Chemical imaging and chemometrics for the analysis of pharmaceutical solid dosage forms

Christelle Gendrin

▶ **To cite this version:**

Christelle Gendrin. Chemical imaging and chemometrics for the analysis of pharmaceutical solid dosage forms. Engineering Sciences [physics]. Université Louis Pasteur - Strasbourg I, 2008. English. NNT : . tel-00341106

HAL Id: tel-00341106
https://theses.hal.science/tel-00341106

Submitted on 24 Nov 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre :

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur

ULP – INSA – ENGEES

THÈSE

présentée pour obtenir le grade de

Docteur de l'Université Louis Pasteur – Strasbourg I
Discipline : Electronique, électrotechnique, automatique
(spécialité traitement d'image)

par

Christelle GENDRIN

Chemical imaging and chemometrics for the analysis of pharmaceutical solid dosage forms

Soutenue publiquement le 13 Novembre 2008

Membres du jury

Directeur de thèse :	M. Ernest Hirsch	Professeur des universités	ULP, Strasbourg
Rapporteur interne :	M. Jihad Zallat	Professeur des universités	ULP, Strasbourg
Rapporteur externe :	M. Dominique Bertrand	Directeur de recherche	INRA, Nantes
Rapporteur externe :	M. Ludovic Duponchel	Professeur des universités	USTL, Lille
Examinateur :	M. Christophe Collet	Professeur des universités	ULP, Strasbourg
Examinateur :	M. Yves Roggo	Docteur	F. Hoffmann-La Roche, Bâle

Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédetection,
UMR-7005 CNRS – ULP
Entreprise pharmaceutique F. Hoffmann-La Roche, Bâle, Suisse

Remerciements

Ce manuscrit, résultat de trois ans de travail, est achevé. Il est temps maintenant de remercier toutes les personnes qui ont accompagné ce travail et ont ainsi contribué, par leur présence, conseils et soutien, à la réussite de ce projet de thèse.

Ce travail a été effectué au sein de laboratoire "Process Robustness Support" de l'entreprise F. Hoffmann-La Roche en collaboration avec le Laboratoire des Sciences de L'image, de l'Informatique et de la Télédetection" (UMR CNRS 7005) de l'université Louis Pasteur de Strasbourg.

J'adresse tout naturellement mes premiers remerciements à Yves Roggo, mon encadrant chez F. Hoffmann- La Roche. Merci pour les conseils, les exigences, qui m'ont poussée à donner le meilleur de moi-même pour ces travaux de recherche. Bien sûr il y a eu des moments plus difficiles où nous n'étions pas d'accord sur la démarche à adopter mais finalement la volonté commune de mener à bien le projet a pris le dessus et au final les résultats sont là !

Ensuite, je remercie Christophe Collet d'avoir bien voulu diriger cette thèse et pour les conseils prodigués tout au long de ces trois ans. Je remercie également Ernest Hirsch d'avoir pris le relais quand cela fut nécessaire.

Cette thèse n'aurait pas été réalisable sans la mise à disposition des moyens financiers et matériels, aussi je remercie Dr. Rolf Altermatt, responsable du contrôle qualité des formes solides, Dr. Anton Fischer responsable du laboratoire PRS, et aussi Dr. Michel Ulmschneider.

Je remercie également Dominique Bertrand, Ludovic Duponchel et Jihad Zallat d'avoir accepté d'être rapporteurs de ce travail.

Des scientifiques de divers disciplines ont pris de leur temps afin de m'expliquer méthodes, algorithmes ou encore procédés. Prof. Pentti Paatero de l'université de Helsinki en fait parti. Je le remercie tout particulièrement pour son accueil chaleureux chez lui, à Helsinki et son aide précieuse pour comprendre les bases de la méthode PMF. Je remercie également sa femme qui nous a préparé de bonnes spécialités de là-bas pour pouvoir poursuivre nos réflexions tout au long de la journée. Puis, je remercie Dr. Anni Pabst, pour ses explications relatives au

développement de nouvelles formes pharmaceutiques, pour son intérêt des nouvelles méthodes d'analyses telles que l'imagerie chimique, et pour les corrections de certaines publications.

Enfin, mon projet de thèse aurait été beaucoup plus fastidieux si l'ambiance de travail n'avait pas été détendue. Je me dois donc de remercier ici tous mes collègues du laboratoire avec lesquels les moments autour d'une pause café ont été des plus agréables. Je remercie donc tout d'abord les personnes présentes à mes débuts : Aurélie et Nadine qui m'ont accueillie au sein du laboratoire et orientée dans les premières semaines et aussi Hélène, stagiaire en même temps que moi. Puis, Merlinda et Caroline pour leur gentillesse et aussi leur disponibilité quand j'avais besoin de mesures supplémentaires pour une étude ou l'autre. Klara pour les discussions "philosophiques" et parties de badminton et enfin, Matthieu, pour les discussions "moins" philosophiques. Merci aussi aux thésards du laboratoire, Pascal et Lene pour les échanges scientifiques et surtout Carmen qui m'a apporté soutien et conseils. Je remercie aussi Jérôme, qui a participé à mon travail en produisant les comprimés nécessaires pour la quantification, pour son travail sérieux et puis sa bonne humeur.

Je voudrais encore remercier toutes les personnes que j'ai croisées chez F. Hoffmann-La Roche et qui m'ont encouragée. Je pense particulièrement à nos collègues du laboratoire infrarouge : André Bubendorf qui nous fait part de sa grande expérience, Monira Siam avec qui j'ai travaillé sur certains projets et qui m'a donné quelques contacts pour la suite, Claire pour nos bavardages intarissables sur la Chine.

Bien sûr je remercie toute ma famille pour le soutien moral.

Enfin et pardessus tout je remercie Alexandre, mon fiancé, qui est là, à mes côtés, me soutient m'encourage, me supporte, me gronde, m'apporte joie et bonheur, me rend la vie plus facile et plus rose.

Résumé français

1. Contexte de la thèse

Des systèmes d'imagerie hyperspectrale ont été mis sur le marché pour l'analyse chimique microscopique et macroscopique. Constitués d'un détecteur plan présentant une matrice de pixels, ou d'un détecteur point lié à une plateforme mobile, ils peuvent acquérir des spectres localisés spatialement (figure 1). Les données générées ont donc deux dimensions spatiales et une dimension spectrale, formant ainsi un cube de données. Elles permettent à la fois l'identification des constituants des échantillons et la visualisation de leur distribution spatiale. L'industrie pharmaceutique s'est vivement intéressée à ce nouveau mode d'analyse chimique. En effet, l'homogénéité des différents constituants est essentielle pour la qualité globale des comprimés pharmaceutiques. Par exemple, un mélange non homogène peut engendrer des problèmes au niveau de la production des comprimés comme une adhérence de la poudre sur les poinçons de la presse, ou produire des comprimés n'ayant pas la quantité nominale en Principe Actif (PA), la molécule soignant la maladie, ou encore amener des différences de dissolution et donc influencer l'efficacité du principe actif. De plus, depuis 2002, l'administration américaine imposant les normes de qualité sur les produits pharmaceutiques et alimentaires, La *Food and Drug Administration* (FDA), promeut la mise en place d'outils de suivi de la production et de contrôle qualité tout au long de la chaîne de fabrication, lançant ainsi l'initiative PAT (*Process Analytical Technology*). Afin d'améliorer la qualité de leur production et de diminuer le nombre de lots rejetés, les industries pharmaceutiques suivent aujourd'hui ces instructions et tendent à implémenter des nouveaux outils de contrôle. Pour ces raisons, les systèmes d'imagerie chimique non destructifs ont toute leur place dans l'industrie pharmaceutique.

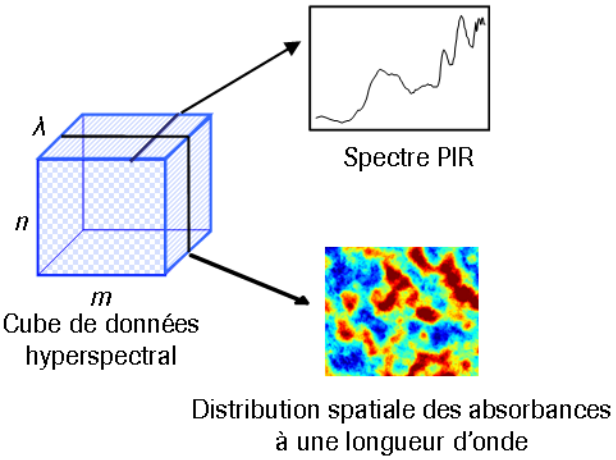


Figure 1. Schématisation du cube de données acquis par imagerie proche infrarouge.

2. Imagerie proche infrarouge

La spectroscopie proche infrarouge (SPIR) fait partie, avec la spectroscopie infrarouge et Raman de la spectroscopie dite vibrationnelle apparaissant entre 750 nm et 1000 µm. Comme son nom l'indique, la spectroscopie vibrationnelle permet la caractérisation d'un échantillon par la détection des fréquences de vibrations des molécules. L'avantage de la spectroscopie proche infrarouge est tout d'abord une mesure non destructive, de plus les échantillons ne requièrent pas de préparation. L'inconvénient est la difficulté de l'interprétation spectrale et la nécessité d'utiliser des méthodes mathématiques et statistiques, ou également appelées méthodes chimiométriques. De plus, un cube de données peut contenir plusieurs milliers de spectres suivant la taille du détecteur. Il faut donc des outils algorithmiques permettant de réduire ces données à une information utile comme la carte de distributions des composés, et des éléments quantitatifs caractérisant l'échantillon (homogénéité, taille de particules, concentration des différentes espèces chimiques).

2.1 Instrumentation

Deux principales techniques d'acquisitions sont utilisées pour obtenir des images hyperspectrales et générer les cubes de données. La plus ancienne et encore couramment utilisée est la technique dite du « mapping ». Chaque spectre est acquis individuellement, les positions spatiales pour les points de mesures étant préalablement définies. L'échantillon est fixé sur une table se déplaçant suivant les directions X et Y. Le principal inconvénient de cette technique est la durée d'acquisition qui peut atteindre plusieurs heures suivant la surface à analyser et la résolution spatiale choisie. De nouveaux détecteurs ont alors été développés depuis une dizaine d'années pour rendre possible l'acquisition simultanée de plusieurs

spectres sur une surface. Ils sont constitués d'une matrice de pixels, chaque pixel agissant comme un simple détecteur point.

Que ce soit point ou plan, ces détecteurs peuvent être alors couplés à tous types de spectromètres classiques. De plus, différents objectifs sont disponibles permettant des analyses micro ou macro métriques. Le spectromètre PIR ayant été utilisé tout au long de cette thèse est le Sapphire (Malvern®). Il est schématisé figure 2. C'est un spectromètre équipé d'un détecteur plan de 256*320 pixels permettant ainsi l'acquisition de 81920 spectres simultanément. L'échantillon est éclairé par des lampes au Tungstène et un filtre à cristaux liquides (LCTF) sélectionne alors les longueurs d'ondes. Les objectifs permettent d'analyser une surface comprise entre $2.3 \times 2.8 mm^2$ et $32.8 \times 40.9\ mm^2$. Il est ainsi possible d'imager une partie d'un comprimé pour une étude microscopique ou plusieurs comprimés en même temps pour faciliter leur comparaison à une échelle macroscopique. Le temps d'acquisition est de quelques minutes suivant les paramètres d'acquisition.

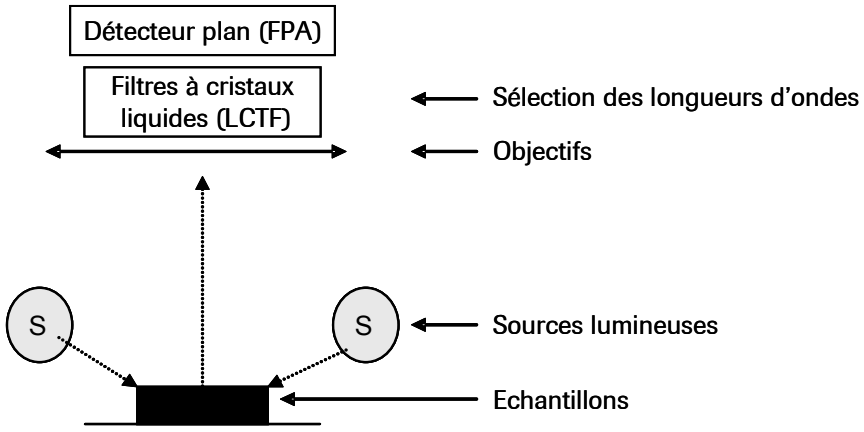


Figure 2. Schématisation du spectromètre utilisé.

2.2 Chimiométrie pour l'analyse de données

Tout d'abord, les spectres sont prétraités pour réduire la ligne de base et les effets de surface. Après le prétraitement des spectres une étape essentielle dans l'analyse de données est leur réduction afin d'extraire les cartes de distributions des composés. Parmi ces méthodes d'extraction, on peut citer la méthode la plus intuitive qui est l'analyse à une longueur d'onde. Elle s'effectue comme suit : les spectres des poudres de références sont tout d'abord acquis afin de trouver les longueurs d'ondes caractéristiques de chaque composé. On choisira alors dans le cube de données les images acquises à ces dernières pour afficher la distribution de ces composés. Cependant, un des inconvénients de la SPIR est la superposition des bandes caractéristiques et il est parfois difficile de trouver une longueur d'onde caractéristique pour chaque espèce chimique. D'autres méthodes d'analyse dites multivariées, qui prennent en compte toute l'information spectrale, ont alors été développées. Deux principales familles

peuvent être distinguées : les méthodes factorielles qui réduisent la dimension des données à quelques facteurs d'intérêts, et les méthodes de classification regroupant les spectres par similitude. Parmi les premières, deux méthodologies sont principalement utilisées : l'Analyse en Composante Principale (ACP) qui réduit les données par maximisation de la variance et l'algorithme des moindres carrés partiels (en anglais PLS : *Partial Least Squares*) pour la quantification. Ce dernier requiert la construction d'un modèle mathématique utilisant un lot de données dont la valeur à quantifier est connue (la plupart du temps la concentration d'un élément). Le modèle doit alors être validé grâce à un nouveau lot d'échantillons avant de pouvoir être utilisé pour prédire de nouvelles concentrations. Parallèlement aux méthodologies classiques, d'autres algorithmes ont été développés en chimie analytique pour extraire des spectres de mélange les informations de concentration des différents constituants ainsi que leur signature spectrale. Ces méthodes sont fondées sur un modèle bilinéaire supposant une additivité pondérée des absorbances de chaque produit pur pour former l'absorbance globale de l'échantillon. La factorisation de la matrice des spectres de mélange en deux sous-matrices de concentration et d'information spectrale n'est pas immédiate et possède une infinité de solutions. L'application de contraintes est nécessaire pour pouvoir en réduire le nombre. La contrainte la plus couramment appliquée est celle de la positivité des spectres et des concentrations. L'algorithme de décomposition le plus fréquemment utilisé en chimie analytique repose sur une extraction alternée de la matrice des concentrations et des spectre purs par moindre carré suivi de l'application des contraintes, il s'agit de la méthode MCR-ALS (*Multivariate Curve Resolution-Alternating Least Squares*). Cependant, dans d'autres domaines scientifiques, d'autres algorithmes pour la factorisation positive ont été proposés, comme par exemple la méthode BPSS (*Bayesian Positive Source Separation*), NMF (*Non-negative Matrix Factorization*) en imagerie et PMF (*Positive Matrix Factorization*) pour l'étude de la pollution de l'atmosphère.

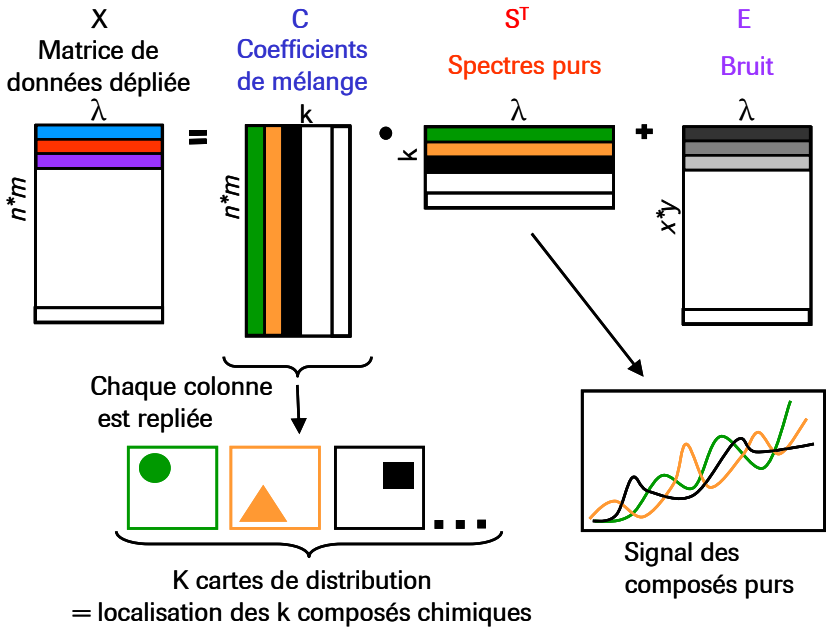


Figure 3. Analyse des données par déconvolution des spectres de mélange.

Enfin, la dernière étape de traitement consiste à extraire, à partir de ces cartes, les caractéristiques permettant de quantifier les différences, comme par exemple la taille des particules, ou encore l'analyse statistique des histogrammes qui représentent la distribution des intensités des pixels.

3. Objectif de la thèse et travail effectué

L'objectif de cette thèse est donc de mettre en place une méthodologie globale du traitement des données hyperspectrales en comparant les avantages et inconvénients de ces différentes approches pour l'analyse de formes pharmaceutiques solides par imagerie proche infrarouge.

3.1 Extraction des cartes des composés et leur caractérisation

Suivant les informations disponibles sur le comprimé, différentes méthodes d'analyses ont été utilisées. Si l'échantillon est bien caractérisé, c'est-à-dire que la nature, le nombre et la concentration des constituants sont connus, comme c'est souvent le cas pour les produits pharmaceutiques, alors une simple analyse à une longueur d'onde caractéristique des composés, ou l'extraction des cartes en utilisant une régression des moindres carrés (CLS: Classical Least squares) sur les spectres purs peuvent s'avérer suffisantes. Nous avons étudié le cas avec un produit en phase de développement. Un intermédiaire de production, les extrudas, constitués d'un excipient (polymère) et d'un actif, et les comprimés finaux contenant cinq excipients et un actif, sont analysés. Quatre lots sont produits avec différents paramètres, conduisant à des intermédiaires et des comprimés finaux de qualités différentes. Après

prétraitement du cube de données, les extrudas sont analysés à une longueur d'onde spécifique du polymère : 2260 nm (Figure 4). L'analyse des histogrammes révèle que l'image des extrudas du lot B est plus contrastée. Le lot B présente une inhomogénéité plus importante que les autres lots.

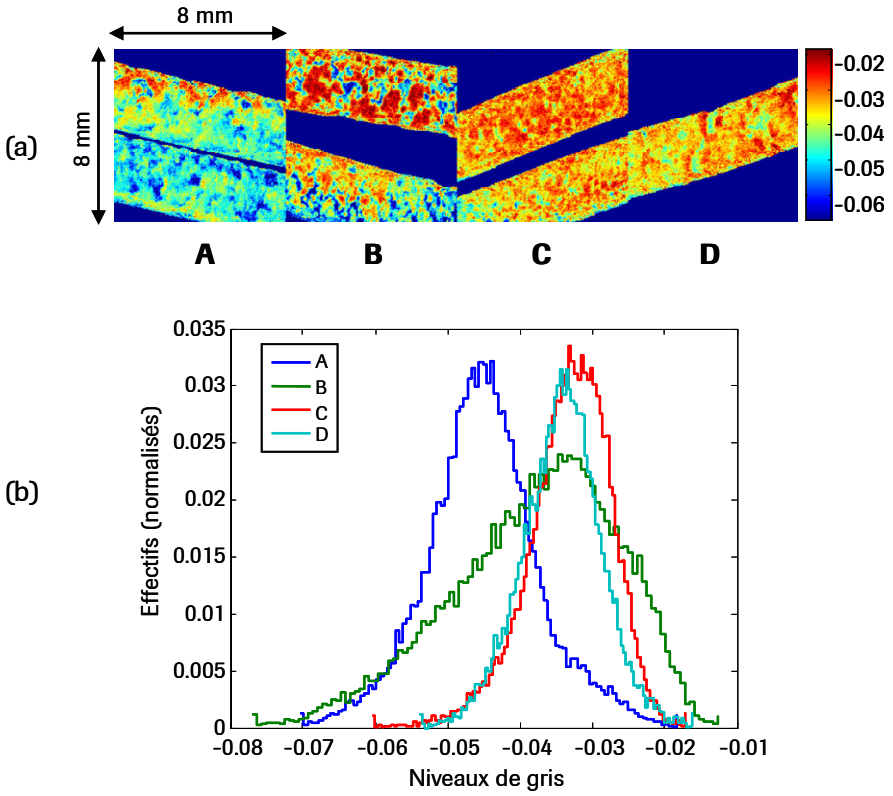


Figure 4. Résultats de l'analyse des extrudas. (a) Images à la longueur d'onde 2260 nm et (b) leur histogramme correspondant.

En ce qui concerne les comprimés, les cartes de distributions sont extraites par une régression des moindres carrés (CLS). Ces cartes apparaissent homogènes mais révèlent une différence de la taille des granulats d'actifs entre les différent lots. Une méthode de traitement d'image afin de détecter de manière automatique la différence de taille des particules est proposée. Elle consiste en une première segmentation des particules par l'algorithme d'Otsu suivi d'un affinement par morphologie mathématique et algorithme dit des *Watershed* qui, grâce à la prise en compte de la topologie de l'image, permet de séparer les particules qui apparaissent accolées à la suite de la première binarisation (figure 5). Une analyse statistique par ANOVA révèle une différence significative des tailles de particules entre les lots. Cependant, la non connaissance de la taille réelle des particules ne permet pas de conclure quand à la précision de leur estimation.

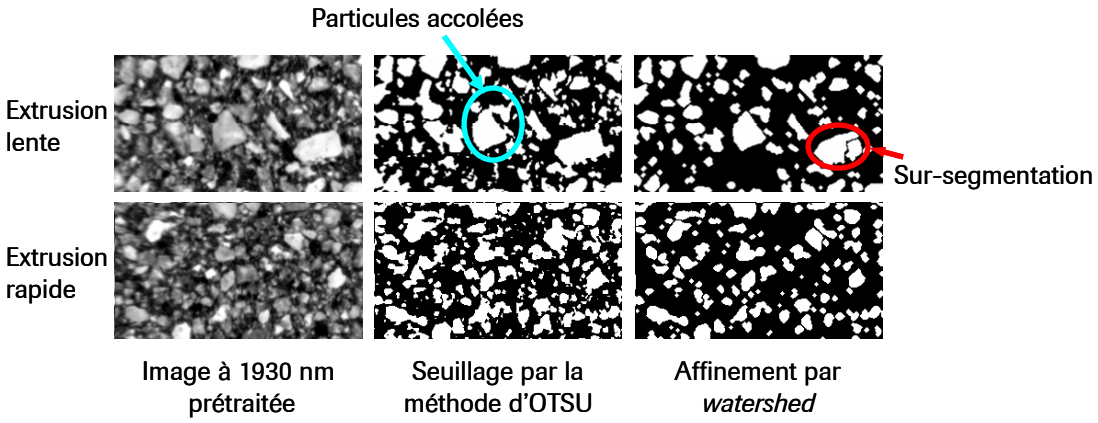


Figure 5. Segmentation des images à une longueur d'onde par seuillage suivi ou non d'un affinement par l'algorithme des "*watershed*".

Cette première étude sur les comprimés issus du développement démontre l'utilité de l'imagerie proche infrarouge pour le développement de nouvelles formulations.
Concernant également l'extraction des cartes de distribution, il est possible qu' un ou plusieurs constituants ne soient pas connus, en cas de contamination ou de contrefaçons par exemple. Il est alors nécessaire d'employer les méthodes qui extraient à la fois l'information de concentration, et l'information spectrale. Une étude bibliographique montre que jusqu'alors elles n'ont pas été utilisées sur des données d'imagerie proche infrarouge. Une comparaison des méthodes de déconvolution a donc été conduite. Un échantillon bien caractérisé comportant trois entités chimiques principales (API : 5%, cellulose 50% et lactose 45%) afin de permettre une comparaison quantitative a été utilisé. De plus, une méthode innovante permettant l'exploration du domaine de solutions en utilisant les outils du logiciel ME2 implémentant la méthode PMF a été proposée. L'étude montre que si la méthode PMF requiert le plus grand nombre de paramètres à optimiser, elle permet la meilleure extraction de l'information spectrale grâce aux méthodes d'exploration du domaine de solutions (figure 6). Les coefficients de corrélation finaux sont supérieurs à 0.93, avec néanmoins une distorsion importante pour le signal de l'actif, et les concentrations finales proches des concentrations réelles (table 1).

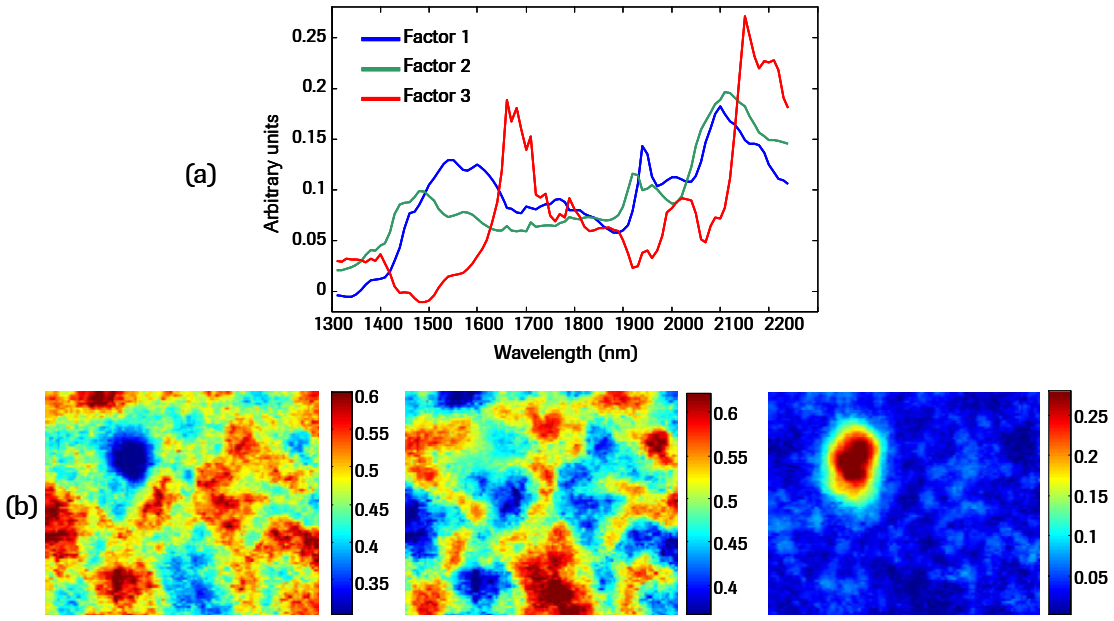


Figure 6. Spectres (a) et cartes de distribution (b) extraits par la méthode PMF et l'application d'une matrice de rotation.

	Corr.	S.D.R.	Conc.
API	0.934	11.1	4.6
Cellulose	0.979	18.73	48.8
Lactose	0.984	20.79	46.6

Table 1 : Résultats finaux obtenus par PMF

3.2 Analyse quantitative des comprimés par imagerie

Apres l'extraction des cartes de distributions des composés et leur caractérisation, l'étude s'est portée sur la quantification du principe actif par imagerie. La teneur en principe actif est un paramètre important pour la qualité du médicament. Sous dosé le comprimé est moins efficace pour soigner le patient ; surdosé il peut être dangereux pour la santé.

Tout d'abord l'analyse a été conduite sur des échantillons binaires contenant un principe actif-(PA) et de la cellulose ; la teneur en PA variant de 0 à 100% par pas de 10%. Pour chaque mélange, 4 échantillons sont comprimés et analysés. Deux méthodes de prétraitement sont comparées: la *Standard Normal Variate* (SNV) et la SNV suivie d'une dérivée seconde.

Idéalement, une méthode de quantification sans a priori permettrait d'avoir des analyses rapides et objectives puisqu'elle ne nécessiterait aucune intervention de l'utilisateur. Les algorithmes de déconvolution MCR-ALS et PMF sont donc testés ainsi qu'une méthode de classification : les k-moyennes. Le pourcentage de pixels appartenant à la classe liée à l'actif

estimerait sa concentration. Les résultats démontrent que la quantification sans apriori n'est pas évidente et ce quelques soient les prétraitements utilisés. Les mélanges PA-cellulose sont constitués de poudres micronisées dont les tailles de particule sont inférieures à la résolution spatiale de l'appareil. Chaque pixel contient donc une information spectrale similaire et la méthode de classification ne peut extraire des classes relatives aux espèces chimiques. Les méthodes de déconvolution quant à elles ne permettent pas l'extraction des spectres de référence à cause des ambigüités rotationnelles. Afin de tester l'influence de la taille des particules une autre gamme d'échantillons binaires contenant du saccharose (taille de particules de l'ordre du micromètre), à la place du PA a été construite. Sur ces données, les deux classes extraites par les k-moyennes sont globalement liées au sucrose et au cellulose (sauf pour les concentrations supérieures à 80% de sucrose où les fines particules de cellulose ne sont plus discriminées). Si les images sont correctement segmentées, en revanche la taille importante des particules pose un problème d'échantillonnage : la surface analysée du comprimé ne rend plus compte de la concentration globale du mélange initial. Les valeurs de références ne sont plus valables. Il apparaît donc clairement que la quantification sans a priori n'est possible que dans des cas limités : lorsque les poudres ne sont pas micronisées (ou lorsqu'elles s'agglomèrent) et que la surface d'analyse est représentative de l'échantillon, ce qui réduit les cas d'applications.

La quantification sans a priori étant difficile, l'utilisation des spectres de référence apparaît comme une possible alternative. En effet, la formulation des comprimés produits est toujours connue et les poudres pures sont disponibles. Les méthodes CLS, PLS-DA peuvent être employées ainsi que MCR-ALS avec augmentation de matrice. Pour cette dernière deux approches sont testées: (1) les premiers facteurs sont fixés égaux aux spectres de références et un facteur complémentaire est libre, (2) plusieurs spectres de références sont ajoutés à la matrice de donnée **X** initiale et sont fixés pendant les itérations. Ces différents algorithmes donnent une précision de quantification similaire sur les spectres dérivés avec une SEP de 7.40%, une pente de 0.97, une ordonnée à l'origine de -4.7. et un coefficient de détermination (carré du coefficient de corrélation) de 0.98 (Figure 7). Pour comparaison, l'algorithme PLS sur les spectres moyens par image donne un SEP de 3% et une bonne linéarité (ordonnée à l'origine de 0.5%, pente de 0.99 et coefficient de détermination de 0.99). Ainsi, sans être aussi précises que la méthode PLS, les approches utilisant les spectres de référence peuvent donner une approximation de la concentration. L'avantage de l'imagerie est la possibilité d'utiliser les prédictions en chaque pixel pour obtenir les cartes de distributions des composés (Figure 8).

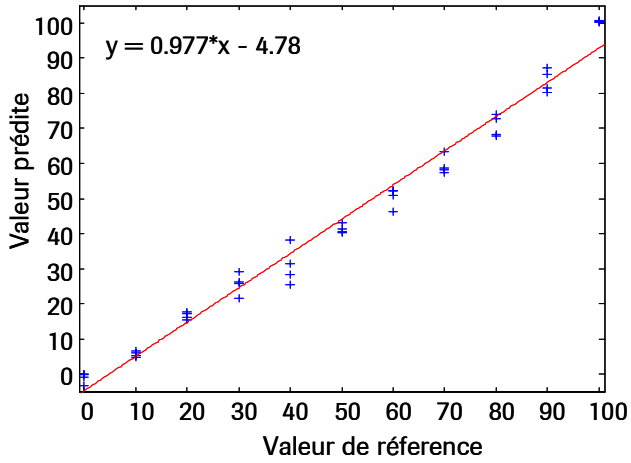


Figure 7. Droite d'étalonnage obtenue sur les spectres dérivés avec un modèle PLS-DA utilisant deux variables latentes

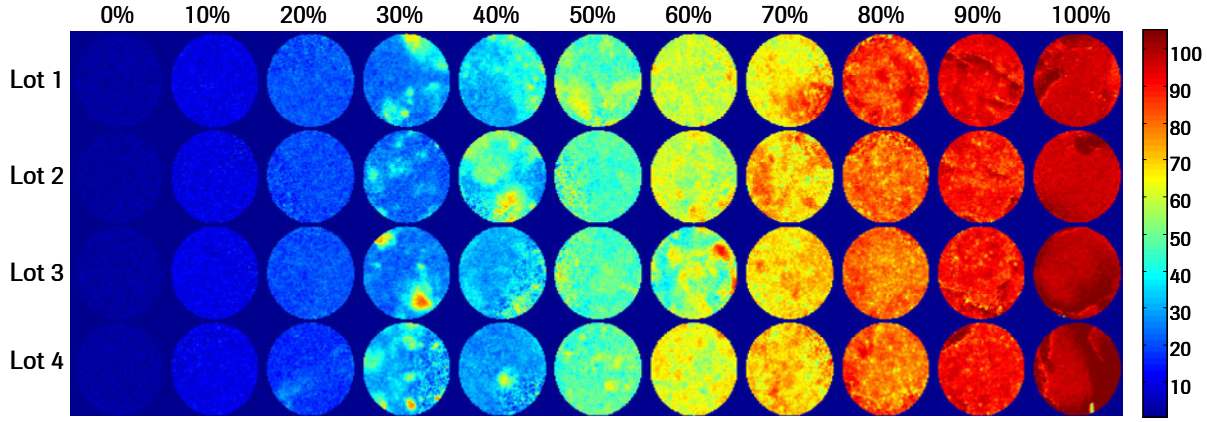


Figure 8. Cartes de distribution du principe actif des comprimés binaires PA-cellulose extraites par régression PLS-DA.

La troisième gamme d'échantillons est basée sur une formulation pharmaceutique réelle comprenant trois constituants majoritaires, deux constituants minoritaires et deux colorants. La concentration en principe actif varie par pas de 1% de 1 à 10 %. D'après les résultats obtenus avec les mélanges binaires, seuls les algorithmes avec a priori sont envisagés pour la quantification des comprimés pharmaceutiques car les poudres utilisées sont micronisées. Puisque les comprimés pharmaceutiques contiennent deux composés minoritaires, deux approches sont utilisées pour l'utilisation des algorithmes CLS et PLS-DA: les spectres de références des espèces minoritaires sont inclus ou non dans les modèles. Les meilleures prédictions sont obtenues avec un modèle PLS-DA incluant les spectres des composés majoritaires et construit avec 5 variables latentes. La SEP est alors de 0.62%, et la droite d'étalonnage présente une pente de 0.93 ainsi qu'une ordonnée à l'origine de 0.11 % (Figure 9). Des agglomérats d'actif sont révélés par les cartes de distributions (figure 10). L'algorithme

CLS donne des estimations négatives pour les premières concentrations (0%, 1% et 2% de PA) et est moins robuste au bruit. L'algorithme MCR-ALS augmenté évite les valeurs négatives mais n'est pas aussi précis que le modèle PLS-DA.

La régression PLS sur les spectres moyens par image et normalisés donne encore les meilleurs résultats avec une SEP de 0.3 %, une droite d'étalonnage présentant une pente de 0.98 et une ordonnée à l'origine de 0.01%.

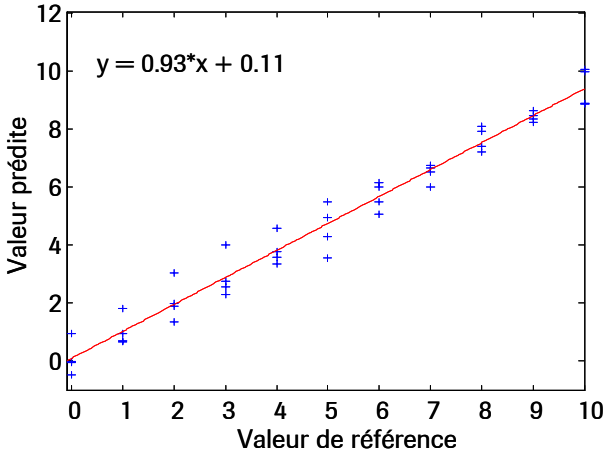


Figure 9. Droite d'étalonnage obtenue sur les spectres dérivés avec un modèle PLS-DA utilisant les spectres de références des trois composés majoritaires et cinq variables latentes

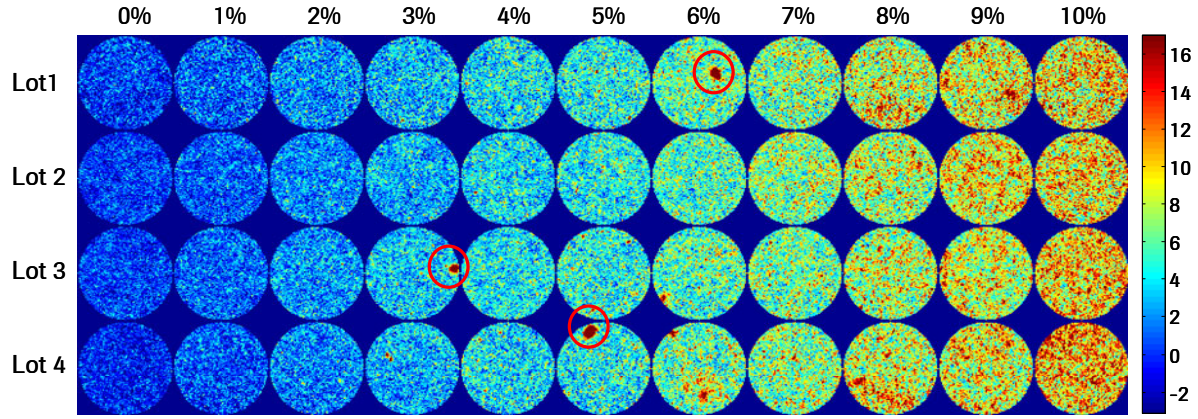


Figure 10. Cartes de distribution du principe actif des comprimés pharmaceutiques extraites par régression PLS-DA.

Pour chaque gamme d'échantillon la méthode classique PLS reste la plus précise pour la quantification du PA, cependant elle nécessite la production d'échantillons dont la concentration en Principe Actif varie. Or, pendant le développement d'un nouveau médicament, la formulation n'est pas fixe et produire une gamme d'échantillonnage est souvent trop couteux en temps et en argent. Si aucun a priori est disponible, la quantification est difficile. Les méthodes de déconvolution sont sujettes aux problèmes de solutions multiples. Les problèmes liés aux propriétés inhérentes de la réflexion diffuse en PIR comme le volume

d'interaction du faisceau incident avec la matière et les effets de matrice rendent difficile une approche par segmentation. Elle peut néanmoins donner une estimation de la concentration si les tailles de particules des poudres sont plus importantes que la résolution spatiale de l'appareil ou en cas d'agglomération. La seule alternative possible est alors d'utiliser les spectres de références. Les résultats démontrent qu'une estimation des concentrations peut être obtenue grâce à l'algorithme PLS-DA.

4. Conclusion

L'imagerie proche infrarouge est un outil puissant pour l'analyse qualitative et quantitative de comprimés pharmaceutiques. Cette technique est un support essentiel pour le développement de nouvelles formulations ainsi que pour l'analyse de produits fini lorsque l'information spatiale est requise.

Abstract

By combining both spectral and spatial information Near Infrared Chemical Imaging (NIR-CI) allows the identification of the chemical species and their localization. Since the distribution of chemical species influences greatly the quality attributes of the medicine this kind of instrumentation has naturally shown to be very useful for the development of pharmaceutical products. However, each analysis generates thousands of spatially resolved spectra which need to be processed for objective comparison of the data.

In the present work, the extraction of distribution maps and their characterization is firstly addressed. In that case the samples' composition is fully known and specific wavelengths or the full reference spectra are used to localize the chemical species. Histogram analysis is performed to assess the homogeneity of the first intermediates revealing a batch with greater inhomogeneity. In the second intermediates, a difference in the particle sizes of two batches is enhanced using a segmentation scheme based on Otsu thresholding and watershed refinement. The usefulness of NIR-CI and image processing to study and compare the quality of intermediates is demonstrated. In a second part, the simultaneous extraction of spectra and distribution maps without a priori information is proven. The accuracy of NMF, BPSS, MCR-ALS and PMF algorithms are compared. The latter proves to extract both spectral profile and concentration with the best accuracy especially when rotational tools are used to investigate the space of feasible solutions.

The last chapter deals with the quantification of API in binary mixtures and pharmaceutical tablets. The quantification without a priori reveals to be quite challenging. With homogeneous sample, the multivariate curve resolution algorithms fail to recover the pure spectra. A segmentation scheme is appropriate only in specific cases if the chemical species particles are larger than the spatial resolution of the device. If a full range of tablets with known concentration is provided, PLS algorithm gives the most accurate quantification results. However, it is demonstrated that with the only knowledge of reference spectra, PLS-DA provides an estimation of the concentration which allows semi-quantitative analysis of samples when the construction of a full range is not possible for instance during the analysis of development samples.

Table of contents

List of abbreviations

ANN:	Artificial Neural Network
API:	Active Pharmaceutical Ingredient
ATR:	Attenuated Total Reflection
BSS:	Blind Source Separation
BTEM:	Band Target Entropy Minimization
CI:	Chemical Imaging
CLS:	Classical Least Squares
2D/3D:	Two-/Three-Dimensional
DA:	Discriminant Analysis
DCLS:	Direct Classical Least Squares
DR:	Diffuse Reflection
EMEA:	European Medicines Agency
EMSC:	Extended Multiplicative Signal Correction
FNNLS:	Fast Non-Negative Least Squares
FPA:	Focal Plane Array
FIR:	Far-Infrared
FOV:	Field Of View
FT:	Fourier Transform
FTIR:	Fourier Transform Infrared
GMP:	Good Manufacturing Practice
HPLC:	High Performance Liquid Chromatography
ICA:	Independent Component Analysis
InGaAs:	Indium Gallium Arsenide
InSb:	Iindium antimonide
IR:	Infrared
ITTFA:	Iterative Target Transformation Factor Analysis
Km:	K-means
KSFA:	Key-Set Factor Analysis

LCTF:	Liquid Crystal Tunable Filter
LDA:	Linear Discriminant Analysis
LUT:	Look-Up Table
MCR-ALS:	Multivariate Curve Resolution–Alternating Least Squares
MIA:	Multivariate Image Analysis
MIR:	Mid-Infrared
MLF-ANN:	Multilayer Feed-Forward – Artificial Neural Network
MLP-ANN:	Multilayer Perception – Artificial Neural Network
MSC:	Multiplicative Scatter Correction
NIR:	Near-Infrared
NMF:	Non-negative Matrix Factorization
OLS:	Ordinary Least Squares
OPA:	Orthogonal Projection Analysis
PARAFAC:	Parallel Factor
PAT:	Process Analytical Technology
PbS:	Phosphate buffered Saline
PCA:	Principal Component Analysis
PLS:	Partial Least Squares
PMF:	Positive Matrix Factorization
RGB:	Red-Green-Blue
ROI:	Regions Of Interest
SA:	Salicylic Acid
SD:	Standard Deviation
SIMPLISMA:	Simple-to-use interactive Self-modeling Mixture Analysis
SMCR:	Self-Modeling Curve Resolution
SNR:	Signal to Noise Ratio
SNV:	Standard Normal Variate
SVM:	Support Vector Machine
WEFA:	Window Evolving Factor Analysis

Introduction

Quality control and Process Analytical Technology (PAT) in the pharmaceutical industry

Since medicine directly interacts with human health, the quality control is a crucial step for the pharmaceutical industry that is performed according to drastic regulatory laws. The pharmaceutical products must undergo several tests during and after their production. These tests include dissolution, stability, content of API, blend uniformity. They are performed following the Good Manufacturing Practice (GMP) rules which describe the devices and the analytical procedures for an optimum quality control. Those rules are updated by governmental organizations such as the Food and Drug Administration (FDA) in the United States or in Europe the European Medicines Agency (EMEA). The pharmaceutical industries are regularly inspected to ensure that the GMP rules are strictly followed. The quality control tests are mainly performed in a destructive manner using for instance High Performance Liquid Chromatography (HPLC) or dissolution apparatus. Those methods are time consuming and only a few entities of a batch that can encompass thousands of individual pharmaceutical forms are checked. Therefore the actual necessary quality control meets two major drawbacks : it drastically delays the release of the batches to the market (until six months) and does not test each medicine.

In order to overcome the limitations of the current quality control, the FDA has launched since 2000 the Process Analytical Technology (PAT) initiative which encourages the development of non destructive analytical methods to monitor the process. The "*quality cannot be tested into products; it should be built-in or should be by design*"[1]. The objective of PAT is double: as a first step it should enable a better understanding of the process and on a long time period 100% quality control should be reached. Besides a better quality of the products, full control will shorten the time to market and decrease the number of rejected batches saving time and money for the industry.

In this context, Near Infrared (NIR) spectroscopy has found out a renew of interest in the pharmaceutical industry since it is a fast and non-destructive analytical method which can be mounted in-line.

Near Infrared spectroscopy and chemometrics

The first discovery of radiation after the visible spectral range was made in 1800 by Sir William Herschel by the help of mercury-in-glass thermometer. He found that the thermometers heated at maximum when placed after the red colour of the visible spectrum, revealing then absorption of energy. Further research and development of instrumentation led to the first Mid infrared (MIR) commercial spectrometer with dispersion technology put on the market as early as 1913 by Adam Hilger. Whereas MIR region was first considered to be the most of interest because it can give accurate information about the molecule structures, NIR spectroscopy has found practical applications as late as the 1970s. These applications were quantitative analysis of water, protein and oil of agricultural products such as grains and soya beans. K. H. Norris of the U.S Department of Agriculture has developed the method and is now considered the pioneer of NIR spectroscopy. With the increase of the computing power, chemometrics have been further developed to allow the statistical analysis of NIR spectra enabling this spectroscopic method to be a method of choice for in-line analysis first for agricultural product and nowadays for pharmaceutical applications.

