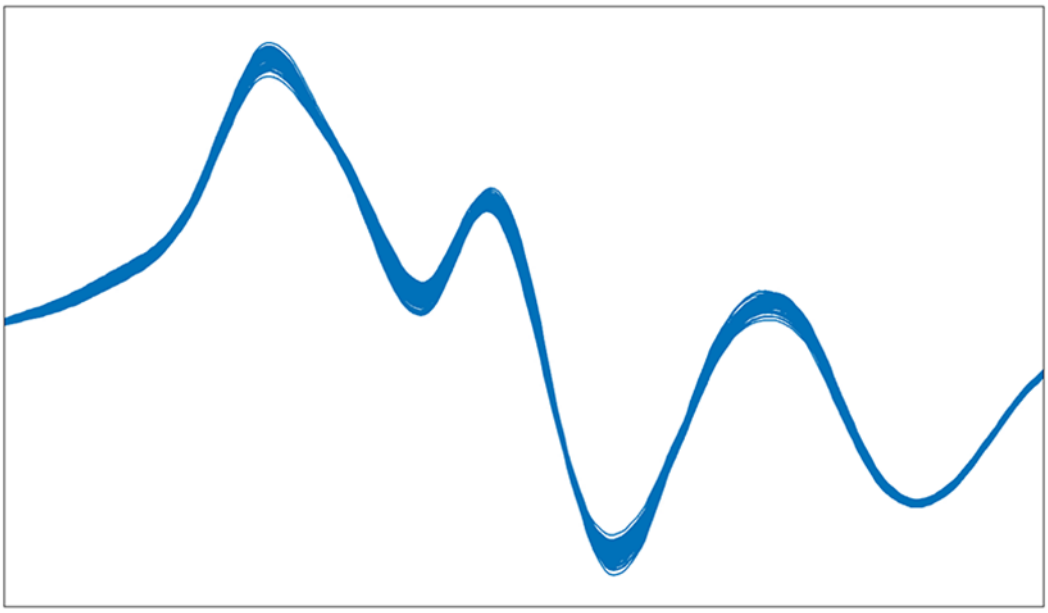
Théorie OMNIS NIR



Manuel d'utilisation

8.0600.8101FR / v9 / 2025-10-10

Metrohm

Metrohm AG
Ionenstrasse
CH-9100 Herisau
Suisse
+41 71 353 85 85
info@metrohm.com
www.metrohm.com

Théorie OMNIS NIR

Manuel d'utilisation

8.0600.8101FR / v9 /
2025-10-10

Technical Communication
Metrohm AG
CH-9100 Herisau

La présente documentation est protégée par les droits d'auteur. Tous droits réservés.

La présente documentation est un document original.

La présente documentation a été élaborée avec le plus grand soin. Cependant, des erreurs ne peuvent être totalement exclues. Veuillez communiquer vos remarques à ce sujet directement à l'adresse citée ci-dessus.

Exclusion de responsabilité

Les défauts résultant de circonstances dont Metrohm n'est pas responsable, par exemple, stockage inapproprié, utilisation non conforme etc., sont expressément exclus de la garantie. Les modifications non autorisées du produit (par exemple, transformations ou ajouts) excluent toute responsabilité du fabricant pour les dommages qui en résultent et leurs conséquences. La documentation du produit Metrohm fournit des instructions et des remarques à respecter strictement. Dans le cas contraire, la responsabilité de Metrohm est exclue.

Table des matières

1 Aperçu 1

1.1 Introduction ... 1

1.2 Cadre conceptuel ... 1

1.3 Informations concernant la documentation ... 2

1.4 Informations complémentaires ... 2

2 Lumière proche infrarouge et spectres 3

2.1 La lumière et son interaction avec les matières ... 3

2.2 Fondements mathématiques ... 6

2.2.1 Loi de Beer-Lambert ... 6

2.2.2 Régression linéaire ... 7

2.3 Transformation de la lumière en spectre ... 9

3 Configuration de l'appareil 13

3.1 Calibrage de la longueur d'onde ... 14

3.2 Standardisation de référence ... 15

3.2.1 OMNIS NIR Analyzer ... 16

3.2.2 2060 The NIR ... 17

3.3 Tests de performance d'appareils ... 26

3.3.1 Externe Tests de performance d'appareils (OMNIS NIR Analyzer) ... 29

4 Développement de modèles 31

4.1 Échantillons physiques ... 33

4.2 Analyse en composantes principales (ACP) ... 36

4.3 Préparation des données ... 40

4.3.1 Prétraitement des données ... 40

4.3.2 Gammes de longueurs d'onde ... 49

4.3.3 Valeur spectrale atypique ... 51

4.3.4 Valeur de référence atypique (quantification) ... 58

4.3.5 Répartition des ensembles de données ... 59

4.4 Quantification ... 61

4.4.1 Régression MCP ... 61

4.4.2 Validation des modèles de quantification ... 64

4.4.3 OMNIS Model Developer (OMD) ... 72

4.4.4 Correction de pente/d'ordonnée à l'origine ... 73

4.5 Identification et vérification ... 76

4.5.1 Machine à vecteurs de support (SVM) ... 76

4.5.2 Prédiction de l'appartenance d'un échantillon à un produit ... 79

4.5.3 Validation des modèles d'identification 81

4.6 Qualification 83
4.6.1 Calcul de modèles de qualification 83
4.6.2 Validation de modèles de qualification 83

5 Prédiction 85

5.1 Quantification 85
5.1.1 Valeurs atypiques et contrôle du résultat 85

5.2 Identification et vérification 87

5.3 Qualification 88

6 Annexe 89

6.1 Exemple d'une régression linéaire 89

6.2 Algorithme ACP 93

6.3 Algorithme MCP 95

6.4 T^2 de Hotelling et résidus Q 96

6.5 Valeur spectrale atypique - algorithme 98

6.6 Valeur de référence atypique - algorithme 100

1 Aperçu

1.1 Introduction

La spectroscopie proche infrarouge (spectroscopie NIR) est une méthode d'analyse non destructive, rapide et sans réactifs qui convient à un large spectre d'échantillons. Elle peut analyser plusieurs paramètres simultanément et déterminer les propriétés tant physiques que chimiques d'un matériau. Il s'agit entre autres des concentrations d'analyte, de la densité, de la taille des particules ou de la viscosité intrinsèque.

La spectroscopie NIR permet en outre d'identifier des échantillons inconnus (à partir du logiciel OMNIS 4.0) et de vérifier des échantillons (à partir du logiciel OMNIS 4.2).

La possibilité de mesurer des échantillons à distance et de manière non destructive est d'une importance capitale dans le contrôle de la qualité et la surveillance des processus.

Ce mode d'emploi décrit les techniques et algorithmes d'acquisition, de traitement et d'analyse de spectres en proche infrarouge et, leur implémentation dans le logiciel OMNIS. Le chapitre 2 explique brièvement comment les signaux de mesure sont convertis en spectres d'absorption. Le chapitre 3 traite du calibrage de l'appareil. Le chapitre 4 décrit le développement de modèles capables de prédire le paramètre d'intérêt (quantification) ou l'association de produits (identification). Le chapitre 5 est consacré à la prédiction des inconnus. Le chapitre 6 regroupe les annexes avec des explications sur différents algorithmes.

1.2 Cadre conceptuel

Les procédés présentés s'intègrent dans le cadre suivant :

1. **Calibrage, standardisation et tests de performance**
   Garantie de portabilité et de fiabilité des spectres d’absorption enregistrés par l’appareil.
2. **Développement de modèles**
   Processus d'élaboration d'un modèle pour prédire un paramètre quantitatif ou identifier des échantillons.
   Le développement est basé sur des échantillons dont les paramètres d'intérêt sont connus ou dont on sait qu'ils sont présents dans le produit.

3. **Analyse d'échantillon**
   Un spectre est enregistré à partir de l'échantillon analysé. Sur les fondements du spectre, un modèle de quantification fournit une prédiction quantitative ou un modèle d'identification identifie ou vérifie l'échantillon.
4. **Contrôle**
   Le contrôle du modèle et de l'appareil confirme que le système convient à une analyse plus approfondie.

1.3 Informations concernant la documentation

Conventions de représentation

Les lettres majuscules en gras représentent les matrices, les lettres minuscules en gras les vecteurs. La transposée d'une matrice ou d'un vecteur est indiquée par un t en exposant, par exemple $\mathbf{x}^{\mathbf{t}}$.

Les scalaires sont représentés par des lettres minuscules. L'accent circonflexe (^) représente les valeurs (calculées) estimées, par exemple $\hat{y}$. Un trait horizontal représente les valeurs moyennes, par exemple $\bar{y}$.

Un scalaire peut représenter une variable dépendant de la longueur d'onde, par exemple l'absorbance A. La notation scalaire peut être utilisée de manière interchangeable avec la notation vectorielle.

1.4 Informations complémentaires

Les pages suivantes contiennent des informations supplémentaires sur le produit :

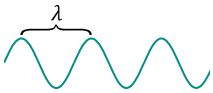
- Site Internet Metrohm *https://www.metrohm.com* – Aperçu de la famille de produits, des documents PDF, des mentions des accessoires et des informations sur les applications.
- Aide du logiciel OMNIS *https://guide.metrohm.com* – Informations filtrées par thème sur le logiciel OMNIS.

2 Lumière proche infrarouge et spectres

2.1 La lumière et son interaction avec les matières

Un spectromètre mesure l'interaction d'un échantillon avec la lumière. La lumière peut être absorbée ou diffusée à des degrés divers. L'interaction dépend des propriétés de la lumière, notamment sa longueur d'onde, et des propriétés du matériau, notamment sa structure moléculaire.

Longueur d'onde



La lumière est un rayonnement électromagnétique. La lumière se déplace dans l'espace sous forme d'onde caractérisée par deux champs oscillants, l'un électrique et l'autre magnétique. Les ondes se propagent dans l'espace et dans le temps. Une onde est donc caractérisée par sa longueur d'onde λ (p. ex. en nanomètres = 10^{-9} mètres) et sa fréquence (Hz).

La longueur d'onde est inversement proportionnelle à la fréquence de l'onde. Des ondes dont la fréquence est élevée (plus d'oscillations par seconde) ont des longueurs d'onde plus courtes et inversement. Du fait de cette relation, l'onde peut être décrite soit par la longueur (nm), soit par la fréquence (Hz).

La lumière peut échanger de l'énergie en unités quantiques discrètes, les photons. L'énergie d'un photon donné, *E*, dépend de sa fréquence *f* ou de sa longueur d'onde λ :

$$E = hf = h\frac{c}{\lambda}$$

Ici, *h* correspond à la constante de Planck et *c* à la vitesse de la lumière.

La *figure 1* montre différentes gammes de rayonnement électromagnétique.

