voisins constituent quasiment des approches universelles en régression pure puisque l'erreur d'entraînement vaut 0 pour $k = 1$. A l'opposé, la solution obtenue tend vers les moindres carrés lorsque l'on considère un voisinage arbitrairement grand ($N_k(x) \to \infty$).

3.2.1c Classification supervisée

Le principe peut s'étendre aux problèmes de classification supervisée. On cherche alors à prédire non plus Y (dans $\Re$) mais une variable qualitative $\Gamma = \{\gamma_1, \gamma_2\}$ pour un problème à 2 classes (illustré *Fig. 3.4*). On définit une fonction g et le risque associé.
Afin de pénaliser les erreurs de prédiction, une fonction de coût spécifique du problème de classification *(3.9)*, composée typiquement de 0 et de 1, est utilisée. L'espérance est estimée compte tenu de la distribution jointe $P(\Gamma|X)$. Lorsque l'on cherche à minimiser le risque, on obtient finalement l'expression *(3.10)* qui correspond à classer dans la classe la plus probable en exploitant la distribution conditionnelle discrète $P(\Gamma|X)$ (classification de Bayes). En pratique, il existe de très nombreuses méthodes pour estimer les probabilités $P(\Gamma|X)$, notamment toutes les méthodes linéaires de classification.

$$H(g) = E[\, L(\Gamma, g(X))] \qquad (3.9)$$

$$g(X) = \gamma_k \text{ si } P\left(\gamma_k \middle| X = x\right) = \max_{\gamma \in \{\gamma_1, \gamma_2\}} P\left(\gamma \middle| X = x\right) \qquad (3.10)$$

La figure suivante (*Fig. 3.4*) illustre deux approches très différentes du même problème. La première est basée sur l'analyse discriminante linéaire (LDA, *Linear Discriminant Analysis*)

[⊥] Un problème des moindres carrés est un problème d'optimisation convexe (voir annexe1).

tandis que l'autre correspond à un modèle de classification SVM qui représente une généralisation des approches à noyaux. Cet exemple sera repris et détaillé au paragraphe *3.4.1*.

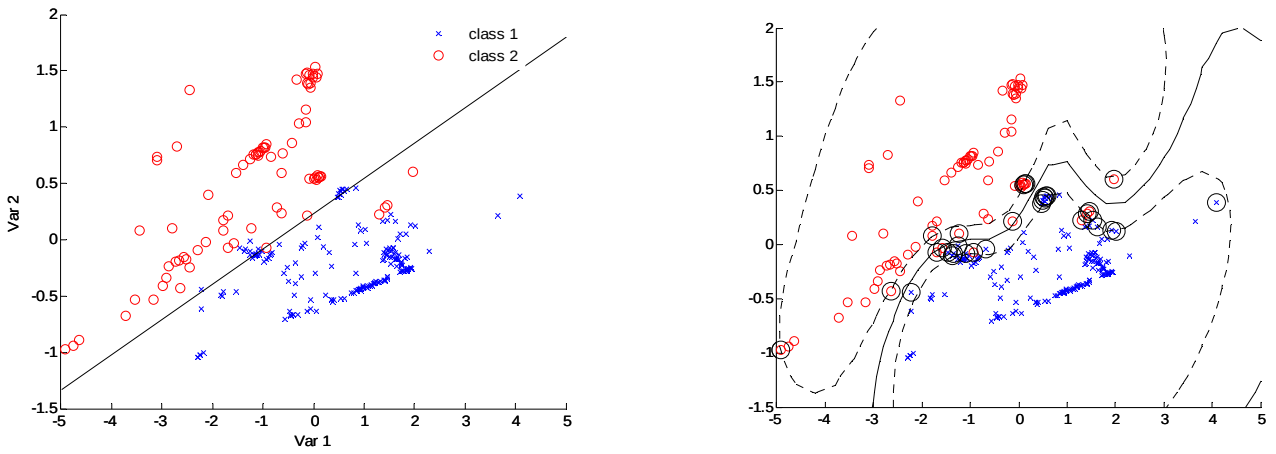


Fig. 3.4 Un problème à deux classes pour des données à deux dimensions : frontières calculées en analyse discriminante linéaire (à gauche) et par machine à vecteurs de support (à droite).

3.2.2 Complexité des modèles prédictifs

■ Le choix d'un modèle résulte d'un compromis entre les contributions de biais et de variance à l'erreur de prédiction. Toutes les méthodes sont concernées par le choix de modèles parcimonieux mais la complexité s'exprime de façon différente pour chaque méthode.

Contrôler la complexité d'un modèle revient à optimiser le compromis entre le biais et la variance pour minimiser le risque. Pour les fonctions de coût quadratiques, on montre d'ailleurs facilement que le terme d'erreur de prédiction s'écrit comme la somme d'un terme de biais (au carré), d'un terme de variance et d'un terme irréductible de bruit. Un modèle trop simple peut présenter un biais important, c'est-à-dire une erreur traduisant l'inadéquation du modèle à la distribution des données que l'on cherche à ajuster. Au contraire, plus un modèle est complexe, plus il est flexible au sens capable de s'ajuster aux données d'apprentissage. Néanmoins, un tel modèle peut s'avérer défaillant en généralisation si la contribution de la variance est surestimée. Cela signifie que la construction du modèle dépend trop des données d'apprentissage (du lot d'exemples particuliers) dont on dispose pour décrire la distribution $P(X,Y)$ inconnue. Tout autre ensemble d'exemples (*iid* de cette distribution) conduirait à un résultat différent. Cette situation correspond à un risque important de surapprentissage (*overfitting*) des données. En pratique, on cherchera donc à contrôler la complexité des modèles en minimisant l'erreur de prédiction (risque empirique) sur des données indépendantes des données d'apprentissage. On notera également que la problématique de la sélection de modèles, c'est-à-dire la détermination

du niveau de complexité d'un modèle adapté à l'échantillon, est un domaine de recherche très actif en statistique non paramétrique.

La complexité d'un modèle dépend du nombre de paramètres que celui-ci intègre et donc du nombre de variables d'entrée. Le cas le plus simple est celui de la régression linéaire puisque le nombre de paramètres est proportionnel au nombre de variables. Néanmoins, cette dépendance est plus complexe pour les méthodes factorielles de régression ou, à plus forte raison, pour les méthodes non linéaires. Dans ce cas, toute fonction passant au plus près des couples de points du lot d'apprentissage est solution. Cet ensemble de solutions doit donc être restreint, en réduisant la taille du voisinage considéré par exemple. C'est l'idée des méthodes *k*-NN et des méthodes d'apprentissage à noyaux. D'autres approches décrivent les comportements locaux de manière plus implicite, par les fonctions sigmoïdes de la couche cachée pour les réseaux de neurones artificiels comme nous l'avons observé précédemment (voir paragraphe *3.1.1*).

3.2.2a Le fléau de la dimension

■ En grande dimension, même les lots de données de grande taille n'échantillonnent l'espace d'entrée qu'avec une faible densité.

Le premier constat est assez intuitif : la complexité des fonctions de *p* variables croit très rapidement avec la dimension *p*. Par contre, le phénomène qui en découle, appelé *explosion de la dimension* ou *fléau de la dimension* (*curse of dimensionality*), est plus difficile à appréhender. L'idée est la suivante. Si on souhaite estimer sans fixer de loi *a priori* des fonctions en grande dimension avec la même précision qu'en dimension restreinte, alors la taille du lot d'entraînement doit croître de la même manière.

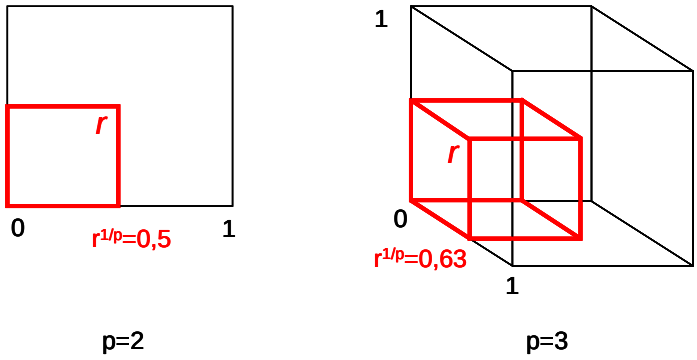


Fig. 3.5 Fléau de la dimension en apprentissage statistique.[7]

Sur l'exemple proposé ci-dessus (*Fig. 3.5*), on considère une quantité d'information unitaire contenue dans un hypercube de $\Re^p$ avec *p* variables continues à valeurs dans l'intervalle

[0,1]. Quelle est la taille du voisinage à considérer pour capturer une fraction *r* de l'information (c'est-à-dire du volume unitaire) qui reste constante lorsque *p* augmente ?

Pour couvrir uniformément 25 % (*r* = 0,25) de la variabilité en dimension 2 (le carré rouge, *Fig. 3.5*), il faut échantillonner 50 % des valeurs possibles pour chacune des 2 variables, soit $r^{1/2}$. En dimension 3, le même voisinage (un cube représentant 25 % du volume du cube unitaire) amène à couvrir ~ 63 % des valeurs possibles pour chaque variable, soit $r^{1/3}$. Plus généralement, la taille du voisinage à considérer pour capturer une fraction *r* de l'information dans $\Re^p$ est en $r^{1/p}$, ce qui devient très vite pénalisant (dans notre exemple, pour *p* = 10 et *r* = 0,25, $r^{1/p}$ = 0,87 ! C'est le phénomène appelé *fléau de la dimension*).

La question peut être posée d'un autre point de vue. Pour une densité constante de la distribution des échantillons, comment se traduit le passage en dimension *p* plus élevée ? L'exemple est tout aussi parlant, si *n* = 10 représente un échantillon dense pour *p* = 1, il faudra pouvoir disposer de 100 exemples pour *p* = 2, de 1000 exemples pour *p* = 3, *etc*. Toutefois, même si cette approche pose clairement le problème, elle ne doit pas entraîner de conclusions trop hâtives. En particulier, on remarquera que les *p* variables considérées dans l'exemple sont des variables non corrélées (orthogonales) et à valeurs dans tout l'intervalle [0,1], ce qui correspondrait à une situation extrême pour des données spectroscopiques.

Rappelons enfin le consensus exprimé par les méthodes d'analyse factorielle (voir paragraphe 2) et valable plus largement en chimiométrie : les données (spectroscopiques) ne sont pas réellement de grande dimension bien qu'elles soient généralement exprimées dans un espace très grand. Cet espace peut être (doit être) synthétisé dans un espace de dimension réduite, en adéquation avec la dimension chimique du problème et respectant la nature des variables chimiques, sans perdre trop d'information. Dès que la structure des données est complexe, la recherche d'un sous-espace adapté au problème d'apprentissage supervisé s'avère délicate.

3.2.2b La Dimension de Vapnik-Chernovenkis

Les travaux de Vapnik-Chernovenkis[5,6] (VC) en théorie de l'apprentissage ont formalisé la notion de complexité des modèles. En particulier, la *dimension de VC* (notée *h*) peut être définie par le plus grand nombre de points pouvant être séparés par des classes de fonctions définies, dans une configuration d'apprentissage donnée.