Recently, the development of Focal Plan Array (FPA) detectors added a new dimension to NIR spectroscopic studies. FPA features a matrix of pixel acquiring spatially resolved spectra. By combining both spectral and spatial information the fast identification and localisation of the chemical compounds are now possible and this imaging modality is also named chemical imaging (CI). Among others, distribution of compounds is an undeniable factor influencing solid dosage form physical attributes. For example heterogeneous compound distribution can decrease the rate of tablet dissolution or lead to process troubleshooting such as bad powder flow properties or tablet sticking to the press punches. Thus, chemical imaging appears as an adequate tool to solve these issues and to answer the PAT initiative. Therein, the number of publications dealing with hyperspectral imaging has exponentially increased since years 2000.

CI has also raised new challenges for data processing. The resulting image stacks formed three dimensional matrices also called data cubes where two dimensions are related to the spatial dimensions whereas the last one to the spectral dimension. The challenge is double. Firstly, the processing tools employed in classical spectroscopy may be applied to each single spectrum. Secondly, it can be viewed as images and the introduction of image processing tool may lead to the extraction of spatial information. Thus the techniques developed in the image processing, signal processing and spectroscopy fields must be considered, definitely making chemical imaging a multidisciplinary subject.

In this context, the objective of this thesis is to provide data analysis schemes for the qualitative and quantitative analysis of pharmaceutical solid dosage forms by NIR chemical imaging.

Organization of the manuscript

The present manuscript is divided into three chapters. The first chapter introduces the theoretical background about near infrared spectroscopy and imaging, the chemometric methods available to process the data and draw up the state of the art of NIR-CI in the pharmaceutical industry.

The second chapter deals with the extraction of the compound distribution maps of solid dosage forms and their characterization for the objective comparison of samples. Firstly the usefulness of NIR-CI to check the homogeneity and particle sizes of samples produced during the pharmaceutical development is demonstrated. Secondly, the feasibility of Multivariate Curve Resolution algorithms applied to NIR-CI data is proven and especially a novel approach to investigate rotational ambiguity is proposed. The last chapter aims at providing different schemes for the quantification of API in tablets by NIR-CI especially without a priori or with the only knowledge of the reference spectra.

Chapter 1

Near infrared imaging in the pharmaceutical industry

I. Introduction

The aim of this first chapter is to present the theoretical framework for Near Infrared hyperspectral imaging. It is divided into three main parts.

The first part deals with the principles of NIR spectroscopy and imaging. A short insight into the physical models and laws explaining the interaction of the NIR radiation with the matter is first given. The required instrumentation for generating NIR images of a sample is subsequently introduced. Several important aspects of imaging, such as acquisition time, calibration, and resolution are also addressed.

When the data are acquired, they must be treated. Thus, the second part draws up the state of the art of chemometrics for the analysis of spectra and images. The methods that aim at extracting relevant information from spectral data are numerous but they can be sorted in different groups depending on their main features to facilitate their comparison. For example there are univariate vs. multivariate methods, linear vs. non-linear or factorial vs. clustering methods.

Lastly, after the state of the art of theory and data processing it is important to draw up the state of the art of NIR-CI applications in the context of pharmaceutical industries. Several issues such as blending, counterfeit, content uniformity, process monitoring have been studied by many authors. The third part covers all these points through an exhaustive list of publications.

II. Near infrared spectroscopy and imaging: theory and practice

II.1. Theory of vibrational spectroscopy

II.1.1. Electromagnetic radiation

Spectroscopy is the science which aims at characterizing matter by its interaction with an electromagnetic radiation. An electromagnetic radiation can be described by two models: the wave and the particle model [2]. In the wave model, the main characteristic is the frequency ν of oscillation of the light. The wavelength λ is the covered distance by the light during one oscillation. It can be linked to the frequency using Equation 1

$$\lambda = \frac{c}{\nu} \qquad \textbf{Equation 1}$$

where c is the speed of the light = 3.10^8 $m.s^{-1}$, λ the wavelength and ν the frequency

The particle model is needed to explain the interaction between light and matter. In this model, the light conveys particles called photons. The energy transported by one photon is given by Planck's equation:

$$E = \frac{h.c}{\lambda} = h.\nu \qquad \textbf{Equation 2}$$

where h is the Planck's constant h = $6.62606876.10^{-34}$ J.s

Depending on the energy of the incident radiation different kinds of interactions with the matter occur, and different spectral regions are defined in the electromagnetic spectrum as shown in Figure 1. Vibrational spectroscopy uses the particular spectral range from the electromagnetic spectrum (13000 cm^{-1} to 10 cm^{-1} or 0.76 to 1000 μm depending on the units employed) where vibrations between chemical bounds in molecules appear. It is the Infrared region and is adjacent to the visible region. Within this spectral region three sub-divisions are defined: the far-infrared (FIR: 400-10 cm^{-1}or 26-1000 μm), the mid-infrared (MIR: 4000-400 cm^{-1} or 2.6-26 μm) and the near infrared (NIR: 13000-4000 cm^{-1} or 0.76-2.6 μm) named in relation with the visible region. By convention in chemistry, the unit chosen in the MIR is

cm^{-1} whereas in the NIR range, nm is preferred. The relationship between the two units is given by:

$$[\mathrm{cm}^{-1}] = \frac{1}{[\mathrm{nm}] \times 10^{-7}} \qquad \textbf{Equation 3}$$

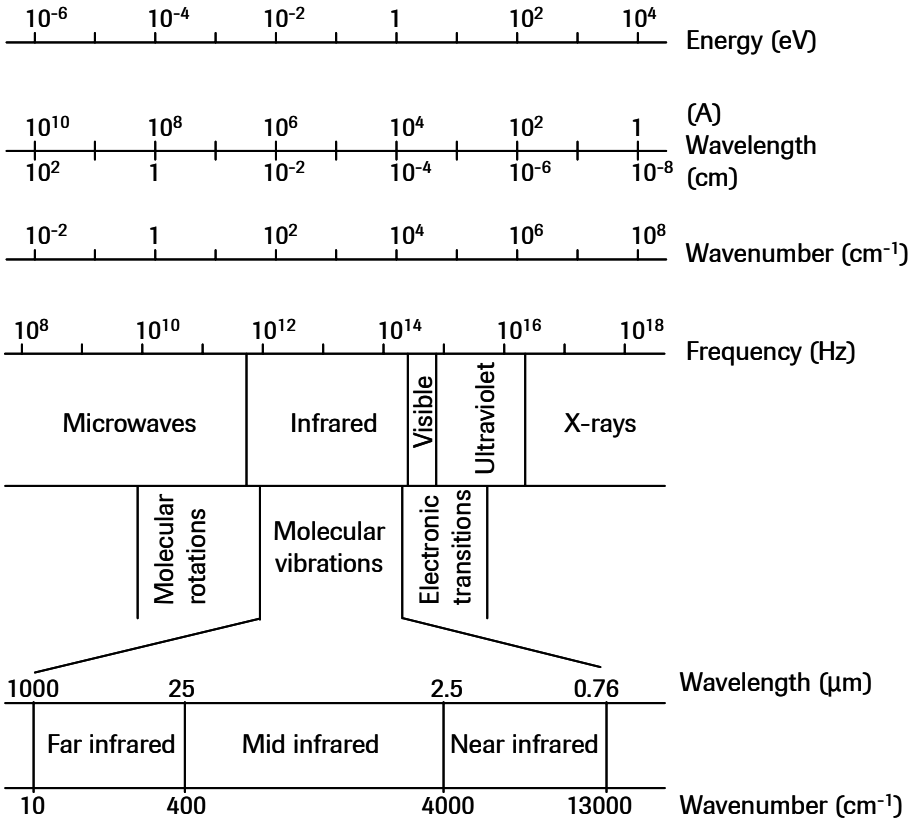


Figure 1. Electromagnetic spectrum adapted from [3], NIR spectroscopy occurs in the spectral range: [760 – 2600] nm.

II.1.2. Molecular vibrations

When lighted by an IR radiation, the atoms in a molecule are not static, they vibrate. In order to explain the interaction, the physicists have modelised the chemical bound by a spring model. First, the model for a diatomic molecule is considered because it is simpler, afterwards it is extended to polyatomic molecules.

II.1.2.1. Diatomic molecules

Harmonic oscillator

A bond between two atoms may be modelised as depicted in Figure 2 by a spring which elongates and shrinks around an equilibrium position (r_{eq}). This is the harmonic oscillator.

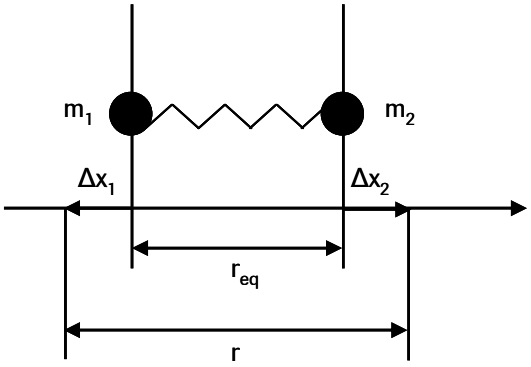


Figure 2. Spring model of a diatomic chemical bond. The spring elongates and shrinks around an equilibrium position.

Potential energy of this system is given by:

$$V(r) = \frac{1}{2}k(r - r_{eq})^2 \qquad \textbf{Equation 4}$$

where r = distance between the two atoms, r_{eq} = distance between the two atoms at equilibrium, k = strength constant.

The solution of Equation 4 leads to the frequency of the harmonic oscillator given by Hook's law:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \textbf{Equation 5}$$

with $\mu = \frac{m_1 * m_2}{m_1 + m_2}$ the reduced mass of the two atoms, m_1, m_2 the mass of each respective atom.

However, the energy state of an atomic system is not continuous but quantized with discrete levels. The possible energy of vibrations E_v for one diatomic molecule are in fact the solutions of the Schroedinger equation and are given by Equation 6.

$$E_v = \left(v + \frac{1}{2}\right) * h\nu \quad \text{with } v=0,1,2 \qquad \textbf{Equation 6}$$

With v the quantum number

Only transitions between two quantum levels are possible. The energy (ΔE) allowing the system to grow from one state of energy to the other one is given by Equation 7:

$$\Delta E = E_{v+1} - E_v = h.\nu \qquad \textbf{Equation 7}$$

When a photon of energy ΔE strikes the diatomic molecule, it will be absorbed and this absorption can be measured.

The transitions from the ground state V=0 to the state V=1 are considered the fundamental vibrations and occur in the mid-infrared. Transitions from V=0 to V= 2,3… are called overtones and appear at higher energy, thus, in the Near Infrared. With the harmonic model, overtones are forbidden, they exist because of anharmonicity and Fermi resonance.

Anharmonic oscillator

Actually, the harmonic oscillator may only explain slight shifts around the equilibrium position. Atomic repulsion limits the approach of the nuclei during the compression step, and the atoms may dissociate at high stretches. The interaction between the atoms is then more truthfully described by the anharmonic oscillator such as depicted in Figure 3.

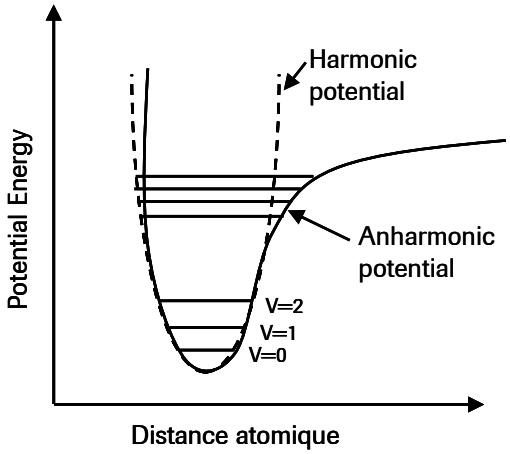


Figure 3. Potential energy of a diatomic molecule as a function of the atomic displacement.

Terms of higher degree appear in the energy equation which is transformed as depicted by the following equation:

$$E_v = \left(v + \frac{1}{2}\right).h.\nu - \left(v + \frac{1}{2}\right)^2.h.\nu.x_e \qquad \textbf{Equation 8}$$

with x_e the anharmonicity constant.

It must be noted that, because they are normally forbidden, overtone transitions are between 10 to 100 times less probable as fundamental bands.

II.1.2.2. Polyatomic molecules

After considering the diatomic case, the model can be extended to polyatomic molecules. A molecule with N atoms have 3N degrees of freedom. Three degrees represent translation motion along x, y and z axes and three other rotational motions about the same axes. The remaining 3N-6 degrees of freedom represent the remaining possible motions between atoms in the molecules, i.e. their vibrational modes. Each of the vibrational modes has a fundamental frequency and overtones. By analogy with the diatomic molecule, the energy associated with a state of a polyatomic molecule can be described by Equation 9.

$$E_v = \sum_{i=1}^{3N-6}\left(v_i + \frac{1}{2}\right) * h\nu \qquad \textbf{Equation 9}$$

With $v_i = 0, 1, 2$ the quantum number

In polyatomic molecules, combinational bands occurs. They consist in subtraction or addition of two fundamental vibrations forming then a single band. The combinational bands involve less energy than for the overtone bands and are therefore situated at longer wavelength.

II.1.3. Comparison of MIR and NIR spectroscopy

As described by the quantum theory, when a molecule is excited, it grows from a state of energy to another state by absorbing a quantity of energy ΔE. A spectrum plots the absorption of a sample as a function of the wavelength. The theory about vibrations of the molecules allows to determine the spectral regions where fundamental, overtones and combinations occur and also their probability of appearance.

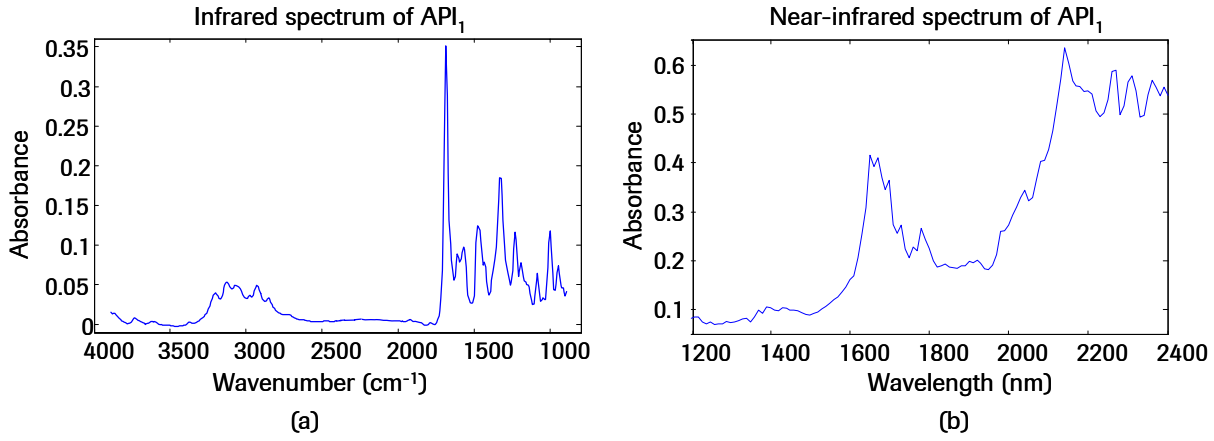


Figure 4. Mid-Infrared spectrum (a) and Near Infrared spectrum (b) of an API. Contrarily to MIR spectrum NIR presents broad and overlapping peaks.

Fundamental and overtone vibrations occur in the Mid-Infrared range from [4000 to 400] cm^{-1}. As an example, Figure 4 (a) depicts the Mid-Infrared spectrum of one API. In this region, the peaks of absorbance are numerous and fine, because of the high probability of absorbance due to fundamental transition. They allow the identification of the functional group and thus the molecule. On the other hand, a large fraction of the light is absorbed. Consequently, for transmission measurements, the samples must be diluted into non-absorbing matrix to allow the signal detection. Liquid might thus be prepared as diluted solutions in a cell. For solid analyses, the sample is dispersed into a Kbr disk or Mull. Attenuated Total Reflexion (ATR) [4] has been developed to avoid preparation of the sample. But then a more sophisticated device set up is needed to perform measurements in this configuration, making it difficult to be used for in-line analysis of solids.

Combination and overtone vibrations are registered in the NIR range. The spectral range is narrower compared to the Mid-IR range and peaks overlap, thus clear identification of the chemicals based on the study of peak positions is not possible. As an example, Figure 4 (b) depicts the near infrared spectrum of one API. Statistical analysis of the spectra is required to identify the molecule and for quantification. However, since the absorbance is weak, because of low probability of occurrence of overtones and combinational vibrations, the sample does not need to be prepared: the analysis is non destructive. Moreover, the analysis through glass and plastic is possible. The development of computer resources as well as easier handling of the device (detector which does not need to be frozen) and safer use (no laser beam) make it a spectroscopy of choice for in-line applications.

II.2. From classical spectroscopy to imaging

With classical spectroscopy one spectrum reflects the integrated information of the sample surface. Therefore the scientist acquires only mean information about the sample.

Since the end of the 20^{th} century the technical improvements have allowed the development of new detectors or methods to acquire simultaneously spectral and spatial information. This technique is referred as "chemical imaging". Spectra are spatially located and it is possible to identify the chemical species inside the samples and also to map their distributions.

II.2.1. Hyperspectral Data cube

A chemical imaging experiment generates a three-way matrix called a data cube as shown in Figure 5. Two dimensions n and m depict the spatial localizations of the chemical species and one dimension λ allows their identification. Several thousands of spectra can be acquired with

chemical imaging depending on the detector size. Those spectra contain more than a hundred wavelengths, that is why data are called hyperspectral.

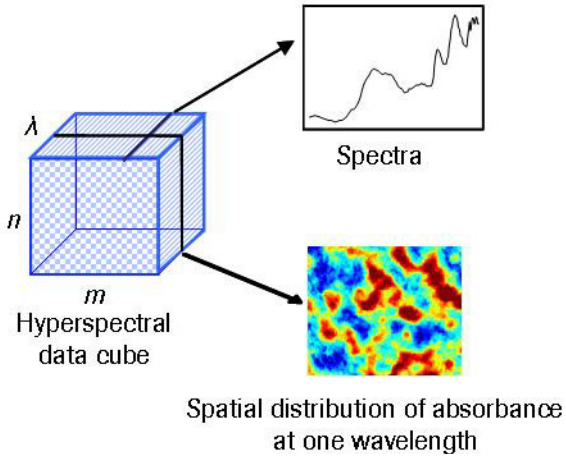


Figure 5. Data cube generated during one chemical imaging experiment. Two dimensions depict the spatial distribution of the compounds, the third one, the spectral dimension, allows their identification.

II.2.2. How to generate an Hyperspectral cube?

They are three techniques to generate an hyperspectral cube as depicted in Figure 6. The first strategy is point mapping (Figure 6 (a)) and until years 2000 this technique was the most widespread one. It consists in a spectrometer combined with a microscope and a moving stage [5]. The user defines a regular grid of spatial positions above the surface of the sample. A spectrum is measured at one position, the sample moves to the next measurement point of the grid and another spectrum is registered. The process is iterative for all the positions in the area which defines the image. Today, nearly every constructors set up the possibility of using point mapping for microscopic spectroscopy.

The second method is called "line imaging" (Figure 6 (b)). The detector allows to acquire simultaneously one spatial dimension and the spectral dimension [6]. As with point mapping, the system acquires spectra according to predefine spatial positions and the line is moved right to left and/or up to down to cover the whole surface. A second configuration of line imaging experimental set up exists for process control. The detector is fixed and the samples, placed on a conveyor belt, are moving.

The third method uses Focal Plane Array (FPA) detectors (Figure 6 (c)) [7, 8]. FPA are composed of several thousands of detector elements forming a matrix of pixels. They are optical detectors placed at the focal plane of spectrometers. It allows the acquisition of

thousands of spectra at the same time [9]. The popularity of such systems has increased these last 10 years and several systems are now available.

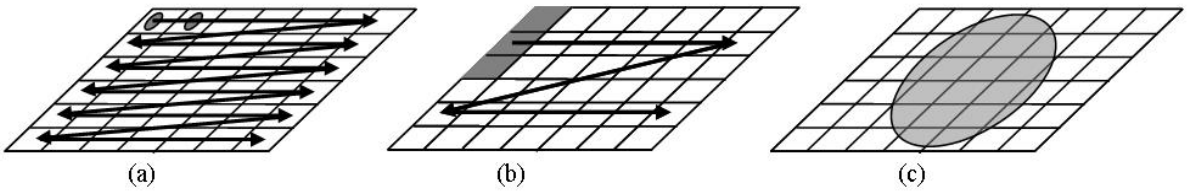


Figure 6. Three approaches exist to register one hyperspectral imaging data cube: (a) point mapping, (b) line mapping and (c) widefield approach, the latter one being the fastest one.

II.2.3. Comparison of the acquisition processes

The strongest argument in favor of widefield apparatus is the rather fast acquisition time in comparison with other mapping techniques, which makes it an instrument of choice when dozens of samples have to be analyzed. In [10], acquisition time for NIR mapping and widefield instruments are compared. The mapping technique took 12.5 hours to acquire 2500 spectra on a 1 mm^2 wide area with a spatial resolution of 20*20 μm^2 and spectral resolution of 16 cm^{-1} by the help of a Fourier Transform-NIR (FT-NIR) spectrometer. With a global illumination instrument, it is possible to acquire one hundred wavelengths and thousands of spectra in 5 minutes. Speed of line-scan spectrometer is somewhere between global illumination and mapping instrument depending on the number of pixels in the detector but the number of wavelengths that can be acquired is fixed by the detector size.

On the other hand, because of its smaller area, a single pixel element of a 2D detector integrates less signal than a single-element detector used in a mapping study. Additionally other sources of noise due to the optics or the non uniform illumination of pixels have to be taken into account when performing a global imaging experiment. As example, the picture in Figure 7 depicts the median count of each pixel computed over the wavelength of a data cube. The data cube was acquired while a high reflectance standard was placed on the stage. The instrumentation employed is described later in this section (II.3.1). A circle-like shape might be detected in the center of the image revealing the optic effects. Moreover, each pixel of the detector has a different response due to the inherent variability of construction [6, 11-13]. The calibration step (cf II.3.3) will correct this artifact but spatial noise may remain. Thus, point mapping experiment gives more accurate spectral resolution and might be a better approach to study the distribution of minor compounds. Moreover, it is more appropriate to analyze samples with a rough surface, because focus can be tuned at each spatial position.

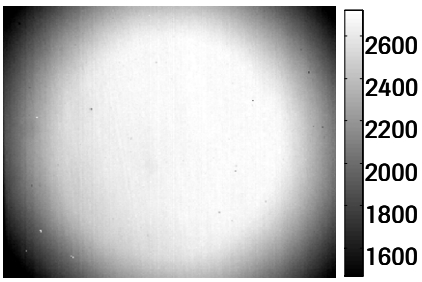


Figure 7. Spatial instrument response curve (median values computed over wavelengths) registered using a high reflectance standard. Non-uniform illumination of the detector due to the optics of the camera is depicted.

In fact, every authors agree to say that the choice of one or the other techniques depends on the applications. If a small surface has to be analyzed to find out minor compounds (Ex: impurity detection) one should use point mapping. If an overview of the sample is required and a lot of samples must be analyzed then global imaging is more appropriate. Line scanning apparatus appears to be more suitable for in-line applications, to analyze samples on a moving conveyor belt [6]. Choosing one or the other techniques is a matter of compromises.

II.3. Instrumentation, sampling and calibration

II.3.1. Instrumentation

The aim of this paragraph is to describe the instrumentation necessary for hyperspectral imaging.

For micro-spectroscopy and imaging, the systems are constituted of four main parts [14]:

- A source to generate the light beam
- One system which allows the separation of the wavelengths. This is the main part of one spectrometer
- Several optics for the selection of the spatial resolution
- A detector to register the signal

A large panel of instruments is commercially available for NIR micro-spectroscopy or global imaging, each of them having their advantages and drawbacks. The spectrometers are classified according to the employed technologies to separate and record the wavelengths. On one side there are systems that acquire the wavelengths one by one: the sequential apparatus such as filter. On the other side the multiplexed spectrometers record several wavelengths at the same time, the most popular being the Fourier Transform (FT) spectrometer. A second

subdivision separates the non dispersive from the dispersive elements such as gratings. A whole description of all the apparatus is out of scope of this manuscript and has already been covered in the literature [15]. The emphasis is rather put on the very instrumentation used all along this thesis work and schematized in Figure 8. It is the Sapphire marketed by Malvern.

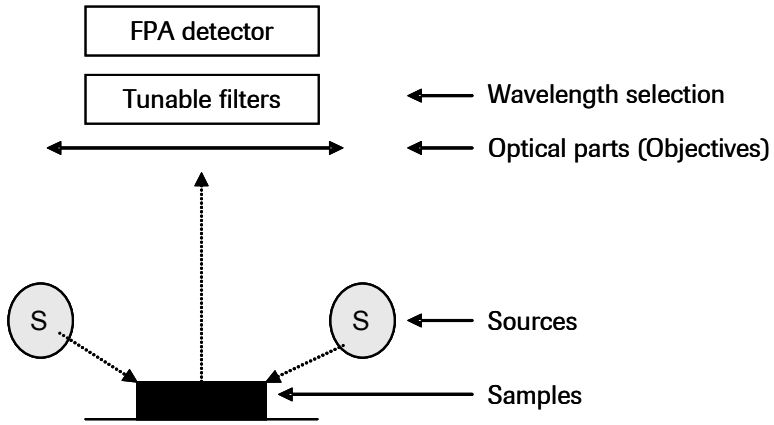


Figure 8. Schematic view of a NIR spectrometer using tunable filter. Adapted from [16].

II.3.1.1. Sources

The sources of Mid-IR or NIR spectroscopy are in most cases polychromatic thermal. An inert solid heated electrically to a temperature of 1500 to 2200 K irradiates uniformly in the infrared spectral range. On the Sapphire, four tungsten lamps generate the NIR radiation. One polarizer is screwed on the top of each lamp to avoid glare from shiny sample.

II.3.1.2. Tunable filters

Tunable filter (TF) is the technology employed to separate the wavelengths in our NIR global imaging spectrometer (Figure 8). A review of TF has been written by Gat in [17]. TF is a device whose spectral transmission can be electronically controlled by applying voltage. Two types of tunable filters are available for imaging: Acousto-optic tunable filter (AOTF) and Liquid Crystal Tunable Filter (LCTF), the latter allowing to reach higher spectral resolution.

A LCTF is built using several Lyot filters such as depicted in Figure 9. For one Lyot filter element the incoming light is polarized when it passes through the first polarizer. Then the birefringent crystal introduces phase difference (δ) between the rays of light. Finally the light passes through the second polarizer named "analyzer" which allows only the transmission of one wavelength. The Lyot filter elements assembled ones behind the others form a stack and specific wavelengths can be selected [9, 18, 19]. This is the most widespread technology for global imaging.

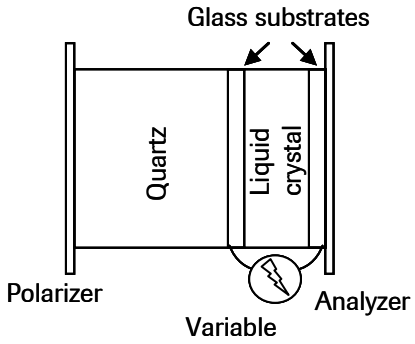


Figure 9. Schematic view of a Lyot filter element

II.3.1.3. Optics

The microscope is fitted with optical elements for selecting spatial resolution. The choice of magnification levels depends on the study aim. Typically 6×, 15×, and 32× objectives are used on a Mid-IR or NIR microscope.

Our device is a macroscopic camera. Several objectives allowing several magnification levels are available. They are easily interchangeable. The Table 1 depicts the different magnifications available with the corresponding field of view and working distance.

Pixel size (µm)	Magnification	Field of view (mm)	Working distance (cm)
9.06	4.7x	2.3×2.8	2.4
20	2.2x	5.1×6.4	3.5
39.2	1.1x	10×12.5	9.5
79.4	0.54x	20.3×25.4	14.5
128	0.33x	32.8×41	25

Table 1. Parameters of the objectives

The choice of the magnification levels will be guided by the aim of the study. The largest field of view allows to analyze several samples at the same time and can be used for an easier comparison of tablets. If one tablet has to be analyzed the objective 40 µm/pixel is usually preferred because it gives the best overview of the sample.

II.3.1.4. Detector

In NIR, lead sulfide (PbS), indium antimonide (InSb), and uncooled indium gallium arsenide (InGaAs) detectors are commonly used. A stirling-cooled InSb detector is mounted on the Sapphire. The detector size is 256*320 pixels enabling the acquisition of 81920 spectra in one experiment. Figure 10 depicts the instrument response curve as a function of the wavelength (median values computed over the pixels) when a high reflectance standard is placed on the sample stage. It can be seen that the sensitivity of the detector is reduced at longer (after 2200

nm) and shorter wavelengths (before 1300 nm). Thus, data acquired in those spectral ranges might be more noisy and therefore taken with caution. The best choice should be to remove those noisy channels. However, it might happen that chemical species have strong signal in those regions, which are therefore of interest.

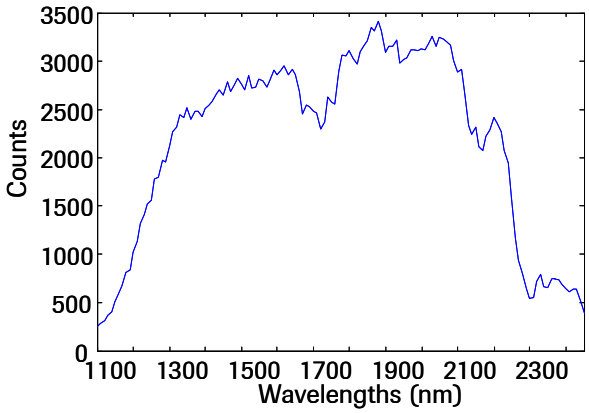


Figure 10. Spectral instrument response curve (median values computed over pixels) registered using a high standard reflectance.

II.3.2. Sampling and spatial resolution

Diffuse reflection (DR) is the sampling technique used to analyze the sample. It is also the most widely employed method in NIR spectroscopy. In DR a photon entering the sample will interact with the particles and several reflections will appear before it exits the sample and reaches the detector such as depicted in Figure 11. This figure also shows the pathlength of photons in this configuration. The gray area indicates the volume of light-matter interaction. The penetration depth is the measure of light entry into the sample. The shorter the wavelength, the higher the energy of the incident beam. High energy leads to deeper penetration depth and thus to an increased volume of interaction. Consequently, penetration depth is the limiting factor of the spatial resolution for NIR diffuse reflectance imaging contrary to micro-spectroscopy where the Rayleigh criterion is the limiting factor [20]. In [21] the authors measured a penetration depth of several thousands of micrometer (up to 700 µm) for shorter wavelength (1100 nm) and around 100 µm for longer wavelength (2380nm) by the help of layer of cellulose paper on a substrate. In [20] the experiments employing first a layer of polystyrene and then a commercial aspirin tablet led to the conclusion that the maximum resolution achievable with NIR diffuse reflectance is not less than 30 µm because of the penetration depth of the NIR radiation.

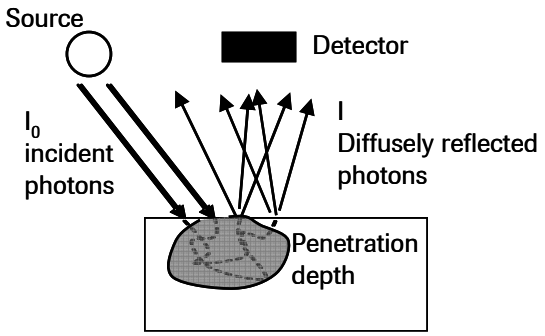


Figure 11. Illustration of the diffuse reflectance and penetration depth. Adapted from [20].

II.3.3. Calibration

The parameter of interest in NIR spectroscopy is the fraction of light which is absorbed by the sample. The absorbance is indirectly calculated using the intensity of the incoming radiation (the background signal I_0), and the intensity I of the light recorded either in transmission or diffuse reflexion. As with classical spectroscopy, calibration of the apparatus is required before the analysis. Especially spatial noise due to the non uniform illumination as well as noise from the detector must be removed when a FPA is used. In diffuse reflectance NIR spectroscopy, the background (I_0) is acquired by the help of a high reflectance standard such as a ceramic. I_0 serves as the reference signal. The absorbance spectrum is then calculated using the Equation 10.

$$A = -log(I / I_0) \qquad \textbf{Equation 10}$$

However, in our experiment configuration the InSb detector features a dark current D which is wavelength-dependent [13]. The dark current is the response of the detector when no signal is recorded and it is necessary to remove it from the raw signal. Dark current is measured by blocking the camera lens, and reflectance is subsequently calculated as in the following equation:

$$R = (I - D)/(I_0 - D) \qquad \textbf{Equation 11}$$

where D is the dark response of the detector

And finally the absorbance A is given in Equation 12 :

$$A = -log((I - D)/(I_0 - D)) \qquad \textbf{Equation 12}$$

In our study, the calibration with the acquisition of a dark cube and 100% standard reflectance is used. This kind of calibration assumes a linear response of the detector between 0% and 100% of reflectance. However, this supposition may be mistaken and finding true 100% reflectance standard may be difficult. More robust calibrations are thus described in the literature. In [13] the authors proposed the use of four standards with 2%, 50%, 75% and 99% of reflectance. To perform the correction, linear or quadratic regressions were first fitted at each pixel position using the four reflectance values. The raw cubes were subsequently corrected using the fitted values. In this study, quadratic regression gave better fit, suggesting a non-linear response of the detector. The same group, in [22], employed six standards (dark, 2%, 25%, 50%, 75%, 99%) and considered three methods of calibration:

- Global or pixelwise data subset selection; in global selection a median spectrum is computed for each reflectance standard hypercube, regression is performed, and each pixel is corrected using the same coefficients; in pixelwise section, each pixel is individually corrected.
- Secondly the model for the regression is either linear or quadratic as previously proposed [13].
- Finally they compared the possibility of either first scanning each standard and then the sample (external calibration) or integrating the standards in the same field of view with the sample under investigation (internal calibration). External calibrations correct pixel to pixel variance whereas internal calibrations may correct time-dependent variations due to temperature or power change.

Experiment 1 tested the utility of scanning several standards for calibration. The results clearly demonstrated that the association of a range of spectral standards with a quadratic model fitting and a pixelwise correction improved the quality of the calibration. The second experiment simulated lamp aging by decreasing the lamp power. Performing internal correction led to a reduction of the measurement errors.

Another standardization method has been recently proposed for line-scan NIR imaging systems [6]. The calibration was first performed following Equation 11. Then six reflectance images of four spatially uniform samples were acquired. Since the sample spectral response was uniform, most of the pixel noise was presumed due to illumination unevenness or difference among detectors. To correct these artifacts, data cubes ($x \times y \times \lambda$) were first reduced to an ($x \times \lambda$) 2D image (termed "average line image" by Liu et al) by averaging the y dimension, thereby relating each value of the 2D image to a detector sensor. The average spectrum of each data cube was then calculated and used as a reference value. The

spatial/spectral values of $l_{x\lambda}$ (average line image at the spatial/spectral coordinate position x and λ) for all average line images were plotted on the same graph (24 average line images in this case) against their reference value, revealing a linear relationship from which slope and bias were calculated. Each spectral/spatial position of the detector was thus associated with two correction coefficients (slope and bias) which were then used to correct new data cubes. This method is easy because only homogeneous standards are required and the 'true' reflectance does not need to be known. The standardization performed in this study significantly decreased the Standard Deviation (SD) at each wavelength band.

III. Chemometrics for the analysis of hyperspectral data

After their acquisition, the data need to be processed. One main issue arises when working with hyperspectral imaging: how can the relevant information being extracted from the huge amount of data? Visual inspection of each image over the wavelengths is time consuming and probably impossible when attempting to compare several data cubes. It is therefore necessary to reduce the data to several image planes and quantitative parameters. Figure 12 shows the classical processing workflow of an hyperspectral data cube. The first step is the preprocessing. The aim is to reduce baseline and scattering effects. The next step is to locate spatially each compound. Postprocessing can also be applied to enhance contrast or reduce noise. The final task is to determine parameters, such as particle sizes and shapes and the distribution of pixel intensity, which enable the images to be objectively compared. As hyperspectral imaging is an application spanning spectroscopy and image processing, most of the processing methods directly derive from these fields. The methods which are used in the two other chapters are explained with more details.

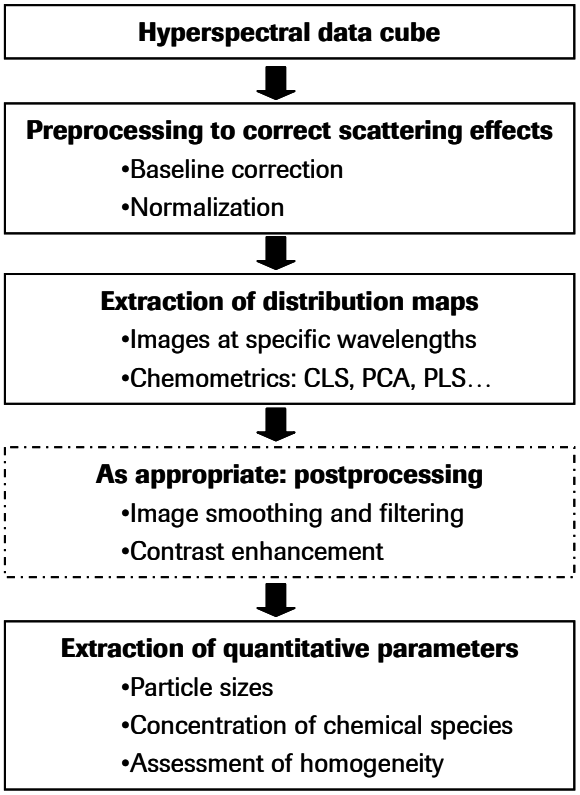


Figure 12. Processing workflow of a chemical data cube to extract information of interest to the analyst.

III.1. Preprocessing

Data preprocessing is an important step for hyperspectral data analysis, both spectral and spatial artifacts such as rough surfaces, optical effects and detector noise must be removed. Hyperspectral raw images mostly feature the same patterns depending on wavelength and reveal physical differences on the tablet surface rather than chemical variations. After appropriate preprocessing it is possible to visualize the distribution maps of the compounds [23]. Spectral pretreatments of hyperspectral imaging data mainly employ the same algorithms as in classical single point spectroscopy and consist of baseline correction and smoothing, normalization [24, 25] and scattering effects removal [26].

III.1.1. Baseline correction

It is mostly assumed that a spectrum is the sum of a background signal which corresponds to uncontrolled variations (= baseline), and a signal which contains the chemical observations. Baseline correction reduces the background effect.

In NIR spectroscopy, the classical method to correct the baseline consists in spectrum derivation. Savitzky-Golay derivation is mostly used [27]. The drawback of this latter is that it enhances noise.

III.1.2. Smoothing

To reduce noise the spectrum can be smoothed. The simpler method is the moving mean. The mean value of a spectral segment is computed. The central point of this segment is replaced by the calculated mean. The segment is then centered on the next spectral position and so on.

Another technique uses Savitzky-Golay filters [27] where a polynomial curve is fitted by least squares regression at each spectral value x_λ.

III.1.3. Normalization

Normalizing the spectrum consists of dividing each spectral value x_λ of a spectrum $\mathbf{x}$ by a representative number a [28]:

$$\text{xnorm}_\lambda = \frac{x_\lambda}{a} \quad \text{for}\, \lambda = 1 \ldots \lambda_{max} \qquad \textbf{Equation 13}$$

a can be computed as follows:

- the maximum value

$$a = max(x_\lambda) \quad \text{for } \lambda = 1...\lambda_{max} \qquad \textbf{Equation 14}$$

- the sum of all variables from the spectrum **x**

$$a = \sum_{\lambda=1}^{\lambda_{max}} (x_\lambda) \qquad \textbf{Equation 15}$$

This normalization is also called unit area.

- the sum of squares of all variables from the spectrum **x**

$$a = \sqrt{\sum_{\lambda=1}^{\lambda_{max}} (x_\lambda)^2} \qquad \textbf{Equation 16}$$

This normalization is also called unit length.

When there is an unknown variation (offset) between the spectra it can be useful to subtract the mean value $\bar{x}$ of the spectra **x** from each spectral variable x_λ. With NIR spectroscopy a Standard Normal Variate [25] is mostly applied (Equation 17) where each spectrum is also scaled by its own standard deviation.

$$xnorm_\lambda = \frac{x_\lambda - \bar{x}}{\sqrt{\frac{\sum_{k=1}^{p} (x_k - \bar{x})^2}{p-1}}} \quad \text{where} \quad \bar{x} = \sum_{\lambda=1}^{\lambda_{max}} x_\lambda \qquad \textbf{Equation 17}$$

Additionally Multiplicative Scatter Correction has been developed to remove multiplicative and additive effects due to the scatter of light in the sample [26]. It allows to a certain extent the linearization of the data. For MSC computation, an ideal spectrum representative of a data set **X** must first be chosen. The mean spectrum ($\bar{\mathbf{x}}$) is a good approximation. For each spectrum (observation) $\mathbf{x_i}$, a model is computed by least squares regression according to the following equation:

$$x_{i\lambda} = a_i + b_i \overline{x_\lambda} + e_{i,\lambda} \qquad \textbf{Equation 18}$$

Then the corrected data are obtained by :

$$xcor_{i\lambda} = \frac{x_{i,\lambda} - \hat{a}_i}{\hat{b}_i} \qquad \textbf{Equation 19}$$

Classical spectral preprocessing is thus applied to NIR hyperspectral data. However, imaging has an advantage over classical spectroscopy: comparing the remaining pixel-to-pixel spectral variations can give more information about the sample or about the ability of the preprocessing method to clean the spectra. For example, in a study to determine which spectral method was most effective at reducing baseline and scattering effects due to different sizes of salt and sugar particles [29] each sample was scanned individually. Kubelka Munk, SNV and absorbance tranforms, unit length and unit area normalization, first and second derivative, and several MSC variants were then applied to reduce the scatter effects in the hyperspectral NIR images. MSC was first applied to individual images, then to all images simultaneously, thus leading to different representative spectra for use in correction. Principal Component Analysis (PCA, cf § III.2.2.1) showed that global piecewise MSC removed all variation due to particle size within spectra, whereas derivative did not remove particle size dependency. Next, a Partial Least Squares (PLS, cf § III.2.2.1) model was computed to predict sugar particle sizes. PLS predictions based on the mean spectra provided accurate results, but predictions made on individual pixels located differences in particle size within a single tablet. Although global piecewise MSC was best at minimizing particle size effects, PLS was still able to find sufficient correlated variance for accurately determining particle size, as it led to broader distribution. Predictions based on mean spectra did not yield such information.

III.2. Extraction of distribution maps

The extraction of the distribution maps allows the localization of the chemical compounds of the sample. The extraction must be as accurate as possible to avoid misclassification of pixels. Many methods have been developed (Figure 13, next page), most of which derive directly from classical spectroscopy or image processing fields. Depending on the information available such as the spectral signatures of the pure compounds in the system under study or the experimental noise, one or the other method should be more appropriate. The first subdivision of methods separates univariate and multivariate analysis.

III.2.1. Univariate analysis

Univariate analysis is the simplest method to obtain the distribution maps. Each chemical entity in a mixture absorbs the NIR light at specific wavelengths depending on its chemical bounds. By choosing the image at those wavelength positions, one should obtain the localization of the specific compounds depicted by the pixels with the higher intensities. If this method is the most straightforward in revealing chemical localization, it also requires the system under study to be well characterized. All compounds should be known and the raw powders must be scanned to characterize their spectral signatures. In pharmaceutical studies every constituents of the medicine are known in the majority of studies and univariate analysis might be sufficient [10]. However, overlaps in complex systems may prevent identification of specific wavelengths for each compound, especially in the NIR range. Moreover some cases may arise where a constituent is not known (in case of a contamination for example) or its low concentration does not allow to extract reliable distribution map by univariate analysis. In those cases multivariate analysis is more adapted for characterizing fine spectral variations and for localizing compounds [30].

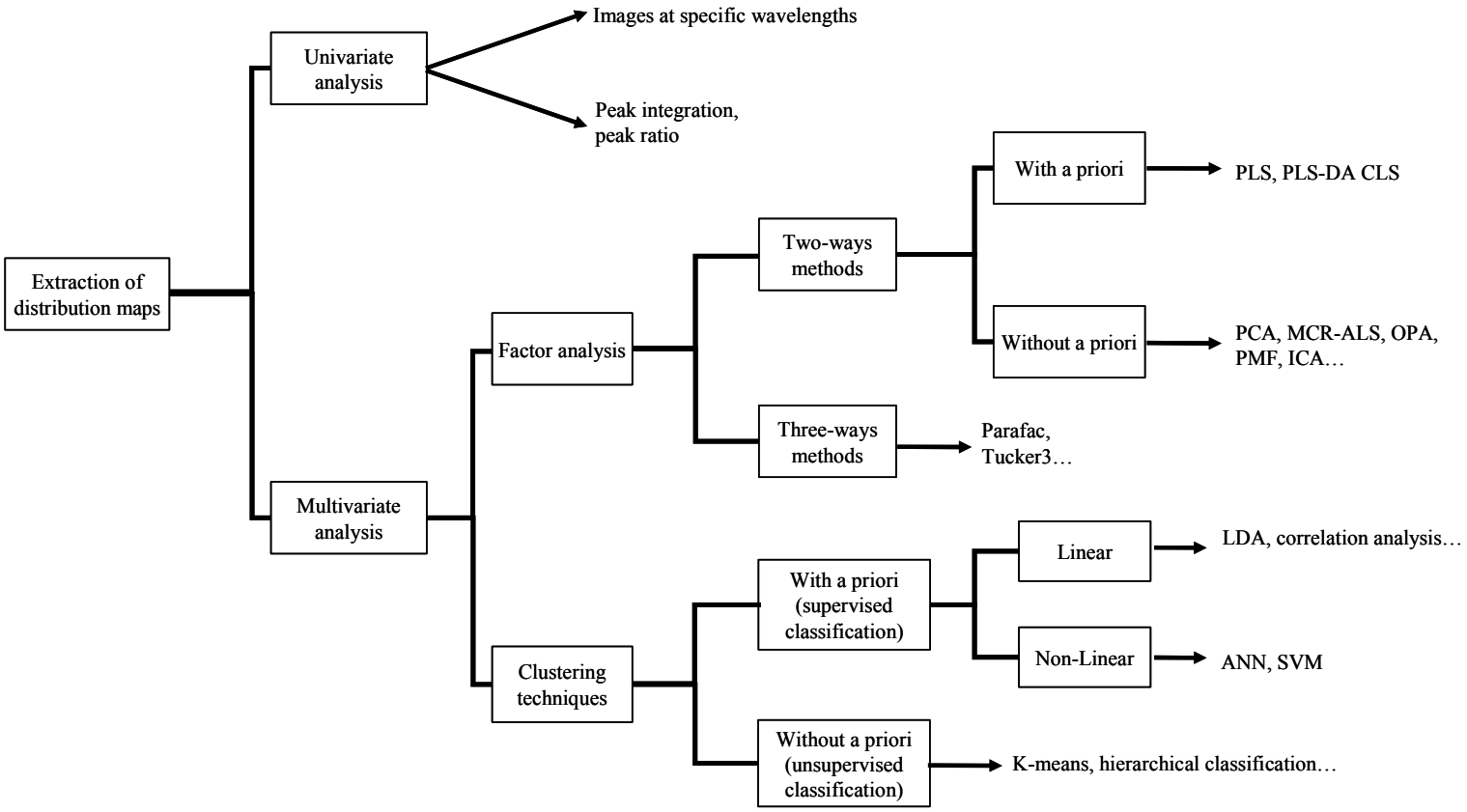


Figure 13. Classification of methods for extracting distribution maps and some examples.