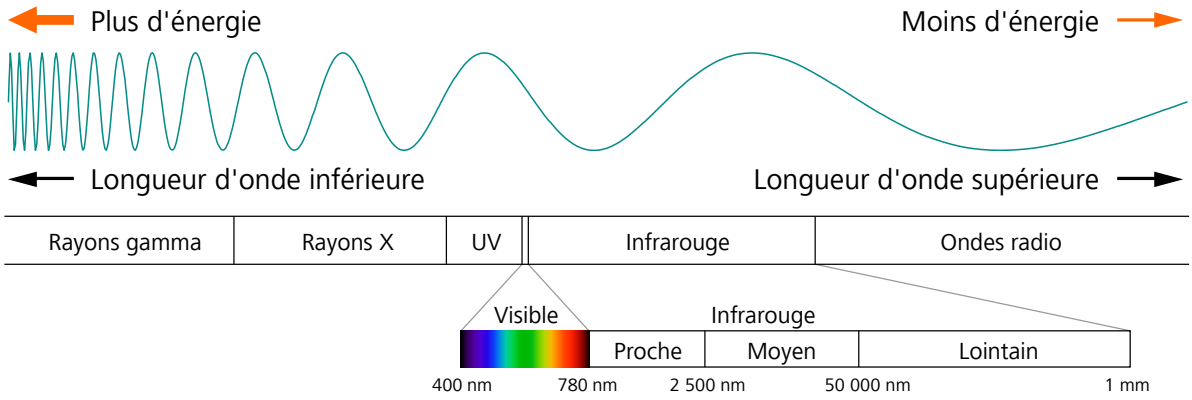


Figure 1 Gammes de rayonnement électromagnétique. Le domaine proche infrarouge (NIR) se situe à côté de la gamme de la lumière visible. Le proche infrarouge comprend des longueurs d'onde de 780 nm à 2 500 nm.

Sources de rayonnement

Les différentes sources de rayonnement émettent dans différentes gammes de rayonnement. Le **rayonnement thermique** est la source de la gamme NIR. Une caméra infrarouge détecte, par ex., le corps humain, car sa température est plus élevée que celle de son environnement.

Intensité

L'amplitude de l'onde électromagnétique détermine l'intensité de la lumière. Plus l'amplitude est élevée, plus l'intensité est grande. Dans le cas de la lumière visible, l'intensité est perçue comme une luminosité.

Interaction de la lumière et de la matière

Le processus d'absorption de la lumière par les molécules est présenté ci-dessous.

La manière dont la lumière interagit avec la matière dépend de la gamme du rayonnement électromagnétique. Par ex., lorsque de l'énergie de la lumière visible est transférée à des molécules, certains de leurs électrons peuvent passer d'un niveau d'énergie inférieur à un niveau d'énergie supérieur (transition électronique). Des liaisons intramoléculaires sont le siège de **transitions vibratoires** dans le proche infrarouge. Les liaisons chimiques, groupes fonctionnels, et molécules peuvent vibrer de différentes manières, par ex. vibration d'étirement, vibration de déformation ou vibration de torsion.

Les molécules ne peuvent prendre que des états vibratoires discrets. À l'ambiante, la plupart des molécules sont à l'état vibratoire fondamental (niveau 0). Les transitions entre l'état vibratoire fondamental et les états excités sont qualifiées selon le schéma suivant :

Transition vibratoire *i* ➔ *j*	Nom
0 ➔ 1	Transition fondamentale
0 ➔ 2	Première transition harmonique
0 ➔ 3	Deuxième transition harmonique

Un état vibratoire plus élevé correspond à un niveau d'énergie plus élevé. Pour passer de l'état *i* à l'état excité *j*, la molécule doit absorber une certaine énergie de transition ΔE_{ij}.

La lumière peut échanger de l'énergie par quantités élémentaires ou quanta $E = hf$, où f est la fréquence de la lumière. Il y a absorption de lumière si l'énergie du photon hf est égale à l'énergie de transition ΔE_{ij}.

Les états vibratoires permis dépendent notamment de la force des liaisons et de la masse des atomes impliqués. Par conséquent, un certain type de liaison peut être associé à des énergies de transition c.-à-d. à des longueurs d'onde absorbées caractéristiques.

D'autres conditions doivent être satisfaites pour que la lumière puisse être absorbée. La transition vibratoire doit déplacer la répartition des charges de manière à modifier le moment dipolaire électrique de la molécule. La probabilité que l'énergie soit absorbée dépend de l'ordre de grandeur de la modification du moment dipolaire le long de la liaison chimique concernée.

Une transition vibratoire peut provoquer une modification du moment dipolaire aussi bien pour les molécules polaires que non polaires et les groupes fonctionnels. Les molécules diatomiques homonucléaires telles que N_2 n'absorbent aucune lumière infrarouge.

La durée de l'état vibratoire excité est limitée. Si la molécule revient à un état vibratoire plus bas, l'énergie est transformée en chaleur.

La gamme spectrale NIR

Les longueurs d'onde correspondant aux transitions fondamentales se situent dans l'infrarouge moyen. Le domaine proche infrarouge englobe les transitions harmoniques et les bandes de combinaison. La *figure 2* présente les bandes de longueurs d'onde absorbées par de nombreuses molécules et de nombreux groupes fonctionnels.

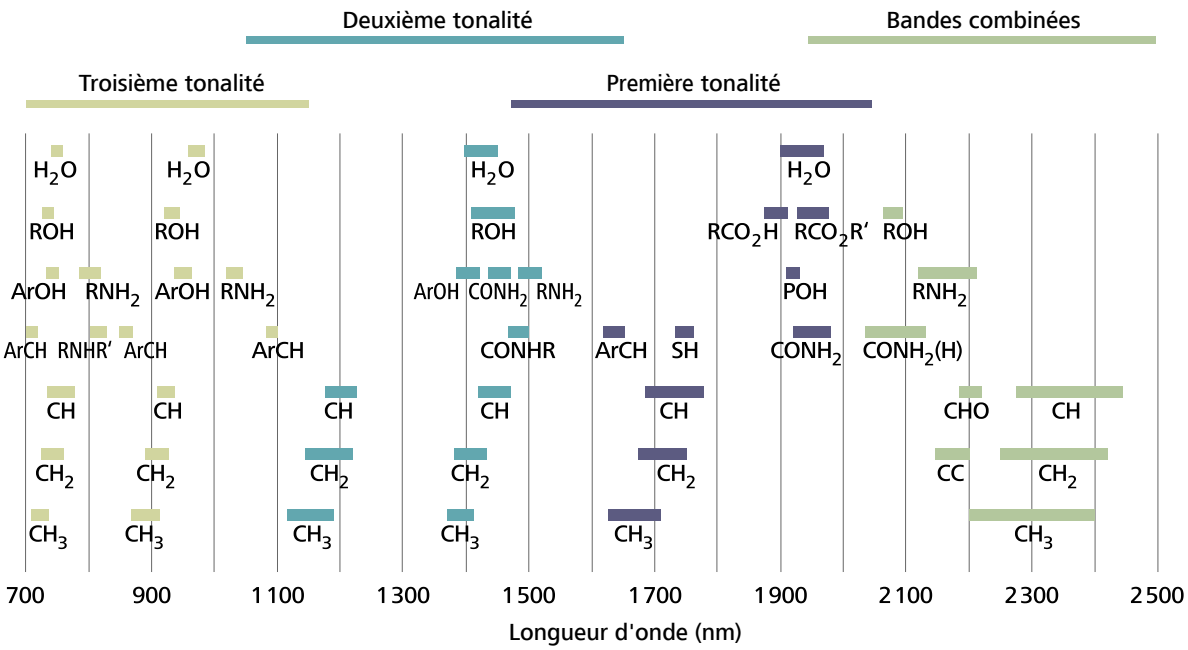


Figure 2 Bandes d'absorption NIR

La transition fondamentale est la transition la plus probable. C'est celle qui se produit le plus souvent. Les transitions harmoniques sont moins probables. Par conséquent, la transition fondamentale absorbe plus de lumière que les transitions harmoniques. En général, l'absorption diminue à chaque harmonique. Les harmoniques conviennent ainsi aux molécules fortement absorbantes.

Il est possible d'exciter simultanément deux vibrations fondamentales ou plus avec une seule fréquence de la lumière qui correspond aux fréquences combinées des vibrations fondamentales. Les bandes d'absorption correspondantes sont appelées **bandes de combinaison**. Certaines bandes de combinaison se situent dans la gamme NIR entre 1 900 et 2 500 nm.

2.2 Fondements mathématiques

2.2.1 Loi de Beer-Lambert

La loi de Beer-Lambert décrit comment l'absorption de la lumière par un échantillon homogène dépend des propriétés d'une substance absorbante dans l'échantillon.

$$A = \varepsilon \cdot c \cdot l$$

Ici, A correspond à l'absorbance, ε au coefficient d'extinction molaire de l'absorbeur (L/mol/cm), c à la concentration de l'absorbeur (mol/L) et l à l'épaisseur de la couche de l'échantillon (cm).

Le coefficient d'extinction molaire ε est une constante qui caractérise l'absorbtion par une substance. Le coefficient d'extinction molaire est

spécifique à une longueur d'onde donnée *i* et à une substance donnée *j*. L'absorbance globale d'un mélange est la somme de l'absorbance de toutes les substances contenues dans le mélange :

$$A_i = \sum_{j=1}^{N} \varepsilon_{ij} c_j \, l$$

où A_i est égal à l'absorbance à la longueur d'onde *i*, *N* au nombre de substances dans le mélange, ε_{ij} au coefficient d'extinction molaire pour la longueur d'onde *i* et la substance *j* et c_j à la concentration de la substance *j*.

La loi de Beer-Lambert établit une relation linéaire entre l'absorbance et la concentration et une relation linéaire entre l'absorbance et le coefficient d'extinction molaire. Ce comportement linéaire s'applique à de nombreuses situations.

La loi de Beer-Lambert permet aux mesures spectroscopiques d'absorption de mettre en évidence les éléments suivants :

- Variations de la concentration d'un absorbeur.
  C'est l'application la plus fréquente.
- Variations des facteurs ayant une influence sur les coefficients d'extinction molaires.
  La température, la viscosité, le pH ou la constante de diélectrique du solvant peuvent influer sur les coefficients d'extinction molaires. Dans certains cas, cela peut être utilisé pour des mesures spectroscopiques.

La diffusion de la lumière n'est pas liée à la loi de Beer-Lambert. La diffusion peut parfois être utilisée pour détecter par ex., une variation de taille des particules.

2.2.2 Régression linéaire

Une longueur d'onde

Dans le cas le plus simple, un mélange a un seul absorbeur. Selon la loi de Beer-Lambert, l'absorbance d'une certaine longueur d'onde est une fonction linéaire de la concentration de l'absorbeur.

Chaque point de la figure 1 représente un échantillon de concentration connue de l'absorbeur (axe x) et de l'absorbance mesurée (axe y). Une régression linéaire donne la droite de régression **A**.