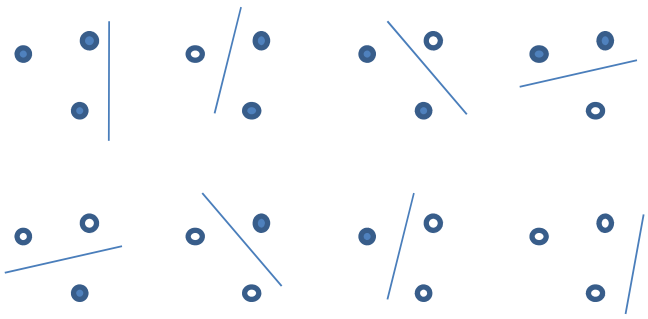


Fig. 3.6 Dimension de VC pour la classe des lignes droites dans un espace de dimension p = 2.

Par exemple, pour un problème de classification à deux classes dans un espace de dimension 2, la *dimension de VC* est $h = 3$ pour la classe des fonctions linéaires. Cela signifie que, quel que soit l'étiquetage de 3 points dans un espace de dimension 2, on peut trouver une droite qui les sépare (*Fig. 3.6*). On dit que cet ensemble de 3 points est pulvérisé par la classe des fonctions linéaires.

Plus généralement, la famille des hyperplans dans un espace de dimension p possède une dimension $h = p + 1$. On pourrait penser que plus les modèles sont décrits par un grand nombre de paramètres, plus leur *dimension de VC* est élevée, et inversement. Notre intuition est là encore contredite pour les fonctions non linéaires. Ainsi, la classe des fonctions $sin(\alpha x)$ qui repose sur un paramètre α unique dans $\Re$ possède une dimension h qui tend vers l'infini.

D'un point de vue plus formel *(3.11)*, connaître la dimension h permet d'estimer la valeur théorique asymptotique du risque réel $R(f)$ (l'erreur de généralisation) à partir du risque empirique $R_{emp}(f)$ (l'erreur d'apprentissage) pour un modèle f et un lot de données de taille n, au niveau de confiance $1-\eta$.

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h\left(\log(2n/h)+1\right) - \log(\eta/4)}{n}} \qquad 0 < \eta < 1 \qquad (3.11)$$

Plus la base de données d'apprentissage est grande (n grand), plus le risque empirique est un bon estimateur du risque réel (le dernier terme tend vers zéro si la dimension h est bornée). On retrouve également le compromis entre biais et variance et le principe de parcimonie qui pénalise les grands modèles qui sur-modélisent les données. Si une série de modèles est entraînée sur un même ensemble d'apprentissage (n constant) avec pour objectif de minimiser l'erreur empirique, alors le modèle sélectionné doit être celui dont la somme du risque empirique et du terme associé à la dimension h est minimale.

Les concepts fondamentaux exprimés par cette équation présentent surtout l'intérêt de rationaliser certaines approches heuristiques mises en place en chimiométrie. Le dernier terme, positif, permet notamment de définir le concept de pénalisation. La pénalité s'ajoute au risque empirique et tient compte de la structure du modèle et des données. D'autre part, le problème revient à borner la différence entre le risque réel et le risque empirique… mais, en pratique, on ne connaît pas le risque réel. Ce constat est à l'origine des approches de validation des modèles en apprentissage supervisé et en chimiométrie. Ainsi, la validation croisée, qui consiste à réaliser successivement plusieurs découpages du lot de données, et plus généralement les méthodes de rééchantillonnage (*bootstrap* en anglais), peuvent être vues comme des alternatives au calcul de la *dimension de VC*. Il s'agit d'approches pragmatiques du problème d'optimisation de la complexité du modèle. Par ailleurs, on remarquera que le risque empirique mesuré sur une base d'échantillons de test (*iid* de la même distribution) est un bon estimateur du risque réel, contrairement au risque empirique estimé sur le lot d'apprentissage qui est bien souvent un estimateur très optimiste. Le problème évoqué précédemment se ramène alors finalement à borner la différence entre deux risques empiriques, en apprentissage et en test, ce qui justifie là aussi les méthodologies utilisées en chimiométrie.

3.3 Méthodes avancées de classification et de régression

Les SVM sont des algorithmes d'apprentissage basés sur la recherche de l'hyperplan optimal, celui qui sépare correctement les données tout en étant le plus éloigné possible des observations pour assurer de bonnes capacités de généralisation. Cette définition se comprend facilement pour des données séparables. Dans le cas de la discrimination de données non séparables, deux astuces de calcul vont permettre de ramener la recherche de surfaces séparatrices non linéaires à un problème de classification linéaire. La première idée consiste à définir l'hyperplan séparateur comme solution d'un problème d'optimisation sous contrainte. La fonction de coût possède alors la particularité de ne s'exprimer qu'à l'aide de produits scalaires. La deuxième idée réside dans l'introduction d'une fonction noyau dans le produit scalaire. Cela induit implicitement une transformation non linéaire de l'espace initial des données vers un espace intermédiaire de plus grande dimension dans lequel est résolu un problème de classification linéaire.

Le principe fondateur des SVM est lié aux développements récents en apprentissage statistique, en particulier le fait que la capacité de généralisation d'un modèle puisse être optimisée en contrôlant sa complexité, c'est-à-dire finalement le nombre de vecteurs de support. Les méthodes SVM ont d'abord été définies pour la discrimination puisque le modèle de classification SVM est une généralisation de l'algorithme du perceptron. Elles ont ensuite été étendues en régression. La notion clé est ici la notion de marge, au sens d'une marge de sécurité prise lors de la décision ; les échantillons doivent non seulement être bien classés mais également se trouver à une distance suffisante de la frontière.

3.3.1 Hyperplan séparateur optimaux (cas des classes séparables)

■ Introduire la notion de marge dans le critère d'apprentissage impose non seulement que les exemples soient bien classés, mais également qu'ils soient le plus possible éloignés de la frontière.

On considère la situation la plus simple en classification, le cas de la recherche de l'hyperplan séparateur optimal pour deux classes séparables. On se place comme précédemment dans le cas d'un échantillon statistique $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ de loi inconnue avec X dans $\Re^P$ et Y dans $\{-1,1\}$.

Le problème se pose comme la recherche d'une frontière de décision, fonction discriminante $f(x)$, dans l'espace des valeurs de X. Cet hyperplan est défini par l'équation *(3.12)* où β vérifie $||\beta|| = 1$. Le plan séparateur optimal est celui qui sépare les deux classes tout en maximisant la marge (la distance au point échantillon le plus proche pour chaque classe). On

exprime ainsi le meilleur compromis entre la capacité d'ajustement et la capacité de généralisation d'un modèle.

$$\{ x : f(x) = x^T \beta + \beta_0 \} \qquad (3.12)$$

Dans le cas de classes séparables, la règle de classification revient à trouver une fonction $f(x)$ telle que $y_i f(x_i) > 0$ pour tout i. De plus, la valeur absolue fournit une indication sur la confiance à accorder au résultat du classement puisque $f(x)$ est proportionnelle à la distance algébrique de n'importe quel point x à la frontière $f(x) = 0$.

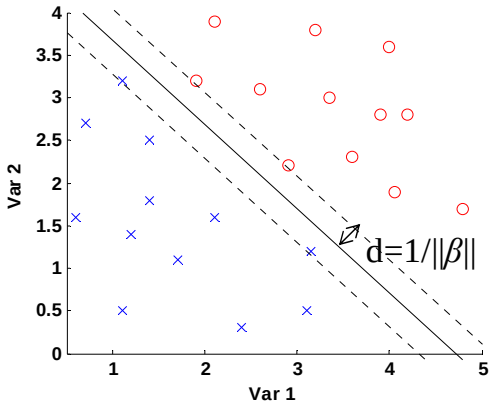


Fig. 3.7 Classification sur la base de l'hyperplan séparateur de marge maximale (cas séparable).

Trouver l'hyperplan qui réalise la plus grande marge de largeur $2d$ (voir *Fig. 3.7*) par rapport aux points d'entraînement pour les classes {-1,1} s'écrit donc sous la forme du problème d'optimisation *(3.13)*.

$$\underset{(\beta, \beta_0, \|\beta\|=1)}{Maximiser}\ d \qquad (3.13a)$$

$$soumis\ à\ \ y_i(x_i^T \beta + \beta_0) \geq d \qquad \forall i = 1, ..., n \qquad (3.13b)$$

Il est possible de s'affranchir de la contrainte $\|\beta\| = 1$ en remplaçant la condition *(3.13b)* par *(3.14)* (cela amène uniquement à redéfinir β_0).

$$y_i(x_i^T \beta + \beta_0) \geq d \times \|\beta\| \qquad \forall i = 1, ..., n \qquad (3.14)$$

Si β et β_0 vérifient ces inégalités alors n'importe quel facteur d'échelle positif les vérifie également et on peut poser arbitrairement $\|\beta\| = 1/d$. L'équation *(3.15)* est donc finalement la formulation à retenir. Elle exprime le problème de recherche de l'hyperplan séparateur de marge maximale *(3.13)* comme un problème d'optimisation convexe (optimisation d'un critère quadratique sous contrainte d'inégalité linéaire, voir annexe 1).

$$\underset{(\beta,\beta_0)}{Minimiser}\ \frac{1}{2}\|\beta\|^2 \qquad (3.15a)$$

$$soumis\ à\ \ y_i(x_i^T\beta+\beta_0) \geq 1 \quad \forall\, i=1,...,n \text{ et } d=\frac{1}{\|\beta\|} \qquad (3.15b)$$

Les étapes principales des calculs pour la résolution de ce problème peuvent être omises en première lecture et sont proposées en annexe (annexe 2). Le résultat est l'équation *(3.16)* définissant l'hyperplan optimal.

$$\beta=\sum_{i=1}^{n}\alpha_i\, y_i x_i \qquad (3.16a)$$

$$\sum_{i=1}^{n}\alpha_i\, y_i x^T x_i+\beta_0=0 \qquad (\alpha_i>0) \qquad (3.16b)$$

$$G(x)=signe(x^T\beta+\beta_0) \qquad (3.16c)$$

La solution est une combinaison linéaire de points appelés vecteurs de support qui sont des points échantillons disposés sur les bords de la marge, c'est-à-dire les points pour lesquels les contraintes sont actives ($\alpha_i > 0$). Cette solution est globale et déterministe. L'identification des vecteurs de support requiert l'utilisation de toutes les données d'entraînement. Néanmoins, la description de la solution peut se faire à l'aide de ces seuls vecteurs de support *(3.16a)*. Ces points correspondent donc finalement aux données les plus importantes du lot d'entraînement. De plus, la complexité du modèle construit ne dépend que du nombre et de la position de ces points. Intuitivement, l'approche semble donc robuste puisque l'hyperplan optimal ne se focalise que sur les points qui comptent dans la décision. Tous les autres points à l'intérieur des classes peuvent être supprimés, voire déplacés (à condition de ne pas franchir la frontière), sans modifier le modèle construit. L'équation *(3.16c)* correspond à la fonction de classification qui permet la prédiction en généralisation pour une observation *x* non apprise.