CLS: Classical Least Squares, PCA: Principal Component Analysis, MCR-ALS: Multivariate Curve Resolution-Alternating Least Squares, PMF: Positive Matrix Factorization, NMF: Non negative Matrix Factorization, BPSS: Bayesian Positive Source Separation, PLS: Partial Least Squares, PLS-DA: Partial Least Squares–Discriminant Analysis. KNN: K Nearest Neighbor, LDA: Linear Discriminant Analysis, ANN: Artificial Neural Network, SVM: Support Vector Machine.

III.2.2. Multivariate analysis

Multivariate analysis takes into account all the spectral information contained in the data cube. As shown in Figure 13 several sub-divisions may be drawn up. On one hand there are the factorial methods which aim at decreasing the dimension by using an underlying multivariate distribution [31, 32]. The measured data are modeled such as to be a linear combination of the factors plus a term of noise. On the other hand, there are clustering techniques which aim at classifying the spectra into different groups of same features. Clustering techniques may be applied to the full spectra but also after factor analysis, in the reduced space [33].

III.2.2.1. Factor analysis

Factor analysis of hyperspectral data is based mainly on algorithms derived from the classical spectroscopy field. Those methods process two dimensional matrices where each row corresponds to one spectrum. However, as stated above, imaging instruments generate a stack of images, thus a three dimensional matrix. A first step of unfold is therefore required to apply classical factorial methods. The data cube **D** of dimension $N * M * J$ is thus transformed into a two-dimensional matrix $\mathbf{X} = ((N * M) * J) = I * J$ as depicted in the next figure.

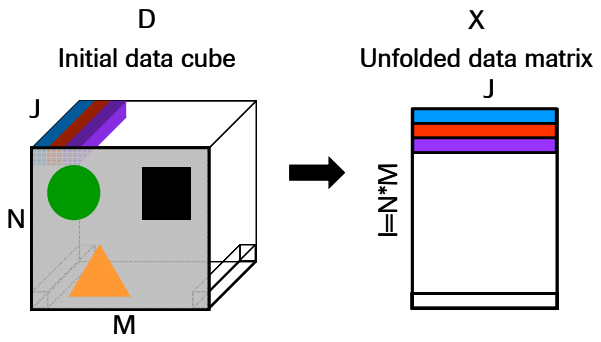


Figure 14. In order to apply classical factorial analysis, the three dimensional data cube must be unfolded into a two-dimensional matrix.

The processing of these unfolded cubes will be first discussed. However, methods analyzing three-way or higher dimensionality matrices have been developed in recent decades by the chemometric community and will be subsequently considered.

Two way processing

Classical factor analysis : Principal Component Analysis (PCA) and Partial Least Squares (PLS).

PCA and PLS [31] are the two algorithms mostly used in the chemometric community for extracting chemical information.

- *PCA*

The main goal of a PCA analysis is to reduce the dimensionality of a matrix by removing correlation between variables. The data are projected in a new space of principal components which are iteratively computed. The first PC is constructed such as to explain the maximum of variance of the data. The second PC is constrained to be orthogonal to the first PC and to explain the residual variance not taken into account by the previous PC and so on (cf Figure 15).

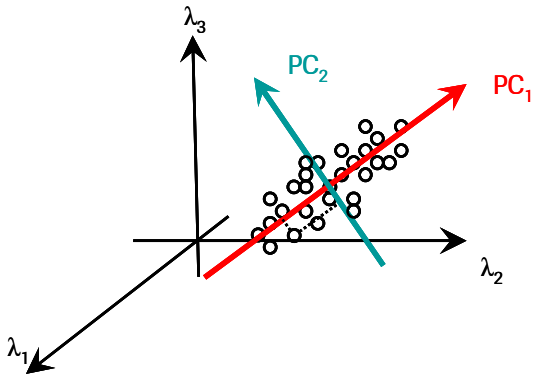


Figure 15. Principal Component Analysis. The principal components are constructed such as to explain the maximum of variance and to be orthogonal to each other.

The principal components are called loadings and the projection of the data into the PC are called scores. The first loadings are considered to explain the most useful information whereas the last ones mostly describe noise. The reduction of the dimensionality is thus performed by discarding those noisier components. After PCA transformation each column of the matrix of scores is folded back to form an image which depicts the variability of the pixels along the corresponding loadings and could eventually be linked to a chemical. PCA has been extensively used for the analysis of hyperspectral imaging to determine compounds distribution in NIR-CI [30]. However, even if it is useful to explain variance, it does not have chemical meaning and it may be difficult to link the PCA loadings with the chemical compounds.

- *PLS and PLS-DA*

PLS [34, 35] aims at predicting variable(s) of a matrix Y based on the observations of variable(s) of a matrix X. PLS constructs the following linear relations:

$$X = TP' + E \qquad \textbf{Equation 20}$$

$$Y = UQ' + F \qquad \textbf{Equation 21}$$

$$\text{Where } \mathbf{u}_a = b_a \mathbf{t}_a \qquad \textbf{Equation 22}$$

Y is projected into latent variables ($t_1, t_2, \ldots t_{kmax}$) which are linear combinations of $x_1, x_2, \ldots x_{\lambda max}$. The PLS method can be used to predict one variable (in this case Y is a vector and is noted y): this is the PLS1 or to predict simultaneously several variables: this is the PLS2 algorithm.

A first calibration step is required to construct the mathematical model linking matrices X and Y. The model is built by the help of a set of samples from which matrices X and Y are clearly known. As example, in case of NIR quantification of chemical compounds, the spectra form the rows of the matrix X and the known concentrations, determined by a reference method, the rows of the matrix Y. In this dissertation the PLS regression is used with the algorithm of calibration proposed by Wold [34].

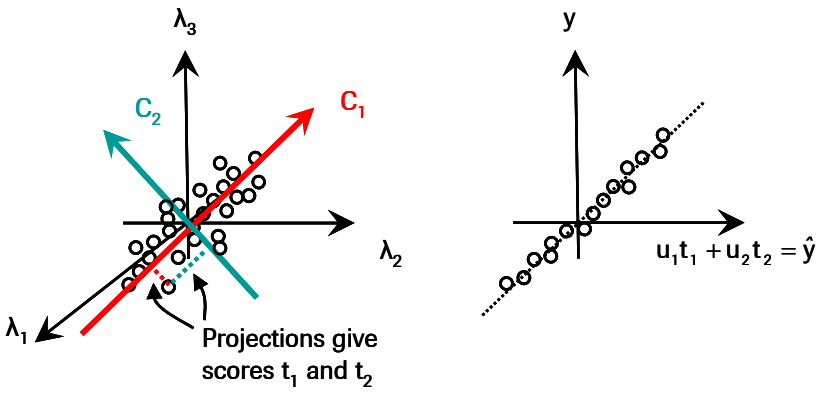


Figure 16. Construction of a Partial Least Squares model. Latent variables (t) are used to link concentration to spectral values.

Once developed, the model accuracy must be checked by the help of a new set of known values before being able to predict unknown samples. This is the validation step.

Imaging technique has several advantages over classical spectroscopy when using PLS algorithm for quantification. For example, when the model is employed to predict pixel to pixel concentrations, the distribution maps of the compounds might be visualized afterwards [36, 37] by folding back the matrix of predictions. Secondly, a data cube might contain several thousands of spectra which could be divided into calibration sets and validation sets. The calibration models may be computed using the mean (or median) spectrum of individual images. The user might also consider the mean spectrum of several spatial Regions of Interest (ROI) of one data cube including thousands of spectra as a calibration set [36]. The data can then provide several configurations of calibration spectra sets to optimize the model.

In case of hyperspectral imaging, it is also possible to construct a PLS-Discriminant Analysis (PLS-DA) classification method using pure reference spectra to extract the distribution maps of the compounds. The aim is to identify latent variables that will enable class separation by taking into account the class membership of observations [38]. This method has been used for NIR-CI images in [39] for the prediction of binary mixtures but also in [10, 23, 40], to display the distribution of compounds. However, when a complex matrix is under study or when the compounds are homogenously distributed, or present at low concentration, the extraction of distribution maps by PLS-DA classification may not be straightforward because of differences of the training library (the pure spectra) and the samples to be predicted [23].

Bilinear modelling

The two classical factorial methods PCA and PLS employ mathematical constraints to extract the factors which carry the most relevant information. As stated above, if they are useful to explain variance, however it might be difficult to interpret the loadings and relate them to chemical species. Thus alternative methods have been developed to unravel from the mixed spectra the pure spectra of the chemicals and their respective concentration. The analysis is based on the bilinear model.

According to the bilinear model, the absorbance of a sample results from the sum of the absorbance of its component chemical species. Thus, the mixed spectrum may be viewed as the weighted sum of each of the spectrum of the pure materials plus the experimental noise. This phenomenon is mathematically described in Equation 23:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \qquad \textbf{Equation 23}$$

Such as depicted in Figure 17, **X** is a two-way matrix of observed signal; **C** is a column wise matrix (I*K) of concentration/abundance of the chemical species; $\mathbf{S}^{\mathrm{T}}$ is a row wise matrix (K*J) of pure spectra and **E** represents the residual noise.

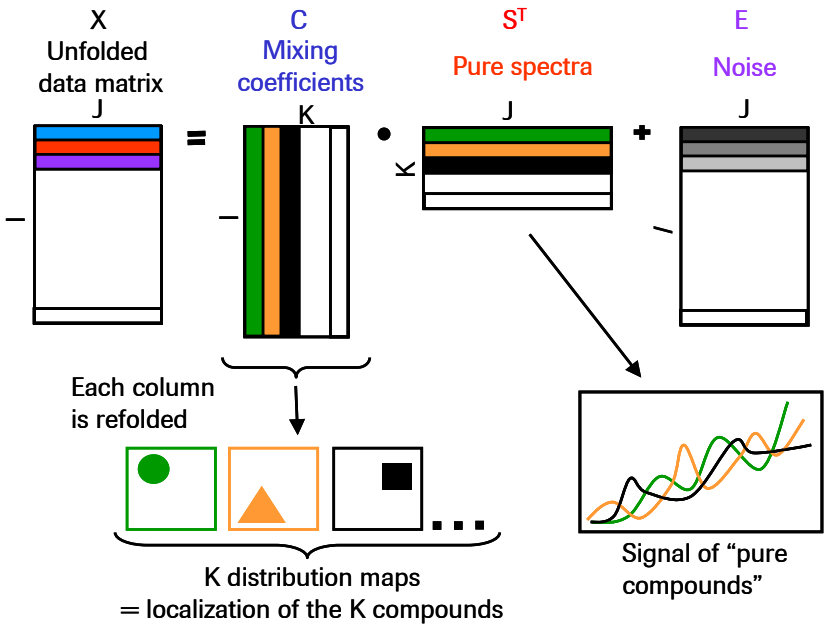


Figure 17. Bilinear modelling: the matrix of mixed spectra is factorized into two matrices related to concentration and pure spectra.

➢ *Bilinear modelling with a priori information*

If the matrix of the pure spectra S^T is available, Direct Classical Least Squares algorithm (DCLS), also called Ordinary Least Squares (OLS) may recover the distribution maps. This method consists in minimizing the sum of the square errors i.e minimizing $\left\|X - CS^T\right\|^2$. C is estimated by the pseudo–inverse:

$$C = XS(S^TS)^{-1} \qquad \textbf{Equation 24}$$

If a set of concentrations is available CLS can also be used to estimate the pure spectra:

$$S^T = (C^TC)^{-1}C^TX \qquad \textbf{Equation 25}$$

To a certain extent DCLS may give information about the relative concentration of constituents because it uses the scaled spectra [23]. However, differences in background, nonlinearity in spectra, especially in the NIR range, and noise prevent accurate concentration determination. Nevertheless, it enables to extract in certain cases reliable distribution maps for well-characterized samples, such as those from pharmaceutical development [9, 41].

➢ *Bilinear modelling with sparse a priori information*

If sparse information is provided for the analysis, then matrices **C** and $\mathbf{S}^T$ must be simultaneously extracted. The factorization of the matrix **X** into two positive matrices **C** and $\mathbf{S}^T$ can have an infinity of solutions due to rotational ambiguity. Mathematical constraints or constraints based on physical phenomena must be introduced to reduce the space of possible solutions [42, 43]. Often, the two kinds of methods are combined together, the first one serving as an initial estimate for the other one. In the analytical community the methods which unravel mixed spectra are referred to Self Modelling Curve Resolution (SMCR) whereas in signal and image processing the term Blind Source Separation (BSS) is employed. Lawton and Sylvester in the 1970s have introduced SMCR methods [44].

o *Introduction of mathematical constraints*

In order to find out loadings more representative of chemical information, methods have been developed that start in the space of principal components and attempt to find an appropriate rotation of the PC. For example, Key-Set Factor Analysis (KSFA) [45] constructs the set of key factors that are the most orthogonal to each other. Iterative Target Transformation Factor analysis (ITTFA) [46] tests whether the factor produced by a rotation of the PC might be a true factor. Recently Band Target Entropy Minimization (BTEM) algorithm [47] has been proposed and applied to Raman image of pharmaceutical tablets [48] to recover pure spectra of minor compounds. Entropy is a measure of disorder of a system, if the entropy value is low, the system is organized and simple. Since pure spectra are assumed to be the simplest underlying patterns, the minimization of the entropy appear to be an appropriate choice to retrieve them. In the proposed procedure, each pure spectrum is estimated by rotating the principal components. Starting with a PCA, the matrix of rotation is optimized by minimizing an objective function that includes two terms: the first minimizes entropy, while the second ensures non-negativity of the spectra.

Other methods work directly with the matrix **X**. From these methods the most popular are maybe SIMPLISMA which was proposed by Winding and Guilement [49] and Orthogonal Projection analysis (OPA) [50]. SIMPLISMA finds out the purest columns in the matrix **X** using a criterion calculated from the mean and standard deviation values of the columns. The purest columns formed the matrix **C** and the matrix $\mathbf{S}^T$ is calculated using Equation 24. OPA identifies the spectra showing most dissimilarity in the **X** matrix. This method is used later in the manuscript and therefore is further explained. It consists in the following steps [51]:

Step 1: Normalization of the spectra to unit length

Step 2: Selection of a reference spectrum (for instance mean spectrum: $\mathbf{x}_{ref}$)

Step 3: Projection of the normalized vector, $\mathbf{x}_i$ onto the reference spectrum using Gram-Schmidt orthogonalization:

$$\mathbf{d}_i = \mathbf{x}_i - \left(\mathbf{x}_i^T . \mathbf{x}_{ref}\right). \mathbf{x}_{ref} \qquad \textbf{Equation 26}$$

Step 4: The norm of each spectrum $\|\mathbf{d}_i\|$ is calculated and is referred as the dissimilarity. The spectrum giving the largest dissimilarity is retained as the new reference spectrum and so on.

Finally, Independent Component Analysis (ICA) [52] has become very popular in the last past years and must also be mentioned. ICA optimizes a criterion which assumes that the sources are mutually statistically independent. However, the basic assumption of independence may not be fulfilled in this case because pure spectra may exhibit correlation (especially NIR spectra) [53, 54]. Moreover, correlation from neighbor pixels may also prevent good separation by ICA.

- *Introducing constraints based on physical meaning*

The above procedures state hypotheses about the data or the pure spectra to derive a mathematical resolution of the bilinear model. The solution is unique in most cases but it may happen that not all the spectroscopic data confirm the hypotheses. They also lead to unrealistic solutions such as negativity of spectral profile or concentration. Therefore methods have been developed that force the solutions into consistency with physical meaning. At the core of such methods is the definition of a criterion to be minimized under specific constraints. The constraints reduce the space of feasible solution and might differ depending on the spectroscopic technique considered. The most widespread is the positivity of both concentration and spectral profiles because it is applicable to all kinds of spectroscopy. Positivity is usually set for hyperspectral imaging with eventually the knowledge of how many compounds are present at one pixel position. Other constraints that might be applied are the spectral specificity such as unimodality or the sum of concentration equal to a constant.

Those methods will be used in the other parts of the present manuscript, therefore, they are more thoroughly explained.

- *Multivariate Curve Resolution-Alternating Least Squares*

Nowadays, the method widely employed in analytical chemistry to deconvolve hyperspectral imaging spectra is the Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) [43]. This method optimizes the criterion Q_1 given in Equation 27.

$$Q_1 = \sum_{i=1}^{I}\sum_{j=1}^{J}\left\| x_{i,j} - \sum_{k=1}^{K} c_{i,k} s_{k,j}^{T} \right\|^2 \qquad \textbf{Equation 27}$$

The minimization is achieved by alternatively calculating matrices C and S^T by least squares fit while the other matrix is fixed. After estimation of one of the matrix product, the constraints are applied. For example constraint of positivity might be forced by setting after the estimation of matrices S^T or C negative values to zeros or applying Fast Non Negative Least Squares [55] algorithm for a smoother constraint. The different steps of the procedure are:

Step 1: Initialization of the matrix S^T

Iterate:

Step 2: Given X and S_j^T calculate $C_{j+1}^{*} = XS_j\left(S_j^T S_j\right)^{1}$ (estimation of C by a least squares resolution). Constrain $C_{j+1}^{*} \rightarrow C_{j+1}^{c}$ (matrix resulting from the constraints application).

Step 3: Given X and C_{j+1}^{c}, calculate $\left(S^T\right)_{j+1}^{*} = \left(\left(C^T\right)_{j+1}^{c} C_{j+1}^{c}\right)^{-1}\left(C^T\right)_{j+1}^{c} X$ (estimation of S^T by a least squares resolution), constrain $\left(S^T\right)_{j+1}^{*} \rightarrow \left(S^T\right)_{j+1}^{c}$ (matrix resulting from the constraints application)

Step 4: Check whether the product $C_{j+1}^{c}\left(S^T\right)_{j+1}^{c}$ reproduces X satisfactorily.

Where the indice j denotes the iteration.

When the stop criterion is fulfilled, the resolution process is ended. If not, the iterative optimization continues going to step 3. The iterations might also begin with the initialization of the matrix C, in that case S^T is first estimated and constrained and so on. The convergence criterion is based on comparison of the model fit between two consecutive iterations.

- *Non negative Matrix Factorization*

In 1999 Lee and Seung published in Letters to Nature [56] the Non negative Matrix Factorization (NMF) for the separation of the sources. The algorithm which minimizes Q_1 is relatively simple, the positivity of spectra and mixing coefficient being ensured by the positivity of the initial values and the update rules given in Equation 28:

$$S_{j+1}^{T} \leftarrow S_j^{T} \frac{\left(C^T X\right)_j}{\left(C^T C S^T\right)_j} \quad (1)$$

$$C_{j+1} \leftarrow C_j \frac{\left(XS\right)_j}{\left(CS^T S\right)_j} \quad (2) \qquad \textbf{Equation 28}$$

It is proven in the paper that the minimization of the criterion converges toward a locally optimal matrix factorization. The algorithm consists in positive random initializations of the matrices **C** and $\mathbf{S}^T$ and iteratively updates **C** and $\mathbf{S}^T$ using (1) and (2) until convergence.

A modified version of the NMF called constrained NMF (cNMF) which includes constraint on the minimum amplitude of the recovered spectra to take into account noise, has been applied to raman imaging spectra [57] and demonstrated good extraction ability.

- *Positive Matrix Factorization (PMF)*

Another alternative algorithm is Positive Matrix Factorization (PMF) developed by Paatero [58] which minimizes the criterion Q_2 (Equation 29).

$$Q_2 = \sum_{i,j} \frac{\left(\mathbf{X} - \mathbf{C}\mathbf{S}^T\right)^2_{i,j}}{\sigma^2_{i,j}} \qquad \textbf{Equation 29}$$

$\sigma_{i,j}$ denotes the estimate of uncertainty of the $(i,j)^{th}$ variable. Introducing such a weighting allows that the most precise variables have more influence on the minimization of the function Q_2. Positivity constraints are applied by the help of penalty functions and the optimization is performed by a conjugate gradient algorithm. Several softwares have been implemented to solve the PMF model (PMF2 and PMF3 software [58]) but recently the Multilinear Engine [59] proposed a more flexible scheme with the possibility of defining specific constraints. By the help of equations the user may define any kind of multilinear problem as well as any kinds of constraints. Especially, tools allow to investigate the domain of possible solutions in order to find more accurate ones based on the user knowledge, such as known values in the matrix **C** or $\mathbf{S}^T$, zeros concentration values, or matrices of rotation [60, 61]. The aim is to introduce the as much knowledge as possible in order to extract the solution which has the most physical sense.

- *Bayesian Positive Source Separation (BPSS)*

The aim of Bayesian approaches is to incorporate a priori knowledge about the statistical distribution of the sources and the concentrations. Recently, Saïd Moussaoui has proposed a method called Bayesian Positive Source Separation (BPSS) for the demixing of spectral data based on Bayesian theory [53, 62]. The basis of the methodology starts with a well-known theorem of the field of probability: the Bayes theorem which allows to write:

$$p\left(\mathbf{S}^T, \mathbf{C} \middle| \mathbf{X}\right) \propto p\left(\mathbf{X} \middle| \mathbf{S}^T, \mathbf{C}\right) * p\left(\mathbf{S}^T\right) * p(\mathbf{C}) \qquad \textbf{Equation 30}$$

Where $p(X|S^T,C)$ is the likelihood, $p(S^T)$ and $p(C)$ are respectively the probability density function of S^T and C. The independence between C and S^T is assumed.

The noise is assumed independent and identically distributed, stationary and Gaussian with zero mean and variances $\{\sigma_i^2\}_{i=1}^{i_{max}}$. Thus:

$$p(E|\theta_1) = \prod_{j=1}^{J}\prod_{i=1}^{I} N(e_{(i,j)};0,\sigma_i) \qquad \textbf{Equation 31}$$

Where $\theta_1 = \{\sigma_i^2\}_{i=1}^{i_{max}}$ and $N(z; \mu, \sigma_2)$ represents the normal distribution with mean μ and variance σ^2. The information about the noise distribution allows to write the likelihood:

$$p(X|C,S^T,\theta_1) = \prod_{j=1}^{J}\prod_{i=1}^{I} N(x_{(i,j)};\sum_{k=1}^{K} c_{(i,k)}s^T_{(k,j)},\sigma_i^2) \qquad \textbf{Equation 32}$$

Now it is necessary to estimate the statistical distributions of the sources and the concentrations. They are estimated by a gamma function of the form:

$$G(y;a,b) = \begin{cases} \dfrac{b^a}{\Gamma(a)} y^{a-1}\exp[-by] & \text{for } y \geq 0 \\ 0 & \text{for } y < 0 \end{cases} \qquad \textbf{Equation 33}$$

Where $\Gamma(a)$ is the gamma function.

Upon hypothesis of mutual independence of signal and mixing coefficients the prior densities of S^T and C are expressed by:

$$p(S^T|\theta_2) = \prod_{j=1}^{J}\prod_{k=1}^{K} G(s^T_{(k,j)};\alpha_k,\beta_k) \qquad \textbf{Equation 34}$$

$$p(C|\theta_3) = \prod_{i=1}^{I}\prod_{k=1}^{K} G(c_{(i,k)};\gamma_k,\lambda_k) \qquad \textbf{Equation 35}$$

where $\theta_2(k) = (\alpha_k, \beta_k)$ are parameters of the statistical distribution of the s_k^T row of the matrix S^T and $\theta_3(k) = (\gamma_k, \delta_k)$ parameters of the statistical distribution of the c_k column of the matrix

C. The Gamma function ensures positivity of sources and concentrations. Hyperparameters θ_1, θ_2, θ_3 are not known a priori and must also be estimated. Thus Equation 30 becomes:

$$p(\mathbf{S}^T, \mathbf{C}, \theta | \mathbf{X}) \propto p(\mathbf{S}^T, \mathbf{C} | \mathbf{X}, \theta) * p(\theta) \propto p(\mathbf{X} | \mathbf{S}^T, \mathbf{C}) * p(\mathbf{S}^T) * p(\mathbf{C}) * p(\theta) \qquad \textbf{Equation 36}$$

From Equation 32, Equation 34, Equation 35 and Equation 36, $\mathbf{S}^T$, C and hyperparameters θ_1 θ_2, θ_3 are simulated via Markov Chain Monte Carlo (MCMC) methods. MCMC is a class of stochastic simulation tools for generating random variables. At the end of the simulation the M last simulations are averaged in order to give an estimate of $\mathbf{S}^T$, C and θ. Saïd Moussaouï draws a parallel between BPSS method and PMF. He demonstrates that the resulting criterion to be minimized in BPSS presents similarities with the PMF criterion, the difference resides in that noise values and terms imposing non negativity in BPSS are estimated by statistical modelling.

For computing MCR algorithms, matrices C or $\mathbf{S}^T$ might be initialized by random values or by the results given by mathematical solutions. OPA, SIMPLISMA and also PCA are the methods mostly used as initial estimates to unravel hyperspectral spectra. Iterative methods that introduce constraints based on physical meaning might then be viewed as a refinement of a preliminary solution given by mathematical constraints.

Several studies have been conducted to compare the performances of these algorithms for hyperspectral imaging demixing. In [63], OPA, SIMPLISMA, PMF and MCR-ALS are evaluated for the decomposition of Mid-IR hyperspectral imaging, as well as several initializations of MCR-ALS. A synthetic hyperspectral data cube containing six polymers was generated together with different levels of noise and spectral shifts to test the robustness of the methods. The best results were obtained using OPA extraction followed by MCR-ALS, this method was reported to be less sensitive to the signal to noise ratio. Andrew et al [64] tried also OPA/ALS and PCA/ALS on Raman image. The two algorithms have the same ability to extract the pure spectra and distribution maps but OPA/ALS was faster. However, no publications were found out that report unmixing of NIR hyperspectral imaging data. In this manuscript a comparison of MCR-ALS, PMF, NMF and BPSS is therefore undertaken.

- *Introduction of spatial constraints*

When dealing with imaging, introducing spatial constraints may lead to a more accurate pixel classification in homogeneous region especially when noise is present. With chemical imaging data sets, spatial information can be used in order to determine the number of species present

in one pixel such as proposed by Window-Evolving-Factor Analysis [65] procedure. In WEFA, adjacent pixels included in a "window" centered on one pixel are selected. A PCA is performed with the spectra included in the window and allows to determine the number of species present. The value is assigned to the centered pixel. The window is subsequently centered to the next pixel and so on. When the procedure has been iterated on all pixels of the data cube, a so called rank-image is obtained. Determining the number of chemical species in one pixel allows to refine the factorization by incorporating zeros in the matrix of concentration. However, additional knowledge, for example about the reference spectra, is needed in order to know which species are present or absent in the pixels. In two publications [66, 67] De Juan et al applied this method for raman imaging data sets.

In [68] a modified version of alternating least squares is presented. This version introduces a matrix of weights to improve the extraction. This matrix may be set by different ways: constant values, linear change with the iteration number, percentage of $\mathbf{S}^T$ and C at each iteration. In [69] the matrix of weights is constructed by employing a probabilistic class partition using Bayesian Discriminant clustering. Thus, at each pixel the probability of presence of the compounds is determined during the computation and acts as a spatial constraint.

Three-way modelling for hyperspectral imaging

A survey of all two-way analysis methods has been previously made, now three-way modelling might also be considered. Three-way modelling exists since the 80s and may be an alternative solution to take into account both spectral and spatial information. Using the multi-way notation a hyperspectral data cube may be viewed as an OOV matrix where O depicts an object mode (i.e observation) and V a variable mode (i.e. the spectral information).

Two models have been particularly studied in the chemometric community to deconvolve multi-way matrices. The first model is the parallel factor analysis (PARAFAC) model. PARAFAC factorizes the matrix into three sub-matrices using the same number of factors such as depicted in Equation 37.

$$x_{n,m,j} = \sum_{r=1}^{R} a_{n,r} b_{m,r} c_{j,r} \qquad \textbf{Equation 37}$$

The solution is found using alternating least squares: two matrices are fixed while the third one is estimated. Constraints of positivity might also be introduced. The advantage of PARAFAC model is the uniqueness of the solution.

Regarding images analysis three-way PARAFAC was used in one publication [70]. The study was about the diffusion of CO_2 into water. Two-ways PCA and three-ways PARAFAC [71] algorithms were compared. PARAFAC was able to explain the two features of the experiment. However, when complex spatial forms are present decomposing spatial dimensions into low-dimensional linear decompositions is not straightforward. A drawback with PARAFAC algorithm is that the same number of components is used for both O and V directions, and decomposing spatial dimensions into lower-linear dimensions may not yield appropriate decomposition for images.

The second model is Tucker3 [72] algorithm which uses different numbers of underlying components for each dimension of the matrix by introducing a fourth matrix G which reflects the importance of the interaction between factors.

$$x_{i,j,k} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} a_{ir} b_{jq} c_{kr} g_{pqr} + e_{i,j,k} \qquad \textbf{Equation 38}$$

Tucker3 might be calculated by alternating least squares [72]. Tucker3 suffers from rotational ambiguity.

In [73], Tucker3 algorithm was employed for hyperspectral imaging decomposition. However, with OOV matrix, it is not possible to reconstruct images by selecting loadings. The data cube must be reconstructed using the fourth sub-matrices yielding a de-noised data cube containing the same number of data points. The conclusion of the study was thus that unfolded methods (i.e. two-way models) would be better suited for explorative data analysis and classification purpose of OOV data such as hyperspectral data whereas multi-way methods would be more appropriate for the analysis of OVV data.

For these reasons, three-ways modelling was no further investigated in the present work, because the focus was on explorative analysis of the data cube and a reduction of the number of image plane.

III.2.2.2. Clustering techniques

Clustering techniques are the second kind of methods for multivariate analysis. The aim is to group pixels which present similar features according to a criterion. A group of pixels is designated by the term class or cluster. Such classes or clusters can be created using a criterion based on distance, probability, and a priori information. Two families of methods among clustering techniques might be distinguished. The first family needs a set of samples which are already labeled, that is to say already assigned to one class. This set will serve as a training set to compute a model. The model will subsequently be used to classify unknown samples. Such

methods are referred to supervised clustering. On the other hand, unsupervised clustering techniques aim at classifying pixels having similar characteristics without the help of references.

Up to now, the use of clustering methods for NIR-CI is quite limited, the method of choice to define class membership being PLS-DA algorithm. However we can note the use of supervised approaches.

The first linear method being used for supervised classification is Multivariate Image Analysis (MIA) [74]. In MIA, a PCA or PLS is first applied to the data sets. Then a two-dimensional score plot is generated. The pixels which have the same spectral features are localized in the same part of the score plots. MIA uses manual selection to define the boundaries of the classes. Once the score plot regions have been correctly identified and linked to image feature, pixels of new images may be classified using the model. MIA has been employed in [75, 76] for real time process monitoring.

Non-linear approaches such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) have also been considered. Non-linear methods may take into account non-linear behaviors of NIR spectroscopy, that's why they have been of particular interest. Van den Broek et al in [77] used multi-layer feedforward artificial network (MLF-ANN) for the identification of plastic material by NIR-CI. In two publications Pierna et al [78, 79] employed SVMs to classify compound feeds by NIR-CI. In the first publication they compared SVM to ANN and PLS classification and found out that SVM was more accurate with their data. In [80], SVM classification was compared to MIA for the discrimination of split and knot of lumber by NIR-CI. SVM provided an accurate classification.

In the present work, we were interested in a method that could classify pixels without user interaction. Maybe the most popular and simplest method for unsupervised clustering is the K-means (Km) algorithm, K being the number of clusters. Km is a hard classifier where a pixel is assigned to only one class. The aim is to minimize the sum of within-cluster variations [81]. Suppose that n pixels have to be clustered into K class $\{C_1, C_2,\ldots C_k\}$ and C_k has n_k number of spectra. The center of the cluster C_k is defined by the following equation;

$$m^k = \left(\frac{1}{n_k}\right)\sum_{i=1}^{n_k} x_i^k \qquad \textbf{Equation 39}$$

Where x_i^k is the i^{th} pixels belonging to cluster C_k

The within cluster variation for cluster C_k is defined by the sum of squared Euclidean distances between each spectrum (Equation 40).

$$e_k^2 = \sum_{i=1}^{n_k} \left(x_i^k - m^k\right)^T \left(x_i^k - m^k\right)$$

Equation 40

Where x_i^k is the i^{th} pixel belonging to cluster C_k

The sum of within-cluster variations is thus given by:

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

Equation 41

The steps of the Km algorithm is described below:

Step 1: Choice of the initial cluster center (ex. random initialization) and first partition.

Step 2: Generates a new partition by assigning each pixel by its closest cluster center.

Step 3: Computation of the new centers.

Step 4: Repeats 2 and 3 until convergence.

The advantages of Km algorithm are its simplicity and computational efficacy but different initializations may lead to different partitions.

Other unsupervised algorithms commonly used in chemometric are hierarchical techniques which group the spectra according to a distance criterion and organize the data in a form of dendogram or tree. Another approach is unsupervised Bayesian learning, which estimates the distribution of the features by probability functions but is computationally demanding.

III.3. Extraction of quantitative parameters

After the extraction of distribution maps, it is necessary to develop methods for the interpretation of the images. The perception of pattern in images is user-dependent and applying image processing techniques to enhance and describe patterns of interest allow to extract user-independent information.

There are several groups of processing techniques such as color and contrast enhancement, segmentation of the image into homogeneous area, edge detection and texture classification.

III.3.1. Image enhancement

Typically, images are digitalized using 256 steps from low intensity pixels to high intensity pixels. To each of these 256 values a color is assigned according to a Look Up Table (LUT).

For example, a gray level LUT assigns black for low intensity pixels and white for high intensity pixels. Intermediate shades of gray are used to display intermediate values. A "jet" LUT assigns blue for low intensity pixels and red for high intensity pixels. Intermediate colors of the visible spectrum are used to display intermediate values. Usually, the assignment of colors is made in a linear fashion. Nevertheless, to enhance the visibility of regions with higher intensities or regions with lower intensities, square root or log transfer functions may be used. Other techniques such as histogram equalization may also provide better contrast. In [82], contrast adjustment performed on the histogram of the image and mathematical filtering are applied on single band image. Mathematical filtering changes the value of each pixel depending on the values of its neighbors. The different filters might create smoothing, or edge enhancement. The advantages and drawbacks of each technique are explained and the authors suggest that depending on the application, the spectroscopist might find one or the other methods more appropriate to its image data sets.

Another useful tool to synthesize results is RGB reconstruction. In such image, each pixel has a value between 0 and 255 for red, green and blue channels, generating 255^3 possible colors. By assigning to each image plan of the RGB image the distribution map of one compound it is possible to display simultaneously the localization of three compounds.

III.3.2. Histogram analysis

Histogram is an important tool for image analysis. It plots the number of pixels as a function of their intensity values. One relative maximum of one histogram is called a mode.

The histogram of one distribution map may give information about the homogeneity of the sample. An histogram which exhibits a symmetric distribution with a narrow base and a sharp peak is representative of an image with a low contrast therefore an homogeneous sample. On the other hand an asymmetric histogram with a large base and flatter peak or having several modes is representative of a contrasted image therefore heterogeneous sample. Four metrics are typically used [16] to characterize the shape of the distribution of the image histogram.

First the mean value of the distribution might be estimated:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{Equation 42}$$

where n represents the total number of pixels within the image and x_i the value of the pixel i.

Secondly the variance describes the variation about the mean:

$$\hat{\sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2$$ **Equation 43**

Thirdly the skew measures the asymmetrical tailing of the distribution (Equation 44). A positive skew indicates tailing toward higher values and negative skew tailing toward lower values (cf Figure 18).

$$\hat{s} = \frac{\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^3}{\hat{\sigma}^3}$$ **Equation 44**

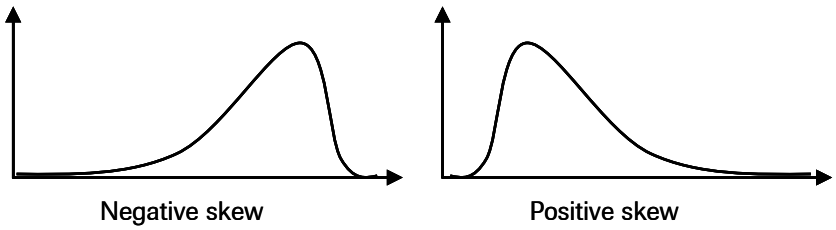


Figure 18. Skewness of a distribution

Finally the kurtosis (Equation 45) gives information about the shape of the histogram peak. The kurtosis for a Gaussian distribution is three. Therefore a kurtosis higher than three indicates sharper peaks with long tail whereas kurtosis lower than three describes flatter peaks with smaller tail such as depicted in Figure 19.

$$\hat{k} = \frac{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^4}{\hat{\sigma}^4}$$ **Equation 45**

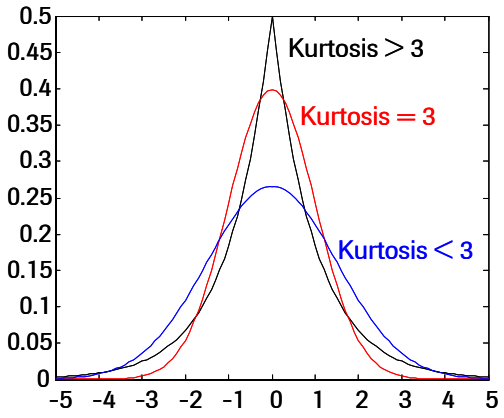


Figure 19. Kurtosis of a distribution. The kurtosis for a Gaussian distribution is equal to 3.

III.3.3. Image binarization

Image binarization aims at separating object of interest (foreground) from the background. In some cases the patterns to be segmented have nearly the same pixel intensities resulting in one mode of the histogram. Threshold values can be set between histogram modes. The pixels falling between two threshold limits belong to the object of interest and the pixels outside are rejected as background pixels. The threshold may be fixed by user interaction, which is the most widespread technique employed in the papers dealing with chemical imaging for pharmaceutical applications. The main disadvantage of this method is that fixing the threshold is subjective and the threshold value would be different among users. Thus, they are automatic methods such a Otsu's method [83] which might be an alternative solution.

III.3.3.1. Otsu's threshold

Otsu uses the first two cumulative moments of the histogram to determine the threshold K. Let L be the number of gray levels, n_i the number of pixels at level I, and N the total number of pixels. The histogram of the image is normalized and regarded as a probability distribution thus:

$$p_i = \frac{n_i}{N}, \quad p_i > 0, \quad \sum_{i=1}^{L} p_i = 1$$

Equation 46

Where p_i is the estimated probability of having a pixel of gray level i

The pixels have to be separated into two classes C_0 and C_1. K is the threshold, thus C_0 denotes pixels with levels [1….k], C_1 pixels with levels [k+1,…,L]. The mean, ω_0 and ω_1 of each of the class is given by:

$$\omega_o = \sum_{i=1}^{k} p_i = \omega(k)$$
$$\omega_1 = \sum_{i=k+1}^{L} p_i = 1 - \omega(k)$$

Equation 47

And the variances μ_0 and μ_1 are calculated using the following equations:

$$\mu_o = \sum_{i=1}^{k} \frac{i * p_i}{\omega_0} = \frac{\mu(k)}{\omega(k)} \quad \text{where} \mu(k) = \sum_{i=1}^{k} i * p_i$$
$$\mu_1 = \sum_{i=k+1}^{L} \frac{i * p_i}{\omega_1} = \frac{\mu_T - \mu(k)}{1 - \omega(k)} \quad \text{where} \mu_T = \sum_{i=1}^{L} i * p_i$$

Equation 48

Otsu chose to maximize the between-class variance, s^2, given by Equation 49 in order to find the threshold k.

$$s^2 = \omega_0\omega_1(\mu_1 - \mu_0)^2 \qquad \textbf{Equation 49}$$

which gives, using Equation 47 and Equation 48 :

$$s^2 = \frac{[\mu_T\omega(k) - \mu(k)]^2}{\omega(k)[1-\omega(k)]} \qquad \textbf{Equation 50}$$

In order to find out K, s^2 is computed for each gray level and the value which gives the maximum is chosen as a threshold limit.

The problem of image binarisation by histogram is that many images have no clear separate modes or one mode may not correspond to a distinct structure in the image. Otherwise, most of the clustering techniques (k-means, Gaussian mixtures, Neural Network etc…) presented above may also be applied to classify pixel into foreground and background leading to a segmented image.

Most of the time, image binarisation is not that accurate because of noise, and false pixel classification might occur. It is thus necessary to refine the segmentation by mathematical morphology for instance. Mathematical morphology [5] aims at changing pixel settings by a simple operation based on neighboring pixels. The two simplest tools are erosion which shrinks foreground objects in size by eroding their boundaries (set background pixels to 0) and dilation which enlarges boundaries of foreground objects (set background pixels to 1) as presented in Figure 20.

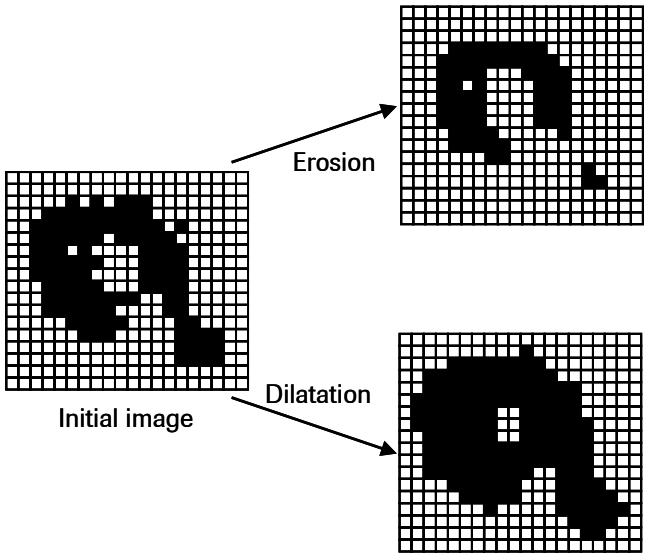


Figure 20. Principle of erosion and dilatation.

One of the major difficulties is to separate features that are touching in a binary image. Watershed [5] belongs as well to mathematical morphology and has been especially developed to separate convex objects, and thus might be of interest for particle detection.

III.3.3.2. Watershed

Watershed uses image topology where the pixel values are interpreted as heights. For example, the topology of the image depicted in Figure 21 (a) is the surface drawn on the right Figure 21 (b). The idea of watershed is to segment the image by flooding the minima ('catchment basins') of the topological function [6]. The watershed function finds the catchment basins and the ridge lines in a gray scale image.

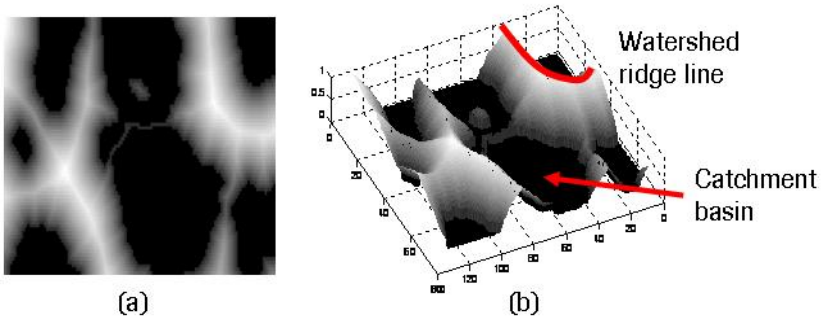


Figure 21. Grayscale image (a) and its corresponding topology (b).

Watershed are usually applied to image gradient. But because of noise, lots of minima are present in the image gradient leading to over-segmentation. In order to avoid that, the marker control watershed procedure has been developed. The aim is to define in the image the minima to be flooded. A first step will detect all foreground objects to be segmented and the second step the background. The foreground and background are imposed as minima of the gradient image and watershed transformation is applied to this new image. Figure 22 illustrates the marker control watershed procedure used later in this manuscript. The different steps of the segmentation are described below:

Step 1: Image preprocessing: image normalization, to span the pixel values in the [0 1] range

Step 2: First threshold obtained by Otsu method.

Step 3: Computation of the image gradient.

Step 4: Marked Foreground (Fgm) objects (particles to be segmented) using mathematical morphology.

Step 5: Marked Background (Bgm) objects using image obtained after step 2 and mathematical morphology such as distance function.

Step 6: Impose Fgm AND Bgm as minimum in the gradient image.

Step 7: Apply watershed algorithm to the image obtained after step 6.

Step 8: Final image obtained by fusing images of step 2 AND image of step 7.