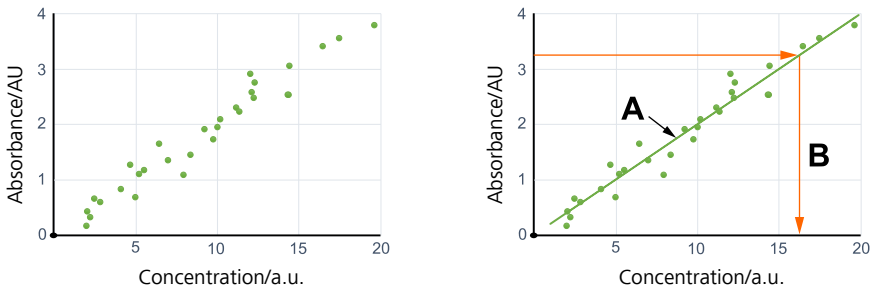


Figure 3 Relation entre valeurs d'absorbance et concentrations

Pour un échantillon dont la concentration de l'absorbeur est inconnue, la concentration peut être déterminée comme suit :

1. Mesure de l'absorbance pour la longueur d'onde donnée.
2. Utilisation des droites de régression pour la détermination de la concentration (**B**).
   La droite de régression est un modèle de quantification simple pour la prédiction du paramètre d'intérêt, par ex. la concentration de l'absorbeur.

Cette procédure échoue toutefois lorsque le mélange contient plusieurs absorbeurs de concentration différente.

Plusieurs longueurs d'onde

Au lieu d'une mesure de l'absorption à une seule longueur d'onde, il est possible de mesurer l'absorption à différentes longueurs d'onde, ce qui donne alors un spectre. Comme ci-dessus, il est possible de déterminer la relation entre les spectres et le paramètre d'intérêt à l'aide d'une régression linéaire. Une régression linéaire multiple (RLM) est nécessaire dans le cas de plusieurs longueurs d'onde.

Il est démontré que la régression linéaire multiple peut prédire le paramètre d'intérêt même si le mélange contient plusieurs absorbeurs de concentrations différentes *(voir "Exemple d'une régression linéaire", Chapitre 6.1, page 89)*.

Cependant, pour une régression linéaire multiple, il faudrait un nombre d'échantillons supérieur au nombre de longueurs d'onde. D'autres méthodes, telles que les méthodes ACP ou MCP, peuvent être utilisées dans le cadre des prédictions spectroscopiques.

2.3 Transformation de la lumière en spectre

Un spectromètre (ou spectrophotomètre) est composé d'une source de lumière et d'une unité de détection. La source de lumière émet une lumière avec un large spectre de longueurs d'onde, en d'autres termes une lumière polychromatique. La lumière interagit avec l'échantillon. Ensuite, le spectromètre enregistre la lumière restante en fonction de la longueur d'onde.

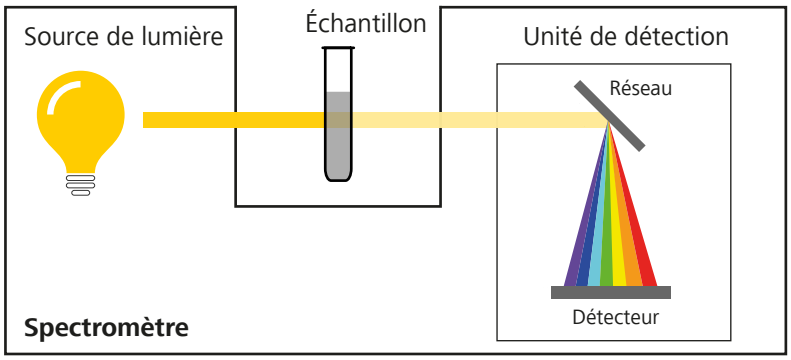


Figure 4 Un spectromètre avec source de lumière et unité de détection.

Dans le spectromètre, la lumière est décomposée en différentes longueurs d'onde à l'aide d'un réseau. Un détecteur mesure la lumière, où chaque longueur d'onde atteint un élément - ou pixel - différent d'un capteur linéaire.

Un **scan** est une mesure sur tous les pixels. Chaque pixel produit un signal photoélectrique proportionnel à l'intensité lumineuse reçue. Les signaux peuvent être rapportés aux pixels.

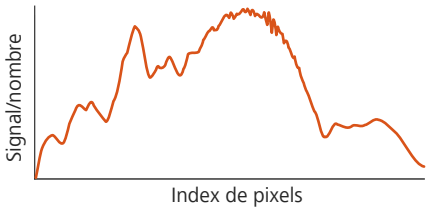


Figure 5 Spectre du signal de détection en fonction du pixel.

Temps d'intégration

Le temps d'intégration est la durée pendant laquelle le détecteur recueille la lumière. Un temps d'intégration plus élevé augmente le signal.

Un temps d'intégration trop long sature le détecteur et entraîne une perte d'information. Un temps d'intégration trop court diminue le signal et donc dégrade le rapport signal/bruit.

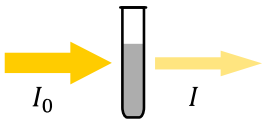
Un **temps d'intégration automatique** garantit un éclairage optimal, c'est-à-dire un rapport signal/bruit optimal sans entraîner de saturation. Plusieurs mesures sont effectuées avant chaque scan d'échantillon et chaque scan de référence. Le temps d'intégration est réglé de façon à ce que le signal atteigne environ 90 % de la gamme détectable avec l'intensité la plus élevée.

Les différences dans les temps d'intégration sont prises en compte dans les calculs ultérieurs.

- **OMNIS NIR Analyzer**
  Le temps d'intégration est toujours réglé automatiquement.
- **2060 The NIR**
  Le temps d'intégration peut être réglé manuellement ou automatiquement.
  Le temps d'intégration manuel peut réduire le temps de mesure pour les mesures similaires répétées. Pour éviter de saturer le détecteur avec un temps d'intégration trop long, il faut régler le temps d'intégration avec une marge suffisante *(voir « Saturation », page 49)*.

Absorbance

Les mesures spectroscopiques déterminent la quantité de lumière absorbée ou diffusée par un échantillon. L'échantillon est exposé à un faisceau lumineux. Un détecteur mesure la lumière émise par une source de lumière et la lumière restante après interaction avec l'échantillon.



L'absorbance A est définie comme le logarithme décimal du rapport entre l'intensité lumineuse avant l'interaction du faisceau lumineux avec l'échantillon (I_0) et l'intensité lumineuse après cette interaction (I) :

$$A = \log_{10} \frac{I_0}{I}$$

Une absorbance de 1 signifie que 10 % de la lumière passent à travers l'échantillon tandis qu'une absorbance de 2 signifie que 1 % passe.

L'absorbance est une grandeur sans dimension et donc dépourvue d'unité.

Scan de référence et scan de l'échantillon

Deux scans sont nécessaires pour le calcul d'un spectre d'absorbance selon la formule ci-dessus. Les scans mesurent le signal photoélectrique S de chaque pixel :

- Un **scan de référence** mesure le signal S_0 avant l'interaction du faisceau lumineux avec l'échantillon.
- Un **scan de l'échantillon** mesure le signal S après l'interaction du faisceau lumineux avec l'échantillon.

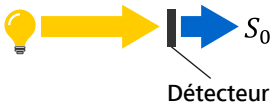
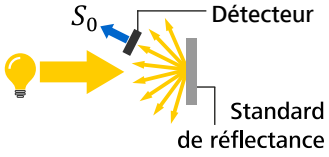
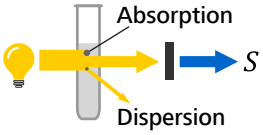
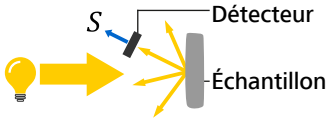
Le signal photoélectrique mesuré d'un pixel est proportionnel à la valeur moyenne de l'intensité lumineuse sur la surface du pixel. Par conséquent, $S_0/S = I_0/I$. Les signaux photoélectriques peuvent ainsi être utilisés pour calculer l'absorbance.

$$A = \log_{10}\frac{I_0}{I} = \log_{10}\frac{S_0}{S}$$

Transmission et réflexion

La lumière qui traverse l'échantillon est mesurée en **mode transmission**. S_0 est mesuré s'il n'y a pas d'échantillon. S est mesuré à partir de la lumière qui a traversé l'échantillon.

La lumière que l'échantillon réfléchit est mesurée en **mode réflexion**. Un standard de réflectance est utilisé comme référence à la place de l'échantillon. Le standard de réflectance réfléchit idéalement 100 % de la lumière. Une partie de la lumière réfléchie est dirigée vers le détecteur et émet le signal S_0. Le signal S est mesuré de la même manière, mais avec l'échantillon qui réfléchit la lumière.

	Transmission	Réflexion diffuse
Sans échantillon		
Avec échantillon		

L'absorbance A calculée représente toute la lumière qui n'atteint pas le détecteur. Par conséquent, A ne contient pas seulement la lumière absorbée par l'échantillon, mais également

- la lumière qui n'atteint pas le détecteur, car elle est diffusée trop loin du détecteur ;
- la lumière diffusée par erreur vers le détecteur.

Spectre d'absorbance

Le spectre d'absorbance est calculé en fonction du scan de référence (signal S_0) et du scan de l'échantillon (signal S) selon la formule ci-dessus.

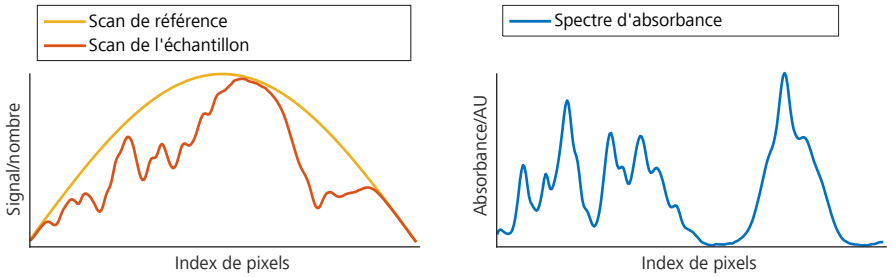


Figure 6 Scan de référence et scan de l'échantillon (à gauche) et le spectre d'absorbance calculé (à droite) en tant que fonction de l'index de pixel.

Le calcul ci-dessus suppose que le scan de référence et le scan d'échantillon utilisent les mêmes chemins optiques ou des chemins optiques avec des propriétés optiques similaires. On utilise un environnement de référencement à plusieurs niveaux dans les environnements de processus *(voir "2060 The NIR", Chapitre 3.2.2, page 17)*.

Des pixels aux longueurs d'onde

L'échelle des pixels est convertie en échelle de longueurs d'onde. L'appareil attribue une longueur d'onde précise à chaque pixel, par exemple :

Pixel 6 ➔ longueur d'onde 1 009,4 nm

La longueur d'onde précise de chaque pixel est calculée à l'aide du le calibrage de la longueur d'onde *(voir "Calibrage de la longueur d'onde", Chapitre 3.1, page 14)*.

Conversion de l'échelle de longueurs d'onde

Le spectre est converti par interpolation sur l'échelle standard de longueurs d'onde :

1 000,0 nm, 1 000,5 nm, 1 001,0 nm, etc.

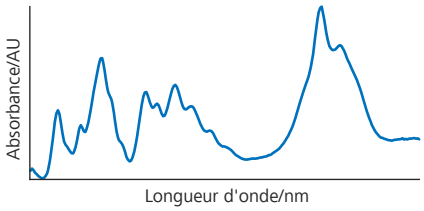


Figure 7 Le spectre d'absorbance sur l'échelle de longueurs d'onde.