On remarquera enfin que la classification SVM est à l'opposé des méthodes plus classiques de discrimination linéaire. En effet, dans ce cas, tous les points d'apprentissage, même ceux très éloignés de la frontière, sont considérés. La frontière calculée n'est alors optimale que lorsque l'hypothèse de classes gaussiennes de même covariance est strictement vérifiée.

3.3.2 Hyperplan séparateur optimaux (cas des classes non séparables)

■ La marge est optimisée en pénalisant, dans l'expression des contraintes, la présence d'échantillons mal classés.

L'approche précédente peut être généralisée aux cas pour lesquels les classes ne sont pas séparables, au sens d'une séparation linéaire, c'est à dire lorsque les distributions d'échantillons se recouvrent dans l'espace des variables (*Fig. 3.8*).

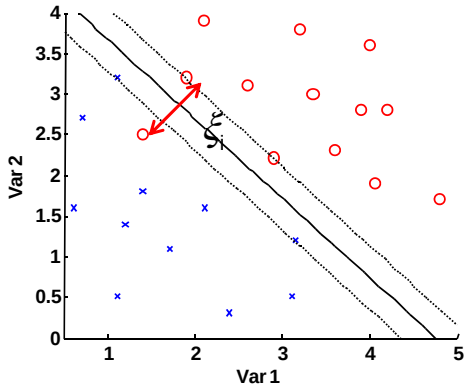


Fig. 3.8 Classification sur la base de l'hyperplan séparateur de marge maximale (cas non séparable).

L'idée est la suivante : continuer de maximiser la marge pour avoir la meilleure généralisation possible, mais tolérer certains points du mauvais coté de la frontière. Un terme pénalisant les erreurs individuelles ξ_i est donc introduit dans l'expression des contraintes *(3.17)*. Ce terme représente la proportion d'échantillons que l'on accepte dans la marge, ou du mauvais coté de la frontière, c'est à dire l'erreur totale de prédiction en apprentissage.

$$y_i(x_i^T\beta+\beta_0) \geq d(1-\xi_i),\ \xi_i \geq 0,\ \sum_{i=1}^{n}\xi_i \leq \text{cte} \qquad \text{et } \forall\, i=1,\ldots,n \qquad (3.17)$$

Une formulation équivalente du problème est donnée par l'équation *(3.18)*.

$$\underset{(\beta,\beta_0)}{Minimiser}\ \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (3.18a)$$

$$soumis\ \text{à}\ y_i(x_i^T\beta+\beta_0) \geq 1-\xi_i \qquad \xi_i \geq 0 \qquad \forall\, i=1,\ldots,n \qquad (3.18b)$$

Il s'agit comme précédemment d'un problème quadratique d'optimisation sous contrainte. On notera l'introduction du terme de pénalité et du paramètre *C*. Il s'agit d'un paramètre d'adaptation (méta-paramètre) qui doit être optimisé lors de l'apprentissage. Plus *C* est grand, plus les erreurs de classification ($\xi_i > 0$) sont pénalisées au profit de l'ajustement mais potentiellement au détriment de la généralisation. A l'inverse, des valeurs faibles de *C* favorisent les marges larges.

Les étapes principales des calculs pour la résolution du problème *(3.18)* sont reprises en annexe 3. L'interprétation des résultats peut être calquée sur celle présentée dans le cas des données séparables. La solution est de la forme *(3.16a)*. Les observations pour lesquelles les contraintes sont actives ($\alpha_i > 0$) sont les vecteurs de support, et β ne s'exprime qu'en fonction d'eux. Parmi ces vecteurs, certains sont disposés sur la marge ($\xi_i = 0$) avec $0 < \alpha_i < C$, les autres correspondent à $\xi_i > 0$ avec $\alpha_i = C$. La fonction de décision *(3.16c)* reste inchangée.

3.3.3 Modèles SVM de classification

■ Les modèles SVM de classification permettent d'établir des frontières non linéaires, par généralisation des concepts de classification linéaire dans un espace multidimensionnel appelé espace des caractéristiques.

Le modèle de classification défini précédemment permet de calculer des frontières linéaires optimales, mais l'hypothèse de classification linéaire est trop rigide pour de nombreuses applications où des frontières plus complexes sont requises. Comme pour toute méthode linéaire, on peut introduire de la flexibilité dans la procédure en utilisant une base de fonctions *h(x) (3.19)* (fonctions polynomiales, fonctions *spline*, *etc*.). L'algorithme de classification linéaire est alors appliqué dans l'espace des caractéristiques engendré par cette base *(3.20)*, espace vectoriel noté И généralisation d'un espace euclidien en dimension quelconque.[⊥] Cet espace se définit comme un espace multidimensionnel dans lequel il est possible de représenter un vecteur x_i associé à un échantillon *i*. La frontière implicite dans l'espace de départ des données est alors non linéaire même si la règle de classification induite par *f(x)* est inchangée. Les machines SVM sont une extension de cette idée. On peut toujours projeter dans un espace suffisamment grand pour que les données *y* soient séparables par un hyperplan.

$$h(x_i) = (h_1(x_i), h_2(x_i), \ldots) \qquad \forall\, i = 1, \ldots, n \qquad (3.19)$$

$$\{ x : f(x) = h(x)^T \beta + \beta_0 \} \qquad (3.20)$$

Le problème d'optimisation posé précédemment dans le cas des classes séparables ne fait intervenir les vecteurs descriptifs des échantillons que part l'intermédiaire de produit scalaires[⊥⊥] (voir également annexe 2). Le calcul peut donc être effectué directement sur les vecteurs projetés dans une nouvelle base *h(x)* avec *h* une application non linéaire de $\Re^P$ dans И muni d'un produit scalaire et de plus grande dimension. C'est ce que traduit l'expression *(3.21)* de la forme duale du Lagrangien reprise ici.

$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{N} \alpha_i \alpha_k y_i y_k \langle h(x_i), h(x_k) \rangle \qquad (3.21)$$

Comme précédemment, la solution peut s'écrire sous la forme *(3.22)* dans laquelle le terme *h(x)* n'est considéré que par le biais de produits scalaires. Cela signifie qu'il n'est pas nécessaire d'expliciter la fonction *h* (ce qui serait souvent impossible en grande dimension) pour faire le calcul. La seule condition est de savoir exprimer un produit scalaire dans l'espace *H* à

[⊥] Un espace hilbertien dit auto-reproduisant et isométrique.

[⊥⊥] Noté indifféremment $u^T v$ ou $\langle u, v \rangle$.

l'aide d'une fonction symétrique positive *k* de $\Re^P \times \Re^P$ dans *И*, appelée fonction noyau ou *kernel* et définie ci-dessous par les équations *(3.23)* et *(3.24)*.

$$f(x) = h(x)^T \beta + \beta_0 = \sum_{k=1}^{N} \alpha_i y_i \langle h(x), h(x_k) \rangle + \beta_0 \qquad (3.22)$$

$$k(x, x_k) = \langle h(x), h(x_k) \rangle \qquad (3.23)$$

$$k(x, x_k) = exp\left(\frac{-\|x - x_k\|^2}{2\sigma^2} \right) = exp\left(-G\|x - x_k\|^2 \right) \quad (3.24)$$

La fonction *kernel* permet de matérialiser une notion de voisinage adaptée au problème de discrimination et à la structure des données. Cette fonction calcule les produits scalaires dans l'espace des caractéristiques. Le *kernel* gaussien *(3.24)* qui possède la plupart des propriétés numériques requises (différentiabilité, décroissance rapide à zéro, séparabilité, *etc.*) est le noyau le plus communément utilisé dans la littérature des SVM. Le paramètre *G* contrôle la largeur du *kernel* gaussien centré sur chaque vecteur de support ; sa valeur doit être optimisée lors de l'apprentissage. A l'instar des méthodes de plus proches voisins, les méthodes à noyaux permettent de spécifier la nature du voisinage et le type de fonction de régularisation appliquée localement. Le noyau correspond à un choix *a priori* sur la notion de similarité entre exemples.

L'astuce numérique appliquée ici (*kernel trick*, anglais) permet la conservation des propriétés d'optimisation lors du passage dans un espace de grande dimension. Les fonctions à noyau $k(x, x_k)$ sont des fonctions mathématiques positives symétriques qui correspondent à des produits scalaires. Il n'est donc jamais nécessaire de calculer explicitement les vecteurs $h(x)$ et $h(x_k)$, simplement d'évaluer leur produit scalaire. On clôturera ce paragraphe consacré aux modèles SVM pour la classification en mentionnant l'approche peut être étendue aux classifications multiples, essentiellement en généralisant à la résolution de plusieurs problèmes binaires.

3.3.4 Modèles SVM de régression

■ Les modèles SVM peuvent être adaptés à la régression pour la prédiction de variables quantitatives. Contrairement aux régressions plus classiques qui minimisent des termes quadratiques de l'erreur, les erreurs de prédiction sont prises en compte linéairement et les coefficients de régression trop élevés sont pénalisés.

L'idée principale des SVM en régression est de minimiser les erreurs de prédiction tout en pénalisant l'amplitude des coefficients de régression, les coefficients importants étant sources de variance élevée en grande dimension. L'équation du modèle de régression linéaire est

rappelée en *(3.25)*, la fonction de coût *(3.26)* étant considérée pour l'estimation des poids β (avec λ un paramètre de régularisation à optimiser).

$$f(x) = x^T \beta + \beta_0 \qquad (3.25)$$

$$\underset{(\beta,\beta_0)}{Minimiser} \sum_{i=1}^{n} L(y_i - f(x_i)) + \frac{\lambda}{2}\|\beta\|^2 \qquad (3.26)$$

Cette fonction de coût traduit la particularité des modèles SVM construits en régression : ne pas tenir compte des erreurs faibles (de taille inférieure à ε) et, à l'opposé, pénaliser les erreurs importantes pour diminuer la sensibilité du modèle aux points aberrants. La fonction qui va être implémentée dans le cas des SVM est donc en général de la forme *(3.27)*.

$$\begin{cases} L_\varepsilon(r) = 0 & si\ |r| < \varepsilon, \\ L_\varepsilon(r) = |r| - \varepsilon & sinon \end{cases} \qquad (3.27)$$

Une représentation schématique de cette fonction est proposée (*Fig. 3.9*). En régression SVM, les points qui sont négligés dans le problème d'optimisation sont ceux pour lesquels les résidus sont faibles (par analogie, en classification SVM les points très éloignés de la frontière n'interviennent pas dans le calcul). On retrouve la notion statistique de régression robuste pour laquelle la fonction de coût quadratique usuelle est remplacée par une fonction linéaire pour les points les plus éloignés de la droite d'étalonnage.