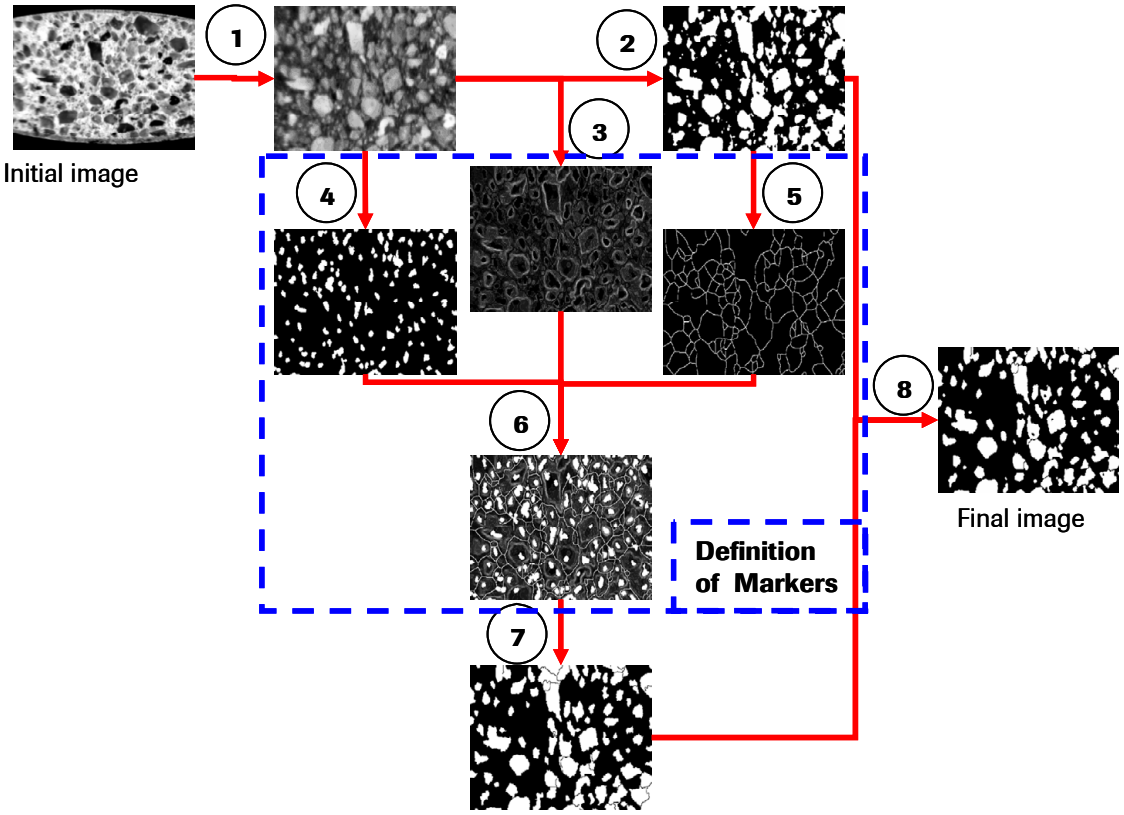


Figure 22. Scheme depicting the marker control watershed procedure used to segment particles. High steps are needed in order to find an appropriate segmentation.

Once the image has been properly segmented, it is possible to extract information about the domain size of the compounds of the samples, such as the mean and standard deviation of particle sizes, percentage of area covered, which may be correlated to the true content of the sample [16, 30, 84].

IV. NIR-CI for pharmaceutical applications

IV.1. Sample preparation and measurement

For a measurement to be as representative of a sample's chemical composition as possible, all physical effects such as roughness of the surface must be minimized. Advice for the preparation of the powders and tablets is given in [85]. Trustworthy images of powder such as raw granulation or capsule fill are difficult to obtain. A press may be employed to generate a kind of pill but it is thus not sure that the compression force does not affect the sample under investigation. For the analysis of tablets, the milling of them is recommended to remove coating and to flatten the surface (Figure 23). However, depending on the tablet hardness and its friability some of the particles might be removed or displaced from the surface of analysis.

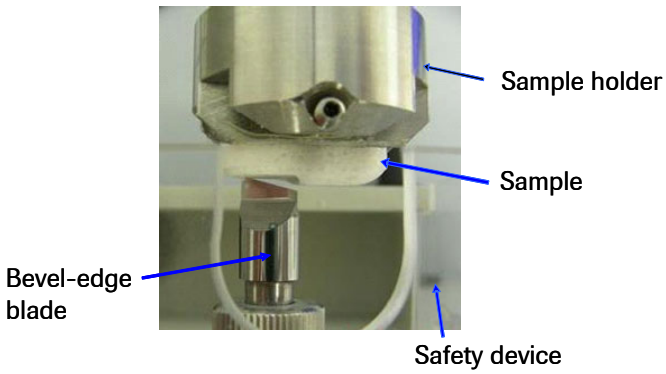


Figure 23. Instrumentation used in the laboratory to flatten the analysis surface of the tablets

The choice of the objective might also be discussed. Is it better to use high spatial resolution for the detection of fine patterns or low spatial resolution to cover a larger surface of analysis. The advice would be first to map the whole surface in a reasonable time and subsequently if minor compounds must be resolved, focusing on particular region with higher spatial/spectral resolution. Of course this depends on the particle sizes of the tablet constituents. Penetration depth of the NIR beam has been proved to be the parameter limiting the spatial resolution of reflectance measurements as explained in the section II.3.2. In [20] a limited spatial resolution of 30 µm/pixel was approximated for NIR diffuse reflectance measurement. As a conclusion, the authors claimed that usefulness of 10 µm/pixel objective might be discussed. For the analysis of pharmaceutical samples with NIR reflectance spectroscopy it should be better to use objective with lower spatial resolution (for instance 40 µm/pixel) which would enable to scan a larger surface and provide a better overview of the sample as objective 10 µm/pixel

could not actually resolves finer pattern. This assumption is supported by the study conducted by Hua Ma et al in [86]. The blend of binary mixtures containing 80% of lactose with a median particle size of 100 µm and 20% of salicylic acid with median particle size of 30 µm were under investigation using several objectives with different pixel sizes: 67.1 µm/pixel, 45.5 µm/pixel and 21.5 µm/pixel. In this publication it was reported that "*higher magnification levels did not provide additional relevant information about the surface features*" even if the spatial resolution of the higher magnification level objective was lower than the median particle size. The choice of the 67.1 µm/pixel objective for the study of this system was thus preferable.

IV.2. Distribution of chemicals

In the literature, several publications deal with the extraction of the distribution maps of the chemicals present in pharmaceutical solid dosage forms. These studies often compare several processing or acquisition modes in order to evaluate the advantages and drawbacks of each methods. Regarding NIR reflectance measurements, the distribution maps of compounds of pharmaceutical tablets have been revealed in [87-89] using single wavelength images, RGB reconstruction or PCA analysis. In particular, in [89] the distribution maps of the compounds of a time release granule have been successfully extracted. The type of tablet was a complex system built up with several layers, each having its own functionality in the release of the medicine. Thus, their distribution is an important factor in the dissolution behavior of the medicine, and may be checked by NIR-CI.

Slobodan Sasic in [90] compared NIR chemical imaging and Raman imaging and found out that, using the same magnification levels, Raman imaging combined with multivariate analysis enabled to detect all five compounds in the tablet while NIR imaging failed to extract the two minor ones. The same author, in a recent publication [91] evaluated Global raman imaging and NIR chemical mapping for the detection of the chemical composition of granulates. It was revealed by both techniques that the granules were mainly a mixture of API and mannitol whereas some granules contained pure compounds. For this formulation Raman imaging allowed a better visualization of the compounds.

IV.3. Blend uniformity

During the pharmaceutical production, the homogeneity of blends must be assured, otherwise it can affect the quality attribute of the medicine, such as dissolution behavior, or lead to tablets which do not contain the right amount of API. The classical method to test the content

uniformity of the blend is to extract by the help of a sample thief an amount of powder and check its content by classical analytical methods such as HPLC or UV-vis-spectroscopy. Since the appearance of chemical imaging, this technique has also been tested as an alternative method to monitor the blend uniformity.

In [92] classical NIR spectroscopy and chemical imaging were investigated as potential noninvasive methods. The three binary mixtures were composed of Salicylic Acid (SA) and Fast-Flow lactose with different amounts of SA. The classical NIR analysis was performed at different positions in the blender through six sapphire windows. For the imaging part, six images taken at different wavelengths which encompassed lactose and SA peaks were acquired. It was necessary to open the slides of the blender to perform the imaging measurements which were done directly into the blender (no sample removal). Reference method which consisted in traditional sampling combined with UV spectroscopy was used in parallel to check the accuracy of the NIR methods. Cluster analysis and moving block Standard Deviation (SD) along wavelength values (spectral space) were employed to analyze the classical NIR spectra. Regarding imaging, the moving block Standard Deviation procedure was applied in the spatial space. In this study it was demonstrated that NIR spectroscopic measurements were in accordance with the reference methods. NIR imaging allowed to sample a larger surface (images of about 15 cm^2) but only the upper layer in the experiment configuration.

In [93] tablets were produced with different mixing times. Those tablets were subsequently analyzed by NIR chemical imaging and classical NIR spectroscopy. Commercial tablets were also used for further comparison with experimental tablets. Histogram analysis of univariate image or of images extracted by PLS method were employed to assess the homogeneity of the constituents in the final tablet. NIR imaging showed better ability to detect slight problems of homogeneity in comparison with classical NIR spectroscopy. Advantage of NIR spectroscopy was that it allowed to investigate micro domains of the tablets whereas classical integrating NIR spectroscopy overlooked slight inhomogeneity.

Recently, a prototype for on-line inspection of blend homogeneity by the help of fiber bundle and chemical imaging has been shortly described in [94] while still in development. An article which appeared in 2005 on the website of in-PharmaTechnologist.com [95] related the development of the device which involved a collaboration between Spectral Dimension (now Malvern) and Pfizer. This kind of equipment would allow an on-line characterization of the blend via chemical imaging.

Other techniques such as histogram analysis for checking uniformity of tablets might also be used. As example, in [96] the wavelet transform combined with PCA was employed to

determine texture of pharmaceutical tablets produced with different amounts of Paracetamol. The detection of the aggregation of Paracetamol was possible with such a processing technique.

IV.4. Content uniformity

The potential of chemical imaging to calculate the concentration of compounds in samples has been investigated in several publications. Jovanovic et al [39] studied first binary mixtures of protein-sugar (lysozyme-trehalose), with concentration of 0, 10, 50, 80, 100 % (w/w) of lysozyme to compare methods for content. In this publication, correlation coefficient between mixture spectra and reference spectra was compared to PLS-DA regression built with the average of reference mixtures to determine the concentration of each of the constituents. The method is subsequently applied on dried and freeze dried samples containing 50% of lysozyme and 50% of trehalose. PLS-DA regression demonstrated better ability to predict concentration in the calibration mixtures, but predictions with dried and freeze dried samples were worst, probably due to extra spectral variations in those samples. Homogeneity of the mixtures was ensured via chemical imaging.

Recently [97] NIR high-throughput approach has been proposed for content of API. Instead of measuring separately samples from calibration sets and samples to be predicted, a large field of view (59.5 mm * 47.5 mm) was employed and tablets were measured simultaneously. A single image at 1600 nm (API peak) featured a trend correlated to the API concentration and thus this wavelength was chosen to check for content uniformity. A calibration curve based on NIR intensities at 1600 nm was computed by the help of calibration samples and unknown samples were subsequently predicted. The analysis is rather fast and showed reliable predictions of API in comparison with the UV measurements. Analyzing at the same time calibration samples and samples to be predicted allows to overcome the problem of day to day repeatability and stability of the instrument (aging of lamp power for example) moreover if one wavelength is sufficient then the measurement is fast and the amount of data to be stored reduced. The authors assumed that by enlarging the FOV, it should be possible to analyze 1500 tablets at once leading then to 100% control.

IV.5. Process understanding, troubleshooting and product design

Chemical imaging plays nowadays an important role for the development of new formulations and to solve process troubleshooting issues. Spatial distribution of the compounds is an important factor for the behavior of the medicine, i.e. its dissolution profile, stability, bio-

availability. Chemical imaging appears as a well suited solution to get insight view of solid dosage forms.

In a book section [10], four applications of NIR point mapping to solve process issues were depicted. The first application dealt with batches with different flow characteristics. The inspection of the blend by NIR chemical imaging revealed that the root cause of bad flow properties was due to the distribution of the lubricant which was not uniform. In the second example the problem was about tablet sticking to the press punches. NIR mapping revealed areas with higher amount of an inorganic filler on sticking tablets. The third example dealt with tablet chipping after a change in the supplier of sugar. The batches generated with the sugar from the supplier 2 depicted a better mixing of sugar with the inorganic filler forming weaker tablets more prone to chipping. The last example checked which excipient of the tablet would absorb water. The experiment showed that it was the lubricant.

In [30] two different sample sets were under investigation. The first sample set presented issues during its manufacture. The NIR images were generated using line mapping and processed using PCA analysis and RGB reconstruction. In the bad samples, larger domain sizes of polymer components were revealed. In the second sample set, it was found that the dissolution properties of tablets depend upon roller compaction force. Image processing revealed increasing cluster size of disintegrant material as a function of compression force.

In [98] NIR chemical imaging has been used to assess the difference of particle sizes of tartaric acid on top and bottom of tablets. Three different batches were produced using different API particle sizes. It was reported that there was a statistical difference regarding particle sizes between top and bottom sides of tablets when produced with larger API particles. In fact, it appeared that the segregation of API occurred during the compression phase, when using larger API particle size. Thus segregation can be minimized by choosing API with fine particles.

Recently, it has been shown [99] that NIR chemical imaging may also detect variations due to differences in tablet compression force and hence be used as an analytical tool to understand physical variations. In [84] blending produced with different API particle sizes were scanned via NIR chemical imaging and it was demonstrated that larger particle sizes led to blends featuring larger API aggregation.

IV.6. Counterfeit and identification

Counterfeit is nowadays a serious problem because it can damage human health by supplying non appropriate substances or products devoid of API. Chemical imaging is an efficient tool to detect rapidly counterfeit. As an example, Figure 24 depicts a suspected counterfeit (left) and a product from F. Hoffmann-La Roche (right). In the pictures taken by a camera in the

visible (Figure 24 (a)), a difference of homogeneity might be detected. The suspected counterfeit figures out large clusters of white powder. After analysis under the NIR camera (Figure 24 (b)) and PCA, the two products were clearly discriminated, PCA scores revealed that the chemical composition was actually different. Advantage of NIR imaging is that a conclusion might be drawn in less than one hour which is fast compared to conventional analysis.

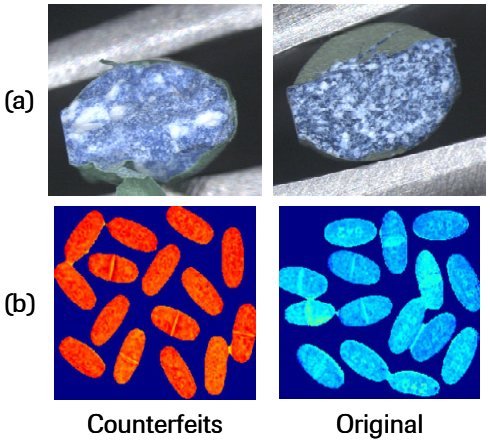


Figure 24. Fast identification of Roche counterfeit by NIR imaging and Principal Component Analysis.

In [100] NIR chemical imaging was demonstrated to be an efficient tool for discrimination between fake and real medicines. Mostly, the samples and the real medicine as a reference were analyzed in the same field of view. PCA score plots revealed samples containing no API or having a different chemical composition. In [101] Multivariate Image Analysis was employed to discriminate the counterfeit from the real product. NIR imaging appeared to be able to detect small differences among tablet uniformity.

Moreover, internet websites proposing medicine with lower price are now more and more widespread. The quality of such products may not be the same as standard products and must be checked out. For example in [102, 103] generic tablets imported from Mexico, India, Thailand, Brazil, Canada and obtained via the Internet were compared to the product of the US innovator by NIR chemical imaging. The study demonstrated that the products imported from the Canada showed similar blend uniformity of API as the US product and thus should be of the same quality. However, tablets from other countries revealed differences in API distribution and particularly API aggregation which may be an indicator of lower drug quality. In another publication [40] NIR chemical imaging was used among other analytical tools to assess quality of different internet drugs. Several internet products failed to fulfill the necessary

quality attributes. Moreover, differences in chemical composition and compound distributions were revealed between the US product and the internet products.

IV.7. Analysis through blister

Because of the spectral window, the beam can penetrate through the blister and thus reveal the capsules. As an example, in Figure 25, the blister is placed under the filter 1930 nm. Empty and filled capsules can clearly be detected.

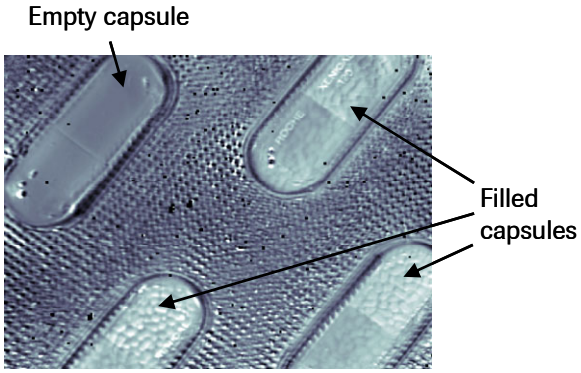


Figure 25. The analysis of capsules through blister is possible with NIR imaging. Empty capsules can be detected.

In [104] the ability of multispectral imaging to evaluate water content and identify thousands of individual tablets through blister was under investigation. For the water content analysis a pinhole was created in the blister and the tablets were exposed to water during 24 hours. 6 tablets were picked up each hour for calibration. Calibration was obtained by measuring 43 tablets at 0.5 m. A standard error of performance of 0.06% was obtained. The second part of the study revealed that about 1300 tablets can be visualized simultaneously, and still discriminate between salicylic acid and acetylsalicylic by PCA analysis which was also stated in [105]. In this last publication, single capsules exposed during different times to atmosphere containing formaldehyde were also analyzed from long distance (0.5km). At such long distance it was still possible to draw a calibration curve for prediction of exposure time to formaldehyde revealing the power of NIR imaging.

V. Conclusions

This first chapter demonstrated that NIR imaging is a powerful tool for the analysis of pharmaceutical solid dosage forms. Thank to the inherent properties of the NIR radiations, the sample needs little preparation, images at the macro-scale can be acquired in a few minutes and give truthfully information about the distribution of the sample's constituents. However, the identification of chemical species is not straightforward and statistical analysis is required. Numerous existing chemometric methods have been developed in the past and have proven to extract reliable distribution maps and to allow compound identification, even for low-dosage tablets. It is then only logical that such a technique has been extensively used to solve pharmaceutical issues when spatial information is relevant. NIR-CI can be applied at each stage of the development of a new formulation, from the characterization of raw materials to process optimisation, process troubleshooting, quality control and counterfeit detection. However, two main issues arise when analysing pharmaceutical samples. The first one is how to characterize objectively sample using image analysis. The second issue is how to determine the percentage of the API or excipients. Those issues are further addressed in the next two chapters of the present manuscript.

Related publications

***C. Gendrin**, Y. Roggo, C. Collet,* Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review, Journal of Pharmaceutical and Biomedical Analysis, **48** (2008), p. **533-553.**

Chapter 2

Compound localization and characterization of samples by NIR-CI.

I. Introduction

The previous chapter has introduced the theory, instrumentation of NIR hyperspectral imaging and chemometrics available to process the data. It has been shown that the applications of NIR imaging to the pharmaceutical industry are various and can be applied when spatial information is relevant. In the present chapter, the qualitative analysis of pharmaceutical solid dosage forms by NIR-CI is studied. The aim is to extract the distribution maps of the compounds and to characterize them for the objective comparison of images.

The first part deals with samples from the pharmaceutical development. In this case, the formulation of the tablets is fully known and strong a priori information, such as the reference spectra, is available and can be used to extract the distribution maps. They are subsequently characterized by the help of image processing tools such as histogram analysis and image binearization in order to evaluate the influence of two process parameters on the homogeneity and particle sizes of two intermediates.

In the second part, the case in which the constituents and thereby reference spectra are not known is addressed. Multivariate Curve Resolution algorithms, namely, MCR-ALS, NMF, PMF and BPSS for the extraction of distribution maps are compared and discussed. Especially a novel approach to investigate the rotational domain with PMF is proposed.

II. NIR imaging analysis of samples issued from the pharmaceutical development

During the development of new pharmaceutical products several parameters such as the mixing time, the particle sizes of the API or excipients and their chemical forms have to be optimized in order to find the right formulation which would allow good bio-availability of the medicine and the most efficient cost production. Each of the batches produced during this campaign is subsequently analysed by classical techniques, for instance sieve analysis to assess the particle size distribution of the intermediate materials, HPLC for content analysis or dissolution test. One of the main issues is the homogeneity of the samples and NIR imaging appears to be an adequate tool for this task because it is rather fast and gives information about the spatial localization of the chemical species in tablets.

II.1. Process and parameters

The present paragraph gives piece of information about the process which led to the production of the intermediates under investigation. However, because of confidentiality reasons, it is not possible to detail it thoroughly and to give precisely the formulation of the samples. Nevertheless, the information which is provided is sufficient for the comprehension of the study.

II.1.1. Process workflow

The process under investigation was based on extrusion which increases the drug solubility by dispersing the API in a polymer matrix. The extrusion is defined by J. Breitenbach [106] as "a process of converting a raw material into a product of uniform shape and density by forcing it through a die under controlled conditions".

In our example, the process workflow was constituted of five main steps and two kinds of process intermediates were taken along the line. Figure 26 depicts the process under investigation. The first step was the mixing in a blender of the API with one poloxamer. The powder blend was then passed through the extruder. The first samples were withdrawn after cooling of the extruded material. They are named the extrudates in the following of the dissertation. The material was subsequently milled. Six excipients were added and the powder was mixed until homogeneity of the material was reached. The mixture was then compressed into cores which were the second process intermediates. Pictures of the two process intermediates taken with a camera are given on the right of the Figure 26. The top right

picture depicts extrudates. They had a tube shape and were simply laid down on a mirror plate for the NIR images acquisition. The bottom right picture depicts the cores. They were cut according to their length by the help of a milling tool (Figure 23) in order to flatten the surface and to avoid problems of focus during the acquisition of the NIR data cubes.

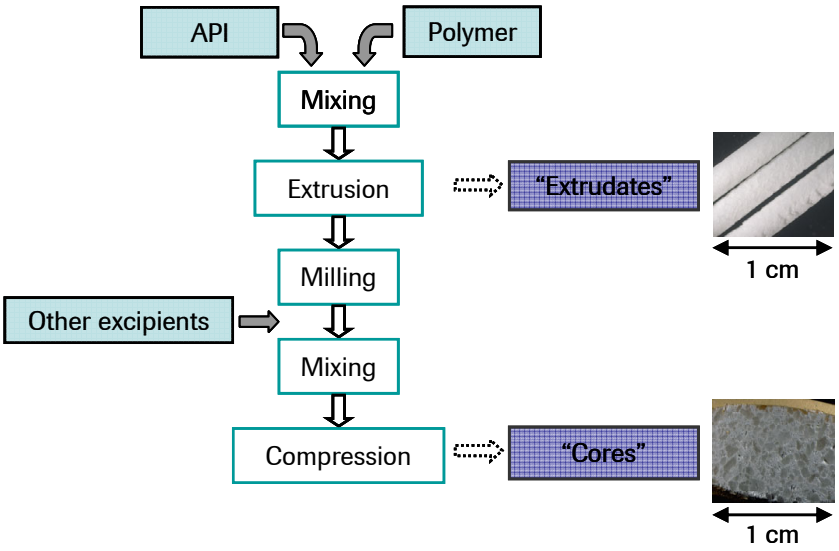


Figure 26. The process workflow employed to produce the tablets. It was constituted of five main steps. The two process intermediates which were studied were the extrudates (top right picture) and the cores (bottom right picture).

II.1.2. Parameters under investigation

In this example four batches were compared. They were produced using one API variant and one process variant (Figure 27). Two API powders with different mean particle sizes were employed. The first API featured large particles whereas the other one presented finer ones. The process variant was introduced during the extrusion. The material flew through the extruder by the help of a screw and two different screw speeds were tested. The batches produced according to this development plan were indicated by the letters A, B, C, D (Figure 27). Batch A was produced using large API particles and low screw speed, batch B was produced using large API particles and high screw speed, batch C was produced using fine API particles and low screw speed and batch D was produced using fine API particles and high screw speed.

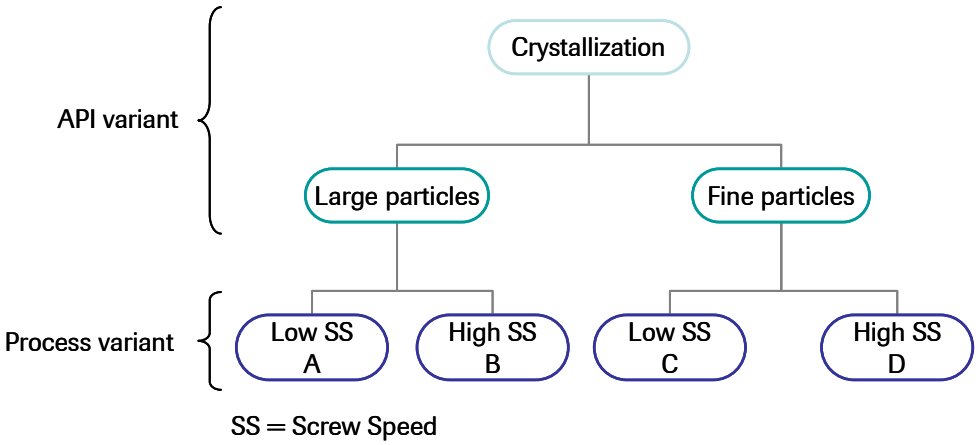


Figure 27. Process variants and corresponding batches. One API variant: the particle size and one process variant: the screw speed were under investigation. Four batches, A, B, C, D were then analyzed by NIR-CI.

II.2. Study of intermediate homogeneity

II.2.1. Material and preprocessing

The samples and the pure powders of the API and each excipient were analyzed by the Sapphire. The acquisition parameters were 16 co-adds and a spectral range of [1100-2450] nm at 10-nm increments. The objective 40 µm/pixel providing a field of view of 1*1.3 cm^2 was employed. With this objective, almost the whole surface of the tablet was imaged.

After the acquisition, the data cubes were preprocessed. First, the spectra were converted in absorbance unit and reduced to the spectral range [1300-2330] nm in order to remove the noisy wavelengths. Then, since the samples did not cover the entire field of view, a mask was applied in order to remove pixels which depicted the mirror and not the sample.

II.2.2. Reference spectra

As stated in II.1.1, besides API and poloxamer, six excipients were added after the milling of the extrudates. Three of these excipients were minor excipients and each of them account for less than 2% of the formulation. It was difficult to detect them with our imaging system. Therefore, the present study discarded the three minor excipients in order to focus on the five major compounds, namely: the API, the poloxamer and the three major excipients.

The reference spectra of these five compounds were acquired by the help of the pure powders. They were normalized by the help of a SNV and are depicted in Figure 28. On this figure, the

absorbance bands are linked to the chemical bounds. The API and Poloxamer absorb between [1600-1800] nm due to CH first overtone and between [2200-2300] nm due to CH, CH_2 and CH_3 combinations. The API also features a band in the spectral range [2000-2180] nm due to NH combinations. Absorption bands of excipient 1 and excipient 2 are mainly due to OH first overtones between [1450-1650] nm, OH combinations between [2000-2200] nm and CH, CH_2 combinations between [2200-2300 nm]. Excipient 1 also absorbs around 1900 nm (OH combinations). The third excipient features absorption around 1700 nm (CH first overtone), between [1880-2000] nm due to CO stretch (second overtone) and also between [2250-2350]nm due to CH_2 bound.

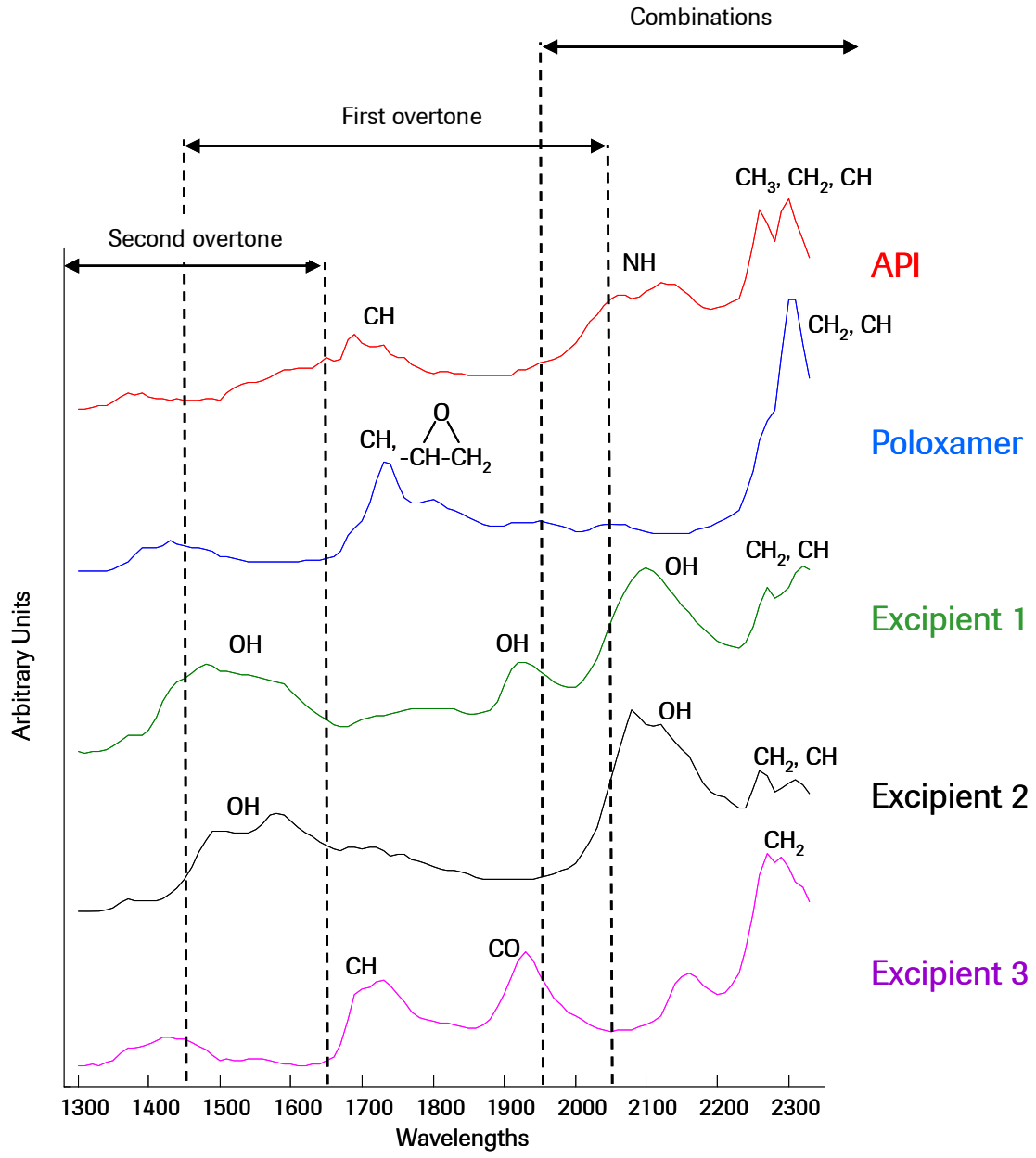


Figure 28. Reference spectra of the five main compounds of the tablets. The spectra were normalized by the help of a SNV. For better visualization, the spectra were offset.

In the NIR range, absorbance bands of the chemical species overlap strongly. Especially, API overlaps with poloxamer; excipient 1 with excipient 2; excipient 3 with excipient 1; API with the poloxamer. In order to enhance the spectral variation due to chemical information it was thus necessary to apply a Savitzky-Golay second derivative with a 9-points window, and a polynomial order of 3 (Figure 29). This preprocessing was applied on all data of our study.

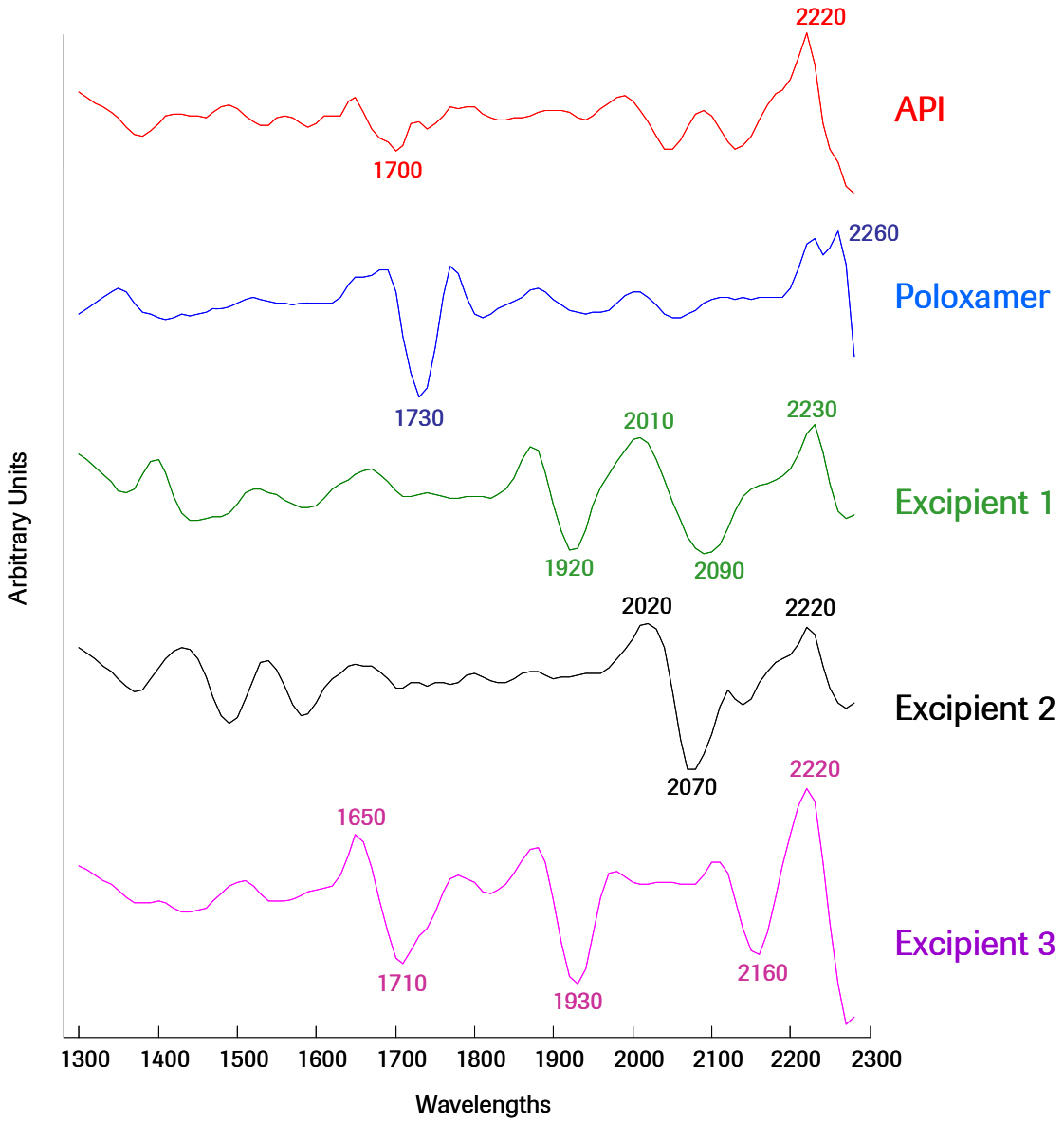


Figure 29. Reference spectra of the five main compounds of the tablets. The spectra were normalized by the help of a SNV and derived by the help of a Savitsky-Golay second derivative (window 9 and polynomial order 3). For better visualization the spectra were offset.

The extrudates contained only two chemical species: the API and the poloxamer. On the second derivative spectra (Figure 29) the wavelength 2260 nm is specific of the poloxamer and

do not overlap with specific wavelengths of the API. Images at this wavelength were therefore chosen in order to compare the homogeneity of these first intermediates.

The cores were constituted of the five compounds and it was difficult to find out specific wavelengths for each of them, even on the second derivative spectra. Especially, the specific wavelengths of excipient 1 overlap with the specific wavelength of the excipient 3 at 1930nm and the specific wavelengths of the excipient 2 at 2020nm, 2070nm and 2220nm. For these reasons, multivariate analysis was more appropriate to extract the distribution maps of the compounds of the cores and has been applied in the present study.

II.2.3. First intermediates: extrudates

After spectral normalization and second derivative, the image at wavelength 2260 nm was extracted from each of the data cube in order to assess the extrudate homogeneity. The four images were displayed according to the same "colorbar" for between image comparison (Figure 30 (a)). High intensity pixels in red depicted areas with higher poloxamer signal, thus, in some extent, higher content of poloxamer. Batch B seemed to be the most heterogeneous, with large poloxamer flakes appearing all over its surface. It also presented the larger contrast with pixel values below -0.06 and above -0.02. The top extrudate in image A featured localized pixels with high intensity (white circle). As well, the bottom extrudates of images B and C presented areas with lower intensity pixel (blue circles). This was probably due to problem of illumination. The extrudates under study were of round shape and even after spectral correction and spectral normalization, illumination unevenness might holds.

In order to quantify the differences in the images, their associated histogram were studied. Since the extrudate did not cover the whole field of view, the total number of pixels which arose from the samples varied from one picture to the other. In order to enable the comparison of the histograms, their normalization was necessary. The gray level bins were divided by the total number of pixels excluded those from the background. The resulting normalized histograms are shown on Figure 30 (b). The green histogram associated with image B, is the most asymmetric with the largest base and the flattest peak, representative of a contrasted image therefore heterogeneous sample. Other histograms were similar with finer peak and smaller base. Histogram A was shifted in comparison with the other ones, revealing maybe a lower content of poloxamer, or a difference of global illumination.

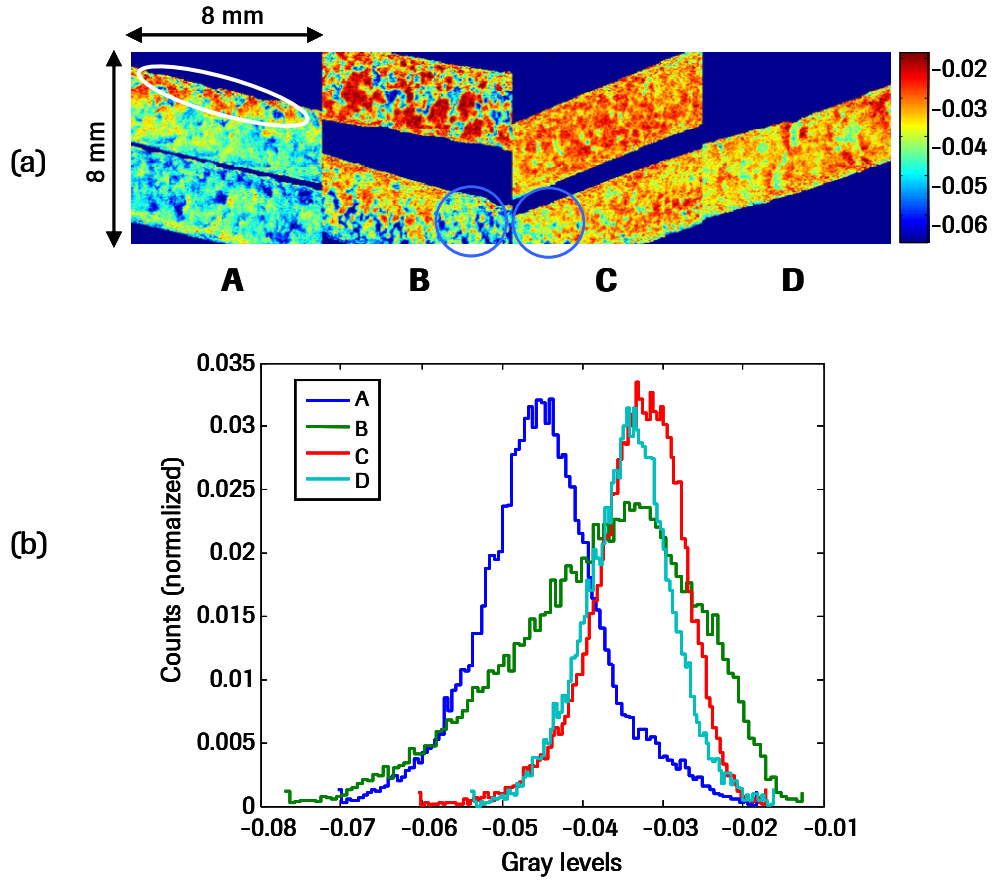


Figure 30. Images extracted at 2260 nm (a), polymer wavelength, and their associated normalized histograms (b). Image B was the most contrasted, its histogram presented the flattest peak, largest base and a tail toward lowest values.

The statistics calculated from the histograms are given in Table 2 (next page). They supported the assumption that sample of the image B was the most heterogenous because histogram B presented the largest variance (13.4 $*10^{-5}$), a negative skew and the lowest kurtosis (2.90). Histograms A and D were quite similar with variance of 6.17 and 3.11, kurtosis of 3.58 and 3.38 respectively. The most negative skew was computed for histogram C revealing a tailing toward lower values. This tailing was certainly due to the problem of illumination reported on the above paragraph (Figure 30 (a), image C, blue circle). Overall, the low variance and high kurtosis of histograms A, C and D revealed lower contrasted images than image B, thus more homogeneous extrudates.

	A	B	C	D
Mean	-0.045	-0.038	-0.033	-0.034
Variance (10^{-5})	6.17	13.4	3.43	3.11
Skew	0.11	-0.51	-0.67	-0.1
Kurtosis	3.58	2.90	4.17	3.38

Table 2. Statistics of the histograms of the Figure 30 (b). The statistics supported the assumption that image B was the most contrasted, because its histogram had the largest variance, a negative skew and the lowest kurtosis .

Histogram analysis revealed that image B was the most contrasted, thus the batch B extrudates were the most heterogeneous. This batch was produced with the largest particles of API and the highest screw speed. However, no differences were revealed in the batches A, C and D by NIR imaging.

II.2.4. Second intermediates: the cores

As stated before (paragraph II.2.2), the cores encompassed five compounds whose spectral signatures overlapped. In this case, multivariate analysis was necessary to extract distribution map of the compounds, especially of the excipient 1. Since the reference spectra were known, either Classical Least Squares or Partial Least Squares-Discriminant Analysis could have been used. The two algorithms were applied for comparison. For CLS the mean spectra of the preprocessed reference data cubes were computed. For PLS-DA a library containing 82 pretreated spectra for each compound (in total 410 spectra) was first built and the PLS-DA model computed with five latent variables. The next two figures provide the distribution maps of one core which were extracted by the CLS (Figure 31) or PLS-DA (Figure 32) algorithms. For those samples the distribution maps were qualitatively quite similar, with identical range of pixel intensities from one algorithm to the other one. In the following, CLS algorithm was employed.

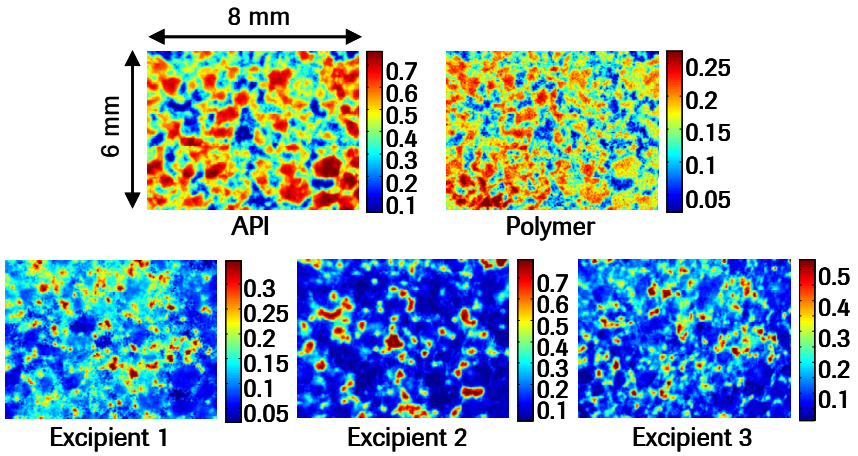


Figure 31. Distribution maps extracted by the help of the CLS algorithm

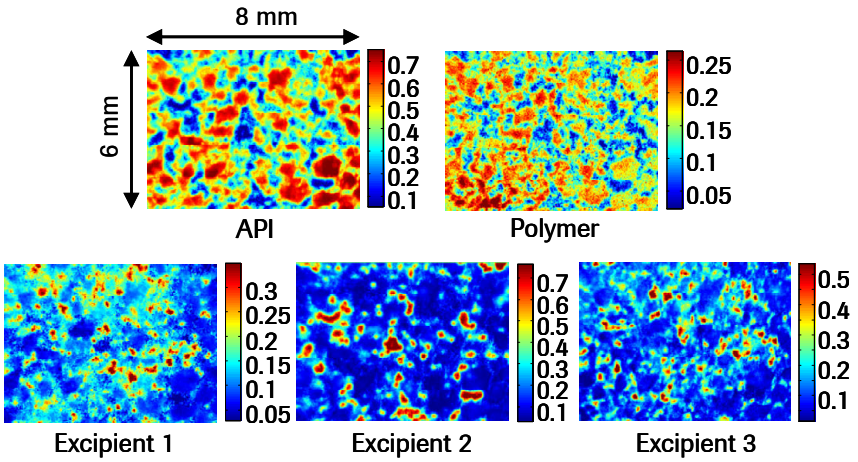


Figure 32. Distribution maps extracted by the help of the PLS-DA algorithm

The distribution maps extracted by the multivariate algorithms revealed that API and poloxamer were localised together in the larger particles. They were thus identified with the particles produced by the milling of the extrudates. In this case study, the NIR imaging device did not provide a spatial resolution fine enough to separate clearly the two compounds. Pixels with higher intensities for the API distribution maps were about 0.7, and 0.25 for the poloxamer. This was in accordance with the concentration which was of 70% of API and 30% of poloxamer in the milled extrudates.

The excipients distribution maps were complementary to each other. The excipients were then present at sparse location in the core. They aggregated with domain sizes of several hundreds of microns.