3 Configuration de l'appareil

Les étapes suivantes permettent de garantir que, pour un échantillon donné, avec une structure de mesure d'échantillon donnée, des spectres identiques sont obtenus indépendamment du fait qu'ils aient été enregistrés à des moments différents et avec le même appareil ou un autre de la famille de produits **OMNIS NIR Analyzer** ou un autre appareil de type **2060 The NIR**.

Il faut tenir compte à la fois de l'axe x et l'axe y des spectres.

- **Axe x** : calibrage de la longueur d'onde *(voir "Calibrage de la longueur d'onde", Chapitre 3.1, page 14)*
- **Axe y** : standardisation de référence *(voir "Standardisation de référence", Chapitre 3.2, page 15)*

En outre, il faut s'assurer que la puissance de l'appareil répond aux exigences.

- Les **Tests de performance d'appareils** doivent être effectués avec succès avant de pouvoir enregistrer des spectres avec l'appareil *(voir "Tests de performance d'appareils", Chapitre 3.3, page 26)*.

Standards

Pour le calibrage de la longueur d'onde et les Tests de performance d'appareils, l'appareil utilise un **étalon de longueur d'onde** interne, métrologique et traçable.

Lorsque le mode réflexion est utilisé, un standard de réflectance est également nécessaire pour la **standardisation de référence** et les Tests de performance d'appareils, en fonction du type d'appareil :

- **OMNIS NIR Analyzer** : standard de réflectance interne
- **2060 The NIR** : standard de réflectance externe

3.1 Calibrage de la longueur d'onde

Le calibrage de la longueur d'onde normalise les valeurs de longueur d'onde, c.-à-d. l'axe x des spectres. Il associe chaque pixel de la matrice de photodétecteurs à une longueur d'onde.

Le calibrage de la longueur d'onde utilise un étalon de longueur d'onde interne, métrologique et traçable. L'étalon de longueur d'onde dispose d'un spectre d'absorbance avec des pics définis et des positions de pic connues.

La fonction **CAL WL** exécute les étapes suivantes :

1. Le spectre d'absorbance de l'étalon de longueur d'onde est enregistré en utilisant un chemin de référence interne sur l'échelle des pixels.
2. Les positions de pic sont identifiées à la précision du sous-pixel dans le spectre enregistré.
3. Une régression polynomiale est réalisée en utilisant les positions de pic mesurées et les positions de pic nominales de l'étalon de longueur d'onde.
4. Le polynôme de régression attribue à chaque pixel sa longueur d'onde respective.

Les coefficients de régression sont enregistrés dans l'appareil.

- **OMNIS NIR Analyzer** : un ensemble de coefficients de régression est enregistré dans l'appareil pour chaque présentation de l'échantillon. Cela signifie que pour l'OMNIS NIR Analyzer Liquid/Solid, un calibrage de la longueur d'onde et une validation doivent être effectués sur les deux groupes fonctionnels.
- **2060 The NIR** : les coefficients de régression sont spécifiques à l'appareil. Le même ensemble est utilisé pour tous les canaux.

Validation du calibrage de la longueur d'onde

Le calibrage de la longueur d'onde doit être validé après son exécution. La fonction **VAL WL** exécute les étapes suivantes :

1. Le spectre d'absorbance de l'étalon de longueur d'onde est enregistré.
2. Validation des longueurs d'onde :
    a. Les positions de pic sont identifiées sur l'échelle de longueurs d'onde dans le spectre enregistré.
    b. Les résidus de longueurs d'onde entre les positions de pic mesurées et les positions de pic connues sont calculés.
    c. Pour chaque pic, le résidu de longueur d'onde doit se situer dans la tolérance pour réussir le test.

3. Validation des largeurs de bande :
   a. Les largeurs de pic sont déterminées dans le spectre enregistré.
   b. Les résidus de largeur de bande entre les largeurs de pic mesurées et les largeurs de pic connues sont calculés.
   c. Pour chaque pic, le résidu de largeur de bande doit se situer dans la tolérance pour réussir le test.
4. L'état global de la validation est réussi si tous les résidus mentionnés ci-dessus se situent dans la tolérance.

Pour pouvoir enregistrer des spectres avec l'appareil, il faut d'abord réussir cette validation.

3.2 Standardisation de référence

La standardisation de référence normalise les valeurs d'absorbance, c'est-à-dire l'axe y des spectres.

Détermination de l'absorbance

Les signaux S_0 (scan de référence) et S (scan de l'échantillon) sont nécessaires pour le calcul de l'absorbance A d'un échantillon *(voir "Transformation de la lumière en spectre", Chapitre 2.3, page 9)* :

$$A = \log_{10} \frac{S_0}{S}$$

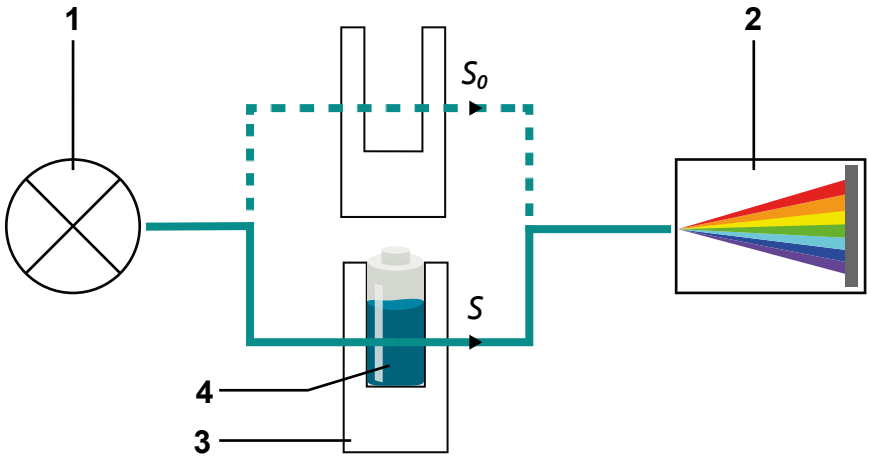


Figure 8 Chemin optique en mode transmission (comme exemple de présentation d'échantillons liquides).

Sur la *figure 8* la lumière provenant de la source de lumière (**1**) passe par un porte-échantillon (**3**) vers le détecteur (**2**).

Le signal de référence S_0 est mesuré sans l'échantillon, le signal S avec l'échantillon (**4**). Dans le cas contraire, les propriétés optiques des deux chemins optiques sont identiques et provoquent une atténuation du

même pourcentage pour les deux signaux. Cela ne change rien au résultat de la formule ci-dessus.

L'équation met S en relation avec la référence S_0. Les deux signaux ont la même importance. Un écart sur l'un des deux signaux entraîne une autre valeur d'absorbance et finalement un autre spectre.

Les variations de l'appareil et les conditions ambiantes influencent aussi bien S que S_0. Pour s'assurer que ces influences s'annulent mutuellement, les deux signaux doivent être mesurés dans un court laps de temps.

L'implémentation de ce principe dépend du type d'appareil :

- **OMNIS NIR Analyzer**
  Le spectre d'absorbance de l'échantillon est calculé avec les signaux S et S_0 *(voir "OMNIS NIR Analyzer", Chapitre 3.2.1, page 16)*.
- **2060 The NIR**
  L'utilisation du même chemin optique pour les mesures de S et S_0 n'est pas pratique dans les environnements de processus. Il faut donc procéder à d'autres mesures *(voir "2060 The NIR", Chapitre 3.2.2, page 17)*.

3.2.1 OMNIS NIR Analyzer

La standardisation de référence s'obtient en mesurant les signaux S_0 et S et en calculant l'absorbance A.

Enregistrer le spectre d'un échantillon

Avant de pouvoir enregistrer des spectres avec un groupe fonctionnel, les Tests de performance d'appareils pour le groupe fonctionnel doivent être effectués avec succès *(voir "Tests de performance d'appareils", Chapitre 3.3, page 26)*.

1. L'échantillon doit être prêt dans la présentation de l'échantillon.
2. Le spectre d'absorbance est calculé à chaque fois avec le dernier spectre de référence enregistré S_0. Pour actualiser la valeur pour S_0, on peut exécuter la fonction **MEAS REF SPEC**.
   Lors de l'utilisation de la présentation d'échantillons solides, l'appareil insère automatiquement un standard de réflectance dans le chemin optique. Ce standard de réflectance ne nécessite aucune correction du signal S_0.
3. L'échantillon est mesuré à l'aide de la fonction **MEAS SPEC**. On obtient le signal S.
4. Le logiciel calcule A, l'absorbance de l'échantillon :
   $$A = \log_{10} \frac{S_0}{S}$$
   Ici, S_0 correspond au signal mesuré sur le chemin de référence et S au signal mesuré par l'échantillon.

3.2.2 2060 The NIR

Les appareils de type **2060 The NIR** requièrent une standardisation de référence externe.

Standardisation de référence externe

La mesure répétée du signal S (avec l'échantillon) et du signal S_0 (sans l'échantillon) par les chemins optiques avec des propriétés optiques identiques est chronophage et source d'erreurs.

C'est pourquoi on introduit deux autres trajets de faisceau *(voir Figure 9, page 17)* :

- Une **référence interne** dans l'appareil. Le chemin de référence interne donne le signal S_{ref} qui peut être mesuré simplement.
- Un autre chemin optique externe dans lequel les fibres optiques sont reliées à un **dispositif de calibrage**. Ce chemin optique donne le signal S_{fiber}.

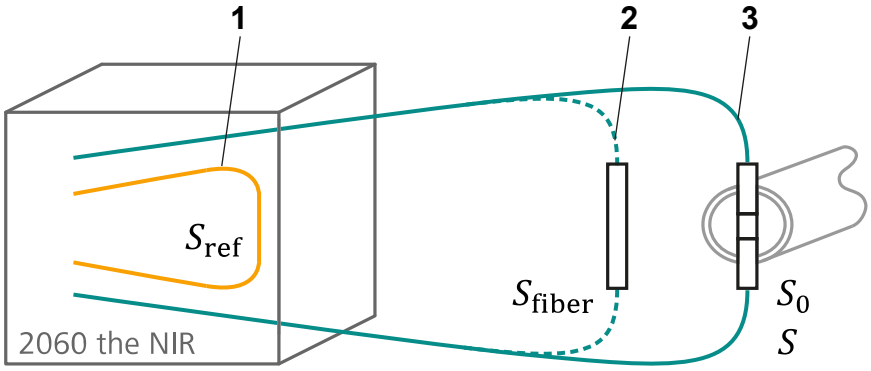


Figure 9 Chemins optiques en exemple comme mode transmission : chemin de référence interne (1), fibres optiques externes liées à un dispositif de calibrage (2) et fibres optiques externes liées à la sonde avec ou sans échantillon (3). Les chemins optiques 2 et 3 représentent la même fibre optique, qui est simplement connectée différemment.