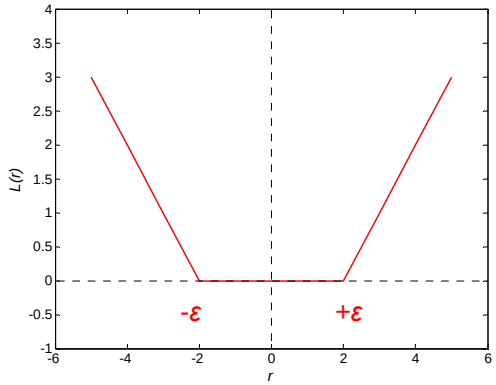


Fig. 3.9 Fonction de coût utilisée par les SVM en régression, $\varepsilon = \pm 2$ pour l'exemple.

Les aspects numériques sont comparables à ceux détaillés en classification. La fonction *(3.27)* peut être réécrite dans le même formalisme *(3.28)*.

$$\underset{(\beta,\beta_0)}{Minimiser} \left\{ \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\left(\xi_i + \xi_i^*\right) \right\} \qquad (3.28a)$$

$$soumis\ à \quad \begin{cases} y_i - \beta x_i - \beta_0 \le \varepsilon + \xi_i \\ -y_i + \beta x_i + \beta_0 \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \qquad (3.28b)$$

Les erreurs de prédiction qui excèdent les valeurs ± ε sont pénalisées linéairement et prises en compte par les variables de relâchement ξ, notées ξ_i (respectivement, ξ_i^*) pour les points situés au dessus (respectivement, en dessous) de la droite de régression. Elles sont caractéristiques de la pénalisation de l'erreur plutôt que de la valeur prédite finale.

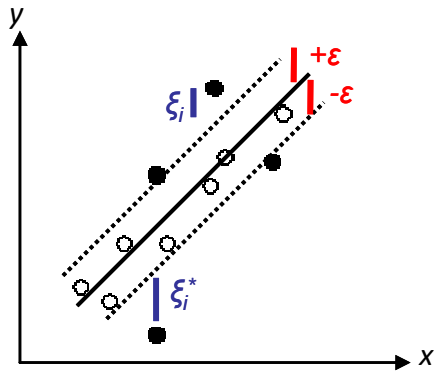


Fig. 3.10 Illustration de la régression SVR (univariée), les points localisés sur ou en dehors de la bande définies par ± ε sont les vecteurs de support.

Une représentation schématique de la régression SVM est proposée sur la figure ci-dessus (*Fig. 3.10*). Si *C* est grand, toute l'attention est portée sur la minimisation de l'erreur, sans se soucier de la taille des coefficients de régression. Si *C* est faible, le meilleur résultat est presque exclusivement dépendant de la taille de ces coefficients. Enfin, le choix de ε est également dépendant du problème pour favoriser des solutions adaptées au traitement de données bruitées. Néanmoins, le choix de trop grandes valeurs de ε peut nuire à la prédiction puisque les vecteurs de support (c'est-à-dire les points pour lesquels les erreurs de prédiction excèdent ± ε) sont alors très éloignés de la droite de régression.

L'équation de la régression peut s'écrire finalement sous la forme *(3.29)* (voir annexe 4), puis (3.30) lorsque le produit scalaire peut être remplacé par des fonctions *kernel*.

$$f(x) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\langle x, x_i^T \rangle + \beta_0 \qquad (3.29)$$

$$f(x) = \sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\langle h(x), h(x_i^T) \rangle + \beta_0 \qquad (3.30a)$$

$$\text{avec} \quad \beta = \frac{1}{2}\sum_{i=1}^{n}(\alpha_i^* - \alpha_i)\, h(x_i) \qquad (3.30b).$$

Les points x_i correspondants à des coefficients de Lagrange non nuls, c'est-à-dire ceux dont l'erreur de prédiction excède ± ε, sont les vecteurs de support.

3.4 Applications des modèles SVM de classification en spectroscopie proche infrarouge

Le choix d'une méthode d'apprentissage est fonction des objectifs analytiques et dépend des données, de leur structure, ou d'aspects numériques. Même si il est difficile d'anticiper les résultats liés à ces choix, les méthodes SVM ont montré leurs performances en généralisation y compris sur des lots d'entrainement de taille limitée relativement à la dimension des données (ce qui est très souvent le cas sur les bases de données des applications industrielles). Un autre aspect important de l'optimisation par des modèles SVM est le caractère déterministe de la solution obtenue pour un ensemble d'apprentissage donné et des valeurs fixées des paramètres. Par ailleurs, l'utilisation de fonctions *kernels* adaptées permet d'ajuster des frontières non linéaires pour suivre la plupart des distributions.

Dans notre communauté, les méthodes SVM sont souvent considérées comme des outils « boîte noire » pour l'optimisation des méta-paramètres et l'interprétation des résultats. Nous proposons de reprendre les résultats obtenus très récemment et de montrer que ces critiques ne sont pas forcément justifiées.

3.4.1 Méthodologie pour l'optimisation des méta-paramètres

Dans le cas de *kernels* gaussiens, l'optimisation des modèles SVM est fonction des valeurs prises par deux paramètres, le paramètre associé à la largeur de la gaussienne (G) et le paramètre de régularisation pour la pénalisation des erreurs (C). L'ajustement se fait par essais successifs sur une grille d'optimisation avec comme critère la minimisation de l'erreur en validation croisée. Pour illustrer l'étendue des solutions proposées lorsque ces paramètres varient, nous reprenons différentes solutions au problème de classification à deux classes en deux dimensions proposé précédemment (voir *Fig. 3.4*).

9-

La Figure (*Fig. 3.11*) permet de visualiser les résultats obtenus pour une grille de 16 nœuds, soient 16 couples de paramètres (C, G). Les valeurs numériques correspondant à chaque modèle sont reportées dans la table (*Tab. 3.1*). Globalement, plus la valeur du paramètre C est grande, plus les erreurs de classification sont pénalisées au profit de l'ajustement de la frontière aux données (y compris aux échantillons atypiques). Cet ajustement s'effectue au détriment de la largeur de la marge et donc potentiellement des capacités de généralisation.

Plus le paramètre G est grand, plus la largeur à mi-hauteur des gaussiennes est faible (plus les gaussiennes sont pointues) et plus la frontière de décision peut être complexe. On notera également que les différents modèles, y compris ceux correspondants à des performances

équivalentes en validation croisée, correspondent à des situations très différentes au regard du nombre de vecteurs de support (points échantillons représentés en gras, *Fig. 3.11*).

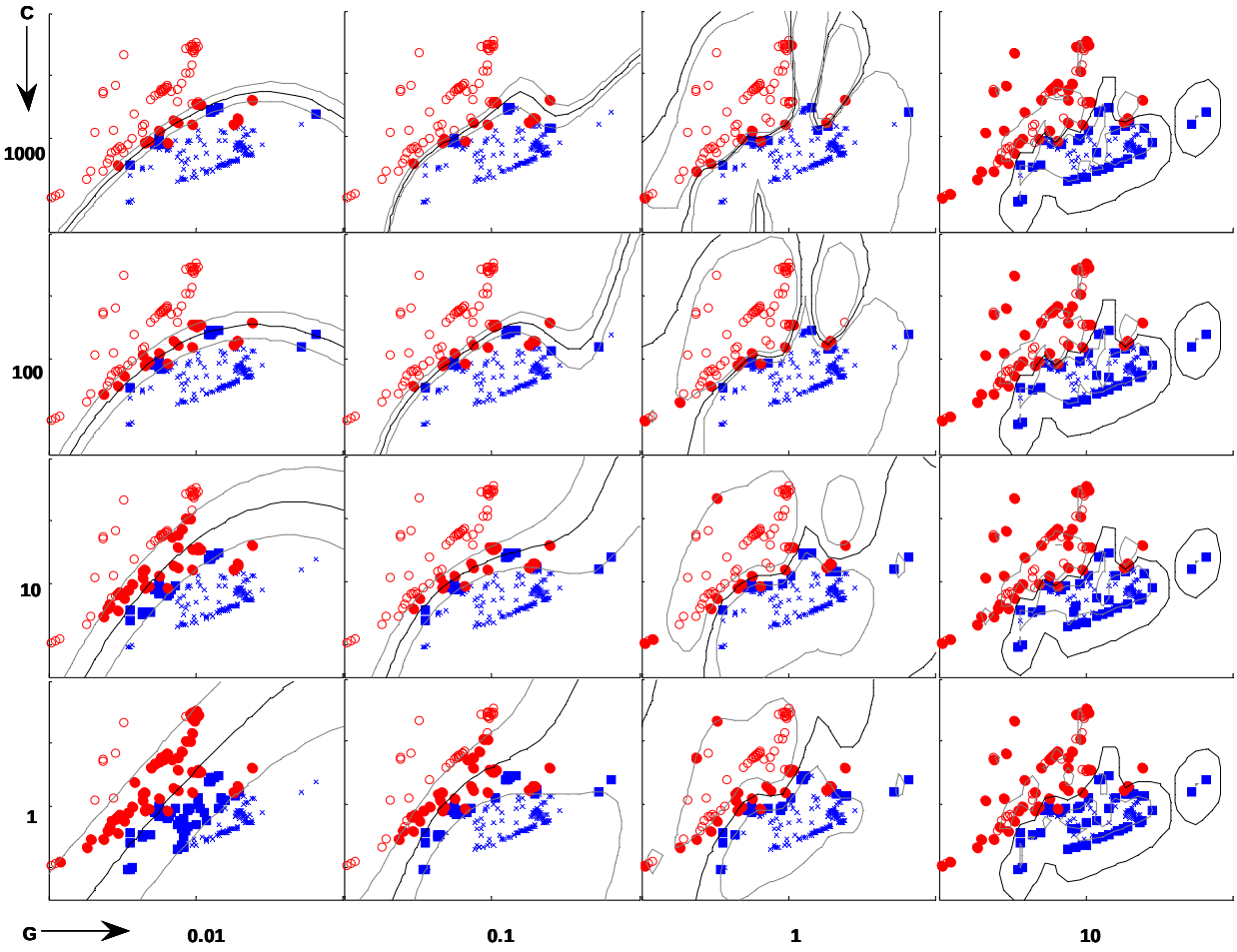


Figure 1. Boundary (black line) and class margins (gray lines) obtained using a RBF kernel and different values of the kernel width (G) and penalty error (C). The support vectors (SVs) are highlighted.

Fig. 3.11 Frontières (traits noirs), marges (traits gris) et points vecteurs de support (en gras) calculés pour différents couples de paramètres (C, G).[70]

Pour les valeurs faibles de *C* et de *G* (par exemple, *C* = 0,01 et *G* = 1, Fig. *3.11*), on retrouve une situation proche de l'analyse discriminante linéaire proposée précédemment (voir *Fig. 3.4*), avec une frontière quasi-linéaire. Dans ce cas, la marge est très large, quasiment tous les points échantillons sont vecteurs de support (et sont donc considérés pour la prédiction) et les fonctions *kernels* centrées sur ces points correspondent à des voisinages très larges, presque uniformes.