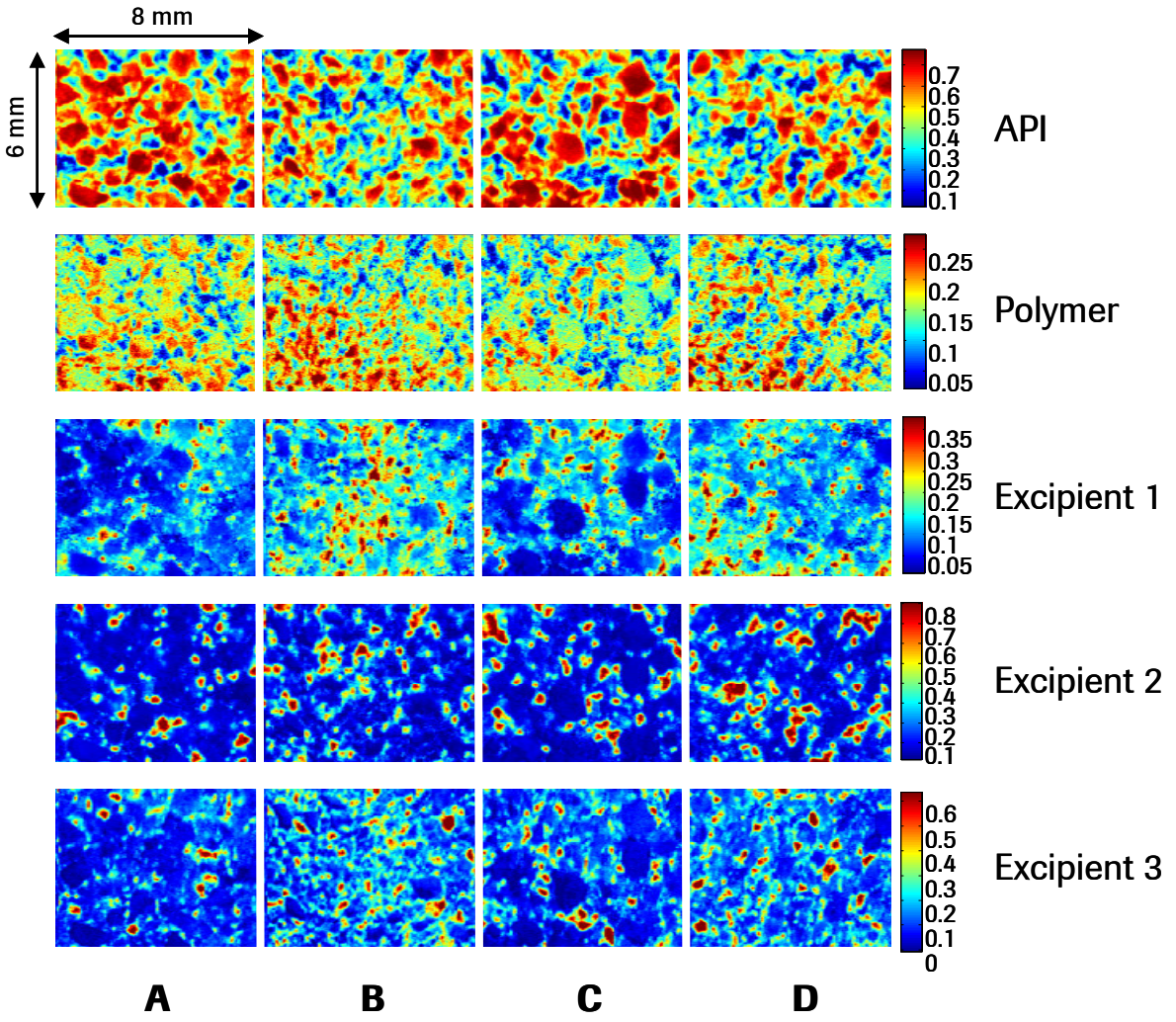


Figure 33. CLS distribution maps of the API, poloxamer, excipient 1, excipient 2 and excipient 3 for each batch A, B, C, D.

CLS algorithm was subsequently applied to each batch. The above picture figures out the distribution maps of the API, poloxamer and the excipients (Figure 33) of one core per batch. A difference in the particle sizes of the milled extrudates (API distribution maps) was qualitatively revealed : batches A and C featured larger particles than batches B and D. Thus, higher screw speed during the extrusion, probably led to extrudates that were more friable thereby producing smaller particles during the milling stage. The distribution maps of the excipients were quite similar from one batch to the other one. This was in accordance with the fact that no process variants were tested after the milling step. The chosen parameters led to regular blending of the excipients.

II.3. Segmentation of particles

With the help of multivariate analysis, the distribution maps of the compounds were extracted, revealing that higher screw speed led to finer API-poloxamer particles. The question was then to know if this difference can be detected by image processing such as binearization.

II.3.1. Material and methods

Two new batches were produced using two different screw speeds. The batch produced with low screw speed was named L_SS and the batch produced with a high screw speed (3 times faster than the low screw speed) was named H_SS. Ten cores of each batch were cut according to their length and analysed in order to have relevant statistics about the particles.

During the previous study, it has been detected that the raw image at 1930 nm also revealed the particles of the milled extrudates. As example, in Figure 34, the left picture is the raw image where large black particles appear on the core surface. The picture on the right is the corresponding distribution map extracted by the help of the CLS algorithm. The area depicted by the distribution map corresponds to the area indicated by the red rectangle on the live image. The black particles of the live image were linked to the particles in red in the distribution map. They were the particles of milled extrudates which were of interest. Thus, it was decided to apply image binearization on the live images in order to provide a fast method of detection that could be eventually applied on-line because it would require the acquisition of only one image at one wavelength.

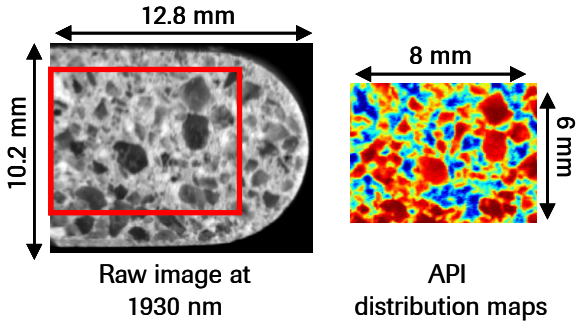


Figure 34. Raw image of the core at 1930 nm and API distribution maps. The "black" particles in the raw image correspond to the milled extrudates.

Two procedures for the segmentation of the particles were tried out. The first method was a threshold using the method of Otsu. It was named method 1. The second method used Otsu followed by marker watershed refinement as explained at the end of Chapter 1III.3.3. It was named method 2.

II.3.2. Image segmentation using Otsu threshold and watershed

Prior to the segmentation each image was preprocessed. The values of the pixels of the image were first normalized so as to be comprised between 0 and 1. The interval [0 1] was divided into 255 steps thus 256 graylevels were available. The images were also inverted. The pictures on the left of Figure 35 depict the preprocessed live images.

The results of the segmentations obtained by the method 1 and method 2 are respectively displayed in the middle and right pictures of the Figure 35. The first row presents the segmentations of one image from batch L_SS whereas the second row presents the segmentations of one image from the batch H_SS.

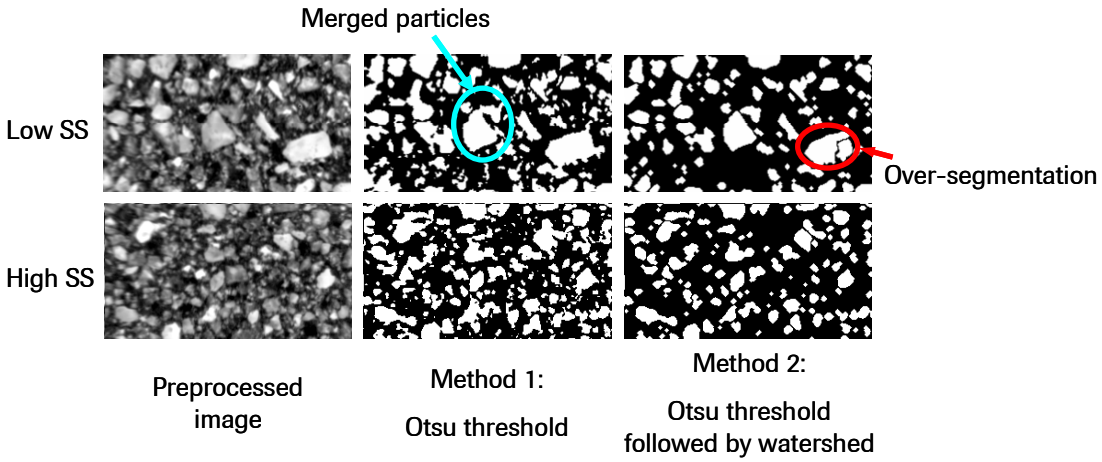


Figure 35. Preprocessed images at 1930 nm (left) and segmentations obtained after a Otsu threshold (middle image) and after refinement by watershed (right image). First row, core produced with high screw (batch H_SS) speed and second row core produced with low screw speed (batch L_SS).

It can be clearly seen that Otsu threshold led to particles that were merged together (blue circle on Figure 35). On the other side, over-segmentation occurred when the segmentation was refined by mathematical morphology and watershed (see red circle on Figure 35) but the particles were better separated due to the fact that the spatial information was taken into account by the watershed procedure. Moreover, the images were also cleaned out with the finer particles disappearing. Thus in order to avoid bias in the comparison of the mean particle size computed by method 1 or method 2, only the particles with an area larger than 50 pixels were considered.

		S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10
Method 1	L_SS	0.54	0.46	0.47	0.40	0.54	0.51	0.43	0.48	0.46	0.61
	H_SS	0.48	0.44	0.49	0.57	0.41	0.38	0.44	0.53	0.40	0.52
Method 2	L_SS	0.32	0.29	0.26	0.25	0.28	0.26	0.27	0.27	0.31	0.32
	H_SS	0.26	0.27	0.26	0.25	0.25	0.22	0.24	0.26	0.23	0.25

Table 3. Mean area of the particles (in mm^2) of each sample of each batch. S. = sample.

The mean particle size computed for each image of batch L_SS and H_SS are given in Table 3. For each sample, the mean particle size was smaller after refinement by watershed which was expected because it separated convex objects in the image. However, it was impossible to know which methods led to the most accurate prediction because the true size of the particles were not known. Nevertheless, the methods were compared by Analysis Of Variance (ANOVA) test. The aim was to determine if the screw speed yields significant different results. ANOVA compares the variation due to the controlled factor (in our case the screw speed) and the variation due to the random error given by the analysis (in our case the image segmentation) [107]. The null hypothesis is that the mean of the variation due to the controlled factor is equal to zero and this assumption is verified by the help of a Fisher test. If the Fisher test is smaller than the tabulated value, the null hypothesis is accepted and the controlled factor does not lead to significant results. If the fraction of variance is smaller than 1, it cannot be compared to the tabulated Fisher value because a Fisher test is always larger than 1, but it can be concluded that the controlled factor is not significant [107]. In our case, the tabulated value of the F-distribution with 1 (two different screw speed are compared) and 18 degrees of freedom (20 measures were done) is 4.41.

ANOVA was first computed with the mean values given by Otsu threshold (method 1). The results are given in Table 4. The Fisher test was equal to 0.69. The screw speed did not lead to difference in the mean particle size.

Source of variation	Sums of squares (SS)	Degrees of freedom	Mean Square	Fisher test
Screw Speed	0.00259	1	0.00259	0.69
Residual	0.06768	18	0.00376	
Total	0.07027	19		

Table 4. ANOVA results on mean particle sizes when Otsu threshold was used.

Then, an ANOVA test was performed with the mean values given by Otsu threshold followed by watershed refinement (method 2). The results are given in the table on the next page. The Fisher text gave a value of 11.73 which was larger than 4.41, the null hypothesis was then rejected, and the screw speed led to a significant difference in the mean particle size.

Source of variation	Sums of squares (SS)	Degrees of freedom	Mean Square	Fisher test
Screw Speed	0.00578	1	0.00578	11.73
Residual	0.00887	18	0.00049	
Total	0.01465	19		

Table 5. ANOVA results on mean particle sizes when Otsu followed by watershed segmentation was used

It has been demonstrated by ANOVA analysis that a simple Otsu threshold led to the conclusion that screw speed did not influence the particle size whereas Otsu threshold followed by mathematical morphology refinement gave the opposite conclusion. A classical sieve analysis was also performed after the milling step and detected a significant difference in the particle size among batches. It is then believed that the refinement of the segmentation by mathematical morphology is a required step for more accurate results about particle sizes.

II.4. Discussion

In this section, the extraction of distribution maps, when the number and the identity of the sample constituents are known, has been addressed. The study was based on the analysis of two process intermediates obtained during the development phase of a new product. The first intermediates were the extrudates. They were composed of only two components: the API and the poloxamer. The univariate analysis at a wavelength specific of the poloxamer was in this case sufficient to extract relevant distribution maps. The second intermediates were the cores which encompassed five major compounds. The reference spectra revealed that one of the excipients overlapped strongly with the other compounds and it was necessary to employ multivariate analysis in order to localize the chemical species. Two algorithms, CLS and PLS-DA were tried out and led qualitatively to the same maps. Those maps revealed that the API and the poloxamer were present at the same position, in the particles of the milled extrudates, whereas the excipients aggregated and formed clusters. The resolution of our NIR imaging system was not fine enough to separate clearly the two first compounds. Since mid-IR or raman microscopy can reach finer spatial resolution (down to 4µ/m and 250 nm respectively) they could provide a better insight of the distribution of the API and the poloxamer in those particles (cf Annex I). But in this case, the field of view might be too narrow to have a surface of analysis representative of the sample.

During the development campaign, four batches, A, B, C, D were produced with different process parameters in order to optimize the final product. Two different API particle sizes and two different screw speeds were tested. To achieve an objective comparison of images and

thereby of the batches, quantitative parameters needed to be extracted. The distribution maps of the extrudates were compared by the help of histogram analysis. The histogram of the image linked to batch B revealed higher variance and lower kurtosis. Extrudates of batch B were more heterogeneous than the other batches. Regarding the cores, the excipient distributions were qualitatively the same whichever the batch. However, it seemed that the particle sizes (the milled extrudates) decreased as a function of the increase of the screw speed. The particle sieving of the milled extrudates also suggested this effect. Two segmentation schemes had therefore been proposed to detect this difference by image analysis. The first procedure used Otsu threshold, and the second procedure Otsu threshold followed by watershed. The refinement of the segmentation by watershed was necessary in order to separate clearly the particles and to find out a significant difference of the mean particle sizes between the two batches produced with the two screw speeds.

It has thus been demonstrated that image processing techniques such as histogram analysis and segmentation can provide quantitative parameters in order to help in the interpretation of images, but those parameters are not absolute measure. For example, the statistical analysis of the histogram led to the conclusion that image B was the most heterogeneous of the four images but did not give information about how heterogenous it was. Moreover, if more images and thereby histograms have to be compared, the discrimination between low contrasted and high contrasted picture might not be that straightforward. In the same way, it was not possible to verify if the estimated particle size was accurate, because the real size was unknown. Besides, the estimation of particle sizes by image segmentation is, in our case, flawed because we have a two dimensional view of particles with different shape and randomly spread in the cores that were cut during its milling. Depending on the position of the cut and of the particles inside the core, the last one might have different apparent surface. The analysis of a large number of images might somewhat overcome this difficulty by introducing statistical variability. In order to compare accurately the algorithms for image segmentation and estimate the bias of the methods, the analysis should be performed on synthetic samples for example polystyrene microsphere as conducted in [108]. However, the absorption of the sample, the energy conveyed by the wavelength, and the scattering effects led to different pathlength and interaction volume of one photon. This might influence the estimation of the particle size: two powders of different chemical species with the same particle size, might have a distinct estimated particle size. The choice of the wavelength must also be optimized.

In fact, in order to be able to determine objectively which is a "good" or a "bad" sample with regard to a property (homogeneity, particle size) by image analysis, the conditions of the image acquisition such as illumination, calibration of the camera, must be first optimized. The parameters of interest have to be chosen and the processing scheme that will enable to extract

the features of interest has to be built. Then discriminative values can be set. Obviously in our study, the number of samples were too small to draw up a general scheme. A data base with several dozens or hundreds of images has to be built. The construction of this data base requires the production of several batches where the parameter of interest, for example the homogeneity, varies from low to perfect. During the optimization of the formulation, several batches are produced and might represent those degrees of variability. If not, the required samples have to be produced in the laboratory, but laboratory batches might not be representative of the production ones due to scaling effect. The data base might also be completed during the production of the first batches.

As a conclusion of this section, it has been shown that NIR imaging was a valuable analytical technique to understand the process and help in the choice of its parameters. The advantages of the imaging technique is that it is fast in comparison with classical analytical tools, it requires little preparation of the samples, and provides information about the localization of the compounds and their identification. Histogram analysis and image segmentation help to the interpretation of the images. For a first insight of the influence of the process parameters on the samples, the analysis of batches obtained during the development is useful in order to find out parameters of interest to be studied. However, to allow objective decision without user interaction, and a possible in-line analysis, a database and a processing scheme specific of the application have to be constructed.

Related publication:

C. Gendrin, *Y. Roggo, C. Spiegel, C. Collet,* Monitoring Galenical Process Development by NIR Chemical Imaging: one case study, European Journal of Pharmaceutics and Biopharmaceutics, Vol **68** (2008), Issue **3**, p. **376-385**

III. Multivariate curve resolution of NIR hyperspectral data

In the previous section, the characterization of samples issued from the pharmaceutical development has been addressed. In this case, the number, proportion and identity of the chemical species inside the solid dosage forms are fully known and the distribution maps are easily extracted by analysis at a specific wavelength or by the help of the reference spectra. However, it might happen that only partial information about the sample is available. In those cases it is necessary to extract simultaneously the factors related to the pure spectra and the distribution maps. The algorithms commonly named Multivariate Curve Resolution are appropriate for this task. The literature survey of the first chapter of this dissertation (Chapter 1III) presented particularly four MCR algorithms : NMF, BPSS, MCR-ALS and PMF. The aim of this section is to compare their ability to factorize NIR hyperspectral data. Especially, a novel approach which introduces rotations to extract more appropriate solutions by the help of the PMF algorithm and user knowledge is proposed.

III.1. Material and methods

III.1.1. Sample and instrumentation

The tablet made in the laboratory contained five compounds: API (5%), cellulose (49%), lactose (44%), talc (0.9%) magnesium stearate (0.4%) and a coloring agent. The objective of approximately 10 μm/pixel was used to give a scanning area of 2.4×3 mm^2. The sample was analyzed in the spectral range [1100-2450] nm with a spectral increment of 10 nm and each channel was scanned 16 times. The pure powder of each of the chemical species was also measured using the same parameters.

Since PMF model required the estimation of the noise, five consecutive measurements of the tablets were performed to enable the calculation of the data point standard deviations (see below, section: III.2.2).

III.1.2. Preprocessing

The spectra were first converted into absorbance. In order to reduce the amount of data, the spatial resolution was decreased by a factor of three. The mean spectrum of a window of 3 by 3 pixels was kept reducing the number of pixels from 256×320 to 86×107, thus 9202 spectra.

The spectra were reduced to the range [1300 2230] nm and 94 wavelengths remained. Then the spectra were normalized using their sum of squares (unit length normalization). The mean of the five repetitions was calculated and used for the factorization.

III.1.3. Implementation

Positivity of both concentration and spectral profiles was the constraint employed during the optimization. After trials with three or four factors, the number of factors to be extracted was fixed to three because it led to the best extractions. Three factors corresponded to the number of the main compounds of the tablet. Because of their low concentration, the minor compounds were not detectable and could not be modelled. Thus, they contributed to the matrix **E** of experimental error.

MCR-ALS and NMF algorithms were implemented using Matlab v7.1. The convergence criterion was based on comparison of the model fit between two consecutive iterations. In this study the stop criterion was a relative change of the lack of fit (Equation 52) between two iterations lower than 10^{-5} %. The Matlab sources provided by Saïd Moussaouï were used to run the BPSS algorithm.

For PMF, the ME2 software [59] developed by Pentti Paatero was used. The error model can be calculated using several strategies [109]. In this study the "standard" error model was chosen. Measurement errors, called C_1 values, were obtained by computing the standard deviation of the five repetitions. Moreover a proportional error of 1%, to account for modeling errors and systematic errors of the data matrix such as impurities or non linear response of the detector[59], was added to the error model (C_3 values set to 0.01). σ_{ij} values were thus computed using the following equation:

$$\sigma_{ij} = (C_1)_{ij} + C_3 * |x_{ij}| \qquad \textbf{Equation 51}$$

The last convergence criterion was a change of 10^{-5} of the criterion Q_2 (cf Equation 29). In order to facilitate the optimization, low limit for concentration and absorbance spectra was set to -0.01. The reader is referred to [109] for more explanations about the implementation of the PMF model via ME2 software and its parameters.

For each of the algorithms, two different kinds of initializations were tested. Firstly, matrices **C** and $\mathbf{S}^T$ were randomly initialized using uniform distribution and ten consecutive runs were launched. Secondly, the matrices **C** and $\mathbf{S}^T$ were initialized by the results given by an OPA extraction (cf p. 38). This kind of initialization has already been reported in the literature and did improve the quality of extraction for mid-infrared and raman spectra [63, 64].

III.1.4. Assessing the quality of the extraction

The quality of the extraction was firstly assessed by determining the lack of fit of the model in comparison with the initial data matrix using the following equation:

$$\mathrm{lof}(\%) = 100 * \sqrt{\frac{\sum_i \sum_j \left(x_{ij} - \sum_k c_{ik} s_{jk} \right)^2}{\sum_i \sum_j x_{ij}^2}} \qquad \textbf{Equation 52}$$

Secondly, the extracted spectra were compared with the reference spectra. Two indices were employed: the correlation coefficient (Equation 53) and the Signal to Distortion Ratio (SDR: Equation 54). The SDR indicates the distortion of the extracted spectra in comparison with the reference spectra: the higher the SDR, the lower the distortion of the extracted spectra.

$$\mathrm{Corr}_k = \frac{\sum_j (s_{jk} - \bar{s}_k)(\mathrm{ref}_{jk} - \overline{\mathrm{ref}}_k)}{\sqrt{\sum_j (s_{jk} - \bar{s}_k)^2} \sqrt{\sum_j (\mathrm{ref}_{jk} - \overline{\mathrm{ref}}_k)^2}} \qquad \textbf{Equation 53}$$

$$\mathrm{SDR}_k = 10 log_{10} \left(\frac{\sum_j \mathrm{ref}_{jk}^2}{\sum_j \left(\mathrm{ref}_{jk} - s_{jk} \right)^2} \right) \qquad \textbf{Equation 54}$$

III.1.5. Assessing the quality of the PMF model

Besides lack of fit, correlation coefficient and SDR, it is possible to assess the quality of the PMF model using different diagnostic tools. First of all the number of outliers should be considered. The outliers are defined by the data points which present a scaled residual, $(X - CS^T)/\sigma$, out of the limits -4 and +4. The fraction of outliers must not be too high, say <1% .

Then the final value of the criterion Q_2 must be checked. If the estimation of the sigma value is good enough and the number of factors correct, then the final Q_2 value should be of the same order as the number of data points in the matrix X because the ratio e_{ij}/σ_{ij} should be one (i.e

the sigma values are equal to the error). In our application, the number of data points of our matrix X was 9202×94 = 864988 thus Q_2 value must be of that order.

The scaled residuals must also be studied. They must feature random signal and, if the model has been correctly constructed, no spectral shape must remain. Their histogram should feature a standard deviation of about 1 and most of the values should fall within [-2 2]. High residuals of data points mean that the associated sigma values were not properly estimated. It can be due to an underestimation of the noise. In that case it is recommended to increase those sigma values by a factor of 2 to 5 [110].

III.2. Results

III.2.1. Reference spectra

The reference spectra were acquired by the analysis of the pure powders. Normalized reference spectra for the three main compounds are displayed in Figure 36. API featured peaks between 1600 nm and 1780 nm due to C-H first overtones, with an additional peak at 2140 nm (N-H combinations). Lactose peaks were located at nearly the same wavelengths as those of cellulose due to O-H combinations. Those spectra had therefore a high correlation of 0.934 and consequently were not orthogonal to each other. Therefore a mathematical method such as PCA failed to recover cellulose and lactose spectra. In this study reference spectra were used as a diagnostic tool to assess the quality of the computation of the matrices C and S^T.

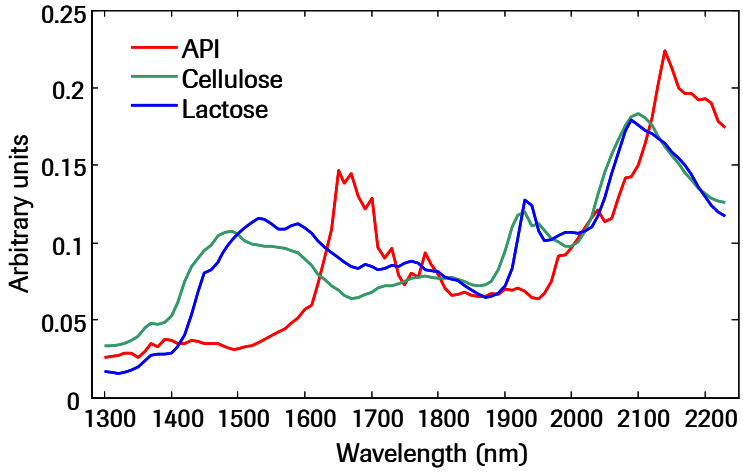


Figure 36. Normalized reference spectra of the three main constituents. The reference spectra were acquired by analysis of the pure powders. They were used to check the extraction ability of the SMCR algorithms.

III.2.2. Estimations of noise and sigma values

The estimation of measurement errors for the PMF model, might not be straightforward. In case of environmental data it is recommended to fix them so as they are a fraction (1% to 5%) of the data points of the matrix **X**, and such an approximation might be recommended for other applications where the sigma values could not be experimentally determined [109]. In our case the measurement performed by the NIR camera is non destructive, only the heating generated by the lamp might dry the samples. Moreover, one measurement is performed in about 5 minutes which is rather fast. Thus, the estimation of the matrix of uncertainties by acquiring successive measurements does not require too much added workload and allows to have an accurate estimation of the experimental errors. In our experiment, as a first estimation, five repetitions were employed to calculate the uncertainties using the standard deviations of each data point.

When dealing with hyperspectral imaging, sources of noise are various and might be due to the detector, optics, environment. Spectral and spatial noises can be studied. Figure 37 (a) represents the median standard deviations at each wavelength. Two wavelengths (1700 and 1940 nm) depicted higher noise values probably due to an artifact of the detector. Standard deviation values increased by a factor of 2 above 2050 nm. At longer wavelengths, the sample absorbs a larger fraction of the signal, thus less signal reaches the detector and noise increases. It can also be due to the sensitivity of the detector which decreases at longer wavelength (cf Figure 10). The peak of Gaussian shape around 1930 nm was due to the loss of water among measurements. Figure 37 (b) shows the median noise values over the wavelengths at each pixel position. A circle shape due to the optics was detected in the middle of the image as well as noise due to non uniform illumination of the sample.

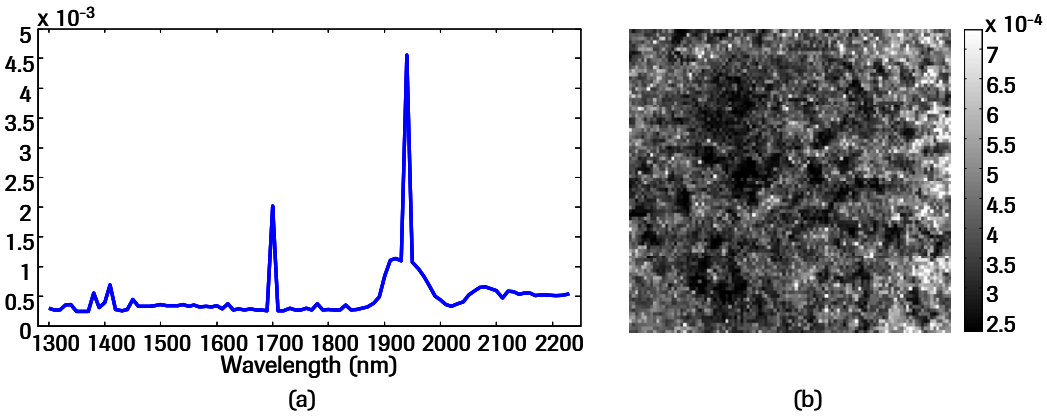


Figure 37. Sigma values of the data matrix. Median sigma values at each wavelength. Artifacts of the detector at 1700 nm and 1940 nm were visible (a). The Gaussian shape around 1930 was due to loss of water among the repetitions. Median sigma values at each pixel. A circle-shape due to optics artifact was visible (b).

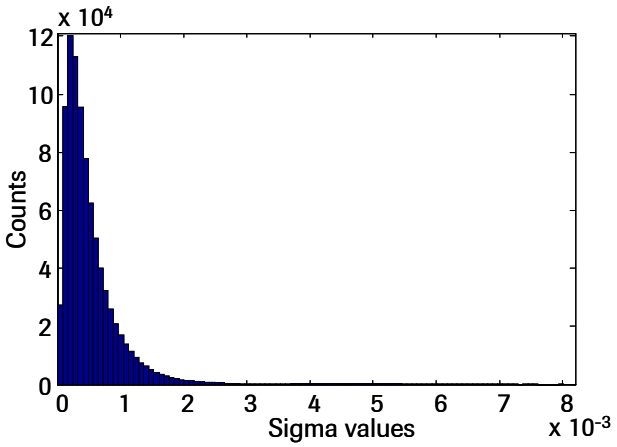


Figure 38. Histogram of the sigma values of all data points of the matrix. The histogram presented a tail toward higher sigma values with suggested data points with higher noise.

Figure 38 shows the histogram of all the noise values of the data cube. The histogram of noise presented a tail toward higher standard deviations which suggested values with higher noise.

It has been therefore demonstrated that noise might be non uniformly distributed across the data matrix. The PMF model allows then to take into account the uncertainties linked to each data point. Moreover, the robust mode of PMF allows to deal with outliers. After each iteration, the program calculates the residuals and identifies those which are significantly larger than the standard deviations [58] (scaled residual outside the range [-4 4]). If they are considered outliers, then their standard deviations are increased accordingly to limit their influence during the optimization. This is a better solution than simply rejecting those values.

In this study, it was chosen to work with the "standard" model for the error modeling as described in Equation 51. This model has proven to be adequate to a lot of problems, especially in environmental science. As it will be explained in the following, this approximation gave a good extraction also with our data set. However, other strategies in order to model the error might be implemented and could be tested in a future work.

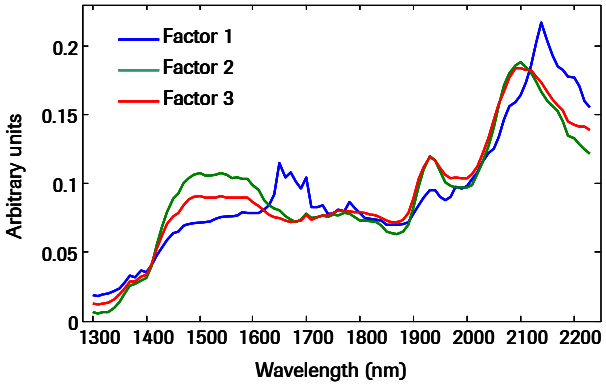
III.2.3. Spectra extracted by OPA algorithm

Figure 39 (a) displays the result obtained after the OPA computation. The first extracted factor featured peak of the API. The second and third factors were quite similar except around 1500 nm, the OH band. They were linked to the excipients but it was difficult to clearly discriminate between cellulose and lactose.

The distribution maps (Figure 39 (b)) were computed using the extracted spectra and CLS algorithm. The color scale on the right of each image links a color level in the image to a predicted concentration, linearly from blue which corresponds to a low concentration to red

which corresponds to a high concentration. The distribution maps were complementary to each other, thus the chemical species were adequately localized. API was agglomerated on the upper left corner revealing a problem of uniformity in the tablet.

(a)



(b)

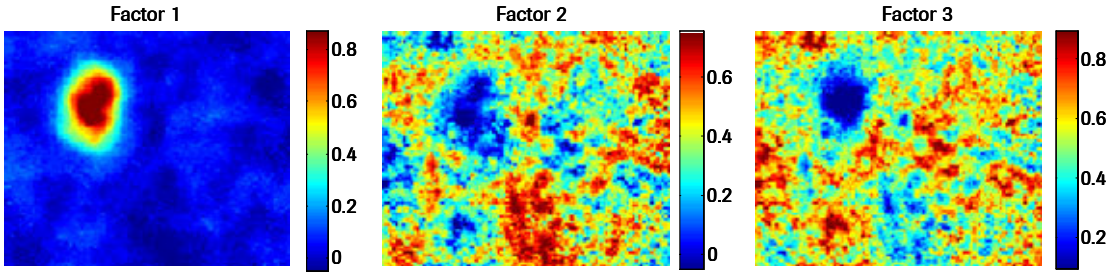


Figure 39. Spectra (a) extracted by OPA algorithm and distribution maps (b) obtained subsequently using CLS algorithm.

III.2.4. NMF, BPSS and MCR-ALS results

III.2.4.1. NMF algorithm

First, ten consecutive runs of NMF algorithm were performed using different random initializations. After convergence, the lack of fit reached 1.9% which was low: most of the data variance has been explained by the model. Figure 40 (a) displays the spectra extracted by several runs. For clarity reasons, only five results are shown. The spectra in blue depicts the result of the factorization with the best global fit to the reference spectra which was given by the sixth run. The best fit led to correlation coefficient of 0.948 for the API, 0.998 for the cellulose and 0.975 for the lactose and SDR of 14.99, 22.31 and 19.72 respectively (Table 6) which was accurate and might be sufficient for identification. However, different initializations led to different factorizations (cf Figure 4 and, Table 6). Positivity constraints were not sufficient to remove rotational ambiguity.

The Figure 40 (b) displays the distribution given by the sixth run (best extraction of spectral profile). The mean values of the estimated concentration (after reweighing so as their sum

equal 100%) were 30%, 34% and 36% for API, cellulose and lactose respectively which was not in accordance with the real concentration. Even if the spectra were well extracted, the algorithm failed to recover the chemical species proportion. Moreover NMF figured out a slow convergence. For the best extraction, the algorithm took more than 9000 iterations to reach the convergence criterion.

When initialized by the results given by OPA algorithm, the global extraction was greatly improved (Table 6, last row). The correlation coefficients to the reference spectra were above 0.94 and the SDR quite high. However, the concentrations were still uncovered, especially for the cellulose (computed concentration of 37.3%) and lactose (computed concentration of 57.7%) . The Figure 41 displays the factors and distribution maps. Compared to Figure 39, the spectra of excipients were now discriminated and associated with the cellulose (factor 3) and lactose (factor 2). The distribution maps of the excipients appeared also smoother than those from Figure 39 (b). Moreover, the number of iterations to reach the convergence decreased to 372.

			Corr	SDR	Conc (%)
Random initialisation	NMF worst extraction	API	0.911	12.986	31.2
		Cellulose	0.893	12.946	39.9
		Lactose	0.775	10.587	28.9
	NMF intermediate extraction	API	0.960	16.078	29.5
		Cellulose	0.868	11.638	31.9
		Lactose	0.978	20.671	40.7
	NMF best extraction	API	0.948	14.993	30.5
		Cellulose	0.998	22.310	33.9
		Lactose	0.975	19.719	35.6
OPA initialisation		API	0.945	14.7	8.9
		Cellulose	0.986	20.67	37.3
		Lactose	0.993	24.9	57.7

Table 6. Fit of the factors to the reference spectra given by the NMF algorithm for the worst (first row), intermediate (second row) and best (third raw) extractions. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration

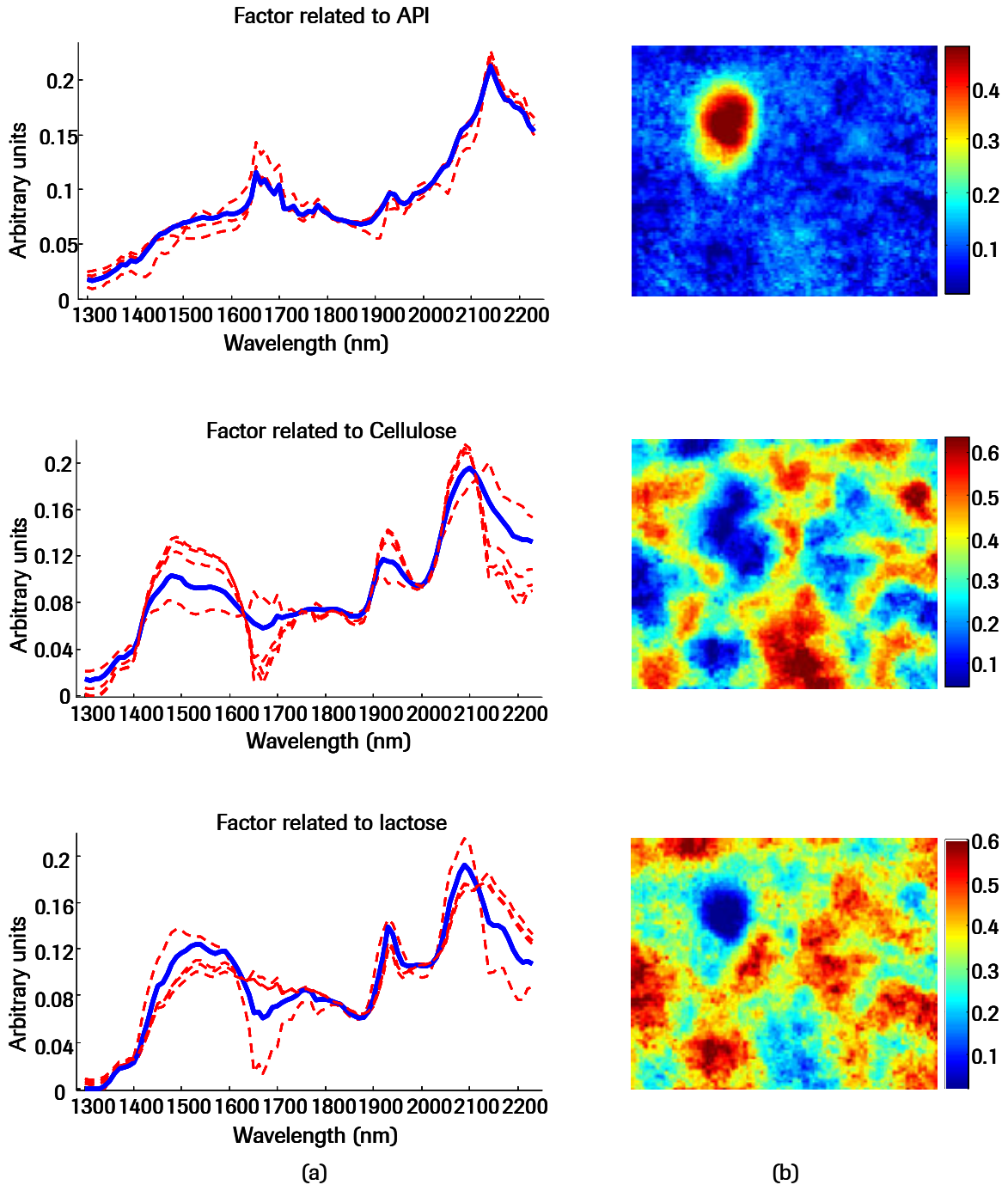
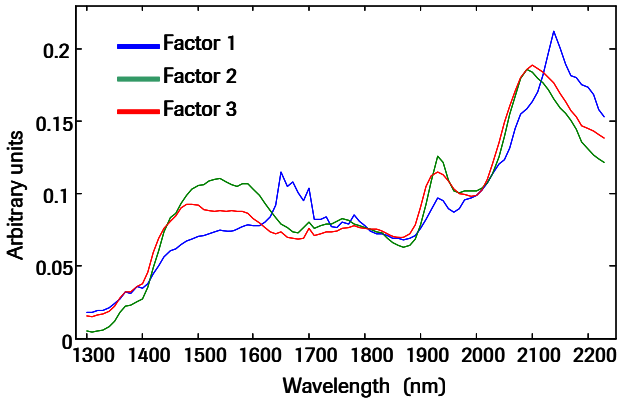


Figure 40. Factors extracted by five initializations of the NMF algorithm (a). The results of the run which gave the best fit to the reference spectra are depicted by the spectra in blue, the results of other extractions are depicted by a red dotted line. Distribution maps of API cellulose and lactose given by the best extraction of NMF algorithm (b).

(a)



(b)

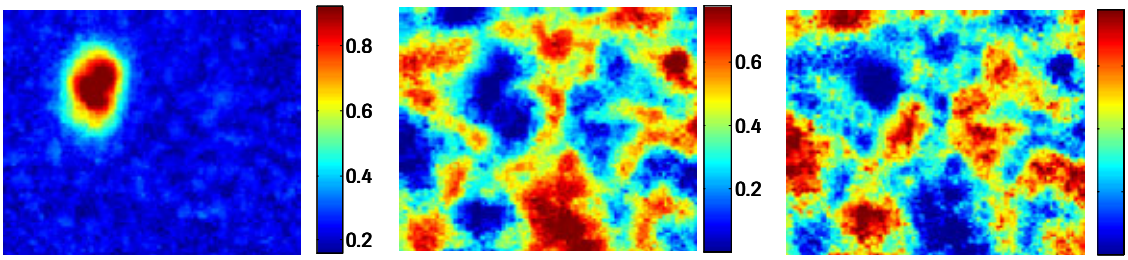


Figure 41. NMF extraction when the algorithm was initialized by OPA results (a) factors, (b) distribution maps.

III.2.4.2. BPSS algorithm

Like with NMF algorithm, ten consecutive runs of BPSS algorithm were performed using different random initializations. After convergence, the lack of fit reached 2.07%. Figure 42 (a) displays the extracted spectra of five runs. The seventh run (blue spectra) gave the best fit to the reference spectra. The best fit led to correlation coefficient of 0.976 for the API, 0.873 for the cellulose and 0.992 for the lactose and SDR of 10.74, 8.25 and 26.3 respectively (Table 7, third row). Cellulose spectrum shown great distortion, the SDR being smaller than 10. Like NMF algorithm, different initializations led to different factorizations and some of them were really poor with negative correlation coefficient and low SDR (Table 7, first and second rows). The Figure 42 (b) displays the distribution maps associated with the best run. The distribution maps were complementary thus we may assume that the chemical species were correctly localized. The mean values of estimated concentration (after reweighing so as their sum equal 100%) were 9.8%, 8.2% and 81% for API, cellulose and lactose respectively which was not in accordance with the real concentration.

Since BPSS contains randomly initialized parameters such as the parameters of the statistical distributions, initialization by OPA algorithm did not improve the extraction ability and several runs also led to different results. For these reasons, the results of BPSS initialized by OPA are not displayed.

BPSS failed to recover accurately the spectra. It is believed that this was due to the a priori given to the statistical distribution of the pure spectra which were approximated by a gamma distribution (cf p 42). This might be true for mid-IR and Raman spectroscopy but not for NIR spectra because of their strong baseline. Figure 43 depicts the statistical distribution (normalized histogram of the absorbance) of the three NIR reference spectra. Those distributions figured out rather multi-modal distributions than a simple gamma distribution. A solution to reach a more accurate extraction could be to change the a priori about the statistical distributions. Another problem was the slow computation because of MCMC algorithm. More than 1700 iterations and nearly one hour were needed to reach the convergence criterion for the best run.

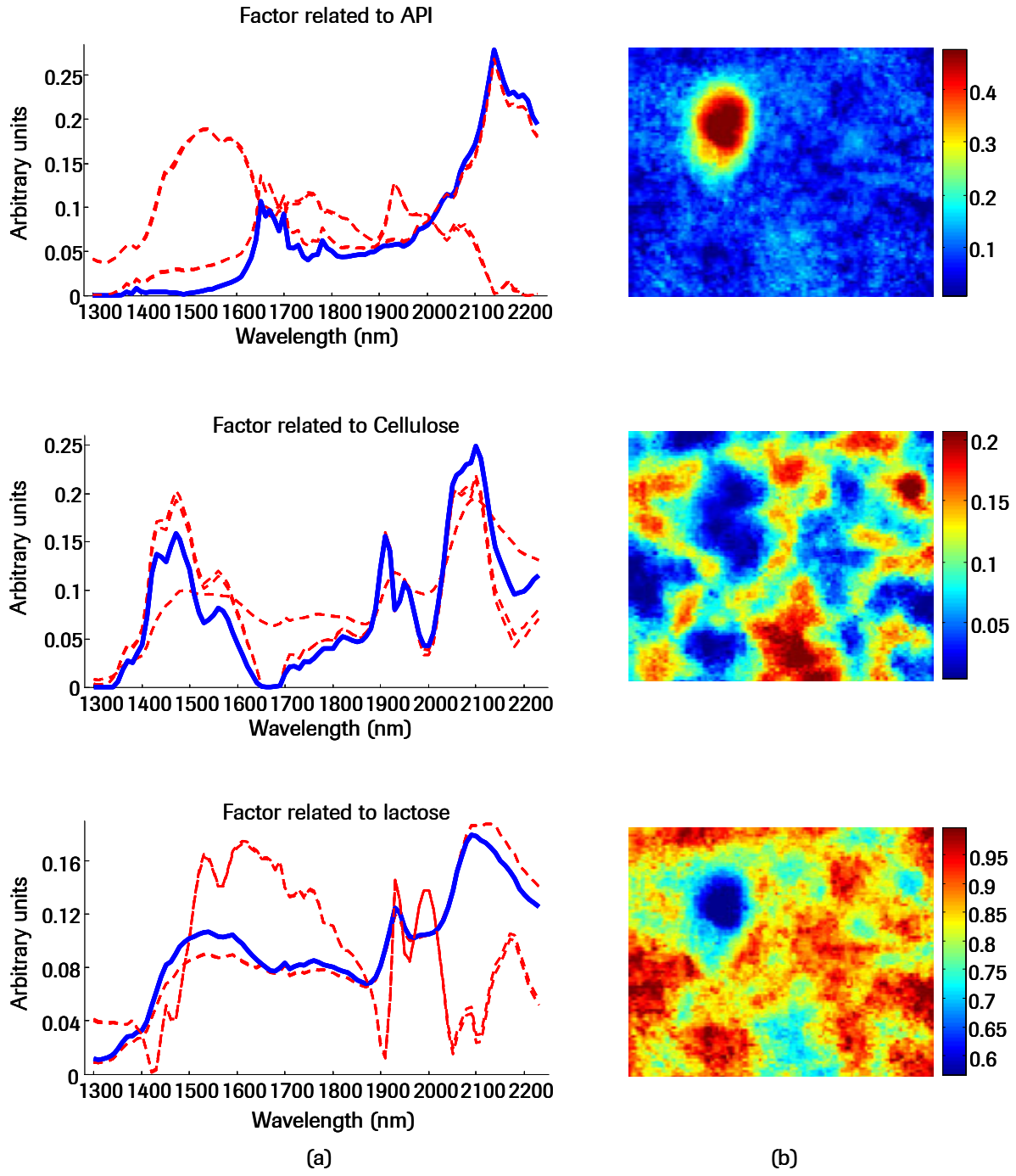


Figure 42. Factors extracted by five initializations of the BPSS algorithm (a). The results of the run which gave the best fit to the reference spectra are depicted by the spectra in blue, the results of other extractions are depicted by a red dotted line. Distribution maps of API, cellulose and lactose given by the best extraction of BPSS algorithm (b).