Le dispositif de calibrage fixe les fibres optiques et constitue ainsi le chemin de référence (**2**). En mode transmission, l'air sert de référence et transmet 100 % de la lumière. En mode réflexion, le dispositif de calibrage enregistre également le standard de réflectance. Dans un premier temps, on part d'un standard de réflectance idéal qui réfléchit 100 % de la lumière.

L'absorbance A de l'échantillon, est calculée à partir des signaux S_0 et S. Le résultat reste inchangé si les deux signaux supplémentaires S_{ref} et S_{fiber} sont ajoutés au numérateur et au dénominateur.

$$A = \log_{10}\frac{S_0}{S} = \log_{10}\left(\frac{S_{\text{ref}}}{S}\cdot\frac{S_{\text{fiber}}}{S_{\text{ref}}}\cdot\frac{S_0}{S_{\text{fiber}}}\right)$$

L'équation peut être convertie en :

$$A = \log_{10}\left(\frac{S_{\text{ref}}}{S}\right) - \log_{10}\left(\frac{S_{\text{ref}}}{S_{\text{fiber}}}\right) - \log_{10}\left(\frac{S_{\text{fiber}}}{S_0}\right)$$

Les 3 termes représentent les valeurs d'absorbance et peuvent être désignés comme suit :

$$A = A_{\text{total}} - A_{\text{fiber}} - A_{\text{window}}$$

La *figure 10* illustre comment les signaux S_{ref}, S_{fiber}, S_0 et S sont mesurés.

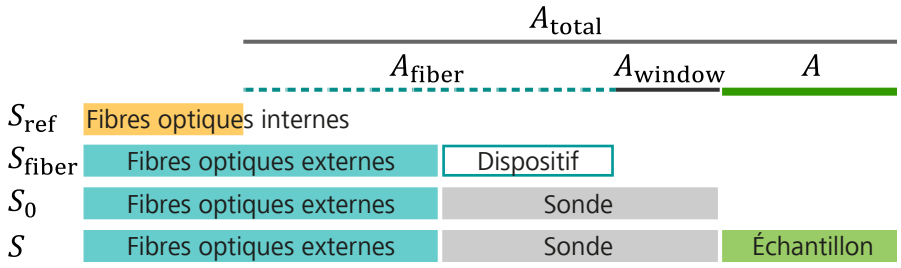


Figure 10 Standardisation de référence externe

A_{total} est l'absorbance de la fibre optique externe, la sonde et l'échantillon par rapport à la fibre optique interne.

A_{fiber} est l'absorbance de la fibre optique externe plus le dispositif de calibrage par rapport à la fibre optique interne.

A_{window} est l'absorbance de la sonde moins le dispositif de calibrage.

Élimination des variations ambiantes

Pour déterminer l'absorbance A de l'échantillon, on mesure 3 valeurs d'absorbance. A est calculé à partir des 3 valeurs selon l'équation ci-dessus :

$$A = A_{\text{total}} - A_{\text{fiber}} - A_{\text{window}}$$

Il en résulte la possibilité de déterminer les 3 valeurs d'absorbance à différents moments. De cette façon, les variations de l'appareil ou des conditions ambiantes peuvent être facilement éliminées pour chacune des 3 déterminations :

- A_{total} est déterminé avec chaque mesure d'échantillon. Les mesures de S_{ref} et S doivent être effectuées sur une courte période afin d'éliminer les variations.
- La **correction de fibre optique** A_{fiber} peut être déterminée plus rarement. Les mesures de S_{ref} et S_{fiber} doivent être effectuées sur une courte période afin d'éliminer les variations.

- La **correction de fenêtre** A_{window} peut être déterminée plus rarement, généralement une seule fois après l'installation. Les mesures de S_{fiber} et S_0 doivent être effectuées sur une courte période afin d'éliminer les variations.

Quand une correction de fenêtre est-elle nécessaire ?

Si le dispositif de calibrage restitue entièrement les propriétés optiques de la sonde, A_{window} est égale à 0. Dans ce cas, il est possible d'ignorer la correction de fenêtre.

Pour le mode transmission, il est généralement nécessaire d'effectuer une correction de fenêtre. En mode réflexion aucune correction de fenêtre n'est nécessaire. Il existe toutefois des exceptions regroupées dans le tableau ci-dessous :

Mode de mesure	Sonde	Standardisation de référence
Transmission	Paire de transmission	Fibre optique + fenêtre
	Sonde de transmission	
	Sonde de transflexion avec fibre unique	
Réflexion	Sonde de réflexion	Fibre optique
	Sonde de transflexion avec MicroBundle	Fibre optique + fenêtre

Pour déterminer la nécessité d'une correction de fenêtre, chaque combinaison de dispositif de calibrage et de sonde doit être examinée séparément. Si le dispositif de calibrage ne restitue *pas* entièrement les propriétés optiques de la sonde, une correction de fenêtre est nécessaire.

Canaux

Un appareil de type **2060 The NIR** offre plusieurs canaux. Chaque canal peut être lié à une autre configuration de fibre optique et de sonde. Il faut pour cela effectuer la standardisation de référence pour chaque canal séparément.

Tous les canaux utilisent le même chemin de référence interne, c'est-à-dire le même signal S_{ref}. Un multiplexeur commute entre la référence interne et les différents canaux de mesure.

Réalisation d'une correction de fibre optique

Après la mise en service ou modification de la configuration de la fibre optique d'un canal, une correction de fibre optique doit être réalisée. Une nouvelle standardisation peut être conseillée après un changement de lampe ou une modification extrême des conditions ambiantes.

Ce procédé requiert un matériel de référence :

- En mode réflexion, il s'agit d'un standard de réflectance. On part d'un standard de réflectance non idéal (p. ex. 99 %). Le standard de réflectance possède un spectre d'absorbance nominal connu $A_{nominal}$.
- En mode transmission, l'air sert de référence. Le spectre d'absorbance nominal est une ligne nulle ($A_{nominal}$ = 0), car on suppose que l'air n'absorbe pas la lumière.

La *figure 11* présente la procédure suivante :

1. Les fibres optiques externes doivent être connectées au dispositif de calibrage.
   En mode réflexion, le dispositif de calibrage est associé au standard de réflectance.
2. La fonction **REF STD** avec l'interface **Fibre optique** exécute les scans suivants :
   a. Un scan de référence interne fournit une valeur pour S_{ref}.
   b. Un scan externe mesure les fibres optiques externes, le dispositif de calibrage et le matériel de référence. On obtient le signal S_{raw}.
3. Le logiciel calcule A_{raw} (**Spectre brut mesuré**) :
   $$A_{raw} = \log_{10} \frac{S_{ref}}{S_{raw}}$$
   Dans ce cas, A_{raw} correspond à l'absorbance des fibres optiques externes, du dispositif de calibrage et du matériel de référence par rapport au chemin optique interne.
4. Le spectre d'absorbance nominal du matériel de référence, $A_{nominal}$, est affiché dans le logiciel comme **Spectre de référence**. Le spectre de référence doit être soustrait de A_{raw} pour obtenir A_{fiber} :
   $$A_{fiber} = A_{raw} - A_{nominal}$$
   Dans ce cas, A_{fiber} correspond à l'absorbance des fibres optiques et du dispositif de calibrage par rapport au chemin optique interne.
   Remarque : en mode transmission, $A_{nominal}$ = 0 et $A_{fiber} = A_{raw}$.
   A_{fiber} représente le **Spectre de correction** de la fibre optique.
5. A_{fiber} reste inchangé jusqu'à ce qu'une correction de fibre optique pour le canal respectif soit à nouveau effectuée.

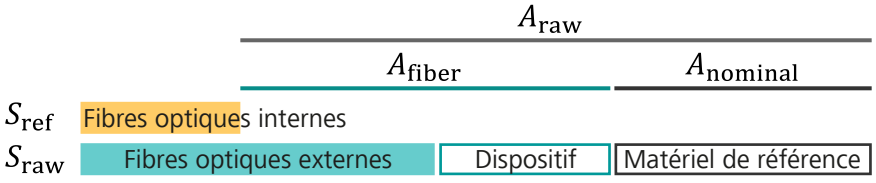


Figure 11 Correction de fibre optique

Validation de la correction de fibre optique

La correction de fibre optique doit être validée avec les mêmes paramètres de mesure et le même dispositif de calibrage.

Ce procédé requiert un matériel de référence :

- En mode réflexion, il s'agit d'un standard de réflectance. On part d'un standard de réflectance non idéal (p. ex. 99 %). Le standard de réflectance possède un spectre d'absorbance nominal connu $A_{nominal}$.
- En mode transmission, l'air sert de référence. Le spectre d'absorbance nominal est une ligne nulle ($A_{nominal}$ = 0), car on suppose que l'air n'absorbe pas la lumière.

La *figure 12* illustre comment les résidus de la validation sont déterminés :

1. Les fibres optiques externes doivent être connectées au dispositif de calibrage.
   En mode réflexion, le dispositif de calibrage est associé au standard de réflectance.
2. La fonction **VAL REF STD** avec l'interface **Fibre optique** exécute les scans suivants :
   a. Un scan de référence interne fournit une valeur pour S_{ref}.
   b. Un scan externe sur le canal respectif mesure les fibres optiques externes, le dispositif de calibrage et le matériel de référence. On obtient le signal S_{raw}.
3. Le logiciel calcule A_{raw} (**Spectre brut mesuré**) :
   $$A_{\mathrm{raw}} = \log_{10} \frac{S_{\mathrm{ref}}}{S_{\mathrm{raw}}}$$
4. Le spectre de correction de la fibre optique corrige A_{raw} pour éliminer l'absorbance des fibres optiques et du dispositif de calibrage :
   $$A_{\mathrm{corrected}} = A_{\mathrm{raw}} - A_{\mathrm{fiber}}$$
   $A_{corrected}$ est affiché dans le logiciel comme **Spectre mesuré corrigé**.
5. Dans le cas idéal, $A_{corrected}$ devrait correspondre avec le **Spectre de référence** $A_{nominal}$. Les différences entre les deux sont calculées comme **Résidus de la validation** :
   $$A_{\mathrm{residual}} = A_{\mathrm{corrected}} - A_{\mathrm{nominal}}$$
   Remarque : en mode transmission, $A_{nominal}$ = 0 et $A_{residual}$ = $A_{corrected}$.

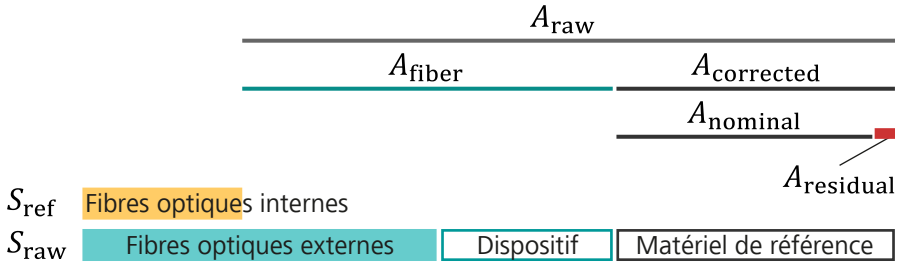


Figure 12 Résidus pour la validation de la correction de fibre optique

Pour étudier les résidus de la validation, la gamme de longueurs d'onde est divisée en plusieurs segments. Pour chaque segment, la moyenne quadratique des résidus sur les longueurs d'onde est le **résidu RMS** (unité : mAU) :

$$A_{\mathrm{RMS}} = \sqrt{\frac{\sum_{i=1}^{f}(A_{\mathrm{residual}_i})^2}{f}}$$

Dans ce cas, f représente le nombre de longueurs d'onde dans le segment et $A_{\mathrm{residual}\ i}$ le résidu de la longueur d'onde i.