C \ G	0.01	0.1	1	10
1000	33 96.8%	23 98.0 %	19 98.8%	67 98.8%
100	48 96.4%	31 97.2%	21 97.6%	68 97.6%
10	78 92.1%	48 95.3%	31 97.2%	71 96.8%
1	110 91.0%	80 92.5%	57 97.2%	90 96.8%

Tab. 3.1 Nombres de vecteurs de support et résultats de classification en validation croisée pour différents couples de paramètres (C, G).[70]

3.4.2 Classification SVM en spectroscopie proche infrarouge

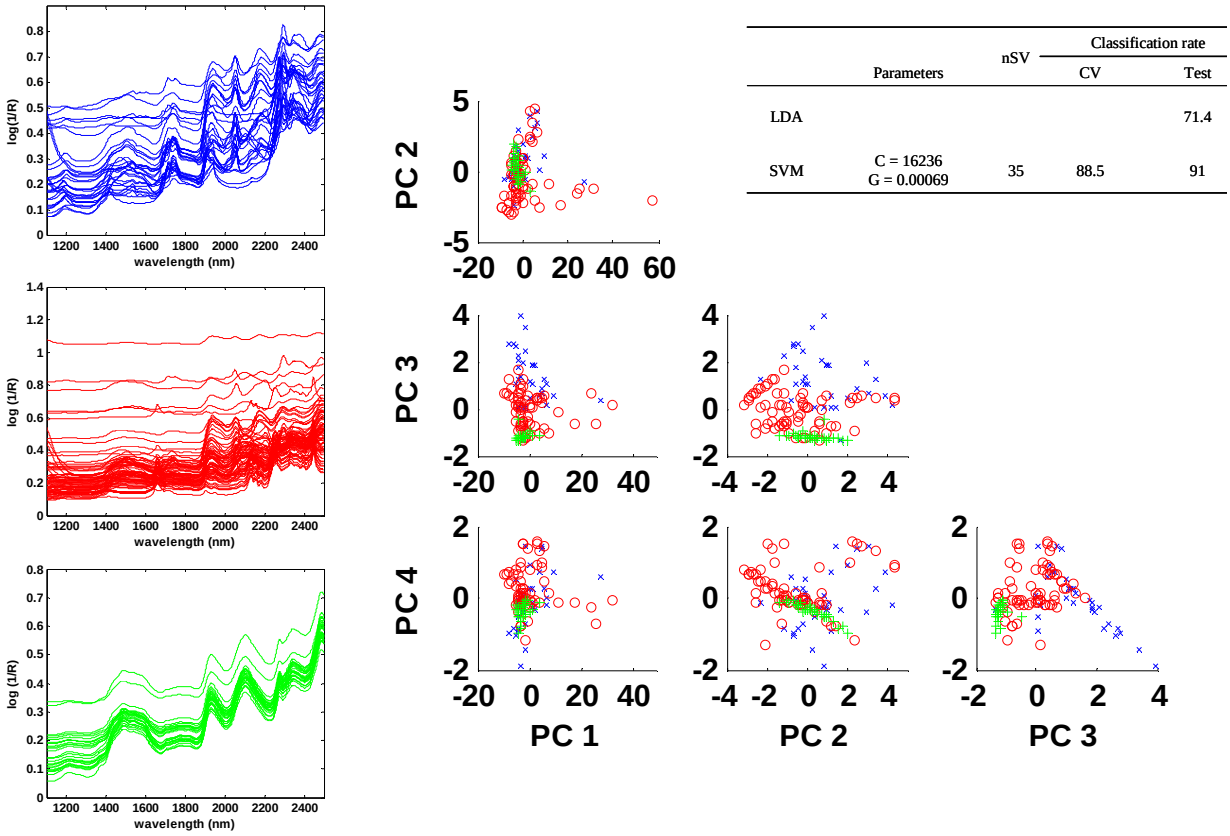
Lorsque que les variables d'entrée sont projetées dans un espace de dimension très élevée, les chances de résoudre le problème de classification non-linéaire augmentent, mais on s'expose également au problème du *fléau de la dimension*. L'idée générale des SVM est donc de restreindre la complexité du modèle relativement à la dimension des données, en cherchant à maximiser la marge. On montre également que si l'hyperplan optimal peut être construit à partir d'un nombre limité de vecteurs support, la capacité en généralisation du modèle sera grande indépendamment de la taille de l'espace. La solution SVM doit donc être la classification la plus parcimonieuse parmi toutes les classifications possibles. Une propriété intéressante supplémentaire concerne le modèle construit qui, au final, est indépendant des échantillons d'entrainement non sélectionnés comme vecteurs de support (ce qui peut représenter la majorité des échantillons).

Comme nous l'avons détaillé précédemment, la formulation duale du problème d'optimisation permet également d'exprimer l'importance de chaque échantillon du lot d'entrainement, les seuls points échantillons pour lesquelles les contraintes associées sont actives ($\alpha_i>0$) étant les vecteurs de support. Pour discuter de l'optimisation et de l'interprétation des modèles de classification SVM, nous reprenons un problème de classification de textiles en spectroscopie proche infrarouge.

10-

On envisage ici la prédiction d'une propriété physique de textiles, de composition et de nature physico-chimique très différentes, à partir du spectre proche infrarouge. Il s'agit d'un problème de discrimination à 3 classes qui est résumé sur la figure (*Fig. 3.12*).

La représentation proposée ensuite (*Fig. 3.13*) permet de visualiser les résultats obtenus pour la grille d'optimisation des paramètres (*C*, *G*) pour l'erreur de validation croisée et le nombre de vecteurs support associé à chaque modèle.



	Parameters	nSV	Classification rate CV	Classification rate Test
LDA				71.4
SVM	C = 16236 G = 0.00069	35	88.5	91

Fig. 3.12 Spectres proche infrarouge caractéristiques des 3 classes, structure des données dans les différents scores plots de l'ACP et résultats des modèles LDA et SVM.[70]

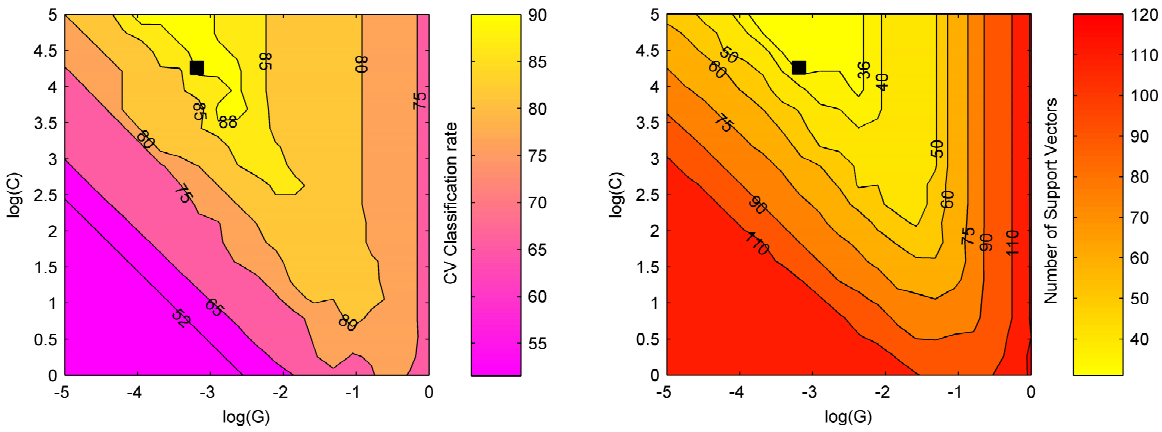


Fig. 3.13 Grille d'optimisation (20×20) des paramètres (C, G) lors de l'apprentissage, les courbes de niveau matérialisent l'erreur en validation croisée (à gauche) et le nombre de vecteurs de support (à droite correspondant aux différents modèles). Le modèle retenu (carré noir) correspond aux paramètres C=16236 et G=0.00069.[71]

Afin d'illustrer graphiquement les résultats pour le modèle retenu, nous proposons ci-dessous (*Fig. 3.14*) de visualiser, dans une représentation des scores de l'ACP, les points échantillons sélectionnés au cours de l'apprentissage ($\alpha_i>0$). Cela permet notamment de se faire une idée de la complexité de la forme de la frontière construite dans un espace de dimension réduite, le sous-espace (PC_2, PC_3) en l'occurrence.

On rappelle ici que l'hyperplan séparateur se caractérise par le fait que le vecteur β s'écrit comme une simple combinaison linéaire des données d'apprentissage *(3.16c)*, la somme étant considérée uniquement sur les échantillons vecteurs de support ; c'est-à-dire les points échantillons qui sont sur la marge, dans la marge ou mal prédits. Les conséquences sont, d'une part, que le nombre de vecteurs de support peut être très petit et, d'autre part, que le modèle de classification ne repose finalement que sur des échantillons déterminés. La classification de données inconnues peut donc être très rapide puisque la seule opération consiste à calculer un produit scalaire. En outre, les points échantillons en question sont directement identifiables par leurs spectres respectifs, ce qui laisse entrevoir des possibilités intéressantes pour l'interprétation, la gestion de redondances ou pour tenter de répondre au problème d'accumulation d'échantillons dans les bases de données.

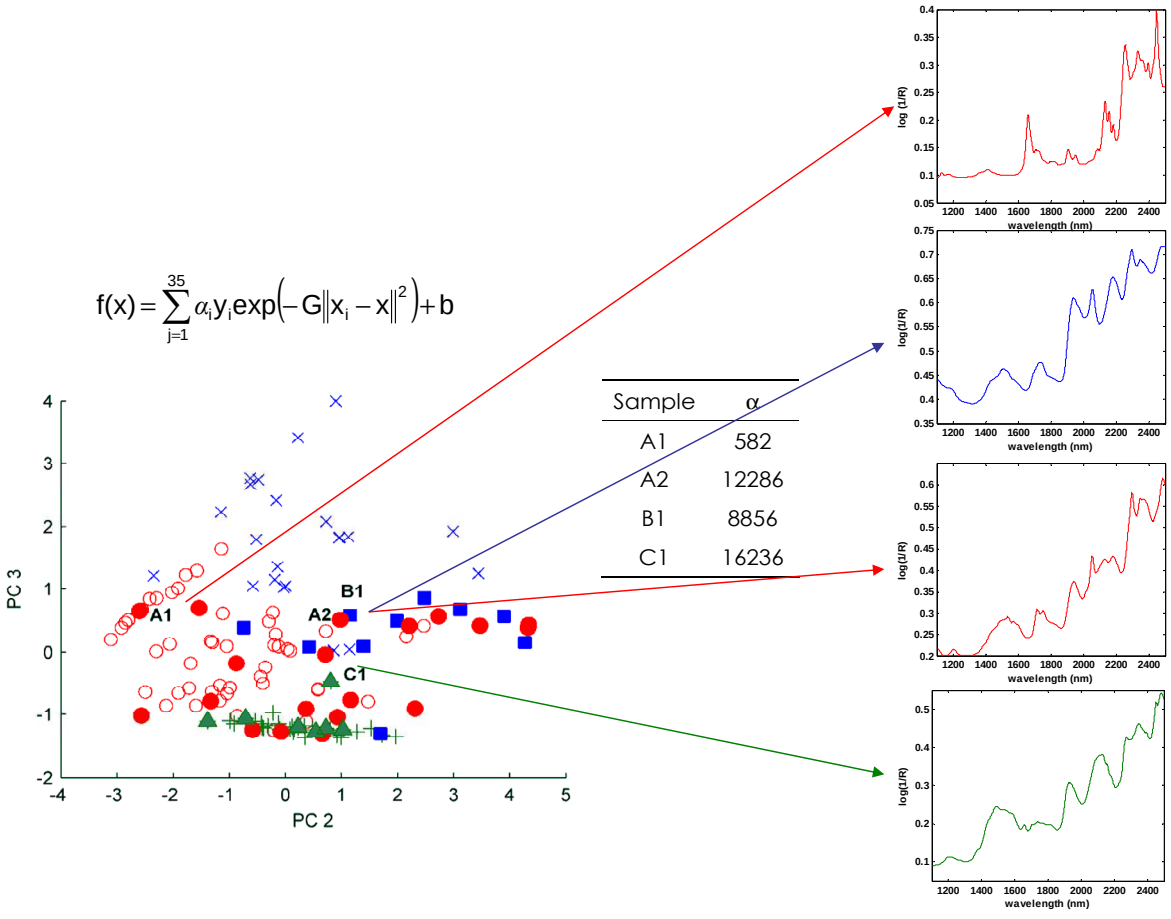


Fig. 3.14 Présentation schématique des résultats d'une classification SVM.[72]