		Corr	SDR	Conc (%)
BPSS worst extraction	API	-0.590	0.891	6.5
	Cellulose	0.655	6.771	8.1
	Lactose	0.966	17.800	85.4
BPSS intermediate extraction	API	0.886	7.871	6.5
	Cellulose	0.986	20.921	84.9
	Lactose	0.286	4.543	9
BPSS best extraction	API	0.976	10.741	9.8
	Cellulose	0.873	8.258	9
	Lactose	0.992	26.303	81.2

Table 7. Fit of the factors to the reference spectra given by the BPSS algorithm for the worst (first row), intermediate (second row) and best (third raw) extractions. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration.

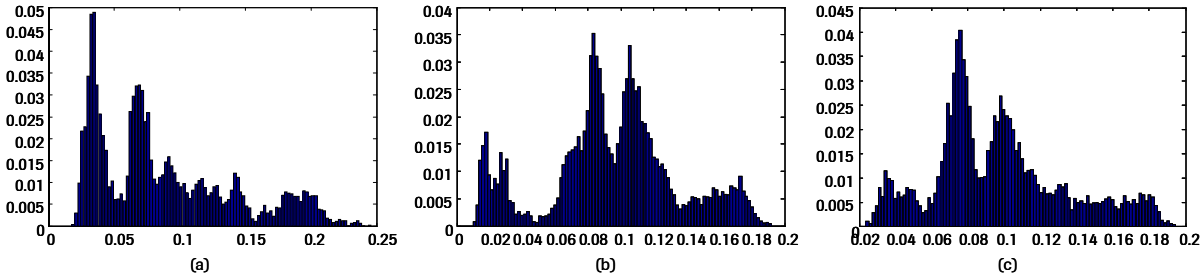


Figure 43. Statistical distribution (normalized histogram of the absorbance) of the API (a), cellulose (b) and lactose (c) reference spectra. They figured out multi-modal distributions.

III.2.4.3. MCR-ALS algorithm

Ten consecutive runs of MCR-ALS algorithm were performed using different random initializations. After convergence, the lack of fit reached 1.2% which was the best fit in comparison with NMF and BPSS algorithms. Figure 44 displays the extracted spectra of five runs. The best global fit to the reference spectra (blue spectra) was given by the sixth run. It led to correlation coefficient of 0.965 for the API, 0.844 for the cellulose, 0.994 for the lactose and SDR of 15.3, 10.7 and 24.1 respectively (Table 8, last row). The cellulose spectrum was the most distorted and depicted inverse peak of API between 1600 and 1700 nm (Figure 44). Once more, different initializations led to different factorizations. Especially correlation coefficient as low as 0.542 (Table 8, first row) was calculated for the lactose spectrum. Positivity constraints were not sufficient to remove rotational ambiguity.

The Figure 4 (d) displays the distribution maps associated with the best run. The distribution maps were complementary. However, the color scale values indicate that the minimum concentration for API was 25% which was unrealistic given the global concentration of API which was of 5%. The mean value of estimated concentration (after reweighing so as their sum equal 100%) is 36%, 53% and 11% for API, cellulose and lactose respectively which was not in accordance with the real concentration. Even if the spectra were well extracted, the algorithm failed to recover the chemical species proportion. However, MCR-ALS algorithm was easy to implement and the convergence was rather fast: the best run took 260 iterations.

When initialized by the results given by OPA algorithm, the global extraction was greatly improved (Table 8, last row). The correlation coefficients to the reference spectra were above 0.95 and the SDR quite high. However, the concentrations were still uncovered, especially for the cellulose and lactose. The Figure 46 displays the factors and distribution maps. Compared to Figure 39, the spectra of excipients were discriminated and associated with the cellulose (factor 3) and lactose (factor 2). The distribution maps of the excipients appeared also smoother athan those from Figure 39 (b). Moreover, the number of iterations to reach the convergence decreased to 130. The extraction was also slightly better than when using NMF algorithm.

			Corr	SDR	Conc (%)
Random initialisation	MCR-ALS worst extraction	API	0.927	13.706	37.6
		Cellulose	0.859	11.286	51.2
		Lactose	0.521	6.049	11.2
	MCR-ALS intermediate extraction	API	0.872	11.699	38.1
		Cellulose	0.960	15.963	46.5
		Lactose	0.887	13.342	15.4
	MCR-ALS best extraction	API	0.965	15.290	36.0
		Cellulose	0.844	10.681	53.0
		Lactose	0.994	24.145	11.0
OPA initialisation		API	0.954	15.5	12.2
		Cellulose	0.993	21.3	30.7
		Lactose	0.996	26.2	57.03

Table 8. Fit of the factors to the reference spectra given by the MCR-ALS algorithm for the worst (first row), intermediate (second row) and best (third raw) extractions. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration

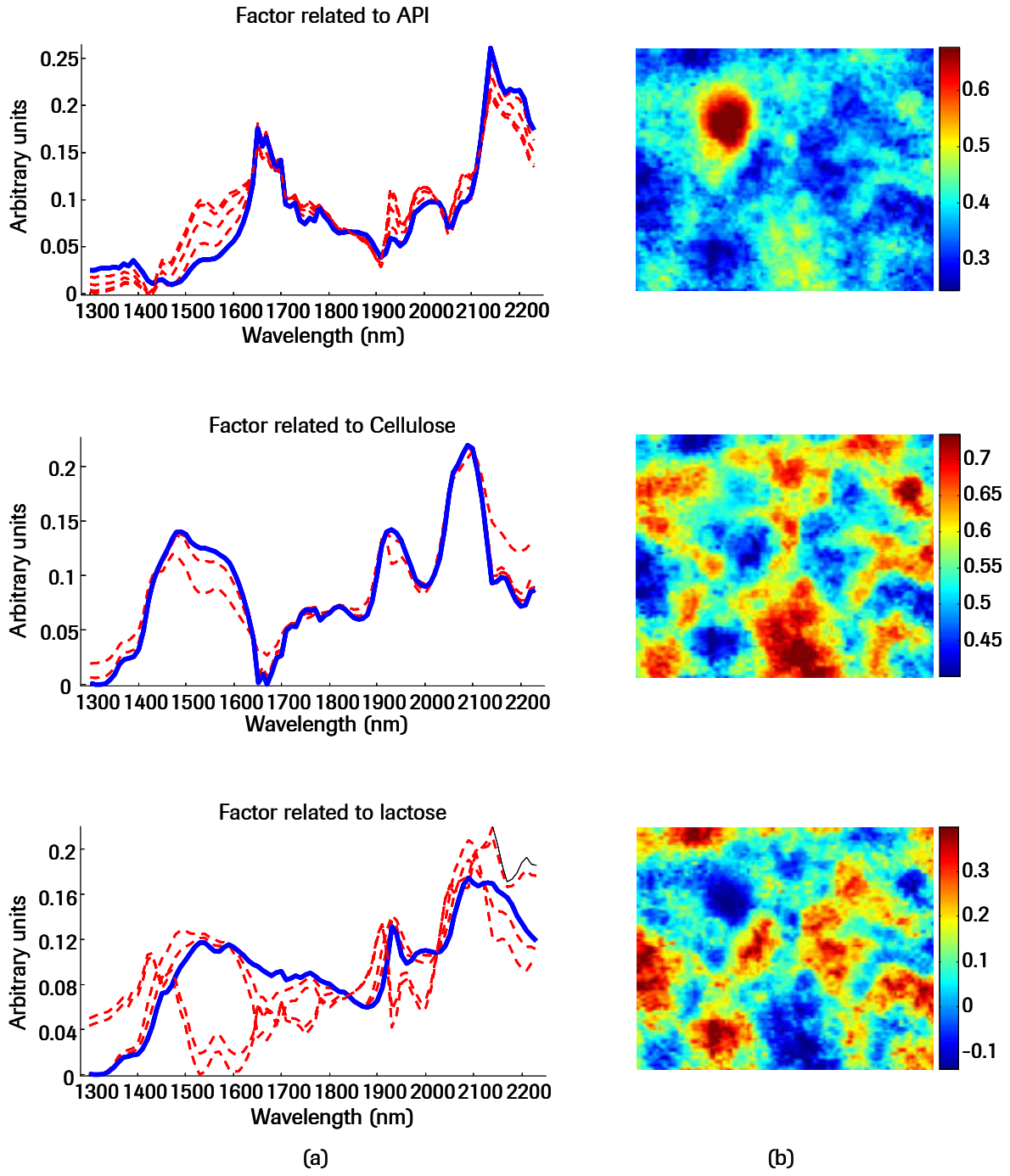
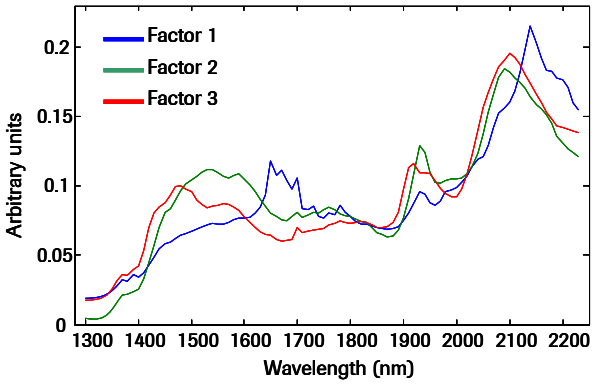


Figure 44. Factors extracted by five initializations of the MCR-ALS algorithm (a). The results of the run which gave the best fit to the reference spectra are depicted by the spectra in blue, the results of other extractions are depicted by a red dotted line. Distribution maps of API cellulose and lactose given by the best extraction of MCR-ALS algorithm (b).

(a)



(b)

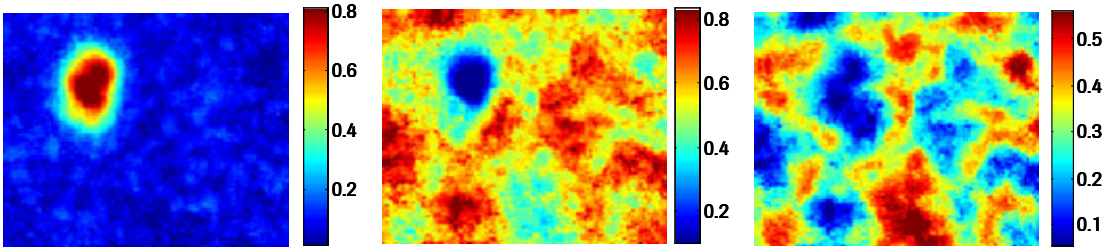


Figure 45. MCR-ALS extraction when the algorithm was initialized by OPA results (a) factors, (b) distribution maps.

III.2.5. Results with PMF model

III.2.5.1. Initial run

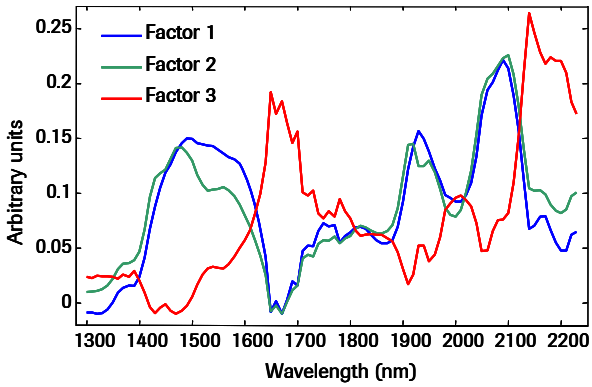
A first run, called run1, of the PMF model led to a Q_2 value of 908541 and approximately 1500 positive and 1500 negative outliers. The lack of fit was 1.30%. The study of the scaled residual revealed higher residuals for the first six wavelengths. This was probably due to a weak spectral absorbance at those wavelengths. The associated sigma values were then increased by a factor of 5 [109] and the algorithm re-run, this run is called run2 in the following.

The run2 gave a Q_2 value of 685163, 302 positive and 335 negative outliers (less than 0.074% of the data points) with a lack of fit of 1.23%. Thus, with the new sigma values, the number of outliers as well as the lack of fit decreased, while the Q_2 value was still acceptable because it approached the number of data points which was 864988. The scaled residuals did not feature higher values at the first wavelengths anymore and depicted a random distribution with an estimated standard deviation of 0.89 and a mean of 0.0027, thus most of the values were comprised in the interval [-2 2]. The second model, with higher sigma values for the first six wavelengths (run 2), was thus the more appropriate to our data set and was utilized in the following of our experimentation. Ten runs of PMF were launched with the same parameters but different initializations. The results were quite similar and we could thus assume that a global minimum has been reached. The extracted spectra are displayed in the Figure 46 and the quality of extraction is given in Table 9.

			Corr.	S.D.R.	Conc.
Random initializations	PMF initial solution (run 2)	API	0.928	11.36	39.8
		Cellulose	0.855	10.62	26.6
		Lactose	0.710	7.99	33.6
OPA initialisation		API	0.968	16.8	14.2
		Cellulose	0.996	22.04	37.6
		Lactose	0.995	25.41	48.16

Table 9. Fit of the factors to the reference spectra given by run2 of PMF. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration.

(a)



(b)

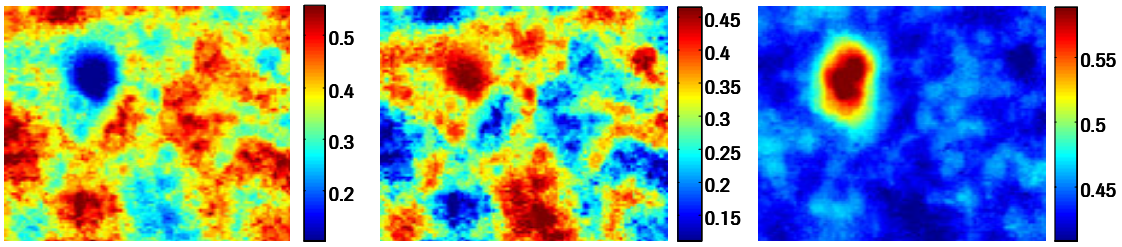


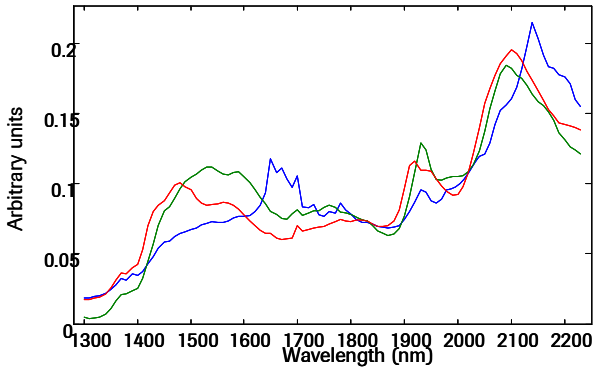
Figure 46. Extraction which was achieved by PMF model (run2), factors (a) and distribution maps (b). The run2 was performed after increasing the sigma values for the first six wavelengths for a better fit. The lack of fit was 1.23%, 637 outliers were detected (0.074% of the data points) and Q_2 value was 685163, which suggested that the chosen error model and the number of factors were appropriate to our data. Extracted factor 1 (related to lactose) and factor 2 (related to cellulose) featured inverse peak of factor 3 (related to API) at spectral ranges [1400-1500] nm and [1850-2100] nm which suggested rotational ambiguity.

Each of the factors could have been linked to a reference spectra: factor 1 to the lactose, factor 2 to the cellulose and factor 3 to the API however they were greatly distorted (SDR values were below 10). In particular it seemed that between wavelengths 1600 and 1700 nm the first and second factors depicted the inverse of the peaks of the third factor. In the same way, factor 3 depicted at spectral ranges [1400-1500] nm, [1850- 2100] nm the inverse of the peaks of factor 2 or 1. Moreover, the mean values of concentration (after reweighing so as their sum equal 100%) were 39.8%, 26.6% and 33.6% for the factor related to API, cellulose and lactose respectively which was not exact. In fact, the solution given by PMF was in the middle of the rotational domain [60].

Like with NMF and MCR-ALS algorithm, when initialized by the results given by OPA algorithm, the global extraction of PMF was greatly improved (Table 9, last row). The correlation coefficient to the reference spectra were above 0.95 and the SDR quite high. However, the concentrations were more accurately estimated. The Figure 47 displays the

factors and distribution maps. Visually, the results were similar to those obtained by MCR-ALS algorithm (Figure 45), except that the second distribution map was more contrasted.

(a)



(b)

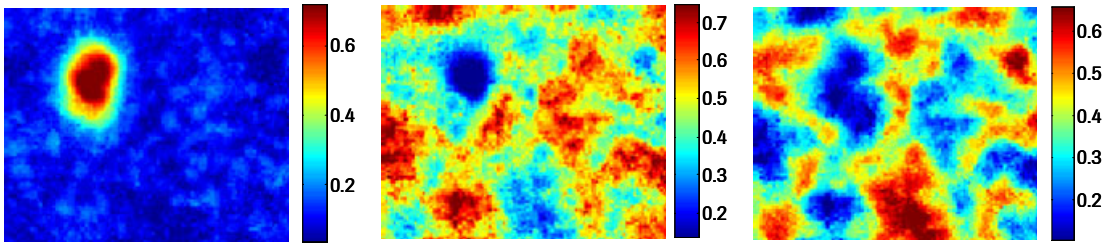


Figure 47. PMF extraction when the algorithm was initialized by OPA results (a) factors, (b) distribution maps.

III.2.5.2. Introducing rotation

The results given by run2 suggested rotational ambiguity. With PMF, it is possible to favor rotations based on user knowledge in order to find out a more appropriate solution. Introducing rotation matrix in the optimization problem has already been discussed by Paatero et al in [60]. In this study we employed a new tool developed recently in the ME2 software. Starting with the factorization resulting from the run2 (random initialization), a matrix of size K×K can be introduced [61].

Given the observations made with the results of the run2 (cf Figure 46), a rotated solution would be to add factor 3 to factor 1 and factor 3 to factor 2 in order to remove the inverse peak of API between the spectral range [1600 1700] nm. These additions can be described by the following equations:

$$\hat{s}_{j,1} = s_{j,1} + r s_{j,3}$$
$$\hat{s}_{j,2} = s_{j,2} + r s_{j,3}$$

Equation 55

An addition among rows of matrix S^T leads to a subtraction between the columns of the matrix C [60]. Thus, the rotation depicted by Equation 55 of the S^T matrix implies the following rotation of the C matrix:

$$\hat{c}_{i,3} = c_{i,3} - r c_{i,1}$$
$$\hat{c}_{i,3} = c_{i,3} - r c_{i,2}$$

Equation 56

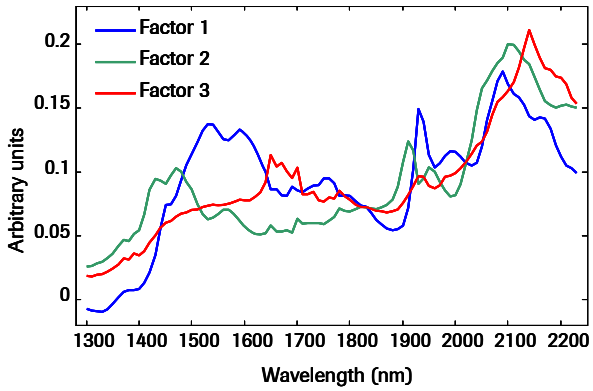
Since the rotation tool given to the software works with rotation of the C matrix, the following form of the rotation matrix (Matrix 1) has been set to create a rotated solution:

	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$
c_1	1	0	$-r$
c_2	0	1	$-r$
c_3	0	0	1

Matrix 1

The run3 has thus been performed by initializing the S^T and C matrices with the results of run2 and applying the Matrix 1 as a desired rotation. The r value is the strength of the rotation. In our case r = 0.5 led to the desired rotation. A value smaller than 0.5 did not change much the initial solution.

(a)



(b)

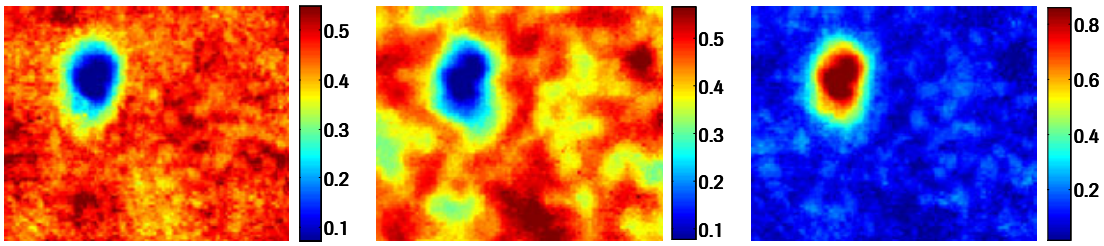


Figure 48. Extraction which is achieved by PMF model using Matrix 1 (run3), factors and distribution maps. The run 3 was performed by initializing the matrix of concentration and spectra by the results of run2 and introducing rotations by the help of Matrix1. Better fit was obtained for cellulose and lactose but the API-like factor still depicted contributions from factor 1 and 2 between 1400 nm and 1600 nm.

		Corr.	S.D.R.	Conc.
PMF first rotation (run 3)	**API**	0.944	14.75	15
	Cellulose	0.965	16.39	41.6
	Lactose	0.931	14.81	43.4

Table 10. Fit of the factors to the reference spectra given by run 3. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration.

Figure 48 displays the results obtained after the run3. It can be clearly seen that the negative peaks of API have disappeared from factor 1 and factor 2 which resembled more to the reference spectra of lactose and cellulose respectively. Table 10 gives the performance of the extraction after this rotation. The correlation coefficients were improved in comparison with the first extraction as well as the SDR (compare Table 9 and Table 10). Moreover the estimated concentration values were 15% 41.6% 43.4% for factor related to API, cellulose, lactose respectively which approached the real values. However, factor 3 had contribution from factor 1 and 2 especially between 1400 nm and 1600 nm. Thus, the run4 has been performed using, as initialization, the results of the run3 and the following matrix:

	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$
c_1	1	0	0
c_2	0	1	0
c_3	0.1	0.1	1

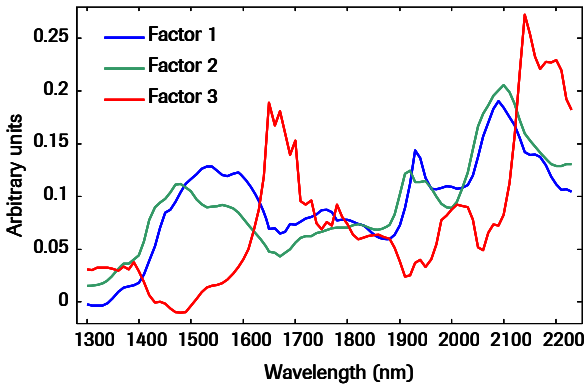
Matrix 2

The results of run4 are displayed in the Figure 49. Cellulose and lactose were well extracted as shown also by Table 11. The correlation coefficient as well as SDR were higher than after the first rotation but API signal was still distorted. Concentration values remained similar to after the first rotation.

		Corr.	S.D.R.	Conc.
PMF second rotation (run 4)	**API**	0.934	11.09	14.5
	Cellulose	0.990	19.8	42.2
	Lactose	0.976	19.6	43.3

Table 11. Fit of the factors to the reference spectra given by run 4. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration.

(a)



(b)

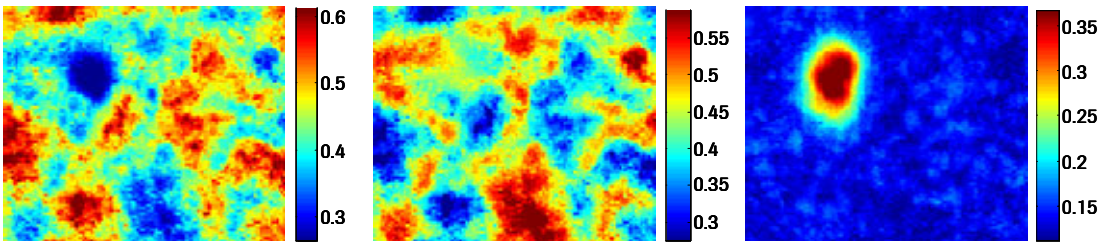


Figure 49. Extraction which was achieved by PMF model using Matrix 2 (run4). (a) Factors and (b) distribution maps. The run4 was performed by initializing the matrix of concentration and spectra by the results of run3 and introducing rotations by the help of Matrix2. Better fit was obtained for API-like factor but it still depicted distortions.

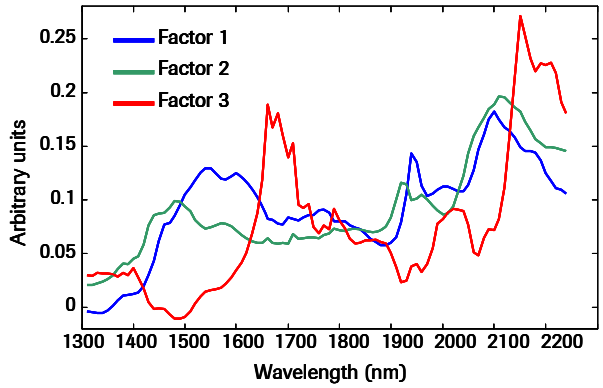
In a third step, the first two rotational matrices, Matrix 1 and Matrix 2 were added to each other to form a third rotational matrix, the Matrix 3.

	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$
c_1	1	0	−0.5
c_2	0	1	−0.5
c_3	0.1	0.1	1

Matrix 3

The run5, was then conducted using the results of the run2 as initialization and introducing the above matrix of rotation. The results are given in Figure 50. The extracted factors were quite similar to those extracted after the two successive rotations, especially it still remained cellulose or lactose contributions in the API like factor. The concentration values extracted using Matrix 3 were really closed to the true values: 4.6% 48.8% and 46.6% were calculated for API, cellulose and lactose respectively (cf Table 12). Thus, it should be hard to improve the extraction of the API and the remaining features from cellulose or lactose in the API spectrum can not be removed. PMF with four factors have been attempted but did not lead to a better extraction, the unwanted features remaining in the API like factor.

(a)



(b)

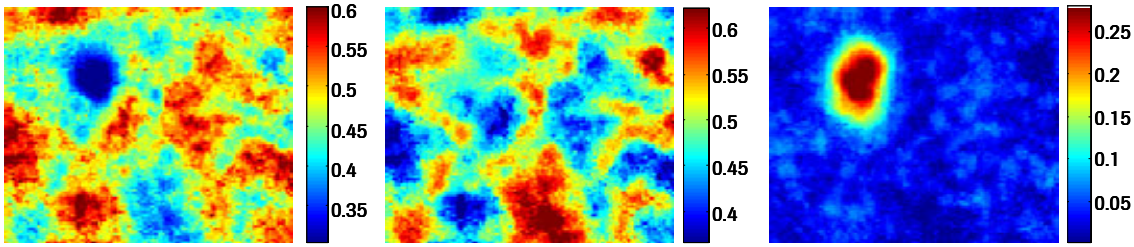


Figure 50. Extraction which was achieved by PMF model using Matrix 3 (run5). (a) Factors and (b) distribution maps. The run5 was performed by initializing the matrix of concentration and spectra by the results of run2 and introducing rotations by the help of Matrix3. The extracted spectra were similar to the results of run4, but the concentrations approached the real values: 46.6%, 48.8 and 4.6% were found out for lactose, cellulose and API respectively. The extraction of the factors could not be improved anymore.

		Corr.	S.D.R.	Conc.
PMF third rotation (run 5)	**API**	0.934	11.1	4.6
	Cellulose	0.979	18.73	48.8
	Lactose	0.984	20.79	46.6

Table 12. Fit of the factors to the reference spectra given by run 5. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration.

III.2.5.3. Introducing target values

Another way of favoring rotations is to introduce target values. When pharmaceutical samples are investigated it is possible that knowledge about the global concentration of some chemicals is known. In this study, three auxiliary equations, 1 per factor (Equation 57), have been defined so as to lead the global concentration of the chemicals to their average values.

Factor 1 was related to lactose which concentration was around 45%. Its related auxiliary equation was then defined by:

$$n * 0.45 = \sum_i c_{1,i}$$

Factor 2 was related to cellulose which concentration was around 50%. Its related auxiliary equation was then defined by:

Equation 57

$$n * 0.50 = \sum_i c_{2,i}$$

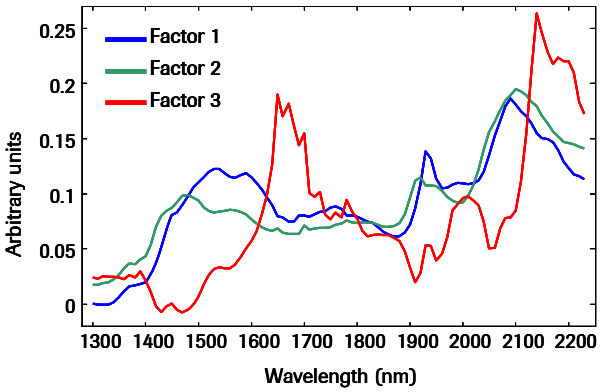
Factor 3 was related to API which concentration was around 5%. Its related auxiliary equation was then defined by:

$$n * 0.05 = \sum_i c_{3,i}$$

With n the total number of pixels (n = 9202 in our case)

Reference [61] can give more explanations about the implementation of pulling equations. The run6 was then performed using the results of the run2 as initialization and Equations 11 as pulling equations. Results of the extraction are displayed in Figure 51 and were quite similar to the results of Figure 50. Cellulose and lactose were accurately extracted with correlation coefficient of about 0.99 whereas API still presented some distortion. Concentration values were found to be 5.1%, 50.2% and 44.7% for API cellulose and lactose respectively (Table 13) which suggested that targeting worked correctly.

(a)



(b)

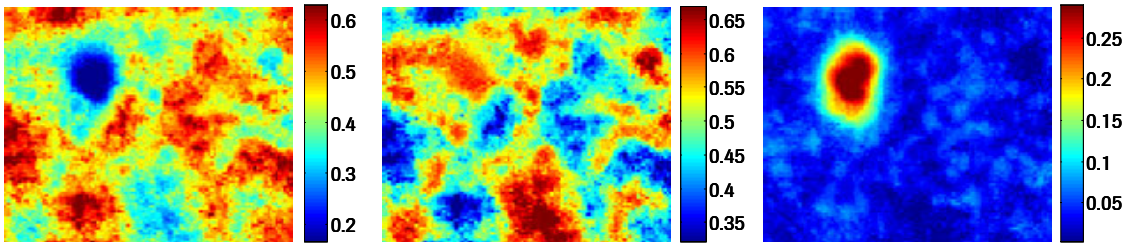


Figure 51. Extraction which was achieved by PMF model using target values (run6). (a) Factors and (b) distribution maps. The run6 was performed by using the global concentration of each compound as target values. The extracted spectra were similar to the results of run5 (Figure 50). The concentrations approached the real values: 44.7%, 50.2 and 5.1% were computed for lactose, cellulose and API respectively.

		Corr.	S.D.R.	Conc.
PMF second rotation (run 4)	**API**	0.933	11.75	5.1
	Cellulose	0.991	20.79	50.2
	Lactose	0.993	23.82	44.7

Table 13. Fit of the factors to the reference spectra given by run 6. Corr = correlation coefficient, SDR = Signal to Distortion Ratio, Conc = estimated concentration.

III.3. Discussion

In this section, the simultaneous extraction of distribution maps and spectra by multivariate curve resolution algorithms has been addressed. Four algorithms, NMF, BPSS, MCR-ALS and PMF were compared. In a first part, the first three procedures were considered. BPSS led to the worst extraction with high distortion of the spectra. This was certainly due to the a priori about the statistical distributions of the spectra. They were approximated by a mono-modal gamma distribution whereas NIR spectra featured multi-modal ones. The initialization of the matrices **C** and $\mathbf{S}^T$ by OPA did not improve the results because of the parameters of the statistical distributions that were randomly initialized. NMF showed an accurate extraction ability, with correlation coefficient larger than 0.94 and high SDR, even with random initialization. However, it needed more than 9000 iterations to reach the global minimum. MCR-ALS was faster and needed 260 iterations for the best extraction. When randomly initialized, it extracted the API spectrum with the most precision (correlation coefficient of 0.965 and SRD of 15.3). Also, the final lack of fit to the initial data matrix reached 1.2% which was the lowest value. Thus, MCR-ALS explained better the variance of the data. The initialization of NMF and MCR-ALS iterations by the results given by an OPA extraction decreased the number of iterations and led to more accurate extractions. However, if NMF and MCR-ALS were able to extract accurately the spectra, none of them was able to recover the concentration of each of the chemical species.

Besides accurate spectral extractions, it has also been demonstrated, that the solutions of NMF, BPSS and MCR-ALS algorithms were not unique and it lacked tools to investigate the domain of possible solutions. This is their main drawback. In the case presented here, the reference spectra were employed to check the accuracy of the extractions. It was then possible to determine the best solution. However, in real case study this information might not been known. Then, how could it be verified which extraction among several runs is the most accurate? Surely, user knowledge such as spectral shape of the spectra might be employed to have an idea about the best results but some ambiguities may reside. One solution to reduce rotational ambiguity, is to initialized the matrices **C** and $\mathbf{S}^T$ by a first estimation. In our study, their initializations by the results of an OPA extraction greatly improve the results of MCR-ALS algorithm. However only one solution would then be possible but it is not sure that it is the best one because such initializations might also, in certain cases, lead to a local minimum. Another possibility is to fix known values in the matrices **C** and $\mathbf{S}^T$ all along the iterations of the MCR-ALS procedure. For example, Window Evolving Factor Analysis (cf p 43) algorithm might provide information about the number of species in each pixel, but the reference

spectra should be known in order to identify the present or absent species. All these solutions reduce rotational ambiguity but do not allow to explore manually the rotational domain.

The second part of this section has then focused on PMF algorithm. When initialized by results of an OPA extraction, PMF algorithm led to accurate extraction of the factor and the most precise extraction of the chemical species proportion. If randomly initialized, PMF led to one global solution. The first run did not result in a better extraction than NMF and MCR-ALS algorithms, nevertheless, as stated above, the advantage of ME2 software is the possibility to explore the rotational domain that can be achieved by the introduction of rotational matrices. The matrix of rotations can be constructed by visually inspecting the factors and determining which subtractions or additions should be done to remove "inverse" peaks of one factor appearing in the other ones. Afterwards, PMF can be run by initializing the matrices **C** and $\mathbf{S}^T$ with a first solution and introducing the matrix of rotations. In our study, the final results led to correlations coefficients above 0.93 and high SDR. The chemical concentrations were also nearly recovered (less than 2% of error). A second solution to rotate the solution is to pull values toward known values. In the study, the global concentration of each of the chemical species was introduced. The final correlation coefficients were of more than 0.99 for the excipient and 0.93 for the API, the API was more distorted but the concentrations were well recovered. It must be noted that global concentration can not been employed when using MCR-ALS procedure. With the latter, the local concentration can be set but it is more challenging to know in advance the precise concentration of one pixel.

The advantage of ME2 software, is that the user knowledge can be introduced in a soft manner. The optimization will fulfill each of the constraints or pulling equations depending on their strength level. For instance, the introduction of rotational matrices or pulling equations in our example led to a weaker normalization of the factors : the final sum of squares value of the factors was about 0.98 instead of 1. The solution was then more accurate whereas the normalization still acceptable. With MCR-ALS for example, the normalization is introduced as a hard constraint and the final factors would have unit length in any case, but this prevents better extractions. With ME it is possible to tune each of the constraints.

Of course, the cost of the flexibility of the PMF approach is the tuning of all parameters and especially the uncertainties which requires expertise before being able to conduct good extractions. The exploration of the data must be done step by step. It is first necessary to set up accurately the error model which might not be straightforward. A change in the initial model, and the solution will be different. Fortunately, the diagnostic tools such as the final Q_2 values and the number of outliers can help to determine if the chosen model is appropriate or not.

Overall, in this section, it is demonstrated that the unmixing of hyperspectral NIR imaging data is possible. Several algorithms are available and if the PMF procedure is the most difficult to put in place it is also the most powerful. It leads to the best estimation of the concentration when using tools to explore the rotational domain.

Related publications:

C. Gendrin, *Y. Roggo, C. Collet*, Self modelling curve resolution of Near Infrared Imaging Data, Proceedings of the ICNIRS 2007, Umea, Sweden, Journal of Near Infrared Spectroscopy, Vol **16**, Issue **3**, p **151-157**

IV. Conclusions

In this chapter, the extraction of distribution maps and their characterization have been considered. Two situations were depicted. In the first one, the number and identity of the constituents were fully known. This is the case of samples issued from the development. If the number of compounds is small and their spectral features do not strongly overlap, univariate analysis is sufficient to have appropriate distribution maps. When spectral overlap does occur, CLS or PLS-DA algorithm are both suited algorithms for this task. The image of the distribution maps might be subsequently processed in order to characterize the samples. Two examples were proposed. On one hand, histogram analysis is useful to study the contrast of images and thereby the homogeneity of the samples. In our case, four batches were under investigation and one batch showed more heterogeneity than the other ones. On the other hand, image segmentation is useful to determine particle sizes. It has been demonstrated that the refinement of the segmentation by watershed algorithm led to a better separation of the particles in the binary image.

In the second situation the reference spectra are not available. In this case, the extraction of distribution maps and pure spectra must be performed simultaneously and might be more challenging. Several multivariate curve resolution algorithms: BPSS, NMF, MCR-ALS and PMF have been compared. NMF and MCR-ALS lead to accurate spectral extractions but the concentrations are badly estimated. Moreover they lack tool to explore the rotational domain. PMF algorithm associated with ME software provides this flexibility. Especially, an approach to investigate the rotational domain has been proposed. Based on the results of a first extraction, rotational matrices can be introduced to find out more appropriate solutions. On the other hand, it is possible to pull values with more or less strength. As example, the global concentration of each of the species has been introduced. These approaches allow to reach a high spectral extraction ability as well as accurate estimation of concentrations. The applications of MCR algorithms for pharmaceutical issues are multiple. They can be employed when some of the compounds in the tablet are not known, for instance when a contamination has occurred or for the analysis of a counterfeit.

Chapter 3
Content uniformity of pharmaceutical solid dosage forms

I. Introduction

After the extraction of the chemical distribution maps and qualitative analysis of pharmaceutical samples, this last chapter focuses on the content uniformity of tablets by means of NIR chemical imaging. Content uniformity is an important topic for pharmaceutical samples because a lower percentage of API than the optimal formulation may decrease drastically the efficiency of the medicine and a higher content may damage human health.

The prediction ability of several algorithms in different situations are investigated. Two different binary mixtures are studied. The first mixtures feature micronized powders and the second one particles of several millimetres. Then, several algorithms using a priori information or not are tried out. After selection of the appropriate algorithms, the API content of pharmaceutical tablets is predicted in the last part.

II. Material and methods

II.1. Samples and measures

Different types of mixtures were produced in order to compare several algorithms for the quantification of API. Firstly, the case of binary mixtures was studied. Using binary mixtures enables to select the best algorithms for the quantification of true pharmaceutical formulation. The first mixtures contained one API, and cellulose. The second mixtures contained the same cellulose but sucrose instead of API. Sucrose particle size was of the order of millimeters. API and sucrose in the binary mixtures ranged from 0 to 100% (w/w) in 10% increments. Both powders were micronized. Four tablets of each mixture were compressed manually. Their characteristics were: a diameter of 12 mm and a height of about 1 mm. They also featured a flat surface.

The 7-compound pharmaceutical mixtures contained API (denoted API_2: 3%), cellulose (50%), lactose (45.4%), talc (0.9%), magnesium stearate (0.4%), and two coloring agents (0.2%, 0.1%). As the emphasis was on API content, the range was constructed by varying API content from 0 to 10% in 1% increments. Concentrations of minor excipients were fixed. Cellulose and lactose concentrations were adjusted proportionally. Tablets were compressed in a rotary press in the development department to mimic the manufacturing process. Batches of about 500 tablets were produced. The characteristics of the tablets were: 8 mm in diameter, 2.2 mm thick, and a weight of 200 mg. The tablets were flat except at the border which was beveled.

Four tablets of each content were measured, thus 44 tablets for each kind of samples. Five replicates of the measurement were performed to compute the noise model for PMF method. The objective 40 µm/pixel was used because it covered the entire surface of the tablets. The tablets were analyzed over the range [1100-2450] nm by increments of 10 nm.

II.2. Algorithms

Several situations for content prediction have been investigated:

1. The first situation assumed that a range of tablets with known concentration was available such as for a calibration method. In that case, PLS algorithm has been employed for content prediction and the mean spectra of each of the data cube were used to build the mathematical model. Since the number of samples was low for a PLS calibration-validation scheme, a leave-one-out cross-validation was performed in

order to determine the number of factors to be used in the model and to assess its accuracy. This method consists in calibrating the model with the whole data set except one sample which will be predicted by the model and serves as a validation sample. The procedure is repeated m times until all samples have been treated as a validation sample. After the construction of the model, each pixel of each data cube was subsequently individually predicted. The global concentration of API of one tablet was calculated by averaging the pixel prediction values.

2. In the second situation, nothing was supposed known about the samples. MCR-ALS and PMF were then tried out using two factors with the binary mixtures. MCR-ALS and PMF were employed using the same parameters as explained in Chapter 2III.1.3. PMF noise model was also estimated using 5 replicates of the measurement. After the iterations, the rows of the concentration matrix C were weighted so as their sum equal 1 and the API concentration was estimated by averaging the pixel predictions. Another possibility for non supervised content prediction was to classify pixels and to estimate the concentration by counting the number of pixels included in the API class. For the segmentation, a k-means algorithm was used. This algorithm was employed using the full spectra.

3. In the last situation the reference concentrations were supposedly unknown but the reference spectra were available. CLS and PLS-DA were then tried out for content prediction. MCR-ALS was also used with the knowledge of reference spectra. Firstly, augmented matrices were constructed: pure spectra were spanned at the end of the unfolded matrix and the corresponding rows of the concentration matrix were fixed while the algorithm iterated (cf Figure 52 (a)). In order to test the influence of the number of added spectra, they represented 0.5%, 1% and 5% of the total number of spectra of the final matrix. Secondly, the reference spectra were introduced in the S^T matrices. Three factors were used: two remained fixed during the iteration whereas the third one was free (cf Figure 52 (b)). MCR-ALS was initialized using the reference spectra. After the extraction of the matrix C its rows were weighted so as their sum equal 1 and the API concentration was estimated by averaging the pixel predictions.

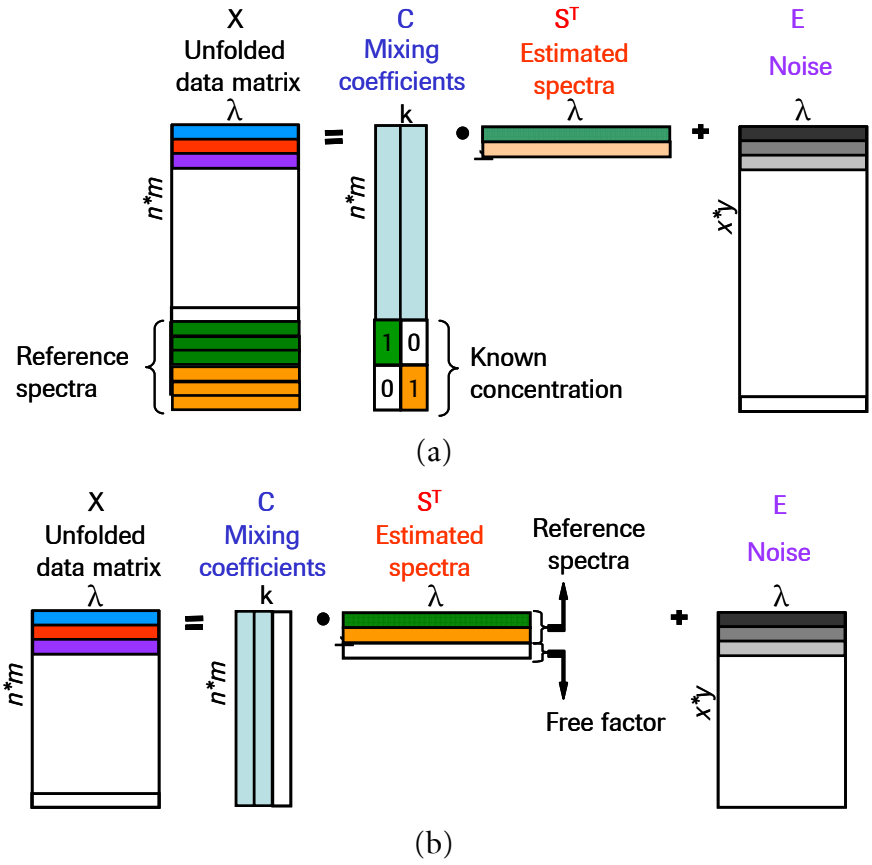


Figure 52. Using known spectral references in MCR-ALS or PMF procedure. (a) augmented matrix, (b) fixing factors in S^T matrix.

II.3. Preprocessing

The bad pixels of the data cubes were first removed. Then the spectra were converted to absorbance. Since the chosen field of view did not cover the whole surface of the tablets, circle-like mask with a 100 pixel radius was applied to the data cubes for the binary mixture tablets in order to remove pixels that represented the mirror. The 7-compound tablets were a bit smaller, thus the radius was of 80 pixels.

The data cubes were subsequently unfolded. The unfolded matrices of the binary mixtures contained then 31417 spectra, and for the 7-compound tablets 20081 spectra. Noisy channels were removed and spectra reduced to the spectral range [1300–2230] nm.