Chaque segment doit respecter une tolérance prédéfinie pour A_{RMS}. Si tous les segments respectent la tolérance, la validation globale est réussie.

La validation doit être effectuée avec succès avant de pouvoir enregistrer des spectres sur le canal concerné avec l'appareil.

Réalisation d'une correction de fenêtre

Si une correction de fenêtre est nécessaire, elle doit être effectuée après la mise en service ou à chaque modification de la configuration de la sonde ou de la fibre optique d'un canal. Une nouvelle standardisation peut être conseillée en cas d'altérations de la sonde, telles que l'encrassement.

Ce procédé requiert un matériel de référence :

- En mode réflexion, il s'agit d'un standard de réflectance. On part d'un standard de réflectance non idéal (p. ex. 99 %). Le standard de réflectance possède un spectre d'absorbance nominal connu A_{nominal}.
- En mode transmission, l'air sert de référence. Le spectre d'absorbance nominal est une ligne nulle ($A_{\mathrm{nominal}} = 0$), car on suppose que l'air n'absorbe pas la lumière.

La *figure 13* présente la procédure suivante :

1. Pour actualiser la valeur pour A_{fiber}, il faut effectuer une correction de fibre optique selon la description ci-dessus.
   Important : la correction de fibre optique doit avoir lieu avant la correction de fenêtre.
2. Les fibres optiques externes doivent ensuite être connectées à la sonde sans échantillon. Si nécessaire un standard de réflectance prend la place de l'échantillon.
3. La fonction **REF STD** avec l'interface **Fenêtre** exécute les scans suivants :
   a. Un scan de référence interne fournit une valeur pour S_{ref}.
   b. Un scan externe sur le canal respectif mesure les fibres optiques externes, la sonde et le matériel de référence. On obtient le signal $S_{\mathrm{échantillon}}$.
4. L'absorbance $A_{\mathrm{échantillon}}$ par rapport au chemin optique interne est :

$$A_{\mathrm{probe}} = \log_{10}\frac{S_{\mathrm{ref}}}{S_{\mathrm{probe}}}$$

5. Le logiciel calcule A_{raw} (**Spectre brut mesuré**) :
   $A_{raw} = A_{probe} - A_{fiber}$
   Dans ce cas, A_{raw} correspond à l'absorbance de la sonde et du matériel de référence par rapport au dispositif de calibrage.
6. Le spectre d'absorbance nominal du matériel de référence, $A_{nominal}$, est affiché dans le logiciel comme **Spectre de référence**. Le spectre de référence doit être soustrait de A_{raw} pour obtenir A_{window} :
   $A_{window} = A_{raw} - A_{nominal}$
   Dans ce cas, A_{window} correspond à l'absorbance de la sonde par rapport au dispositif de calibrage.
   Remarque : en mode transmission, $A_{nominal}$ = 0 et $A_{window} = A_{raw}$.
   A_{window} représente le **Spectre de correction** de la fenêtre.
7. A_{window} reste inchangé jusqu'à ce qu'une correction de fenêtre pour le canal respectif soit à nouveau effectuée.

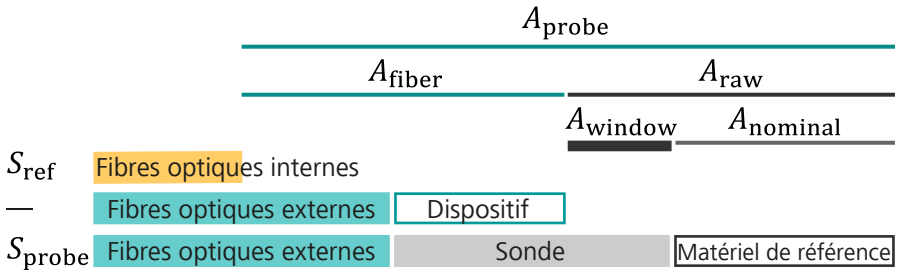


Figure 13 Correction de fenêtre

Validation de la correction de fenêtre

La correction de fenêtre doit être validée avec les mêmes paramètres de mesure et le même dispositif de calibrage.

Ce procédé requiert un matériel de référence :

- En mode réflexion, il s'agit d'un standard de réflectance. On part d'un standard de réflectance non idéal (p. ex. 99 %). Le standard de réflectance possède un spectre d'absorbance nominal connu $A_{nominal}$.
- En mode transmission, l'air sert de référence. Le spectre d'absorbance nominal est une ligne nulle ($A_{nominal}$ = 0), car on suppose que l'air n'absorbe pas la lumière.

La *figure 14* illustre comment les résidus de la validation sont déterminés :

1. Les fibres optiques externes doivent être connectées à la sonde sans échantillon. Si nécessaire un standard de réflectance prend la place de l'échantillon.
2. La fonction **VAL REF STD** avec l'interface **Fenêtre** exécute les scans suivants :
   a. Un scan de référence interne fournit une valeur pour S_{ref}.
   b. Un scan externe sur le canal respectif mesure les fibres optiques externes, la sonde et le matériel de référence. On obtient le signal $S_{échantillon}$.

3. L'absorbance $A_{\text{échantillon}}$ par rapport au chemin optique interne est :

$$A_{\text{probe}} = \log_{10} \frac{S_{\text{ref}}}{S_{\text{probe}}}$$

4. Le logiciel calcule A_{raw} (**Spectre brut mesuré**) :

$$A_{\text{raw}} = A_{\text{probe}} - A_{\text{fiber}}$$

5. La soustraction du spectre de correction de fenêtre de A_{raw} permet d'éliminer les différences d'absorbance entre le dispositif de calibrage et la sonde :

$$A_{\text{corrected}} = A_{\text{raw}} - A_{\text{window}}$$

$A_{\text{corrected}}$ est affiché dans le logiciel comme **Spectre mesuré corrigé**.

6. Dans le cas idéal, $A_{\text{corrected}}$ devrait correspondre avec le **Spectre de référence** A_{nominal}. Les différences entre les deux sont calculées comme **Résidus de la validation** :

$$A_{\text{residual}} = A_{\text{corrected}} - A_{\text{nominal}}$$

Remarque : en mode transmission, $A_{\text{nominal}} = 0$ et $A_{\text{residual}} = A_{\text{corrected}}$.

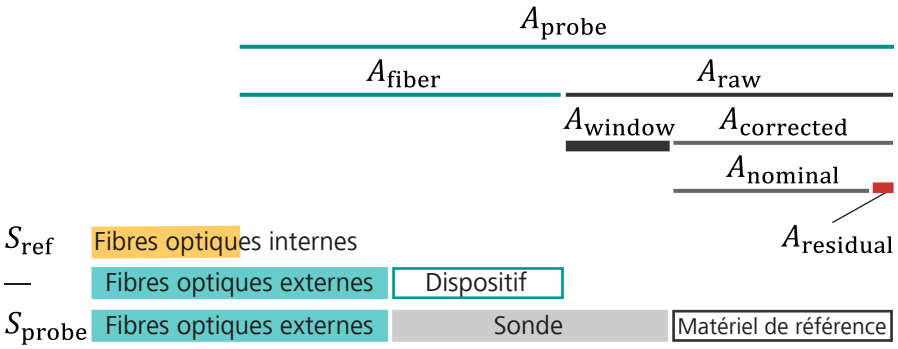


Figure 14 Résidus pour la validation de la correction de fenêtre

Pour étudier les résidus de la validation, la gamme de longueurs d'onde est divisée en plusieurs segments. Pour chaque segment, la moyenne quadratique des résidus sur les longueurs d'onde est le **bruit de fond RMS** (unité : mAU) :

$$A_{\text{RMS}} = \sqrt{\frac{\sum_{i=1}^{f} (A_{\text{residual}_i})^2}{f}}$$

Dans ce cas, f représente le nombre de longueurs d'onde dans le segment et $A_{\text{residual } i}$ le résidu de la longueur d'onde i.

Chaque segment doit respecter une tolérance prédéfinie pour A_{RMS}. Si tous les segments respectent la tolérance, la validation globale est réussie.

Enregistrer le spectre d'un échantillon

Les Tests de performance d'appareils doivent être effectués avec succès sur chaque canal avant de pouvoir enregistrer des spectres avec l'appareil *(voir "Tests de performance d'appareils", Chapitre 3.3, page 26)*.

La procédure d'acquisition du spectre d'un échantillon est présentée sur la *figure 15* :

1. Les fibres optiques externes doivent être connectées à la sonde. Il doit y avoir un échantillon.
2. Le spectre d'absorbance est calculé à chaque fois avec le dernier spectre de référence enregistré S_{ref}. Pour actualiser la valeur pour S_{ref}, on peut exécute la fonction MEAS REF SPEC .
3. La fonction **MEAS SPEC** mesure l'échantillon, y compris sonde et fibres optiques. On obtient le signal S.
4. Le logiciel calcule A_{total}, l'absorbance de l'échantillon, sonde et fibres optiques incluses, par rapport au chemin optique interne :

   $$A_{\mathrm{total}} = \log_{10} \frac{S_{\mathrm{ref}}}{S}$$

5. L'absorbance de l'échantillon est ensuite calculée comme décrit ci-dessus en utilisant le spectre de correction de la fibre optique A_{fiber} et le spectre de la correction de fenêtre A_{window} du canal respectif :

   $A = A_{\mathrm{total}} - A_{\mathrm{fiber}} - A_{\mathrm{window}}$

   A représente le spectre de l'échantillon.

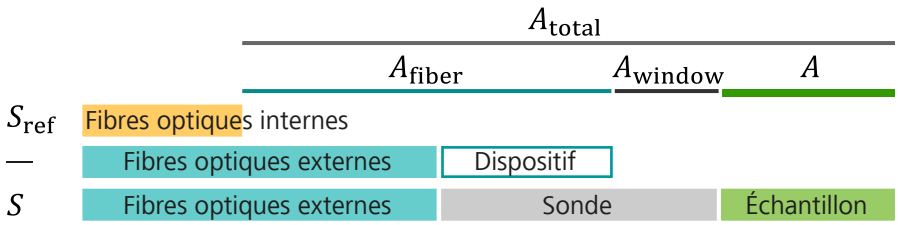


Figure 15 Enregistrer un spectre d'un échantillon

3.3 Tests de performance d'appareils

Les Tests de performance d'appareils peuvent être effectués via des chemins optiques internes et externes.