Sur la représentation proposée (*Fig 3.14*), 4 points vecteurs de support particuliers ont été sélectionnés pour la discussion. Ces points correspondent à différentes situations. Le point A1, caractéristique d'un *cluster* contenant des échantillons de même nature chimique, se voit associé un poids relativement faible (α = 582). Il peut être proposé de réduire la densité de l'échantillonnage dans cette zone. Les points A2 et B1 sont deux points importants (α = 12286 et 8856, respectivement), à la frontière de deux classes et correspondant à des échantillons de composition chimique assez proche (voir les spectres correspondants, *Fig. 3.14*). Enfin, le point C1 possède les caractéristiques d'un échantillon mal prédit ($\alpha = C$) et dont la validité peut probablement être remise en cause.

Pour conclure, on notera une amélioration significative des capacités prédictives des modèles construits (91 % de prédictions correctes en généralisation sur un lot de validation externe) par rapport aux approches plus classiques.

3.4.3 Conclusion et orientation des recherches

Les modèles SVM sont actuellement les algorithmes les plus performants en classification supervisé et, très récemment, l'intérêt s'est porté sur des applications où la dimension des données pose problème, en biologie[73] puis en chimie. Néanmoins, au-delà de ces performances reconnues, de la simplicité numérique des modèles SVM et de leur capacité à gérer des données de grande dimension, l'effort méthodologique et pédagogique pour affiner l'optimisation des paramètres d'une part, et soigner l'interprétation d'autre part, doit être poursuivi (si l'on veut convaincre en dehors de notre communauté).

Comme nous l'avons illustré sur des applications récentes en spectroscopie proche infrarouge, le problème revient à construire un vecteur dual dont chaque composante est associée à un échantillon, les contraintes actives correspondant aux vecteurs de support. Les modèles SVM ne reposent finalement que sur ce petit nombre d'échantillons, ce qui peut permettre de mieux comprendre une structure de données, d'identifier les échantillons les plus importants ou les échantillons critiques. Nous pensons tenir ici une idée intéressante pour tenter de répondre au problème d'accumulation d'échantillons dans les bases de données, un échantillon n'étant significatif que s'il modifie le nombre des vecteurs de support, ou leur distribution.

D'autre part, appliqué au problème primal, le même type d'approche (sommant sur les longueurs d'onde pour tous les échantillons) doit permettre d'identifier les variables d'intérêts et d'éliminer les variables auxquelles sont affectées des poids faibles dans l'hyperplan SVM.[74] Ce problème n'a été que très peu exploré jusqu'ici. En fait, il n'est pas d'un très grand intérêt du point de vue des avantages numériques classiques de la sélection de variables, puisque la complexité de l'entrainement d'un modèle SVM n'augmente que linéairement avec le nombre de variables de l'espace de départ (la complexité augmente très fortement avec le nombre d'échantillons, en $O(n^3)$ d'après la littérature). Par contre, pour l'interprétation des modèles, l'identification de longueurs d'ondes ou le développement de capteurs spécifiques, l'approche ouvre incontestablement des perspectives.

En parallèle, sur la base de ces développements en apprentissage statistique, on peut également envisager le développement de fonction de coût *ad hoc* pour la résolution du problème de sélection de l'information, implémentant des critères basés sur la pénalisation du risque empirique, tenant compte explicitement du nombre de variables considérées pour les approches classiques de régression, ou intégrant l'optimisation de méta-paramètres pour les algorithmes génétiques.

4 Conclusion

La chimiométrie est souvent présentée comme une discipline de recherche appliquée et les difficultés rencontrées sont assez proches de celles des disciplines expérimentales. Au delà de la physico-chimie des systèmes étudiés, la chimométrie couvre autant le domaine des mathématiques appliquées que celui de l'instrumentation.

L'objectif principal reste avant tout une quête de sens chimique. C'est dans cette optique que les méthodes orientées données, qui intègrent l'information physico-chimique à la modélisation ou à la résolution, doivent être privilégiées. L'approche mathématique est quant à elle nécessaire pour construire les modèles et interpréter leurs résultats. Les objectifs peuvent également être plus théoriques, l'analyse des données expérimentales amenant alors à proposer des hypothèses et à les tester par le biais de modèles physiques. Enfin et d'un point de vue plus académique, les approches chimiométriques qui combinent plusieurs sources d'information permettent de porter un regard différent sur les problèmes physico-chimiques étudiés. En ce sens, malgré l'existence de nombreux logiciels, la chimiométrie reste un domaine de recherche de spécialistes capables d'interpréter numériquement et chimiquement les solutions proposées.

Dans ce manuscrit, le parti pris a été d'aborder les différentes méthodes d'un point de vue relativement conceptuel. Nous nous sommes focalisés sur les développements méthodologiques récents en théorie de l'apprentissage statistique, à l'origine des machines à vecteur de support, et sur les méthodes multivariées de résolution des systèmes chimiques évolutifs. Ce choix est sous-tendu par la nature des données analysées en spectroscopie moléculaire, par la complexité des systèmes physico-chimiques étudiés et par le dynamisme des recherches dans ces domaines. Les résultats couvrent une partie des recherches effectuées ces cinq dernières années au Lasir et sont la plupart du temps le fruit d'un travail d'équipe, de collaborations académiques et industrielles. Néanmoins, les solutions proposées ou évoquées dans ce travail n'ont à ce jour certainement pas toutes été complètement exploitées. Elles pourront l'être dans les perspectives de travail à court terme. Concernant les travaux très récents, notamment l'orientation vers les spectroscopies résolues en temps pour l'étude des processus photoinduits, les réponses apportées sont encore incomplètes. Les développements attendus sont essentiellement méthodologiques et multidisciplinaires, tirés par les évolutions de l'instrumentation, par la structure des données et par les systèmes étudiés.

Références

[1]W.H. Lawton, E.A. Sylvestre, *Technometrics* **13** (1971) 617-633.
[2]R. Tauler, A. Izquierdo-Ridorsa, E. Casassas, *Chemom. Intell. Lab. Syst.* **18** (1993) 293-300.
[3]P. Paatero, U. Tapper, *Environmetrics* **5** (1994) 111-126.
[4]R. Tauler, *Chemom. Intell. Lab. Syst.* **30** (1995) 133-146.
[5]V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New-York, 1995.
[6]V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, New-York, 1998.
[7]T. Hastie, R. Tibshirani, J. Friedman, *The elements of Statistical Learning*, Springer-Verlag, New-York, 2003.
[8]K. Pearson, *Phil. Mag.* **2** (1901) 559-572.
[9]H. Hotelling, *J. of Educ. Psy* **24** (1933) 417-441.
[10]R.M. Wallace, *J. Phys. Chem.* **64** (1960) 899-901.
[11]R.J. Hanson, C.L. Lawson, *Solving least-squares problems*, Prentice-Hall, N.J. Englewood Cliffs, 1974.
[12]W.H Lawton, E.A Sylvestre, *Technometrics* **13** (1971) 617-663.
[13]R. Tauler, *J. Chemom.* **15** (2001) 627-646.
[14]P. Gemperline, *Anal. Chem.* **71** (1999) 5398-5404.
[15]M. Maeder, *Anal. Chem.* **59** (1987) 527-530.
[16]W. Winding, J. Guilment, *Anal. Chem.* **63** (1991) 1425-1432.
[17]J. Toft, O.M. Kvalheim, *Chemom. Intell. Lab. Syst.* **19** (1993) 65-73.
[18]M. Amrhein, B. Srinivasan, D. Bonvin, M.M. Schumacher, *Chemom. Intell. Lab. Syst.* **33** (1996) 17-33.
[19]C. Ruckebusch, Conférence présentée à *Chimiométrie06*, Paris, 2006.
[20]C. Ruckebuch, L. Duponchel, J.P. Huvenne, J. Saurina, *Anal. Chim. Acta* **515** (2004) 183-190.
[21]S. Bijlsma, A.K. Smilde, *Anal. Chim. Acta* **396** (1999) 231-240.
[22]A. de Juan, M. Maeder, M. Martinez, R. Tauler, *Chemom. Intell. Lab. Syst.* **54** (2000) 123-141.
[23]M. Maeder, A.D. Zuberbuhler, *Anal. Chem.* **62** (1990) 2220-2224.
[24]G. Golub, V. Pereyra, *Inverse Problems* **19** (2003) R1-R26.
[25]A.K. Smilde, H.C.J. Hoefsloot, H.A.L. Kiers, S. Bijlsma, H.F.M. Boelens, *J. Chemom.* **15** (2001) 405-411.
[26]L. Blanchet, C. Ruckebusch, A. Mezzetti, A. de Juan, J. P. Huvenne, *soumise à Anal. Chem.*
[27]P. Paatero, U. Tapper, *Environmetrics* **5** (1994) 111-126.
[28]J. Lu, L. Wu, *J. Chemom.* **18** (2005) 519-525.
[29]F. Guimet, R. Boqué, J. Ferré, *Chemom. Intell. Lab. Syst.* **81** (2006) 93-106.
[30]R. Tauler, *Anal. Chim. Acta* **595** (2007) 289-298.
[31]R.A. Palmer, *Appl. Spectrosc.* **8** (1993) 26-34.
[32]C. Ruckebusch, Conférence présentée à *ICAVS*, Nottingham, 2003.
[33]C. Ruckebusch, L. Duponchel, B. Sombret, J.P. Huvenne, J. Saurina, *J. Chem. Inf. Comput. Sci.* **43** (2003) 1966-1973. Voir p.99.
[34]L. Blanchet,C. Ruckebusch, J.P. Huvenne, A. de Juan, *Chemom. Intell. Lab. Syst.* **89** (2007) 26-35. Voir p.107.
[35]A. Mezzetti, W. Leibl, J. Breton, E. Nabedryk, *FEBS Letters* **537** (2003) 161-165.