Quantification with NIR spectroscopy has been shown in the literature to be the most successful with SNV and derivative preprocessing, thus these two methods were tried out in the following. The parameters for the Savitzky-Golay derivative were a second derivative with a 9-point window and polynomial order 3. The positivity constraints for MCR-ALS and PMF on the spectra were removed, thus, the sole constraint on concentration remained.

II.4. Statistical indicators

Methods were compared using linear regression. Accuracy of prediction was evaluated by slope, intercept, square correlation coefficient of the linear regression and an additional measure of model error: the Root Mean Square Error of Prediction (RMSEP [111]), calculated as follows:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{C}_i - C_i\right)^2}{n}} \qquad \textbf{Equation 58}$$

Where $\hat{C}$ is the estimated concentration, C the reference concentration, and n the total number of samples (n = 44). When using cross-validation, the Root Mean Square Error of Cross-Validation (RMSECV) is computed using Equation 58 with $\hat{C}$ the estimated concentration during cross validation.

III. Quantification of binary mixtures

III.1. Reference and mean spectra of the tablets

Figure 53 (a) shows the SNV normalized spectra of each of the pure compounds: the cellulose (blue) and the API (green). Figure 53 (b) displays the SNV normalized mean spectra of the eleven data cube of the first set. The mean spectra showed clear spectral variation across API concentration in the binary mixtures. The API signal increased (blue arrows) whereas the cellulose signal decreased (black arrows). The spectral variations due to the different content of API were clearly marked.

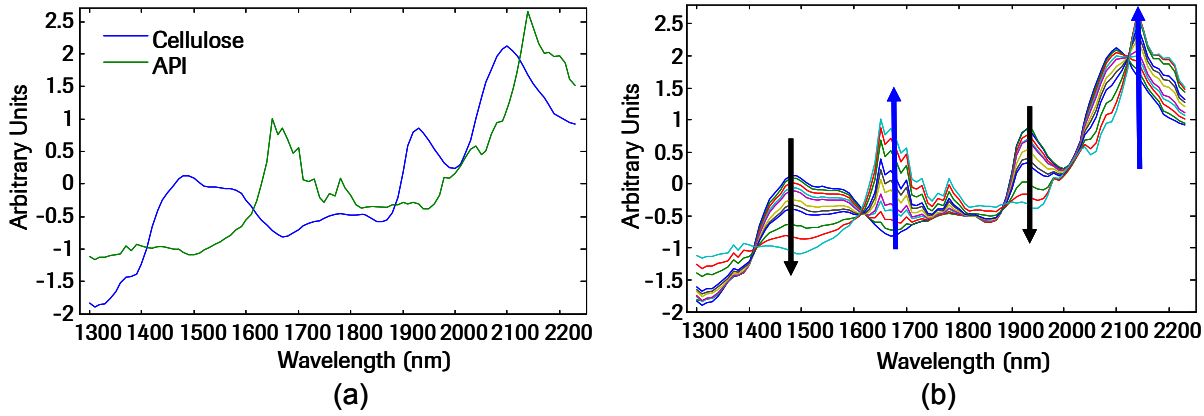


Figure 53. SNV normalized reference spectra (a) and SNV normalized mean data cube spectra across the API concentration range (b), blue arrows demonstrate the increase of the API signal while cellulose signal is decreasing (black arrows).

III.2. PLS calibration

The cross-validation results (Figure 54) indicated that the best models were obtained with four factors for both preprocessing. The corresponding RMSECV were the lowest with a value around 3%. Both preprocessings led to an accurate calibration curve with a slope of nearly one, low intercept of 0.5% and R^2 values of 0.99 (Figure 55 (a)). From the calibration curve it can be suspected that the batches at 30% and 40% were the less homogeneous because the prediction values at those percentages showed the largest variability (Figure 55 (b)).

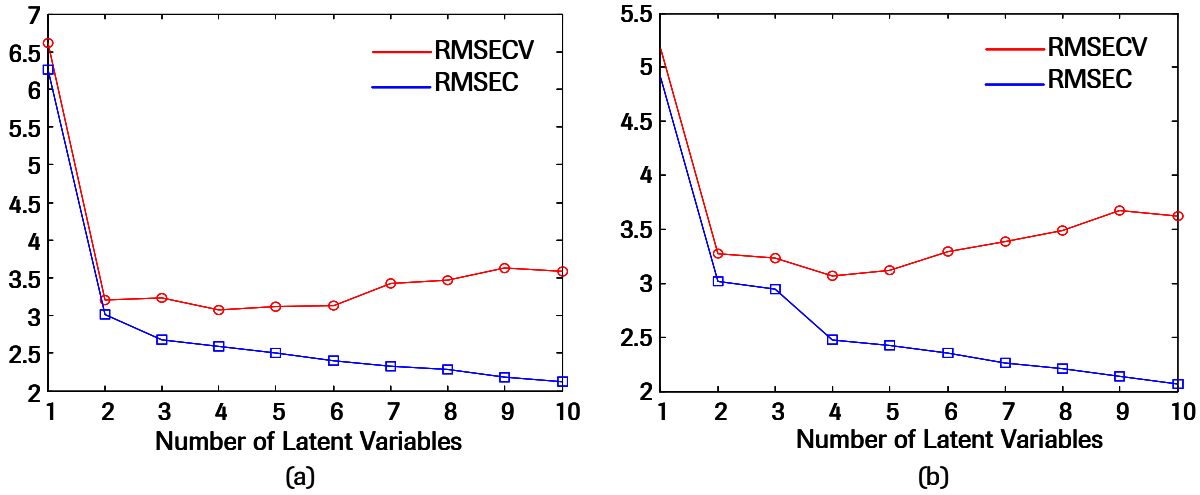


Figure 54. Cross-validation results with (a) SNV normalized spectra and (b) SNV followed by second derivative.

	PLS (4 lv)	
	snv	snv+sg
Slope	0.988	0.990
Intercept	0.55	0.47
R^2	0.991	0.991
RMSECV (%)	3.072	3.068

(a)

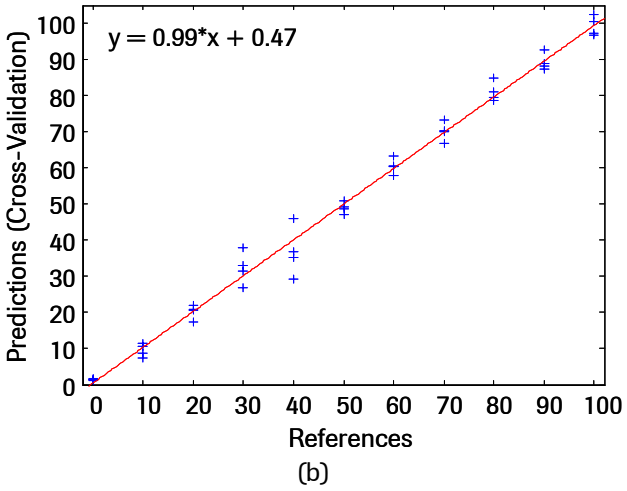


Figure 55. Results of the PLS cross-validation using four latent variables (a) statistical indicators, (b) cross-validation predictions using second derivative spectra.

Figure 56 displays the prediction maps of each tablet. The "color" of each tablet clearly varied from "blue" to "red" and was linked to the increase of the API concentration. At 30 %, 40 % homogeneity problems have been distinguished (red circles), which confirms the observation made with the calibration curve. However, problem of homogeneity could also be detected for example at 60%.

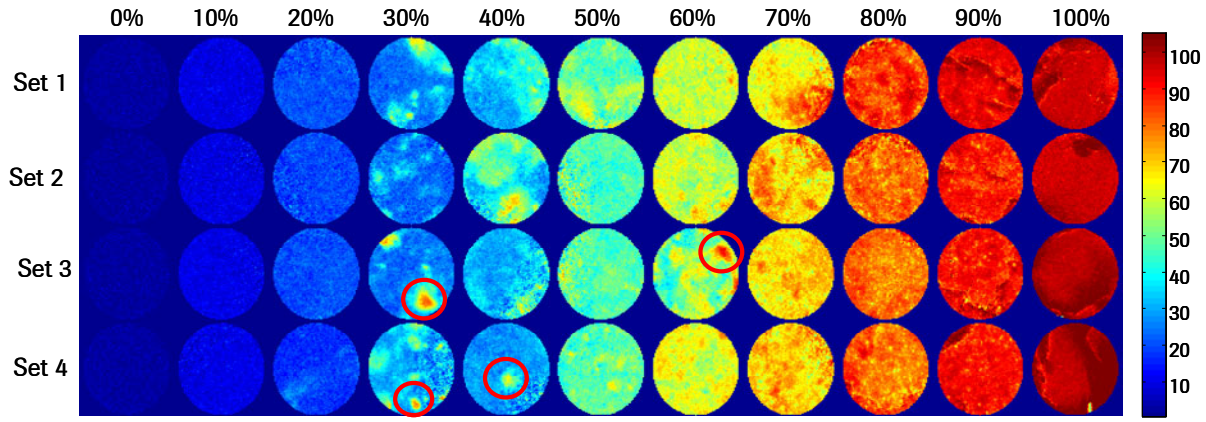


Figure 56. Prediction maps obtained with PLS model using four latent variables and SNV normalized spectra followed by second derivative.

III.3. Quantification without a priori information

PLS calibration has demonstrated high prediction ability for the API content. However, a full concentration range is often not available during an early stage of the development of new pharmaceutical forms: the formulation is often not fixed, it is time consuming and costly to produce ranges with different contents in the production. Moreover, if the formulation requires specific steps such as granulation, it is difficult to produce samples in the laboratory. Therefore, it is of interest to try out other approaches for content predictions. Ideally, one would like to have a method that employs no a priori at all.

	MCR-ALS		PMF	
	snv	snv+sg	snv	snv+sg
Slope	0.096	0.133	-0.043	0.003
Intercept	40.34	53.86	52.81	51.53
R^2	0.030	0.224	0.102	0.002
SEP (%)	25.02	23.10	18.48	19.73

Table 14. Statistical indicators for the quantification without a priori.

In our study, MCR-ALS and PMF were firstly tried for quantitative analysis. As can be seen in Table 14, these two methodologies failed to recover the concentration and even a trend as a function of the increase of API, as demonstrated by a nearly null slope, low R^2 and high intercept (around 40 -50). Rotational ambiguities were mainly responsible for the low results obtained with PMF or MCR-ALS algorithm. The algorithms converged towards the middle of the rotational domain which always gave concentrations around 40-50-60%. Besides, the noise model of PMF algorithm was no more valid with the preprocessing techniques employed for quantification. The final residuals were five times higher than the estimated noise values (σ

values) of the PMF approach. SNV introduced spectral distortion and second derivative enhanced the noise. It has been attempted to change the noise model by increasing the σ values but without success. Some runs did not converge and the computation resulted in high scaled residuals. For these reasons, the PMF method was not employed in the following.

Another remark is that below 30% of API, none of the extracted factors could have been clearly linked to the API. For example, Figure 57 (a) presents the factor extracted with a 20% tablet. The blue factor presented a "shoulder" at the API wavelength around 2150 nm but did not clearly featured API signal. With a 30% tablet the blue factor could have been clearly linked to the API (Figure 57 (b)), however contributions from the cellulose were still present. The extraction of the reference spectra from the matrix of mixed spectra was difficult because in the initial matrix none of the pixel featured pure compounds. In order to achieve a better extraction, one should introduce rotational matrices as explained in the previous chapter, but this requires user's intervention.

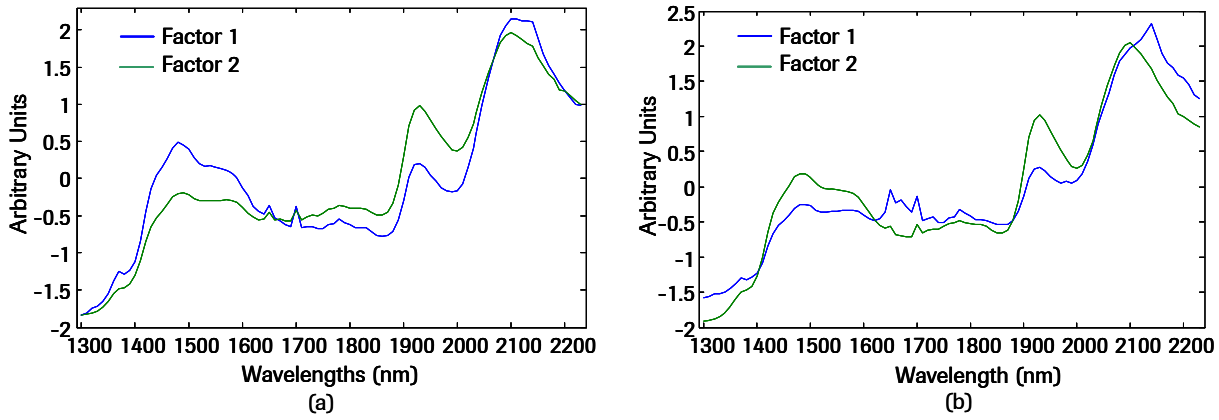


Figure 57. Extracted factors by MCR-ALS algorithm (a) 20% tablet, (b) 30% tablet

The second approach for quantification without a priori is image segmentation. However, with the studied data set, the segmentation maps that were obtained did not feature chemical species distributions except when API agglomerated, for example in the batch 30% and 40% (Figure 58).

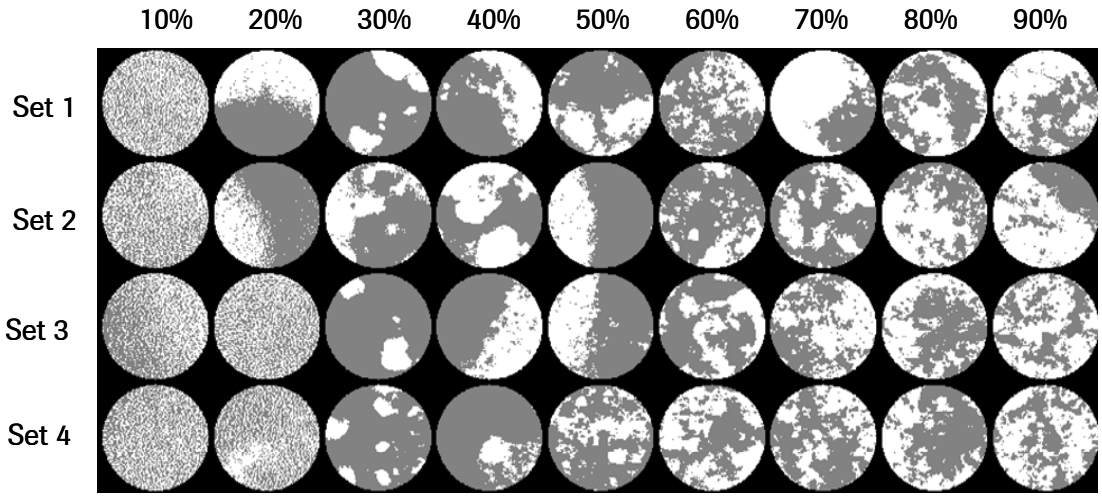


Figure 58. K-means segmentation of the API-cellulose tablets using second derivative spectra.

Because of a lack of pixel variability, K-means segmentation failed to recover the API concentration. In our case, the spatial resolution was larger than the particle sizes. Thus each pixel contained a mean information about both the API and the cellulose and could not have been separated into two classes, each of them clearly related to one of the chemical species. The K-means segmentation discriminated in fact differences due to the surface effects of the tablets and not a chemical change except when the API agglomerated. For example, the center of the two clusters extracted by K-means on two tablets (30% and 50% of the third series) are displayed on Figure 59. The 30% tablet agglomerated and thus one of the centers (Figure 59 (a) green spectrum) depicted a stronger API signal and the agglomeration was correctly segmented (Figure 58). The 50% tablet was homogeneous and the two centers (Figure 59 (b)) depicted similar spectral shapes except around 1900 nm, the water band. The clusters could not have been assigned to API or cellulose and the corresponding segmentation (Figure 58) featured a difference in the sample thickness.

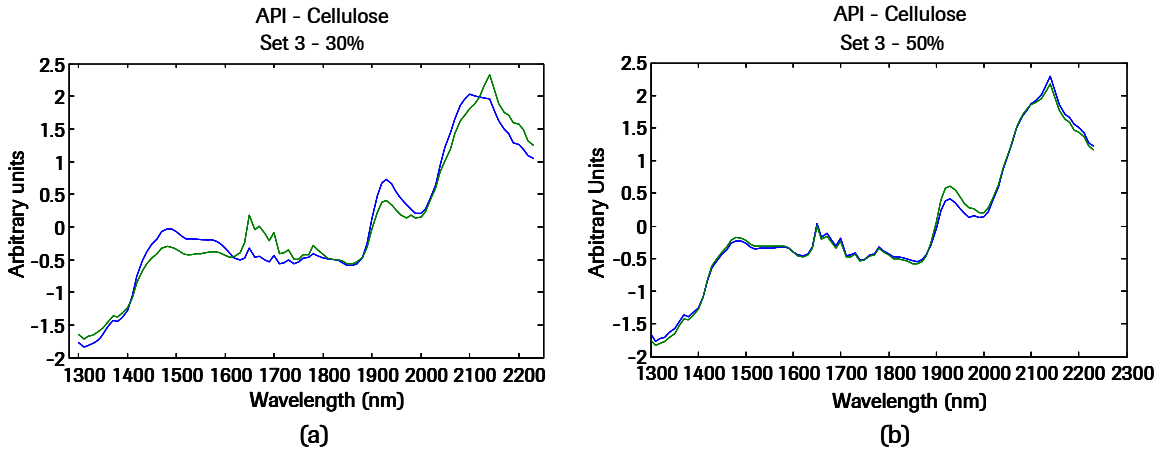


Figure 59. Center of the clusters obtained by K-means segmentation. (a) centers of 30% tablets third series, (b) centers of 50% tablets, third set.

It has just been demonstrated that when the particle size is lower than the spatial resolution of the apparatus, k-means algorithm failed to segment the image into classes related to the chemical species, thereby preventing an accurate API prediction. In order to check if the segmentation is possible with larger particles, a second range of binary mixtures has been prepared using sucrose particles of few millimeters instead of API. The results of the k-means segmentation are given in the figure below, the sucrose particles are depicted in white. Two tablets of 10% did not feature sucrose particles on the surface and the segmentation was not linked to the chemical species. The size of the cellulose particles was finer than the spatial resolution of the device, therefore, at high sucrose concentration (80% and 90%) they surrounded the sucrose particles and could not have been clearly segmented. In the other maps (comprised in the red square) the sucrose particles were correctly segmented. This example demonstrates the feasibility of image segmentation when the particles of interest are not micronized. However because of the large size of the sucrose particles, a problem of sampling has been encountered and the tablets did not represent the actual concentration of sucrose present in the blends. Thus, it lacked the references and a calibration curve could not have been drawn. With a large particle size, the presence or absence of one particle on the surface clearly changes the estimated concentration. When image segmentation is possible, the problem is then to know if the analyzed surface is representative of the batch. However, the employed powders for the production of tablets are often the micronized ones and such an approach might only be seldom used.

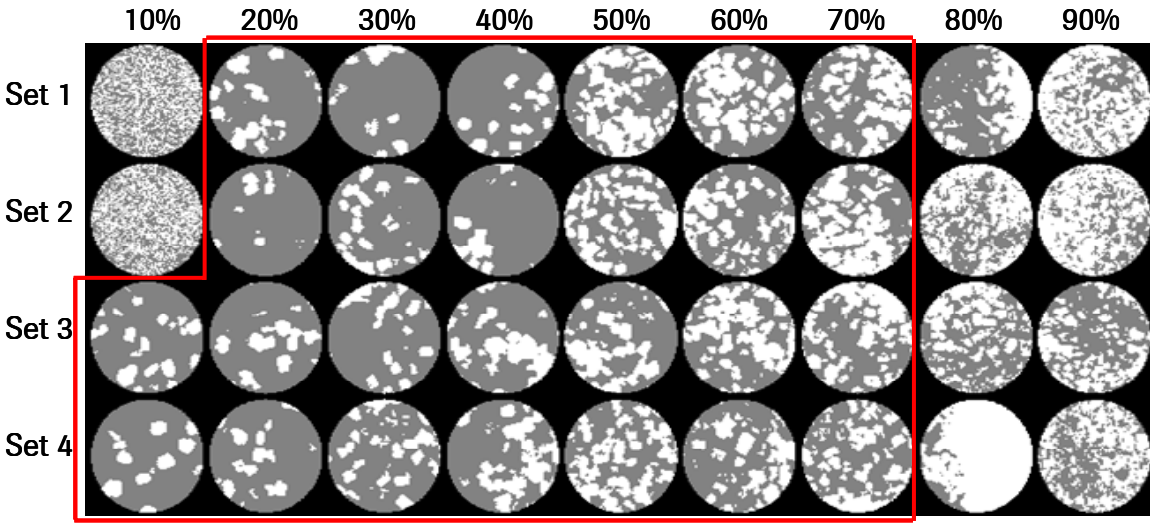


Figure 60. Segmentation of the sucrose-cellulose binary mixtures obtained using k-means algorithm on spectra preprocessed by a second derivative.

III.4. Quantification with the help of reference spectra

The above paragraph has demonstrated that quantification without any a priori is challenging for most of the pharmaceutical formulations. Therefore other strategies must be employed for quantification. If a range of tablets with known concentration is not available, at least, the compounds present in the tablets are very often fully known. Thus, PLS-DA, CLS and augmented MCR-ALS approaches appeared to be appropriate candidates for content prediction.

	PLS-DA (2 lv)		PLS-DA (3 lv)		CLS	
	snv	snv+sg	snv	snv+sg	snv	snv+sg
Slope	0.960	0.977	0.970	0.979	0.966	0.974
Intercept	-7.30	-4.78	-7.28	-7.81	-7.83	-4.70
R^2	0.960	0.980	0.964	0.971	0.960	0.980
SEP (%)	11.25	7.43	10.66	10.40	11.45	7.47

Table 15. Statistical indicators of the quantification obtained using PLS-DA and CLS algorithm. Lv = latent variables

PLS-DA was tried out with two and three latent variables. The best model for PLS-DA was obtained when using two latent variables and second derivative preprocessing (Table 15 and Figure 61). The best number of latent variables corresponded to the number of chemical compounds. The SEP was 7.43% when two latent variables were employed in the model versus 10.40% with three latent variables and the linearity was improved. CLS algorithm on second derivative reached accuracy similar to the one obtained with PLS-DA model constructed with two latent variables (Table 15). The calibration curve depicted on Figure 61 also revealed homogeneity problem within the batches 30% and 40%.

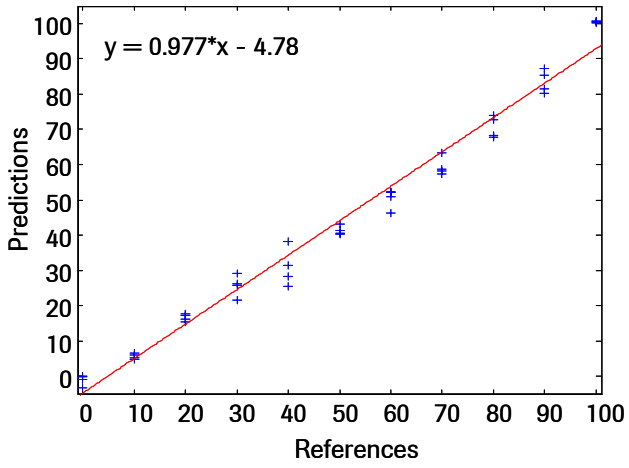


Figure 61. Calibration curves using PLS-DA with spectra corrected by SNV followed by second derivative.

After PLS-DA and CLS quantification, MCR-ALS with reference spectra has been considered for quantification. The first strategy was to fixed two factors equal to the reference spectra and let one supplementary factor free that would take into account the non-modeled spectral variations. The results of this first approach are given Table 16, first columns. The best quantification results were achieved with second derivative spectra but were not more accurate than those given by a CLS (Table 15). The other strategies were to spann at the end of the matrices of mixed spectra reference spectra. Different amounts of reference spectra were added. First they represented 0.5% of the total number of spectra then 1% and 5%. Globally, the results (Table 16) showed that those strategies did not either improve the statistical results in comparison with a CLS algorithm (Table 15). However, the SEP decreased by 1% between the first augmented matrix (0.5%) and the last one (5%), thus the number of spectra spanned at the end of the matrix influenced the results of the quantification.

	MCR-ALS		MCR-ALS		MCR-ALS		MCR-ALS	
	2 fixed factor + 1 free		Augmented 0.5%		Augmented1%		Augmented 5%	
	snv	snv+sg	snv	snv+sg	snv	snv+sg	snv	snv+sg
Slope	0.952	0.959	0.920	0.973	0.921	0.96	0.937	0.94
Intercept	-7.24	-4.01	-5.87	-5.09	-5.82	-4.29	-6.40	-2.93
R^2	0.959	0.981	0.958	0.966	0.959	0.970	0.960	0.978
SEP (%)	11.58	7.50	11.90	8.71	11.75	8.31	11.48	7.55

Table 16. Statistical indicators of the quantification obtained using MCR-ALS with reference spectra

Given the above results, it has been tried to add more reference spectra (they represented 10%, 30% and 50% of the total number of the spectra) to the initial matrix preprocessed by a SNV followed by second derivative. A slight decrease of the SEP (Table 17) that reached a plateau (7.40%) was obtained but still those results were not significantly more accurate than the results given by a CLS algorithm (Table 15).

	MCR-ALS Augmented 10%	MCR-ALS Augmented30%	MCR-ALS Augmented 50%
	snv+sg	snv+sg	snv+sg
Slope	0.943	0.947	0.951
Intercept	-2.95	-3.22	-3.47
R^2	0.980	0.981	0.981
SEP (%)	7.45	7.40	7.40

Table 17. Quantification results obtained with augmented matrices (10% 30 % 50 %), the spectra were preprocessed using SNV followed by second derivative

Globally, none of the approaches of quantification with the reference spectra reached the accuracy of a PLS calibration. However, they were able to approximate the API content which gives an intermediate solution for semi-quantitative analysis when a full range of tablets with different concentrations can not be constructed.

IV. Quantification of pharmaceutical tablets

After the quantification on synthetic binary mixtures, quantification of pharmaceutical tablets was performed. These tablets were made of micronized API, thus, taking into account the previous results, the quantification was performed using PLS, PLS-DA, CLS and augmented MCR-ALS algorithms. Given the low content of API, without a priori information MCR-ALS and PMF failed to recover the API spectrum. Segmentation by K-means was not appropriate because the particle size was below the spatial resolution of the device.

IV.1. Reference and mean spectra of the tablets

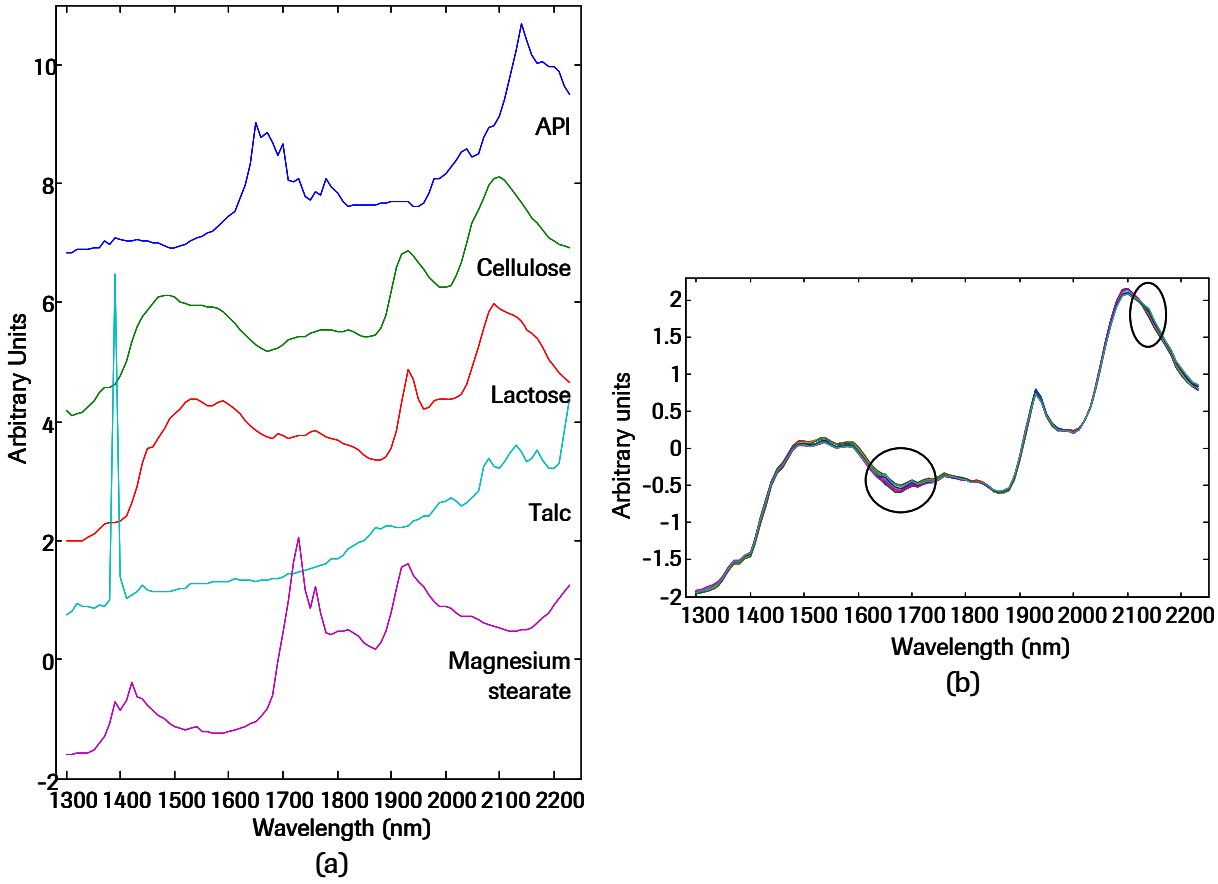


Figure 62. SNV normalized reference spectra (a) and SNV normalized mean data cube spectra across the API concentration range (b), black circles depict the spectral variations due to the increase of the API content.

Figure 62 (a) depicts the spectra of the pure compounds of the pharmaceutical tablets. Figure 62 (b) shows eleven mean spectra across the API concentration range. The black circles underline the spectral variation between [1600 and 1700] nm and [2100-2200] nm due to the increase of the API content. These variations were subtle.

IV.2. Quantification results

IV.2.1. PLS quantification

The cross-validation results (Figure 64) indicated that the best PLS models were obtained with 7 latent variables with SNV normalized spectra; and 8 factors with SNV followed by second derivative spectra. Both models gave an SEP of 0.34 %. Both preprocessings led to an accurate calibration curve with a slope of nearly one, intercept around 0 and R^2 values of 0.98 and 0.99 (Figure 65(a)). The prediction values were evenly distributed across the API concentration range revealing that the PLS model was accurate for prediction.

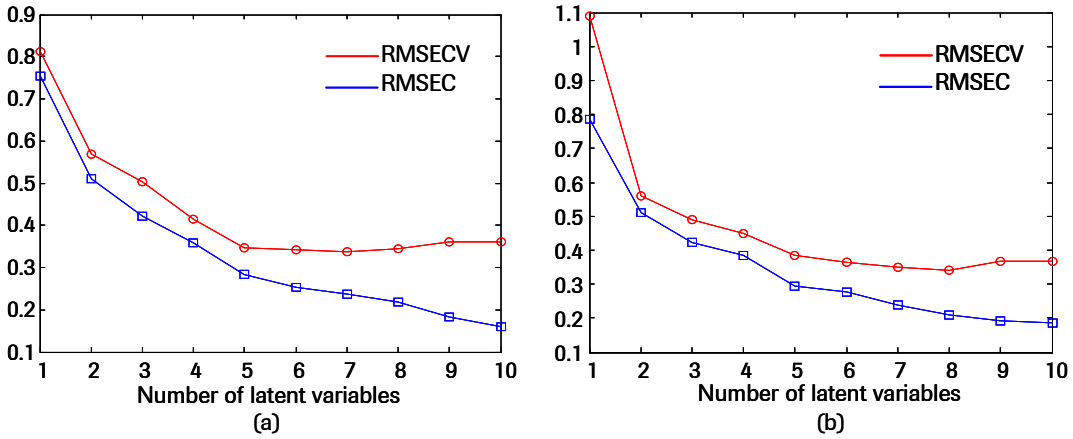


Figure 63. Pharmaceutical tablets: cross-validation results with SNV normalized spectra (a) and SNV followed by second derivative (b).

	PLS	
	snv (7 lv)	snv+sg (8 lv)
Slope	0.999	0.989
Intercept	0.02	0.07
R^2	0.989	0.988
SEP (%)	0.34	0.34

(a)

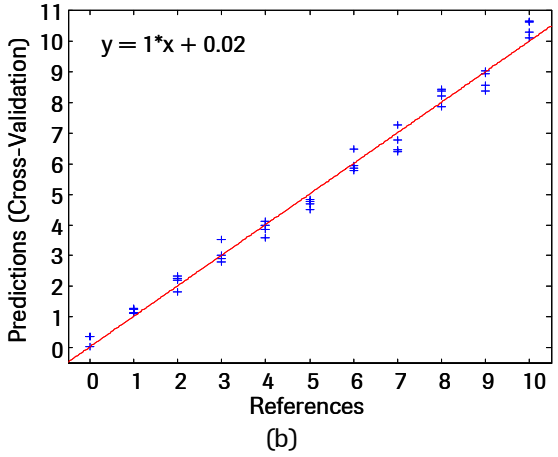


Figure 64. Results of the PLS calibration: statistical indicators (a), calibration curves obtained with SNV normalized spectra (b) (7 latent variables).

The prediction maps given in Figure 65 showed a clear color variation as a function of the increase of API content. Inhomogeneity in the tablets (red circles) can be visually detected.

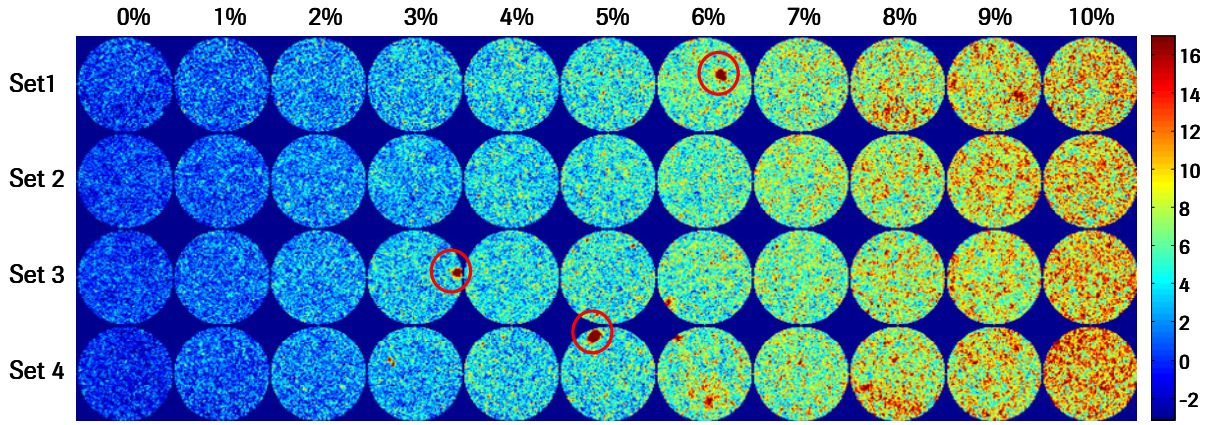


Figure 65. Prediction maps obtained with PLS model using six latent variables and SNV normalized spectra.

IV.2.2. PLS-DA quantification

Different PLS-DA models were tested. Firstly the spectral library was constructed considering three classes (the major compounds). 100 pure spectra of each of the three major compounds (300 spectra) were used to compute the model. The second spectral library considered the minor compounds and hence contained five classes (5×100 spectra). The aim was to test the influence of the minor compounds. Then, different numbers of latent variables were employed. Firstly, the number of latent variables was the same as the number of compounds and it has been increased until the prediction ability worsens.

	PLS-DA (300 sp.) 3 lv		PLS-DA (300 sp.) 4 lv		PLS-DA (300 sp.) 5 lv		PLS-DA (300 sp.) 6 lv	
	snv	snv+sg	snv	snv+sg	snv	snv+sg	snv	snv+sg
Slope	0.647	1.067	0.848	0.928	0.845	0.926	0.867	0.869
Intercept	0.85	-2.30	0.12	0.59	0.27	0.11	-0.43	0.44
R^2	0.955	0.938	0.973	0.954	0.974	0.971	0.974	0.964
SEP (%)	1.51	2.16	0.92	0.72	0.83	0.62	1.25	0.70

	PLS-DA (500 sp.) 5lv		PLS-DA (500 sp.) 6 lv		PLS-DA (500 sp.) 7 lv		PLS-DA (500 sp.) 8 lv	
	snv	snv+sg	snv	snv+sg	snv	snv+sg	snv	snv+sg
Slope	0.685	1.000	0.792	1.011	0.814	0.951	0.796	0.914
Intercept	1.45	-0.97	0.76	-1.97	1.05	-0.85	1.35	-0.53
R^2	0.948	0.968	0.970	0.789	0.970	0.944	0.968	0.965
SEP (%)	1.12	1.12	0.84	2.53	0.75	1.33	0.86	1.14

Table 18. PLS-DA calibrations of the pharmaceutical samples.

The results of PLS-DA quantification are shown on Table 18. Using 300 spectra, the SEP was at minimum when the model was constructed with 5 latent variables and second derivative spectra (gray box). The SEP was of 0.62 and the slope of 0.93. Using 500 spectra, the best

calibration was obtained with 7 latent variables on SNV normalized spectra. However, the prediction ability of this model was not as accurate as the first one: the SEP was a bit higher: 0.75 against 0.62 and the linearity smaller: slope of 0.814 against 0.926. With our tablets, introducing the minor compounds in the PLS-DA model did not improve the content prediction.

The calibration curve using 300 spectra and 5 latent variables is displayed on Figure 66. Prediction values showed more variability for the tablets which contained less than 5% of API.

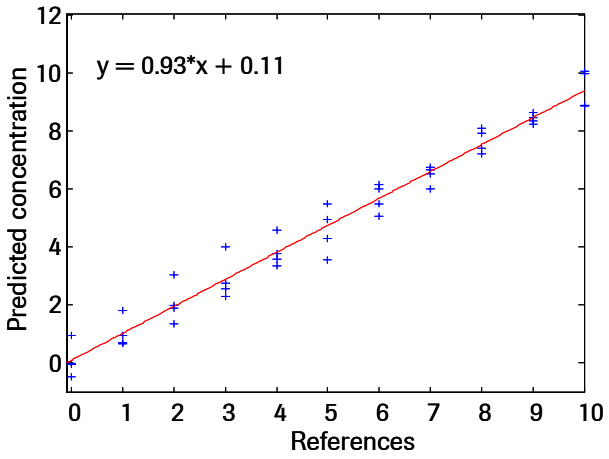


Figure 66. PLS-DA calibration curve obtained using 3 reference spectra, 5 latent variables and second derivative spectra.

IV.2.3. CLS quantification

So as with PLS-DA algorithm, CLS was applied first with the three reference spectra of the major compounds, then with the five reference spectra. The results are given Table 19. The most accurate predictions were obtained using the spectra of the five compounds with second derivative pretreatment: the SEP of 1.2% was the lowest with a good linearity (slope around 1 and high R^2: 0.967). However, the intercept was quite high (-1.13). The calibration curve (Figure 67) showed negative predictions at low content (0%, 1% and 2%). CLS gave the worst prediction ability in comparison with PLS and PLS-DA. Nevertheless, it must be noted that with the same tablets, CLS has shown previously the same prediction ability than a PLS algorithm [112]. Two years have past between the measurements. It is thus believed that the noise of the instrument has increased over time and CLS was no more able to predict accurately the API content in the second series of experiments. CLS algorithm was less robust to measurement errors than PLS or PLS-DA algorithm.

	CLS 3 refs		CLS 5 refs	
	snv	snv+sg	snv	snv+sg
Slope	0.645	1.073	0.685	1.009
Intercept	0.930	-2.331	1.501	-1.128
R^2	0.954	0.940	0.947	0.967
SEP (%)	1.473	2.157	1.121	1.235

Table 19. CLS statistical indicators using three of five reference spectra.

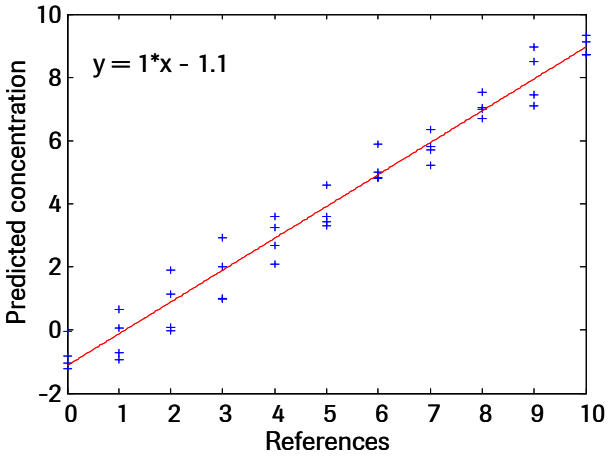


Figure 67. CLS calibration curve obtained using second derivative spectra and five reference spectra

IV.2.4. Augmented MCR-ALS

As with the binary mixtures the two augmented approaches in order to introduce the reference spectra to MCR-ALS algorithm were tested as well as the influence of the minor compound reference spectra. The results given by the first strategy which consists in fixing the first factors equal to the reference spectra and letting on additional factor free are given in Table 20. The best quantification was obtained when fixing the reference spectra of the main compounds and second derivative spectra. This strategy led to SEP smaller than the SEP given by the best CLS quantification (Table 19), however, PLS-DA remained more accurate (Table 18).

	MCR_ALS 3 refs+1free		MCR_ALS 5 refs+1free	
	snv	snv+sg	snv	snv+sg
Slope	0.705	0.899	0.530	0.892
Intercept	2.021	0.768	3.359	2.714
R^2	0.822	0.939	0.879	0.678
SEP (%)	1.498	0.835	1.901	2.937

Table 20. Statistical indicators of the quantification obtained when fixing three or five factors equal to the reference spectra and let one factor free.

Table 21 displays the statistical indicators provided by the second strategy of augmented MCR-ALS which spanned reference spectra to the matrix of mixed spectra. Different amounts of reference spectra were added. The most accurate quantification was obtained when 90% of the total number of spectra in the augmented matrices were reference spectra. The SEP were the smallest: 1.017% when using only the reference spectra of the major compounds and 0.719% when the reference spectra of the minor compounds were also considered. The slopes of the calibration curves were also highest at that amount of reference spectra. Second derivative preprocessing always improved the determination of the API content. Augmented MCR-ALS was globally more accurate than CLS algorithm because it prevented negative predictions at lower content of API (Figure 68), however, PLS-DA remained the best approach (minimum SEP of 0.62% versus 0.72% for augmented MCR-ALS).

	MCR-ALS aug 3 refs 30%		MCR-ALS aug 3 refs 50%		MCR-ALS aug 3 refs 70%		MCR-ALS aug 3 refs 90%	
	snv	snv+sg	snv	snv+sg	snv	snv+sg	snv	snv+sg
Slope	0.456	0.555	0.484	0.676	0.509	0.775	0.536	0.853
Intercept	3.719	0.800	3.128	0.513	2.646	0.292	2.346	0.065
R^2	0.964	0.957	0.961	0.957	0.958	0.955	0.956	0.952
SEP (%)	2.009	2.039	1.748	1.572	1.601	1.217	1.510	1.017
	MCR-ALS aug 5 refs 30%		MCR-ALS aug 5 refs 50%		MCR-ALS aug 5 refs 70%		MCR-ALS aug 5 refs 90%	
	snv	snv+sg	snv	snv+sg	snv	snv+sg	snv	snv+sg
Slope	0.265	0.627	0.299	0.758	0.353	0.860	0.405	0.920
Intercept	2.021	1.093	1.739	0.922	1.455	0.804	1.326	0.698
R^2	0.936	0.965	0.928	0.964	0.935	0.961	0.941	0.959
SEP (%)	2.861	1.459	2.844	0.938	2.726	0.712	2.520	0.719

Table 21. Statistical indicators of the quantification obtained using the second strategy of augmented MCR-ALS.

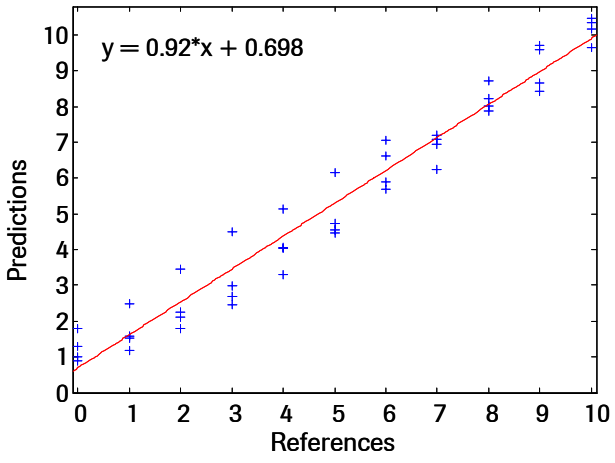


Figure 68. Augmented MCR-ALS calibration curve obtained using second derivative spectra, and reference spectra which accounted for 90% of the total number of spectra in the mixed matrix.

IV.3. Discussion

The present work on binary mixtures has underlined several issues for the estimation of content uniformity of tablets with NIR-CI. Ideally, one would like to have a method that works without a priori information. Two schemes might be used to reach that goal: (1) multivariate curve resolution algorithm to extract the matrix C of concentration or (2) classification that aims at assigning each spectrum of the data cube to classes related to the chemical species and could provide quantitative information by counting the number of pixels belonging to the class of interest. However, several limitations, mainly due to the inherent properties of NIR spectroscopy then appeared. The first ones are the spatial resolution of the apparatus which is limited according to the literature to 30 μm/pixels [20] and the penetration depth which can reach several hundreds of micrometers [21]. Thus, if the chemical species feature micronized powders spread all over the surface of the sample, each spectrum contains similar mixed information and it lacks spectral variability to clearly extract pure spectra. In the same way, clustering techniques failed to extract classes that are related to the pure compounds. When the particles are larger than the spatial resolution, the segmentation is possible but then one should be sure that the surface of the tablet is representative of the whole sample which is not necessarily the case. It is possible that quantification of micronized powder by segmentation might be achieved using other imaging techniques such as Infrared or Raman spectroscopy because the spatial resolution is finer, the spectra are more specific of the chemicals and the beam penetrates less into the sample (see annex).