- **OMNIS NIR Analyzer**
  - **Tests de performance d'appareils internes** (obligatoires) : les tests internes utilisent le chemin de référence de la présentation de l'échantillon correspondante. Les essais vérifient les longueurs d'ondes et le bruit du signal.
    Avant les tests, le calibrage de la longueur d'onde doit être effectué avec succès et validé pour chaque présentation de l'échantillon.
    Les tests internes doivent être effectués avec succès pour pouvoir enregistrer des spectres avec l'appareil sur la présentation de l'échantillon correspondante.
  - **Tests de performance d'appareils externes** (optionnels) : les tests externes permettent la validation selon les pharmacopées telles que USP <856>, Ph.Eur 2.2.40 et JP 2.27. Longueurs d'onde, bruit du signal et linéarité photométrique sont vérifiés *(voir "Externe Tests de performance d'appareils (OMNIS NIR Analyzer)", Chapitre 3.3.1, page 29)*.
- **2060 The NIR**
  Les tests peuvent utiliser le chemin optique interne ou l'un des chemins optiques externes. Les essais vérifient les longueurs d'ondes et le bruit du signal.
  Avant les tests, le calibrage de la longueur d'onde et la standardisation de référence externe doivent être effectués avec succès et validés pour le canal correspondant.
  Les Tests de performance d'appareils doivent être effectués avec succès avant de pouvoir enregistrer sur le canal correspondant des spectres avec l'appareil. Les tolérances prédéfinies doivent être respectées. Les tolérances admissibles dépendent de la configuration de la fibre optique du canal correspondant. Elles sont indiquées dans les données spécifiques de l'appareil (mode de mesure, type et longueur de fibre).

Test de longueur d'onde

Le test de longueur d'onde consiste à examiner l'exactitude et la précision de la longueur d'onde. À cet effet, on utilise un étalon de longueur d’onde qui possède un spectre d’absorbance avec des pics définis et des positions de pics connues :

- **Interne** : le spectre d'absorbance de l'étalon de longueur d'onde interne, traçable métrologiquement est déterminé par le chemin optique interne :

  $$A_{\mathrm{WL}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{ref,WL}}}\right)$$

  Dans ce cas, A_{WL} correspond à l'absorbance de l'étalon de longueur d'onde interne, S_{ref} au signal mesuré sur le chemin de référence interne et $S_{\mathrm{ref,WL}}$ au signal mesuré sur le chemin de référence interne avec l'étalon de longueur d'onde interne.
- **Externe** pour des appareils de la famille de produits **OMNIS NIR Analyzer** : le test de longueur d'onde externe est optionnel *(voir « Test de longueur d'onde externe », page 29)*.
- **Externe** pour des appareils de type **2060 The NIR** :
  Les fibres optiques externes doivent être connectées au dispositif de calibrage pour les fibres uniques et au standard de réflectance pour le MicroBundle.
  Le spectre d'absorbance de l'étalon de longueur d'onde interne, traçable métrologiquement est déterminé par un chemin optique externe :

  $$A_{\mathrm{WL}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{fiber,WL}}}\right) - A_{\mathrm{fiber}}$$

  Dans ce cas, A_{WL} correspond à l'absorbance de l'étalon de longueur d'onde interne, S_{ref} au signal mesuré sur le chemin de référence interne, $S_{\mathrm{fiber,WL}}$ au signal mesuré sur le chemin optique du canal correspondant, les fibres étant reliées au dispositif de calibrage ou au standard de réflectance, et l'étalon de longueur d'onde interne étant inséré dans le chemin optique, et A_{fiber} correspond au spectre de correction de la fibre optique issu de la standardisation de référence.

A_{WL} contient également l'absorbance du standard de réflectance, ce qui n'est pas pertinent pour les calculs suivants. Dans le cas idéal, les positions de pic de A_{WL} sont identiques aux positions de pic nominales de l'étalon de longueur d'onde.

L'exactitude et la précision de la longueur d'onde sont vérifiées comme suit :

1. Une série de spectres d'absorbance de l'étalon de longueur d'onde est enregistrée comme décrit ci-dessus (A_{WL}).
2. Les positions de pic sont identifiées dans les spectres enregistrés.
3. Les statistiques suivantes sont calculées sur les spectres enregistrés pour chaque position de pic :
   a. valeur moyenne (en nm)
   b. écart-type (en nm)

4. **Exactitude** : pour chaque pic, la différence entre la position de pic moyenne et la position de pic nominale doit se situer dans la tolérance prédéfinie.
   ➡ **Remarque** : les positions de pic nominales peut varier sensiblement d'un test à l'autre. Cela est dû au fait que les positions de pic sont corrigées en fonction de la température. Une correction de la température similaire est effectuée lors du calibrage de la longueur d'onde. Elle contribue à ce que les résultats des mesures pour une température déterminée soient comparables sur tous les appareils.
5. **Précision** : l'écart-type doit se situer dans la tolérance prédéfinie pour chaque pic.
6. Si les tolérances des pics sont respectées, l'état global du test de longueur d'onde est réussi.

Test du bruit de fond

Le bruit du signal peut être vérifié de manière interne ou externe :

- **Interne** : le bruit de fond est déterminé comme absorbance du chemin optique interne, par rapport à l'absorbance d'une autre mesure sur le même chemin optique :

  $$A_{\text{noise}} = \log_{10}\left(\frac{S_{\text{ref},1}}{S_{\text{ref},2}}\right)$$

  Dans ce cas, $S_{\text{ref},1}$ et $S_{\text{ref},2}$ correspondent aux signaux mesurés sur le chemin de référence interne.
- **Externe** pour des appareils de la famille de produits **OMNIS NIR Analyzer** : le test du bruit de fond est optionnel *(voir « Tests du bruit de fond externes », page 30)*.
- **Externe** pour des appareils de type **2060 The NIR** :
  Les fibres optiques externes doivent être connectées au dispositif de calibrage pour les fibres uniques et au standard de réflectance pour le MicroBundle.
  Le bruit de fond est déterminé en tant que différence entre le spectre d'absorbance mesuré et le spectre d'absorbance nominal :

  $$A_{\text{noise}} = \log_{10}\left(\frac{S_{\text{ref}}}{S_{\text{fiber}}}\right) - A_{\text{fiber}} - A_{\text{nominal}}$$

  Dans ce cas, S_{ref} correspond au signal mesuré sur le chemin de référence interne, S_{fiber} au signal mesuré sur le chemin optique externe, les fibres étant reliées au dispositif de calibrage ou au standard de réflectance. A_{fiber} correspond au spectre de correction de la fibre optique issu de la standardisation de référence et A_{nominal} correspond au spectre d'absorbance nominal du standard de réflectance.
  Remarque : en mode transmission, $A_{\text{nominal}} = 0$.

Dans le cas idéal, $A_{\text{noise}} = 0$.

Le test du bruit de fond exécute les étapes suivantes :

1. Une série de spectres de bruit de fond est enregistrée comme décrit ci-dessus (A_{noise}).
2. Les spectres de bruit de fond sont répartis dans différents segments de longueur d'onde.
3. 3 valeurs sont calculées pour chaque spectre de bruit de fond et chaque segment :
    a. bruit de fond photométrique (en mAU)
    b. bruit de fond pic à pic (en mAU)
    c. biais de ligne de base du bruit de fond (en mAU)
4. La valeur moyenne est calculée sur les spectres de bruit de fond enregistrés pour chacune des 3 valeurs dans chaque segment.
5. Si toutes les valeurs moyennes se situent dans les tolérances prédéfinies, l'état global du test du bruit de fond est réussi.

3.3.1 Externe Tests de performance d'appareils (OMNIS NIR Analyzer)

Des appareils de la famille de produits **OMNIS NIR Analyzer** peuvent être validés selon les pharmacopées telles que USP <856>, Ph.Eur 2.2.40 und JP 2.27 (à partir de la version logicielle d'OMNIS 4.4). Ces tests nécessitent des standards de référence externes, traçables métrologiquement. Les standards de référence ont des spectres d'absorbance individuels nominaux qui ont été mesurés avec un appareil de référence à température ambiante.

Test de longueur d'onde externe

L'exactitude et la précision de la longueur d'onde sont vérifiées comme suit :

1. L'étalon de longueur d'onde externe (transmission ou réflexion) doit être mis en position.
2. Une série de spectres d'absorbance de l'étalon de longueur d'onde externe est enregistrée :

    $$A_{\mathrm{WL}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{WL}}}\right)$$

    Dans ce cas, A_{WL} correspond à l'absorbance de l'étalon de longueur d'onde externe, S_{ref} au signal mesuré sur le chemin de référence interne et S_{WL} au signal mesuré avec l'étalon de longueur d'onde externe.
3. Les positions de pic sont identifiées dans les spectres enregistrés.
4. Les statistiques suivantes sont calculées sur les spectres enregistrés pour chaque position de pic :
    a. valeur moyenne (en nm)
    b. écart-type (en nm)
5. **Exactitude** : pour chaque pic, la différence entre la position de pic moyenne et la position de pic nominale doit se situer dans la tolérance prédéfinie.

6. **Précision** : l'écart-type doit se situer dans la tolérance prédéfinie pour chaque pic.
7. Si les tolérances sont respectées pour tous les pics, l'état global du test de longueur d'onde est réussi.

Tests du bruit de fond externes

Le bruit du signal est vérifié une fois à faible courant photonique (flux faible) et une fois à haut courant photonique (flux élevé) :

1. Le standard de référence externe (transmission ou réflexion) pour l'essai à flux faible ou à flux élevé doit être placé en position.
2. Une série de spectres de bruit de fond est enregistrée en tant que différence entre les spectres d'absorbance mesurés et les spectres d'absorbance nominaux du standard de référence :

   $$A_{\mathrm{noise}} = \log_{10}\left(\frac{S_{\mathrm{ref}}}{S_{\mathrm{ND}}}\right) - A_{\mathrm{nominal}}$$

   Dans ce cas, S_{ref} correspond au signal mesuré sur le chemin de référence interne, S_{ND} au signal mesuré sur le standard de référence externe et A_{nominal} au spectre nominal du standard de référence.
3. Les spectres de bruit de fond sont répartis dans différents segments de longueur d'onde.
4. 3 valeurs sont calculées pour chaque spectre de bruit de fond et chaque segment :
   a. bruit de fond photométrique (en mAU)
   b. bruit de fond pic à pic (en mAU)
   c. biais de ligne de base du bruit de fond (en mAU)
5. La valeur moyenne est calculée sur les spectres de bruit de fond enregistrés pour chacune des 3 valeurs dans chaque segment.
6. Si toutes les valeurs moyennes se situent dans les tolérances prédéfinies, l'état global du test du bruit de fond est réussi.

Linéarité photométrique

Cet essai vise à démontrer une relation linéaire entre la réflectance (ou la transmittance) et l'absorbance mesurée sur toute la gamme de longueurs d'onde :

1. Des spectres d'absorbance de 5 standards de référence sont enregistrés avec une réflectance (ou transmittance) différente.
2. La relation linéaire entre la réflectance (ou la transmittance) et l'absorbance mesurée est assurée par une régression linéaire sur plusieurs longueurs d'onde.
3. Si les pentes et ordonnées à l'origine de toutes les droites de régression se situent dans les tolérances prédéfinies, l'état global de l'essai est réussi.