[36]L. Blanchet, A. Mezzetti, C. Ruckebusch, A. de Juan, J. P. Huvenne, *Anal. Bioanal. Chem.* **387** (2007) 1863-1873. Voir p.117.
[37]A.K. Dioumaev, *Biophys. Chem.* **67** (1997) 1-25.
[38]I.H.M van Stokkum, *J. Phys. Chem.* **98** (1994) 852-866.
[39]I.H.M. van Stokkum, D.S. Larsen, R. van Grondelle, *Biochim. Biophys. Acta* **1657** (2004) 82-104.
[40]S. Aloise, C. Ruckebusch, L. Blanchet, J. Réhault, G. Buntinx, J.P. Huvenne, *J. Phys. Chem. A* 112 (2008) 24-32. Voir p.129.
[41]C. Ruckebusch, S. Aloise, L. Blanchet, G. Buntinx, J.P. Huvenne, *Chemom. Intell. Lab. Syst. doi:10.1016/j.chemolab.2007.05.007*. Voir p.137.
[42]M. N. Leger, V. M. Montoto, P. D. *Wentzell, Chemom. Intell. Lab. Syst*. **77** (2005) 181-205.
[43]G. D'Agostini, *Bayesian reasoning in data analysis*, World Scientific Publishing, 2003.
[44]H. Chen, B.R. Bakshi, P.K. Goel, *Anal. Chim. Acta* **602** (2007) 1-16.
[45]P.Pernot, *Analyse Bayésienne des données pour la modélisation des signaux spectro-cinétiques, in Réactions Ultrarapides en Solution, Approches Expérimentales et Théoriques*, M. Mostafavi & T. Gustavsson (Eds), CNRS Edition, 2006.
[46]F. Renou, P. Archirel, P. Pernot, B. Lévy, M. Mostafavi, *J. Phys. Chem. A* **108** (2004) 987-995.
[47]B. Soroushian, I. Lampre, J. Bonin, P. Pernot, S. Pommeret, M. Mostafavi, *J. Phys. Chem.* A **110** (2006) 1705-1717.
[48]A. Hyvarinen, E. Oja, *Neural Networks* **13** (2000) 411-430.
[49]M. Pumbley, *IEEE Trans Neural Networks* **14** (2003) 534-543.
[50]S. Moussaoui, D. Brie, A. Mohammad-Djafari, C. Carteret, *IEEE Trans. Signal Processing* **54** (2006) 4133-4145.
[51]S. Moussaoui, D. Brie, A. Mohammad-Djafari, C. Carteret, *Chemom. Intell. Lab. Syst.* **81** (2006) 137-148.
[52]C. Ruckebusch, N. Nedjar-Arroume, S. Magazzeni, J. P. Huvenne, P. Legrand, *J. Mol. Struct.* **478** (1999) 185-191.
[53]C. Ruckebusch, L. Duponchel , J. P. Huvenne, P. Legrand, N. Nedjar-Arroume, B. Lignot, P. Dhulster, D. Guillochon, *Anal.Chim. Acta* **396** (1999) 441-451.
[54]C. Ruckebusch, L. Duponchel , J. P. Huvenne, *Anal.Chim. Acta* **446** (2001) 257-268.
[55]R. Froidevaux, F. Krier, N. Nedjar-Arroume, D. Vercaigne-Marco, E. Kosciarz, C. Ruckebusch, P. Dhulster, D. Guillochon, *FEBS Letters* **491** (2001)159-163.
[56]C. Ruckebusch, B. Sombret, R. Froidevaux, J.P. Huvenne, *Appl. Spectrosc*. **55** (2001) 1610-1617.
[57]L. Dolmotova, C. Ruckebusch, N. Dupuy, J. P. Huvenne, P. Legrand, *Appl. Spectrosc.* **52** (1998) 329-337.
[58]Y. Roggo, L. Duponchel, C. Ruckebusch, J.P. Huvenne, *J. Mol. Struct*. **654** (2003) 253-262.
[59]N. Dupuy N., C. Ruckebusch, L. Duponchel, P. Beurdeley-Saudou, B. Amram, J. P. Huvenne, P. Legrand, *Anal. Chim. Acta* **335** (1996) 79-85.
[60]L. Dolmatova, C. Ruckebusch, N. Dupuy, J. P. Huvenne, P. Legrand, *Chemom. Intell. Lab. Syst.* **36** (1997) 125-140.
[61]C. Ruckebusch, L. Duponchel, J. P. Huvenne, *Chemom. Intell. Lab. Syst.* **62** (2002) 189-198. Voir p.149.
[62]L. Dolmatova, V. Tchistiakov, C. Ruckebusch, N. Dupuy, J. P. Huvenne, P. Legrand, *J. Chem. Inf. and Comput. Sci*. **39** (1999) 1027-1036.
[63]V. Tchistiakov, C. Ruckebusch, L. Duponchel, J. P. Huvenne, *Chemom. Intell. Lab. Syst.* **52** (2000) 93-106. Voir p.159.
[64]L. Duponchel, C. Ruckebusch, J. P. Huvenne, P. Legrand, *J. Mol. Struct.* **480** (1999) 551-556.

[65]L. Duponchel, C. Ruckebusch, J. P. Huvenne, P. Legrand, *J. Near Infrared Spectrosc*. **7** (1999) 155-166.
[66]C.E. Shannon, *The mathematical theory of computation*, Urbana IL, 1949.
[67]A. Durand, C. Ruckebusch, O. Devos, J. P. Huvenne, *Anal. Chim. Acta* **595** (2007) 72-79. Voir p.173.
[68]C. Ruckebusch, F. Ohran, A Durand, T. Boubellouta, J.P. Huvenne, *Appl. Spectrosc*. **60** (2006) 539-544.
[69]A. Durand, *Thèse de doctorat de l'Université des Sciences et Technologies de Lille*, numéro d'ordre 4073, 2007.
[70]O. Devos, Conférence présentée à *FACSS*, Memphis, 2007.
[71]O. Devos, C. Ruckebusch, L. Duponchel, J.P Huvenne, *submitted to J. Chemom.*
[72]C. Ruckebusch, Conférence présentée à *Chimiométrie*, Lyon, 2007.
[73]J.P. Vert, *Kernel methods in computational biology*, HDR, Université Paris 6, 2005.
[74]I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Machine Learning* **46** (2002) 389-422.

Annexes

Annexe 1 Problèmes d'optimisation sous contrainte

Cette annexe pose le problème d'optimisation quadratique sous contrainte, les contraintes limitant les solutions possibles.

Le problème d'optimisation peut être exprimé sous la forme générale suivante :

$$\begin{aligned} &\textit{Minimiser} && f(\boldsymbol{x}) \\ &\textit{Soumis à} && \boldsymbol{a_i} \,.\, \boldsymbol{x} + b_i \geq 0 \quad \text{pour } i = 1,\ldots N, \end{aligned}$$

avec $\boldsymbol{x}$ et $\boldsymbol{a_i}$ des vecteurs de $\Re^m$, b_i dans $\Re$, et f une fonction strictement convexe (dérivable continue).

1er cas. Considérons tout d'abord le problème de minimisation sans contrainte.

$$\begin{aligned} &\textit{Minimiser} && f(\boldsymbol{x}) \end{aligned}$$

La fonction f étant convexe, le point $\boldsymbol{x^*}$ est solution si et seulement si les dérivées partielles de f par rapport aux m coordonnées de $\boldsymbol{x}$ sont nulles, c'est-à-dire si et seulement si $\nabla f(\boldsymbol{x^*}) = 0$, où l'opérateur ∇ représente le gradient de la fonction au point considéré.

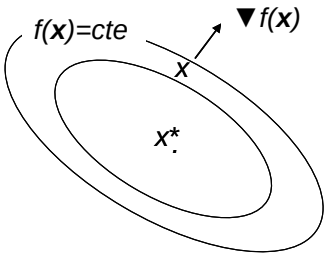


Fig. A1.1 Représentation du gradient d'une fonction convexe.

La figure *Fig. A1.1* permet de rappeler les propriétés du gradient :
- le vecteur gradient pointe dans la direction de variation maximale de f au point $\boldsymbol{x}$;
- le vecteur gradient est orthogonal aux lignes de niveaux ($f(\boldsymbol{x})$ = constante).

2ème cas. Considérons maintenant le problème de minimisation soumis à une contrainte unique.

$$\begin{aligned} &\textit{Minimiser} && f(\boldsymbol{x}) \\ &\textit{sous la contrainte} && \boldsymbol{a} \,.\, \boldsymbol{x} + b \geq 0 \end{aligned}$$

La contrainte appliquée est linéaire, l'ensemble des points $\boldsymbol{x}$ qui la vérifient constitue un demi-espace délimité par l'hyperplan $\boldsymbol{a} \,.\, \boldsymbol{x} + b = 0$ (hyperplan perpendiculaire au vecteur $\boldsymbol{a}$ dirigé vers le sous-espace autorisé, *Fig. A1.2*).

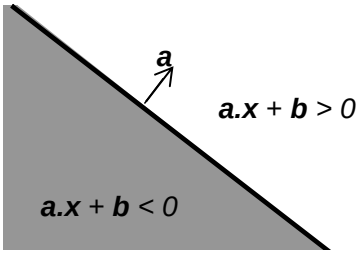


Fig. A1.2 Effet d'une contrainte $\boldsymbol{a} \,.\, \boldsymbol{x} + b = 0$ *(en gris, le demi-espace interdit par la contrainte)*

Notons $\boldsymbol{x}^*$ le point qui minimise $f(x)$ sous la contrainte et $\boldsymbol{x_0}^*$ le point qui minimise la fonction en l'absence de contrainte. Les deux situations possibles sont reprises schématiquement ci-dessous (*Fig. A1.3*), selon la position relative du minimum global $\boldsymbol{x_0}^*$.

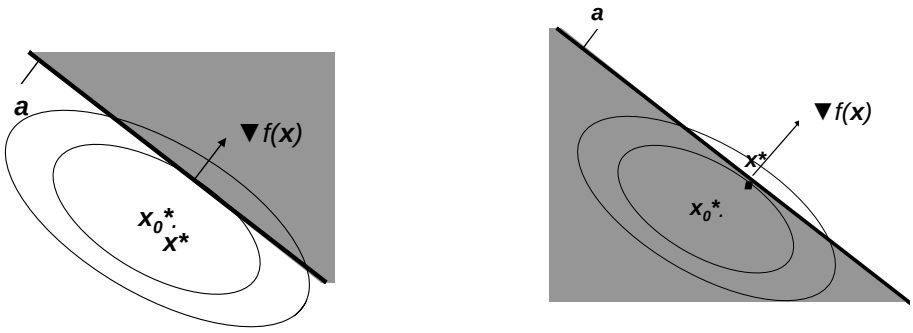


Fig. A1.3 Solutions au problème d'optimisation sous contrainte linéaire en fonction de la position du minimum global de la fonction

- Si $\boldsymbol{x_0}^*$ est dans le demi-espace autorisé, alors $\boldsymbol{x}^* = \boldsymbol{x_0}^*$ et le système suivant caractérise $\boldsymbol{x}^*$:

$$\nabla f(\boldsymbol{x}^*) = 0 \qquad (\boldsymbol{x}^* \textit{ est miminum global})$$

$$\boldsymbol{a} \,.\, \boldsymbol{x}^* + b \geq 0 \quad (\boldsymbol{x}^* \textit{ satisfait les contraintes}).$$

- Si $\boldsymbol{x_0}^*$ est dans le demi-espace interdit, alors $\boldsymbol{x_0}^*$ n'est plus solution du problème sous contrainte et $\boldsymbol{x}^* \neq \boldsymbol{x_0}^*$. La solution au problème d'optimisation sous contrainte est alors :

$$\nabla f(\boldsymbol{x}^*) = \lambda\, \boldsymbol{a},\ \lambda \textit{ positif} \quad \text{(le gradient est parallèle à } \boldsymbol{a})$$

$$\boldsymbol{a} \,.\, \boldsymbol{x}^* + b = 0 \qquad (\boldsymbol{x}^* \textit{ satisfait les contraintes}).$$

Le point $\boldsymbol{x}^*$ est le point sur la frontière de l'hyperplan (ce qui correspond à respecter la minimisation du gradient) pour lequel le gradient est parallèle à $\boldsymbol{a}$, les deux vecteurs étant orientés dans le même sens (le minimum global est dans le demi-espace interdit).