In fact, such as with classical spectroscopy, the best quantification is obtained when a full range of concentration and a PLS model are constructed. However, it might be difficult to produce these ranges of tablets when a pharmaceutical form is under development: the formulation might change, it is time consuming and costly to construct calibration samples and some formulations involving granulation for instance cannot be produced at a laboratory scale. The intermediate solution is then to use the reference spectra of the pure compounds which are often known when analysing pharmaceutical forms. In that case, PLS-DA algorithm is the most precise for quantification in comparison with CLS algorithm which is more sensitive to noise and augmented MCR-ALS. Without reaching the prediction ability of a PLS algorithm, the estimated content suffices to have a semi-quantitative analysis of the samples and to enable the comparison of batches. The advantage of imaging spectroscopy is that the quantitative distribution maps then reveal agglomerations.

V. Conclusions

It has been demonstrated in this chapter that the content prediction of API with NIR chemical imaging is challenging but possible. The best method is still a classical PLS regression but requires the construction of a set with known concentrations that are often not available during the development of a new formulation. Without a priori, MCR-ALS and PMF fail to recover the concentrations because of rotational ambiguity, and at low content the API signal can not be extracted. As well, a segmentation scheme would work only in specific situation where the powder particles are of similar sizes and larger than the spatial resolution of the device or if they agglomerate. The limitations of the unsupervised methods for quantification are due to the inherent limitations of NIR chemical imaging such as penetration depth and matrix effects. A possible alternative is to use PLS-DA algorithm that gives an estimation of the API concentration as demonstrated with the API - cellulose binary mixtures and the pharmaceutical tablets.

Conclusions

The aim of the present thesis was to provide processing schemes of Near infrared hyperspectral data cubes for the analysis of pharmaceutical solid dosage forms. Two main problematics have driven this work:

1) The first one was the extraction and characterization of the distribution maps of the chemical species for objective comparison of different pharmaceutical samples.

2) The second one was the quantification of the active in tablets.

The extraction of distribution maps has been firstly addressed using samples from the development that were produced with different parameters for process optimization. In this case the full composition of the medicine is known. The pure powders are available and the reference spectra of the chemicals can be acquired. If the chemicals feature spectral signature that does not overlap, images at a single wavelength are sufficient to extract distribution maps otherwise PLS-DA or CLS can be employed. In our example, the main issues were the homogeneity of the chemicals and the particle size distribution. Histogram analysis revealed that one batch was significantly more heterogeneous than the others. Regarding particle sizes, a processing scheme using Otsu thresholding and watershed detected a significant difference of the particle sizes between two batches. Thanks to this study it has been demonstrated that NIR-CI is a useful tool for the systematic investigation of samples from the development. Routine analysis which allows to determine which parameters influence mostly the quality of the products, are now performed in the laboratory. Those analysis generate data base that can serve to develop further in-line or at-line control of the pharmaceutical product.

After the study of fully known sample, the extraction of distribution maps without any a priori information has been focused on. In that case both distribution maps and spectral signatures must be simultaneously extracted. Different algorithms, namely NMF, BPSS, MCR-ALS and PMF have been compared as well as two different initializations of the matrices $\mathbf{C}$ and $\mathbf{S}^T$. With a random initialization, each run of NMF, BPSS and MCR-ALS led to different extractions because of rotational ambiguity, PMF reached a global minimum. When the initialization was performed using OPA algorithm, the quality of the extractions were greatly

improved. Globally, BPSS gave the worst extraction. NMF, MCR-ALS and PMF led to similar extractions of the reference spectra. Besides the comparison of the algorithms, a procedure to investigate the domain of rotational ambiguity using PMF and ME2 software has been proposed. By visual inspection of the factors, matrices of rotation can be introduced to achieve a better resolution. Another possibility is to introduce the knowledge about the global concentration of each of the compounds. The PMF extraction has been greatly improved using one of the other techniques and the concentration could have been accurately recovered. It must be noted that only ME2 software provides the possibility of introducing such information. This is a novel approach to overcome rotational ambiguity for the multivariate curve resolution of NIR-CI and a further extension would be to develop a method that would allow to construct the matrices of rotation in a non supervised manner.

Quantification of tablets has been performed, firstly on binary mixtures and then on pharmaceutical formulation. The reported results demonstrated that quantification is a challenging task. In our experiments, none of the alternative approaches to PLS regression could have reached such accurate predictions. When the samples featured micronized powder, the quantification without a priori was impossible because of particle size lower than the spatial resolution of the technique. If the sample was made of larger particles, the segmentation was possible but then sampling problem has been enhanced: the surface which was analyzed was no more representative of the whole sample. Besides, it has been demonstrated that the quantification using the reference spectra with PLS-DA algorithm provides an estimation of the concentration.

All things considered, the usefulness of NIR-CI for pharmaceutical application needs no further demonstration. Currently, NIR-CI is very well suited for macroscopic investigations. However, one should also be aware of its limitations which are mostly inherent from diffuse reflection NIR spectroscopy. Because of the penetration depth, the spatial resolution is limited and matrix effects can prevent accurate extraction of minor compounds or micronized powders. To quote Geladi, "*a pixel in a near-infrared image is not an isolated sample cell*" [113]. The determination of the interaction volume of NIR radiation with the matter surely remains the biggest challenge in order to put forward NIR-CI applications.

Annex A

Comparison between NIR, IR and Raman chemical imaging

Figure A.1. and Figure A.2. display a RGB image of a tablet acquired by the help of a Raman and Mid-IR spectrometer respectively. In the figures, the spectra located at specific pixel positions (black) are compared to the spectra of the pure compounds (see legend for details). The advantage of Mid IR and Raman spectroscopy in comparison with NIR (Figure A.3) is that the spectra present fine and numerous peaks allowing the identification of chemicals without chemometrics. The beam penetrates a few micrometers under the surface and the acquired spectra match nearly the reference spectra. The second advantage is their highest spatial resolution. With a Raman system it goes down to 1 μm enabling the localisation of fine patterns. For Mid-IR analysis of tablets it is necessary to use ATR (Attenuated Total Reflection) accessory to avoid saturation of the detector. ATR consists of a crystal placed at the optical surface of the sample. With ATR, the resolution can go down to 4 μm. The main drawback of these methods over NIR imaging is the relative higher acquisition time and the instrumentation which is more costly and not suitable for on-line analysis. With Raman spectroscopy, the chemicals might fluoresce and hide the peaks . Moreover, the high energy of the laser beam might damage the samples.

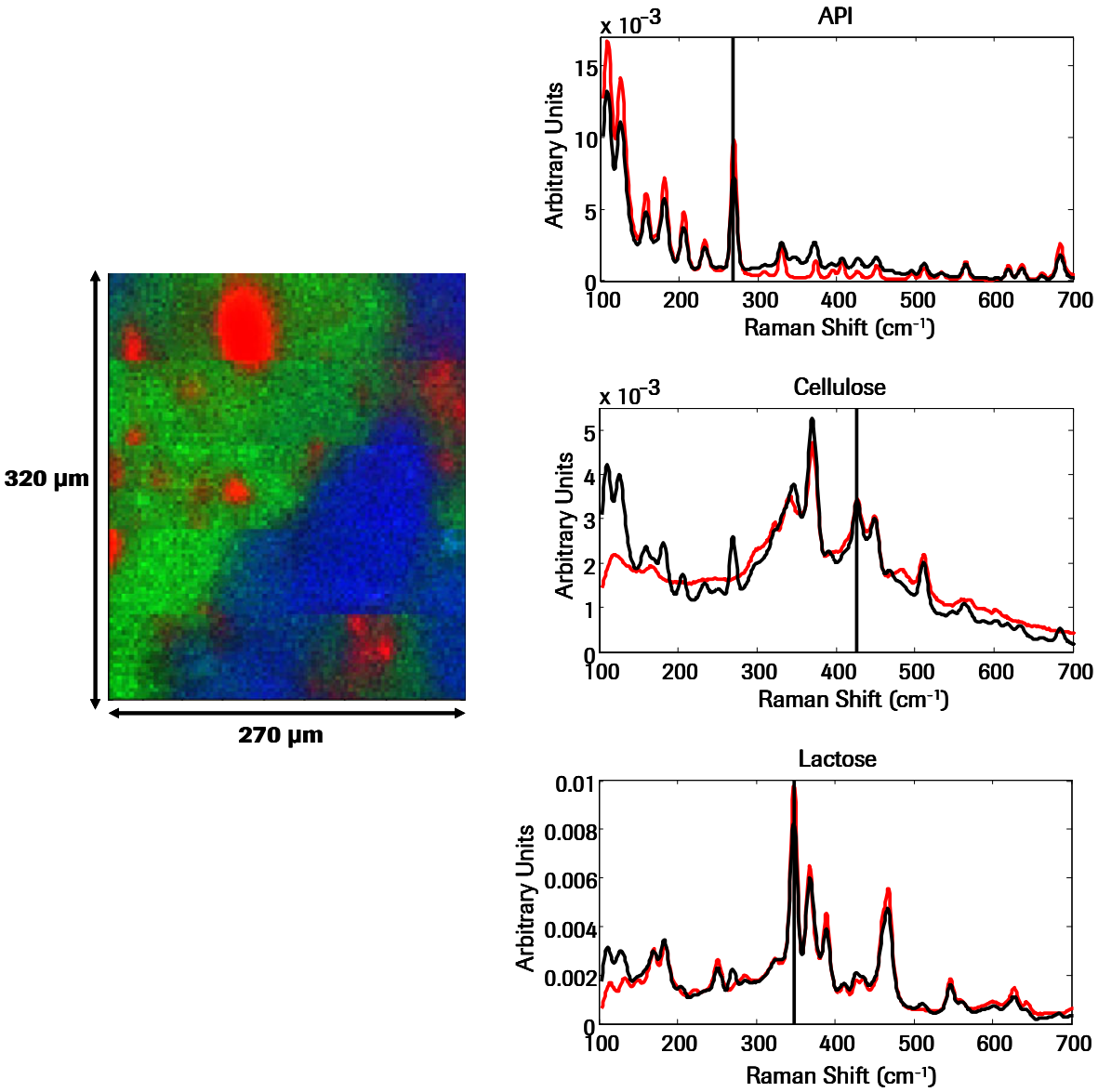


Figure A.1. Micro-scale red-green-blue (RGB) image of a tablet. The data were acquired with a Raman spectrometer (Invia reflex microscope, Renishaw) equipped with a line detector. Red depicts active pharmaceutical ingredient, green cellulose, and blue lactose. The black spectra on the right were extracted from the data cube. API spectrum was extracted from the red area, cellulose spectrum from the green area and lactose spectrum from the blue area. The spectra extracted from the data cube match the reference spectra (red spectra). Vertical lines: the wavelengths chosen to build the RGB image (269 cm^{-1} for the API, 427 cm^{-1} for cellulose, and 347 cm^{-1} for lactose).

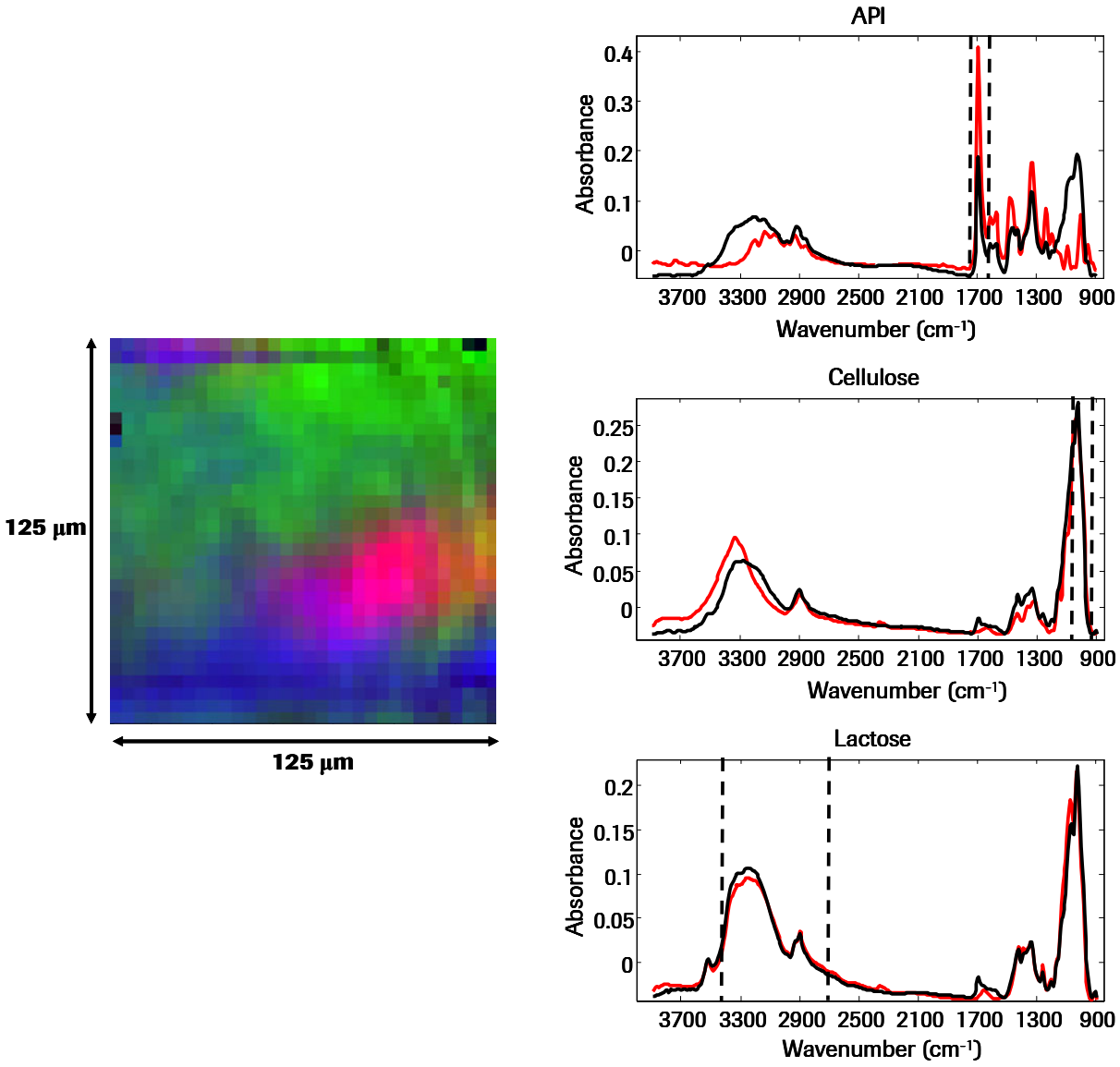


Figure A.2. Micro-scale red-green-blue (RGB) image of a tablet. The data were acquired with a FT-IR spectrometer (Equinox 55 and Hyperion microscope, Bruker) equipped with a 64*64 FPA detector by the help of an ATR accessory. Red depicts active pharmaceutical ingredient, green cellulose, and blue lactose. The black spectra on the right were extracted from the data cube. API spectrum was extracted from the red area, cellulose spectrum from the green area and lactose spectrum from the blue area. The spectra extracted from the data cube match the reference spectra (red spectra). Dotted vertical lines: the spectral regions integrated to build the RGB image ([1743 1635] cm^{-1} region for the API, [1187 941] cm^{-1} region for the cellulose, [3486-2730] cm^{-1} for the lactose).

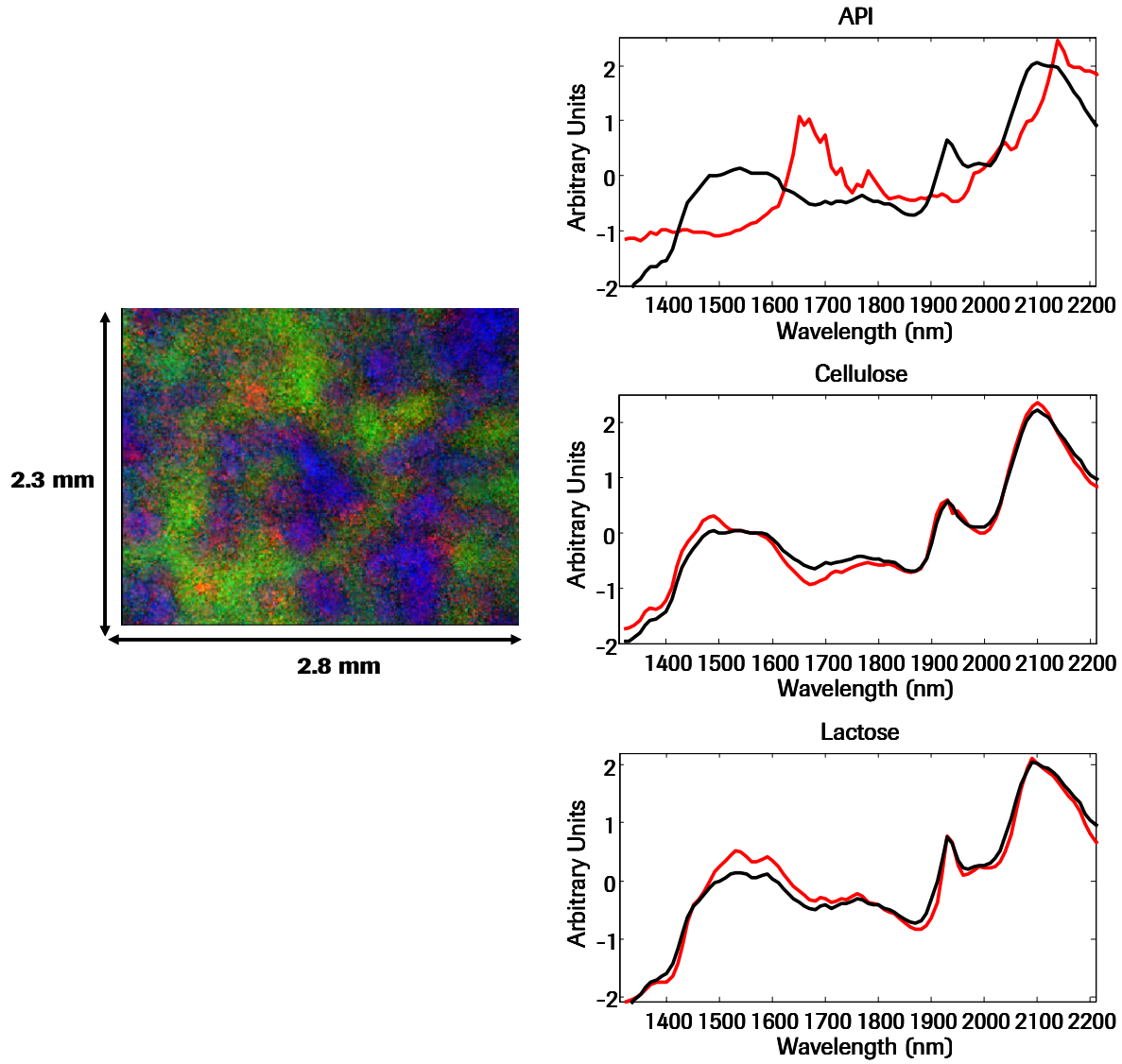


Figure A.3. Micro-scale red-green-blue (RGB) image of a tablet. The data were acquired with a NIR spectrometer (Sapphire, Malvern) equipped with a 256*320 FPA detector and a 10μm / pixel objective. Red depicts active pharmaceutical ingredient, green cellulose, and blue lactose. The black spectra on the right were extracted from the data cube. API spectrum was extracted from the red area, cellulose spectrum from the green area and lactose spectrum from the blue area. The spectra extracted from the data cube does not match the reference spectra (red spectra). Especially, the spectra extracted from a red area ("API") does only feature a shoulder at 2140 nm. Vertical lines: the wavelengths chosen to build the RGB image (2140 nm for the API, 1910 nm for the cellulose, 1930 for the lactose).

Annex B

List of author's publications

Papers in international journals

1. *C. **Gendrin**, Y. Roggo, C. Collet*, Vibrational chemical imaging and chemometrics for pharmaceutical applications: A review, Journal of Pharmaceutical and Biomedical Analysis, Vol **48** (2008), p. **533-553**

2. *C. **Gendrin**, Y. Roggo, C. Collet*, Self modelling curve resolution of Near Infrared Imaging Data, Proceedings of the ICNIRS 2007, Umea, Sweden, Journal of Near Infrared Spectroscopy , Vol **16** (2008), Issue **3**, p. **151-157**

3. *C. **Gendrin**, Y. Roggo, C. Spiegel, C. Collet,* Monitoring Galenical Process Development by NIR Chemical Imaging: one case study, European Journal of Pharmaceutics and Biopharmaceutics, Vol **68** (2008), Issue **3**, p. **376-385**

4. *C. **Gendrin**, Y. Roggo, C. Collet,* Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: A feasibility study, Talanta, Vol **73** (2007), p. **733-741**

Other papers

5. *Y. Roggo, **C. Gendrin**, C. Spiegel*, Intérêt de l'imagerie chimique proche infrarouge pour l'industrie pharmaceutique, Spectra Analyse, 258 (2007), 26-30

Oral presentation

6. *Y. Roggo, **C. Gendrin**,* Near Infrared Spectroscopy and chemical imaging for Process Analytical Technologies, Pharmaceutical analysis Course, Informa conference, February 2008, Dublin, Ireland

7. *Y. Roggo, **C. Gendrin**,* Imaging the future of NIR spectroscopy ?, 12th International Meeting on Recent Developments in Pharmaceutical Analysis (RDPA 2007), September 2007, Island of Elba, Italy.

8. *C. **Gendrin**, Y. Roggo, C. Collet,* NIR technologies for Process Understanding, Informa life sciences, October 2007, Amsterdam, Netherlands

9. *C. **Gendrin**, Y. Roggo, C. Collet,* Blind Source Separation of NIR imaging data, 13th International conference of Near infrared spectroscopy (ICNIRS 2007), June 2007, Umea, Sweden

Posters

10. *C. **Gendrin**, Y. Roggo, C. Collet*, Particle detection in Near-infrared images by image segmentation using mathematical morphology, 11th international conference on Chemometrics in Analytical Chemistry (CAC 2008), Montpellier, France

11. *C. Spiegel, Y. Roggo, **C. Gendrin**, A. Fischer,* Fast identification of Drug counterfeits by Near infrared Imaging, 13th International conference of Near infrared spectroscopy (ICNIRS 2007), Umea, Sweden

12. ***C. Gendrin**, Y. Roggo, C. Spiegel, C. Collet,* Monitoring Galenical Process Development by Chemical Imaging, Phanta9, Düsseldorf, Germany, September 2006

13. *C. **Gendrin**, Y. Roggo, C. Collet,* Analyzing Pharmaceutical Solid Dosage Forms by Chemical Imaging and Chemometrics, Chimiométrie 2006, Paris, France, December 2006

14. *C. **Gendrin**, Y. Roggo, A. Fischer,* Comparison of methods to quantify hyperspectral images, Chimiométrie 2005, Lille, December 2005

Table of illustrations

Figures

Tables

List of References

[1] Guidance for Industry, PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance, 2007, http://www.fda.gov/cder/guidance/6419fnl.pdf

[2] G. Lachenal, Introduction à la spectroscopie infrarouge **in** *La spectroscopie infrarouge et ses applications analytiques*, D. Bertrand and E. Dufour, Tech & doc, 2nd edition, Paris, p. **32-75**, 2006

[3] M. Dalibart and S. Servant, Spectroscopie dans l'infrarouge **in** *Technique de l'ingénieur*, p. **2845**, 2000

[4] R. W. Hannah, Standard Sampling Techniques for Infrared Spectroscopy **in** *Handbook of vibrational spectroscopy*, J. M. Chalmers and P. R. Griffiths, Volume 2, John Wiley & Sons, London, p. **933-952**, 2002

[5] M. A. Harthcock and S. C. Atkin, Imaging with Functional Group Maps Using Infrared Microspectroscopy, *Applied Spectroscopy*, **42**, p. **449-455**, 1988

[6] Z. Liu, H. Yu and J. F. MacGregor, Standardization of line-scan NIR imaging systems, *Journal of Chemometrics*, **21**, p. **88-95**, 2007

[7] C. D. Tran, Infrared Multispectral Imaging: Principles and Instrumentation, *Applied Spectroscopy Reviews*, **38**, p. **133-153**, 2003

[8] P. Treado, I. W. Levin and E. N. Lewis, Indium Antimonide (InSb) Focal Plane Array (FPA) Detection for Near-Infrared Imaging Microscopy, *Applied Spectroscopy*, **48**, p. **607-615**, 1994

[9] P. J. Treado and M. P. Nelson, Raman imaging **in** *Handbook of vibrational spectroscopy*, J. M. Chalmers and P. R. Griffiths, Volume 2, John Wiley & Sons, London, p. **1429-1459**, 2002

[10] S. V. Hammond and F. C. Clarke, Near-infrared Microspectroscopy **in** *Handbook of vibrational spectroscopy*, J. M. Chalmers and P. R. Griffiths, Volume 2, John Wiley & Sons, London, p. **1405-1418**, 2002

[11] C. M. Snively and J. L. Koenig, Characterizing the Performance of a Fast FT-IR Imaging Spectrometer, *Applied Spectroscopy*, **53**, p. **170-177**, 1999

[12] R. Bhargava, B. G. Wall and J. L. Koenig, Comparison of the FT-IR Mapping and Imaging Techniques Applied to Polymeric Systems, *Applied Spectroscopy*, **54**, p. **470-479**, 2000

[13] P. Geladi, J. Burger and T. Lestander, Hyperspectral imaging: calibration problems and solutions, *Chemometrics and Intelligent Laboratory Systems*, **72**, p. **209-217**, 2004

[14] D. Bertrand and V. Baeten, Instrumentation **in** *La spectroscopie infrarouge et ses applications analytiques*, D. Bertrand and E. Dufour, Tech & doc, 2nd edition, Paris, p. **247-301**, 2006

[15] D. E. Pivonka, J. M. Chalmers and P. R. Griffiths, *Applications of vibrational spectroscopy in pharmaceutical research and development*, John Wiley & Sons, Chichester, 2007

[16] E. N. Lewis, Near-infrared Chemical Imaging as Process Analytical Tool **in** *Process Analytical Technology*, K. A. Bakeev, Blackwell Publishing, 1st, Oxford, p. **187-225**, 2005

[17] N. Gat, Imaging spectroscopy using tunable filters: a review, *Proceedings SPIE*, **4056**, p. **50-64**, 2000

[18] H. R. Morris, C. C. Hoyt, P. Miller and P. J. Treado, Liquid Crystal Tunable Filter Raman Chemical Imaging, *Applied Spectroscopy*, **50**, p. **805-811**, 1996

[19] D. L. Wetzel, A. J. Eilert and J. A. Sweat, Tunable Filter and Discrete Filter Near-infrared Spectrometer **in** *Handbook of vibrational spectroscopy*, J. M. Chalmers and P. R. Griffiths, Volume 1, John Wiley & Sons, London, p. **436-452**, 2002

[20] S. J. Hudak, K. Haber, G. Sando, L. H. Kidder and E. N. Lewis, Pratical limits of spatial resolution in diffuse reflectance NIR chemical imaging, *NIR news*, **18**, p. **6-8**, 2007

[21] F. C. Clarke, S. V. Hammond, R. D. Jee and A. C. Moffat, Determination of the Information Depth and Sample Size for the Analysis of Pharmaceutical Materials Using Reflectance Near-Infrared Microscopy, *Applied Spectroscopy*, **56**, p. **1475-1483**, 2002

[22] J. Burger and P. Geladi, Hyperspectral NIR image regression part I: calibration and correction, *Journal of Chemometrics*, **19**, p. **355-363**, 2005

[23] D. Clark, M. Henson, F. Laplant and S. Sasic, Pharmaceutical Applications of Chemical Mapping and Imaging **in** *Applications of Vibrational Spectroscopy in Pharmaceutical Research and Development*, D. E. Pivonka, J. M. Chalmers and P. R. Griffiths, John Wiley & Sons, London, p. **309-335**, 2007

[24] D. Bertrand and E. Vigneau, Prétraitement des données spectrales **in** *La spectroscopie infrarouge et ses applications analytiques*, D. Bertrand and E. Dufour, Tech & doc, 2nd edition, Paris, p. **427-447**, 2006

[25] R. J. Barnes, M. S. Dhanoa and S. J. Lister, Standard Normal Variate Transformation and De-trending of Near Infrared Diffuse Reflectance Spectra, *Applied spectroscopy*, **43**, p. **772-777**, 1989

[26] P. Geladi, D. MacDougall and H. Martens, Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat, *Applied Spectroscopy*, **39**, p. **491-500**, 1985

[27] A. Savitzky and M. J. E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures, *Analytical Chemistry*, **36**, p. **1627-1639**, 1964

[28] D. Bertrand and E. Vigneau, Prétraitement des données spectrales **in** *La spectroscopie infrarouge et ses applications analytiques*, E. Dufour, Tech & doc, 2nd edition, Paris, p. **248**, 2006

[29] J. Burger and P. Geladi, Spectral pre-treatments of hyperspectral near infrared images: Analysis of diffuse reflectance scattering, *Journal of Near Infrared Spectroscopy*, **15**, p. **29-37**, 2007

[30] F. Clarke, Extracting process-related information from pharmaceutical dosage forms using near infrared microscopy, *Vibrational Spectroscopy*, **34**, p. **25-35**, 2004

[31] H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, Chichester, 1991

[32] P. Paatero and U. Tapper, Analysis of different modes of factor analysis as least squares fit problems, *Chemometrics and Intelligent Laboratory Systems*, **18**, p. **183-194**, 1993

[33] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, Wiley-Interscience, Chichester, 2004

[34] S. Wold, M. Sjostrom and L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **58**, p. **109-130**, 2001

[35] H. Martens and T. Naes, Methods for calibration **in** *Multivariate Calibration*, John Wiley & Sons, Chichester, p. **73**, 1991

[36] J. Burger and P. Geladi, Hyperspectral NIR image regression part II: dataset preprocessing diagnostics, *Journal of Chemometrics*, **20**, p. **106-119**, 2006

[37] J. Burger and P. Geladi, Hyperspectral NIR imaging for calibration and prediction: a comparison between image and spectrometer data for studying organic and biological samples, *The Analyst*, **131**, p. **1152-1160**, 2006

[38] L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, Classification and discrimination **in** *Multi- and Megavariate Data analysis, Principles and Application*, Umetrics Academy, Umea, p. **193**, 2001

[39] N. Jovanovic, A. Gerich, A. Bouchard and W. Jiskoot, Near-Infrared Imaging for Studying Homogeneity of Protein-Sugar Mixtures, *Pharmaceutical Research*, **23**, p. **2002-2013**, 2006

[40] B. J. Westenberger, C. D. Ellison, A. S. Fussner, S. Jenney, R. E. Kolinski, T. G. Lipe, R. C. Lyon, T. W. Moore, L. K. Revelle and A. P. Smith, Quality assessment of internet pharmaceutical products using traditional and non-traditional analytical techniques, *International Journal of Pharmaceutics*, **306**, p. **56-70**, 2005

[41] S. Šašic, D. A. Clark, J. C. Mitchell and M. J. Snowden, A comparison of Raman chemical images produced by univariate and multivariate data processing - A simulation with an example from pharmaceutical practice, *The Analyst*, **129**, p. **1001-1007**, 2004

[42] R. Tauler, A. Smilde and B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, *Journal of Chemometrics*, **9**, p. **31-58**, 1995

[43] R. Tauler, Multivariate curve resolution applied to second order data, *Chemometrics and Intelligent Laboratory Systems*, **30**, p. **133-146**, 1995

[44] W. H. Lawton and E. A. Sylvestre, Self Modeling Curve Resolution, *Technometrics*, **13**, p. **617-633**, 1971

[45] E. R. Malinowski, Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra, *Analytica Chimica Acta*, **134**, p. **129-137**, 1982

[46] B. G. M. Vandeginste, W. Derks and G. Kateman, Multicomponent self-modelling curve resolution in high-performance liquid chromatography by iterative target transformation analysis, *Analytica Chimica Acta*, **173**, p. **253-264**, 1985

[47] E. Widjaja and M. Garland, Pure component spectral reconstruction from mixture data using SVD, global entropy minimization, and simulated annealing. Numerical investigations of admissible objective functions using a synthetic 7-species data set, *Journal of Computational Chemistry*, **23**, p. **911-919**, 2002

[48] E. Widjaja, R. Kim and H. Seah, Application of Raman Microscopy and Band-Target Entropy Minimization to Identify Minor Components in Model Pharmaceutical Tablets, *Journal of Pharmaceutical and Biomedical Analysis*, **46**, p. **274-281**, 2008

[49] W. Windig and J. Guilment, Interactive self-modeling mixture analysis, *Analytical Chermistry*, **63**, p. **1425-1432**, 1991

[50] F. C. Sanchez, J. Toft, B. Van den Bogaert and D. L. Massart, Orthogonal Projection Approach Applied to Peak Purity Assessment, *Anal. Chem.*, **68**, p. **79-85**, 1996

[51] F. Cuesta Sánchez, B. van den Bogaert, S. C. Rutan and D. L. Massart, Multivariate peak purity approaches, *Chemometrics and Intelligent Laboratory Systems*, **34**, p. **139-171**, 1996

[52] A. Hyvärinen, J. Karhunen and E. Oja, *Independant Component Analysis*, Wiley & Sons, Chichester, 2001

[53] S. Moussaoui, C. Carteret, D. Brie and A. Mohammad-Djafari, Bayesian analysis of spectral mixture data using Markov Chain Monte Carlo Methods, *Chemometrics and Intelligent Laboratory Systems*, **81**, p. **137-148**, 2006

[54] C. Gobinet, E. Perrin and R. Huez, *Application of non-negative matrix factorization to fluorescence spectroscopy*, European Signal Processing Conference, Vienna, Austria, 2004

[55] R. Bro and S. De Jong, A fast non-negativity-constrained least squares algorithm, *Journal of Chemometrics*, **11**, p. **393-401**, 1997

[56] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, **401**, p. **788-791**, 1999

[57] P. Sajda, S. Du and L. Parra, *Recovery of constituent spectra using non-negative matrix factorization*, SPIE Wavelets x, San Diego, 2003

[58] P. Paatero, Least squares formulation of robust non-negative factor analysis, *Chemometrics and Intelligent Laboratory Systems*, **37**, p. **23-35**, 1997

[59] P. Paatero, The Multilinear Engine - a Table-driven Least Squares Program for Solving Multilinear Problems, Including the n-way Parallel Factor Analysis Model, *Journal of Computational and Graphical Statistics*, **8**, p. **854-888**, 1999

[60] P. Paatero, P. K. Hopke, X.-H. Song and Z. Ramadan, Understanding and controlling rotations in factor analytic models, *Chemometrics and Intelligent Laboratory Systems*, **60**, p. **253-264**, 2002

[61] P. Paatero and P. K. Hopke, Rotational tools for factor analytic models implemented by using the Multilinear Engine, *Submitted to Journal of Chemometrics*, p. 2008

[62] S. Moussaoui, D. Brie, A. Mohammad-Djafari and C. Carteret, Separation of Non-negative mixture of non-negative sources using a Baysian Approach and MCMC sampling, *IEEE transactions on signal processing*, **54**, p. **4133-4145**, 2006

[63] L. Duponchel, W. Elmi-Rayaleh, C. Ruckebusch and J. P. Huvenne, Multivariate Curve Resolution Methods in Imaging Spectroscopy: Influence of Extraction Methods and Instrumental Perturbations, *Journal of Chemical Information and Computer Sciences (now called Journal of Chemical Information and Modeling)*, **43**, p. **2057-2067**, 2003

[64] J. J. Andrew and H. T. M., Rapid Analysis of Raman Image Data Using Two-Way Multivariate Curve Resolution, *Applied Spectroscopy*, **52**, p. **797-807**, 1998

[65] M. Maeder, Evolving factor analysis for the resolution of overlapping chromatographic peaks, *Analytical Chemistry*, **59**, p. **527-530**, 1987

[66] A. De Juan, M. Maeder, T. Hancewicz and R. Tauler, Local rank analysis for exploratory spectroscopic image analysis. Fixed Size Image Window-Evolving Factor Analysis, *Chemometrics and Intelligent Laboratory Systems*, **77**, p. **64-74**, 2005

[67] A. De Juan and R. Tauler, Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis, *Trends in Analytical Chemistry - TrAC*, **23**, p. **70-79**, 2004

[68] J.-H. Wang, P. K. Hopke, T. M. Hancewicz and S. L. Zhang, Application of modified alternating least squares regression to spectroscopic image analysis, *Analytica Chimica Acta*, **476**, p. **93-109**, 2003

[69] T. Hancewicz and J.-H. Wang, Discriminant image resolution: a novel multivariate image analysis method utilizing a spatial classification constraint in addition to bilinear nonnegativity, *Chemometrics and Intelligent Laboratory Systems*, **77**, p. **18-31**, 2005

[70] S. P. Gurden, E. M. Lage, C. G. De Faria, I. Joekes and M. M. C. Ferreira, Analysis of video images from a gas-liquid transfer experiment: a comparison of PCA and PARAFAC for multivariate image analysis, *Journal of chemometrics*, **17**, p. **400-412**, 2003

[71] R. Bro, PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems*, **38**, p. **149-171**, 1997

[72] R. Bro, *Multi-way analysis in the Food Industry*, University of Copenhagen, Denmark, 1998

[73] J. Huang, H. Wium, K. B. Qvist and K. H. Esbensen, Multi-way methods in image analysis--relationships and applications, *Chemometrics and Intelligent Laboratory Systems*, **66**, p. **141-158**, 2003

[74] P. Geladi and G. H., *Multivariate Image Analysis*, John Wiley & Sons, Chichester, 1996

[75] M. H. Bharati and J. F. MacGregor, Multivariate Image Analysis for Real-Time Process Monitoring and Control, *Industrial & Engineering Chemistry Research*, **37**, p. **4715-4724**, 1998

[76] H. Yu, J. F. MacGregor, G. Haarsma and W. Bourg, Digital imaging for online monitoring and control of industrial snack food processes, *Industrial and Engineering Chemistry Research*, **42**, p. **3036-3044**, 2003

[77] W. H. A. M. Van Den Broek, D. Wienke, W. J. Melssen and L. M. C. Buydens, Plastic material identification with spectroscopic near infrared imaging and artificial neural networks, *Analytica Chimica Acta*, **361**, p. **161-176**, 1998

[78] J. A. Fernandez Pierna, V. Baeten, A. Michotte Renier, R. P. Cogdill and P. Dardenne, Combination of support vector machines (SVM) and near-infrared (NIR) imaging spectroscopy for detection of meat and bone meal (MBM) in compound feeds, *Journal of Chemometrics*, **18**, p. **341-349**, 2004

[79] J. A. Fernandez Pierna, V. Baeten and P. Dardenne, Screening of compound feeds using NIR hyperspectral data, *Chemometrics and Intelligent Laboratory Systems*, **84**, p. **114-118**, 2006

[80] J. J. Liu, M. H. Bharati, K. G. Dunn and J. F. MacGregor, Automatic masking in multivariate image analysis using support vector machines, *Chemometrics and Intelligent Laboratory Systems*, **79**, p. **42-54**, 2005

[81] A. K. Jain, R. P. W. Duin and J. Mao, Statistical pattern recognition: a review, *IEEE transaction on pattern analysis and machine intelligence*, **22**, p. **4-37**, 2000

[82] R. Bhargava, S.-Q. Wang and J. L. Koenig, Processing FT-IR imaging data for morphology visualization, *Applied spectroscopy*, **54**, p. **1690-1706**, 2000

[83] N. Otsu, A threshold selection method from gray-level histograms, *IEEE transactions on systems, man, and cybernetics*, **9**, p. **62-66**, 1979

[84] W. Li, A. Woldu, R. Kelly, J. McCool, R. Bruce, H. Rasmussen, J. Cunningham and D. Winstead, Measurement of drug agglomerates in powder blending simulation samples by near infrared chemical imaging, *International Journal of Pharmaceutics*, **350**, p. **369-373**, 2008

[85] F. Laplant, Factors Affecting NIR chemical Images of Solid dosage Forms, *American Pharmaceutical Review*, **7**, p. **16-24**, 2004

[86] H. Ma and C. A. Anderson, Optimisation of magnification levels for near infrared chemical imaging of blending of pharmaceutical powders, *Journal of Near Infrared Spectroscopy*, **15**, p. **137-151**, 2007

[87] E. N. Lewis, J. Schoppelrei and E. Lee, Near-infrared Chemical Imaging and the PAT initiative, *Spectroscopy*, **19**, p. **26-36**, 2004

[88] F. W. Koehler, e. Lee, l. H. Kidder and E. N. Lewis, Near infrared spectroscopy: the practical imaging solution, *Spectroscopy Europe*, **14**, p. **12-19**, 2002

[89] E. N. Lewis, J. E. Carrol and F. C. Clarke, A near infrared view of pharmaceutical formulation analysis, *NIR news*, **12**, p. **16-18**, 2001

[90] S. Šašic, An In-Depth Analysis of Raman and Near-Infrared Chemical Images of Common Pharmaceutical Tablets, *Applied Spectroscopy*, **61**, p. **239-250**, 2007

[91] S. Šašic, Chemical imaging of pharmaceutical granules by Raman global illumination and near-infrared mapping platforms, *Analytica Chimica Acta*, **611**, p. **73-79**, 2008

[92] A. El-Hagrasy, H. R. Morris, F. D'Amico, R. A. Lodder and J. K. Drennen, Near-infrared Spectroscopy and imaging for the monitoring of powder blend homogeneity, *Journal of Pharmaceutical Sciences*, **90**, p. **1298-1307**, 2001

[93] R. C. Lyon, D. S. Lester, E. N. Lewis, E. Lee, L. X. Yu, E. H. Jefferson and A. S. Hussain, Near-Infrared Spectral Imaging for Quality Assurance of Pharmaceutical Products: Analysis of Tablets to Assess Powder Blend Homogeneity, *AAPS PharmSciTech*, **3**, p. **1-17**, 2002

[94] E. N. Lewis, l. H. Kidder and E. Lee, NIR chemical imaging as a process analytical tool, *Innovations in pharmaceutical technology*, p. **107-111**, 2005

[95] Chemical imaging investigated for process monitoring, 2005, www.in-pharmatechnologist.com

[96] O. Svensson, K. Abrahamsson, J. Engelbrektsson, M. Nicholas, H. Wikstrom and M. Josefson, An evaluation of 2D-wavelet filters for estimation of differences in textures of pharmaceutical tablets, *Chemometrics and Intelligent Laboratory Systems*, **84**, p. **3-8**, 2006

[97] E. Lee, W. X. Huang, P. Chen, E. N. Lewis and R. V. Vivilecchia, High-throughput analysis of pharmaceutical tablet content uniformity ny near-infrared chemical imaging, *Spectroscopy*, p. 2006

[98] L. Hilden, C. J. Pommier, S. Badawy and E. M. Friedman, NIR Chemical Imaging to Guide/Support BMS-561389 Tablet Formulation Development, *International Journal Of Pharmaceutics*, **353**, p. **283-290**, 2008

[99] R. B. Shah, M. A. Tawakkul and M. A. Khan, Process analytical technology: Chemometric analysis of Raman and near infra-red spectroscopic data for predicting physical properties of extended release matrix tablets, *Journal of Pharmaceutical Sciences*, **96**, p. **1356-1365**, 2007

[100] J. Dubois, J.-C. Wolff, J. K. Warrack, J. Schoppelrei and E. N. Lewis, NIR chemical imaging for counterfeit pharmaceutical products analysis, *Spectroscopy*, p. 2007

[101] O. Y. Rodionova, L. P. Houmoller, A. L. Pomerantsev, P. Geladi, J. Burger, V. L. Dorofeyev and A. P. Arzamastsev, NIR spectrometry for counterfeit drug detection: A feasibility study, *Analytica Chimica Acta*, **549**, p. **151-158**, 2005

[102] M. A. Veronin, E. Lee and E. N. Lewis, "Insight" into Drug Quality: Comparison of Simvastatin Tablets from the US and Canada Obtained via the Internet, *The Annals of Pharmacotherapy*, **41**, p. **1111-1115**, 2007

[103] M. A. Veronin and B.-B. C. Youan, MEDICINE: Enhanced: Magic Bullet Gone Astray: Medications and the Internet, *Science*, **305**, p. **481**, 2004

[104] I. Malik, M. Poonacha, J. Moses and R. A. Lodder, Multispectral imaging of tablets in blister packaging, *AAPS PharmSciTech*, **2**, p. **1-7**, 2001

[105] S. J. Hamilton, A. E. Lowell and R. A. Lodder, Hyperspectral techniques in analysis of oral dosage froms, *Journal of Biomedical Optics*, **7**, p. **561-570**, 2002

[106] J. Breitenbach, Melt extrusion: from process to drug delivery technology, *European Journal of Pharmaceutics and Biopharmaceutics*, **54**, p. **107-117**, 2002

[107] D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Michotte and L. Kaufman, Evaluation of Sources of Variation in data. Analysis of Variance **in** *Chemometrics: a textbook*, Elsevier Science, Amsterdam, p. **59**, 1988

[108] W. Doub, W. Adams, J. Spencer, L. Buhse, M. Nelson and P. Treado, Raman Chemical Imaging for Ingredient-specific Particle Size Characterization of Aqueous Suspension Nasal Spray Formulations: A Progress Report, *Pharmaceutical Research*, **24**, p. **934-945**, 2007

[109] P. Paatero, User's guide for the Multilinear Engine program "ME2" for fitting multilinear and quasi-multilinear models, 2000

[110] P. Paatero and P. K. Hopke, Discarding or downweighting high-noise variables in factor analytic models, *Analytica Chimica Acta*, **490**, p. **277-289**, 2003

[111] H. Martens and T. Naes, Assessment, Validation and Choice of Calibration Method **in** *Multivariate Calibration*, John Wiley & Sons, Chichester, p. 1991

[112] C. Gendrin, Y. Roggo and C. Collet, Content uniformity of pharmaceutical solid dosage forms by near Infrared hyperspectral imaging: a feasibility study, *Talanta*, **73**, p. **733-741**, 2007

[113] P. Geladi, Are pixels sample cells? Hyperspectral diffuse near infrared imaging experiments with pinholes, *Journal of Near Infrared Spectroscopy*, **16**, p. **357-363**, 2008