On caractérise donc la solution de ce problème d'optimisation sous contrainte par deux jeux d'équations qui peuvent être regroupés. Le point $\boldsymbol{x}^*$ est alors solution du problème

$$\textit{Minimiser} \qquad f(\boldsymbol{x})$$

$$\textit{sous la contrainte} \qquad \boldsymbol{a} \,.\, \boldsymbol{x} + b \geq 0$$

si et seulement si il existe un nombre réel λ^* tel que la paire ($\boldsymbol{x}^*$, λ^*) vérifie le système suivant (théorème de Kuhn-Tucker) :

$$\nabla f(\boldsymbol{x}) = \lambda \, \boldsymbol{a} \qquad \lambda \textit{ positif}$$

$$\boldsymbol{a} \,.\, \boldsymbol{x} + b \geq 0$$

$$\lambda \geq 0$$

$$\lambda \, (\boldsymbol{a} \,.\, \boldsymbol{x} + b) = 0$$

Généralisation : on peut généraliser au cas d'un problème soumis à plusieurs contraintes :

$$\textit{Minimiser} \qquad f(\boldsymbol{x})$$

$$\textit{sous les contraintes} \qquad a_i \, x_i + b_i \geq 0 \qquad \text{pour } i = 1,\ldots N.$$

Chaque contrainte i définit deux demi-espaces (*Fig. A1.4*) séparés par l'hyperplan $a_i \, x_i^* + b_i \geq 0$. L'intersection de tous les demi-plans autorisés représente l'ensemble des vecteurs qui satisfont la contrainte, et le problème est de trouver parmi ceux-ci le point qui minimise f. On peut monter que les résultats précédents sont généralisables au problème à N contraintes :

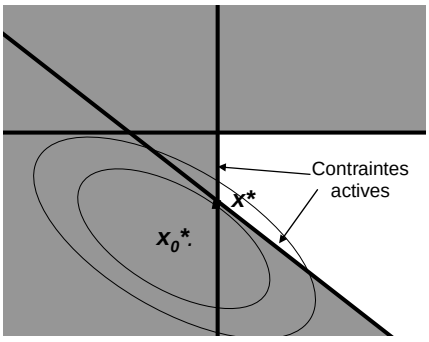


Fig. A1.4 Solution au problème d'optimisation sous contrainte linéaires multiples

Le point $\boldsymbol{x}^*$ est solution du problème

$$\textit{Minimiser} \qquad f(\boldsymbol{x})$$

$$\textit{sous la contrainte } a_i\, x_i + b_i \geq 0 \qquad i=1\ldots N$$

si et seulement si il existe un nombre réel $\lambda_i^* \ldots \lambda_N^*$ tel que la paire $(\boldsymbol{x}^*, \lambda^*)$ vérifie le système suivant (théorème de Kuhn-Tucker) :

$$\nabla f(\boldsymbol{x}) = \lambda_1\, a_1 + \lambda_2\, a_2 + \ldots \lambda_N\, a_N = \Sigma\, \lambda_i\, a_i$$

$$a_i\, x_i + b_i \geq 0 \qquad i=1\ldots N$$

$$\lambda_i \geq 0 \qquad i=1\ldots N$$

$$\lambda_i\, (a_i\, x_i + b_i) = 0 \qquad i=1\ldots N$$

Le théorème précédent caractérise la solution $\boldsymbol{x}^*$ du problème d'optimisation sous contrainte par l'existence d'un vecteur $\boldsymbol{\lambda}$ qui satisfait un ensemble de conditions données.

Pour trouver une paire $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$, on introduit une fonction de Lagrange, fonction de $N+m$ nombre réels, définie par :

$$L(\boldsymbol{x}^*, \lambda^*) = f(\boldsymbol{x}) - \Sigma\, \lambda_i\, (a_i\, x_i + b_i).$$

. La paire $(\boldsymbol{x}^*, \lambda^*)$ qui satisfait les conditions du théorème précédent correspond à un point particulier de cette fonction (le point selle). La valeur de la fonction de Lagrange au point selle correspond au minimum de la fonction *f*.

$$L(\boldsymbol{x}^*, \lambda^*) = f(\boldsymbol{x}^*)$$

L'intérêt des fonctions de Lagrange est qu'elles peuvent en partie être résolues analytiquement par calculs directs. Pour le calcul des $\boldsymbol{\lambda}$, il faut à nouveau recourir à une maximisation sous contrainte. En d'autres termes, on remplace le problème original de minimisation par un autre problème d'optimisation appelé problème dual qui diffère par le fait que :

- l'optimisation est effectuée sur les λ_i, de dimension N, c'est-à-dire le nombre de contraintes (alors que la minimisation de départ est sur les vecteurs $\boldsymbol{x}$ qui ont une dimension m, dimension des observations). Cette formulation est donc particulièrement intéressante lorsque $N < m$, c'est à dire lorsqu'il y peu de contraintes dans une espace de grandes dimensions ;
- les contraintes sur les λ_i sont des contraintes de positivité.

Annexe 2 Solution pour l'hyperplan séparateur de marge maximale (classes séparables)

Pour faciliter sa résolution d'un point de vue numérique, le problème d'optimisation convexe *(3.18)* est reformulé en introduisant les multiplicateurs de Lagrange. Cette méthode consiste à injecter pour chaque contrainte une inconnue scalaire supplémentaire (un multiplicateur de Lagrange). Les contraintes peuvent alors être directement exprimées sur les multiplicateurs de Lagrange. La fonction de coût est alors une combinaison linéaire de la fonction de départ *(3.18a)* et des contraintes *(3.18b)* qui sont les bornes supérieures des valeurs prises par les variables (problème primal). On peut remarquer que les données d'apprentissage ne sont mises en jeu que par l'intermédiaire de produits scalaires ce qui permettra de généraliser aux cas non séparables.

$$L_P(\beta,\beta_0,\alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i y_i \left(x_i^T \beta + \beta_0\right) + \sum_{i=1}^{n} \alpha_i$$

La solution est obtenue en minimisant L_P par rapport à β et β_0 et aux n multiplicateurs de Lagrange α_i ($\alpha_i \geq 0$). En réinjectant les conditions d'annulation des dérivées partielles de L_P dans l'équation, on peut écrire L_D qui correspond à la formulation dite duale du Lagrangien. Cette formulation est particulièrement intéressante ici car elle exprime peu de contraintes dans un espace de grande dimension.

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \ )$$

$$L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{n} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

Trouver la solution consiste à maximiser $L_D(\alpha)$ avec $\alpha_i \geq 0$. La solution, appelée point selle, doit satisfaire les conditions dites de Karush-Kuhn-Tucker

$$\alpha_i \left[y_i (x_i^T \beta + \beta_0) - 1 \right] = 0 \,,\ \alpha_i \geq 0 \qquad \forall i = 1,\ldots,n$$

- Si $\alpha_i > 0$ *alors* $y_i \left(x_i^T \beta + \beta_0\right) = 1 \,, \forall i = 1,\ldots,n$ et les vecteurs x_i pour lesquels la contrainte ci-dessus est active sont sur les hyperplans définissant la marge. Ces vecteurs x_i sont les vecteurs de support.

- Si $\alpha_i = 0$ *alors* $y_i \left(x_i^T \beta + \beta_0\right) > 1 \,, \forall i = 1,\ldots,n$ et les points correspondants les plus proches du plan vérifient l'équation précédente mais sont en dehors de la marge.

Annexe 3 Solution pour l'hyperplan séparateur de marge maximale (classes non séparables)

$$L_P(\beta,\beta_0,\alpha,\mu)=\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left[y_i\left(x_i^T\beta+\beta_0\right)-\left(1-\xi_i\right)\right]-\sum_{i=1}^{n}\mu_i\xi_i$$

Les conditions d'annulation des dérivées partielles du problème primal sont les mêmes que précédemment mais sont soumises aux restrictions.

$$\alpha_i = C-\mu_i \qquad et \qquad \alpha_i,\ \mu_i,\ \xi_i > 0 \qquad \forall\, i$$

L'expression de la forme duale L_D du Lagrangien est identique à celle explicitée précédemment. On cherche le maximum de L_D soumises aux contraintes ci-dessous ainsi qu'à celles imposées par les conditions de Karush-Kuhn-Tucker

$$0 \le \alpha_i \le C$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

$$et$$

$$\alpha_i\left[y_i\left(x_i^T\beta+\beta_0\right)-\left(1-\xi_i\right)\right]=0$$

$$\mu_i\,\xi_i = 0$$

$$y_i\left(x_i^T\beta+\beta_0\right)-\left(1-\xi_i\right)\ge 0$$

Annexe 4 Solution du problème de régression SVM

L'équation posée traduit un problème d'optimisation sous contrainte. Les multiplicateurs de Lagrange α_i* sont calculés numériquement, chacun d'entre-eux étant associé à un exemple du lot de données d'entraînement. Le problème dual consiste à minimiser l'équation ci-dessous soumis aux contraintes.

$$L_D(\alpha) = \varepsilon \sum_{i=1}^{n}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{n} y_i(\alpha_i^* - \alpha_i) + \frac{1}{2}\sum_{i,k=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_k^* - \alpha_k)x_i^T x_k$$

$$0 \leq \alpha_i, \alpha_i^* \leq 1/\lambda$$

$$\sum_{i=1}^{n}(\alpha_i^* - \alpha_i)(\alpha_k^* - \alpha_k)x_i^T x_k$$

$$\alpha_i \alpha_i^* = 0$$

On peut montrer que la fonction solution prend la forme suivante, c'est-à-dire que les coefficients de la régression peuvent s'écrire comme une somme de multiplicateurs de Lagrange multipliés par les données d'apprentissage correspondantes.

$$\beta = \frac{1}{2}\sum_{i=1}^{n}(\alpha_i^* - \alpha_i)x_i$$