4 Développement de modèles

Les types de modèles sont les suivants :

- Un **Modèle de quantification** décrit la dépendance d'un paramètre d'intérêt (p. ex. teneur en eau) des spectres des échantillons enregistrés.
- Un **Modèle d'identification** (à partir de la version logicielle d'OMNIS 4.0) classifie les échantillons en différents produits sur la base des spectres enregistrés (p. ex. différentes sortes de grains de café).
  Un produit désigne une substance chimique particulière ou une substance chimique particulière ayant des propriétés physiques spécifiques (p. ex. taille des particules).

Un spectre de l'échantillon est enregistré pour l'analyse d'un échantillon inconnu. Selon l'application, le spectre est utilisé ainsi :

- Quantification : sur les fondements du spectre, un modèle de quantification crée une prédiction, par exemple pour la teneur en eau de l'échantillon.
- Identification : sur les fondements du spectre, un modèle d'identification identifie l'échantillon comme du café arabica, par exemple.
- Vérification (à partir de la version logicielle d'OMNIS 4.2) : sur les fondements du spectre, un modèle d'identification vérifie, par exemple, si l'échantillon est du café arabica ou non.

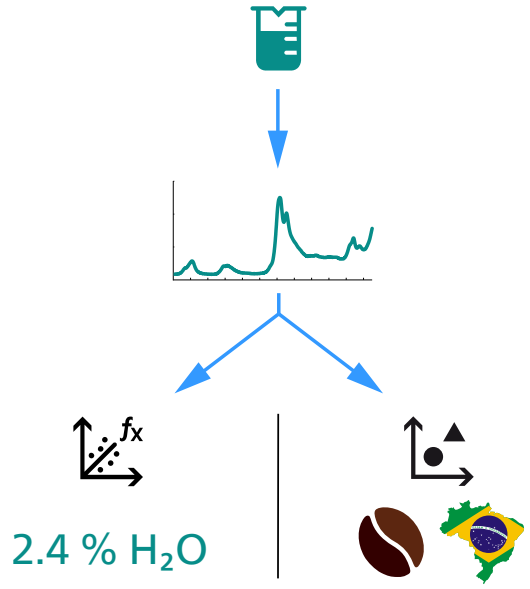


Figure 16 Quantification (en bas à gauche) et identification (en bas à droite).

Le développement d'un modèle et l'analyse d'échantillons comprennent les étapes suivantes :

1. **Échantillonnage**
   Des échantillons physiques sont prélevés et traités.
   a. Un spectre est enregistré pour chaque échantillon.
   b. Pour la quantification, une valeur de référence est mesurée pour le paramètre d'intérêt (p. ex. teneur en eau) avec une méthode de référence (p. ex. titrage). La mesure de référence doit être exacte et précise.
   c. Pour l'identification, l'appartenance du produit à l'échantillon doit être connue.
2. **Développement de modèles**
   Le développement de modèles s'effectue dans un processus itératif qui comprend les étapes suivantes :
   a. Division de l'ensemble de données dans un ensemble de données de calibrage, un ensemble de données validées et un ensemble de données atypiques.
   b. Application d'un prétraitement des données et de gammes de longueurs d'onde appropriés aux spectres.
   c. Calcul d'un modèle sur la base de l'ensemble de données de calibrage.
   d. La validation du modèle garantit qu'il répond aux exigences. La validation s'appuie principalement sur l'ensemble de données validées, qui n'ont pas été utilisées lors du développement du modèle.
      Lors de la quantification, le modèle prédit le paramètre d'intérêt pour les spectres dans l'ensemble de données validées. Les valeurs calculées sont ensuite comparées aux valeurs de référence connues.
      Le modèle associe les spectres à différents produits pour l'identification. Les appartenances de produits prédites sont comparées aux appartenances de produits réelles respectives.
3. **Contrôle**
   Le contrôle du modèle garantit que la capacité prédictive ne diminue pas au cours du temps. Chaque modification du processus ou des échantillons requiert une revalidation.

4.1 Échantillons physiques

L'opération commence par la collecte et l'analyse d'échantillons physiques. Un échantillonnage correct est la condition préalable du développement d'un modèle robuste. À cet égard, plusieurs aspects doivent être pris en compte.

Étendue de la variation

Les échantillons doivent inclure des variations types prévisibles de l'échantillon. Les concentrations de tous les composants chimiques et les tailles des particules doivent couvrir au moins l'étendue de la variation attendue.

Les échantillons doivent couvrir une variation raisonnable des conditions et une période de temps appropriée. Toutes les variations doivent être prises en compte, par exemple, les variations de processus, les variations saisonnières ou les variations des conditions ambiantes.

Les échantillons doivent être répartis uniformément sur l'étendue de la variation. Lors de la quantification, cela inclut la gamme des valeurs de référence. Si la gamme des valeurs de référence est, par exemple, de 1 % à 10 %, les échantillons doivent être répartis uniformément entre 1 % et 10 %.

Échantillons de calibrage et de validation

En général, on utilise 2 ensembles d'échantillons :

- Ensemble de calibrage : échantillons utilisés pour le développement de modèles.
- Ensemble de données validées : échantillons utilisés pour la validation du modèle.

Les deux ensembles doivent couvrir les variations attendues. Pour l'ensemble des données validées, on recueille de préférence des échantillons indépendants pour vérifier la robustesse du modèle. Des scénarios tels que différents opérateurs, différents fournisseurs ou, le cas échéant, différents appareils devraient être envisagés.

S'il n'est pas possible de recueillir des échantillons indépendants pour le calibrage et la validation, les spectres des échantillons disponibles peuvent être répartis en un ensemble de données de calibrage et un ensemble de données validées. Pour garantir l'indépendance de l'ensemble, l'algorithme de répartition automatique doit être utilisé.

Dès qu'un modèle a été développé, l'utilisation d'échantillons explicites de valeurs atypiques peut être envisagée afin de vérifier la détection des valeurs aberrantes.

Tous les échantillons doivent être traités de la même manière. Pour l'acquisition des spectres, la même méthode doit être utilisée avec la même

configuration matérielle et les mêmes paramètres de mesure. Pour la quantification, la même méthode de référence doit être utilisée avec les mêmes paramètres de mesure.

Nombre d'échantillons

Plus il y a de variations de conditions, de composants chimiques ou de tailles des particules à couvrir, plus il faut d'échantillons.

Quantification : pour que l'analyse statistique fonctionne correctement, un minimum d'environ 50 échantillons est nécessaire, l'ensemble de calibrage et l'ensemble de données validées devant contenir chacun au moins 20 à 25 échantillons.

Identification et vérification : pour chaque produit, les échantillons doivent couvrir les variations attendues. Le nombre d'échantillons de chaque produit peut varier, 3 étant le minimum.

Un minimum de 10 à 20 échantillons (selon le nombre de variations) permet de développer un premier modèle sans ensemble de données validées. Si la validation croisée (pour les modèles de quantification) ou la validation interne (pour les modèles d'identification) indique qu'un modèle adéquat peut être créé, des échantillons de calibrage et de validation supplémentaires doivent être recueillis pour le développement du modèle final.

Réplicats

Parfois, il n'y a que très peu d'échantillons dans une certaine gamme de conditions ou de valeurs de référence. Pour compenser, on peut essayer de reproduire ces échantillons. Cela est toutefois problématique. Si des répliques d'un échantillon sont présentes tant dans l'ensemble de calibrage que dans l'ensemble de données validées, les figures de mérite sont trompeuses (trop optimistes). Il faut également éviter les répliques dans le même ensemble.

Méthode de référence (quantification)

Pour la quantification on utilise une méthode de référence afin de mesurer les valeurs de référence. L'**erreur standard du laboratoire (SEL)** sur la méthode de référence utilisée joue un rôle essentiel dans le développement d'un modèle de quantification. La SEL est l'écart-type des différences entre les mesures d'échantillons dupliqués.

La SEL est fréquemment l'erreur la plus importante qui contribue à l'erreur standard de prédiction (SEP) dans la méthode NIR *(voir « SEP – erreur standard de prédiction », page 71)*. La SEL ne devrait pas dépasser 0,7 fois, de préférence 0,5 fois, la SEP requise. La gamme de valeurs de référence devrait être au moins 3 fois, de préférence 5 fois, celle de la SEL.

La SEL peut être réduite en répétant des mesures de référence sur chaque échantillon. La moyenne des valeurs de mesure devrait être définie comme

valeur de référence de l'échantillon. Il convient d'effectuer le même nombre de mesures de référence pour chaque échantillon. Les figures de mérite sont exprimées par rapport à un certain nombre de mesures de référence répétées. Un nombre différent de mesures de référence répétées conduirait à des valeurs estimées erronées des figures de mérite et devrait être évité.

Température de l'échantillon

La température de l'échantillon a une influence essentielle sur les spectres de liquides contenant de l'eau ou d'autres ponts d'hydrogène. Des spectres d'autres liquides polaires peuvent être influencés de la même manière que les spectres de solides contenant de l'eau, de l'humidité ou des solvants. De tels échantillons doivent être mesurés à une température définie.

Valeur atypique

Certains échantillons sont identifiés ultérieurement comme des valeurs atypiques, le cas échéant. Une valeur atypique est un échantillon qui, pour une raison ou une autre, varie de la majorité des échantillons. Pour éviter une influence négative sur le modèle, les échantillons identifiés comme valeurs atypiques ne sont pas pris en compte dans le calcul du modèle.

Il existe différents types de valeurs atypiques :

- **Valeur spectrale atypique**
  Si le spectre acquis d'un échantillon s'écarte de la plupart des autres spectres, l'échantillon est éventuellement reconnu comme une valeur spectrale atypique *(voir "Valeur spectrale atypique", Chapitre 4.3.3, page 51)*.
- **Valeur de référence atypique (quantification)**
  Des valeurs de référence individuelles peuvent présenter des anomalies au cours d'une quantification et être identifiées ultérieurement comme des valeurs de référence atypiques *(voir "Valeur de référence atypique (quantification)", Chapitre 4.3.4, page 58)*.
  L'OMNIS Model Developer (OMD) détecte les échantillons correspondants comme des valeurs atypiques sur la base de la détection des valeurs atypiques selon ASTM D8321-22 *(voir "OMNIS Model Developer (OMD)", Chapitre 4.4.3, page 72)*.

4.2 Analyse en composantes principales (ACP)

Les données spectroscopiques des échantillons de calibrage comportent une quantité importante de variables (longueurs d'onde). Les variables sont fortement corrélées entre elles. Les données sont donc extrêmement redondantes. Pour traiter ce type de données, on utilise des modèles de variables latentes tels que l'ACP et la MCP.

L'**analyse en composantes principales** (**ACP**, en anglais *principal component analysis*) se concentre sur les spectres sans tenir compte des valeurs de référence.

Le logiciel OMNIS utilise l'ACP pour la répartition automatique des ensembles de données et la détection de valeurs spectrales atypiques pendant le développement de modèle.

Étape de préparation

Les étapes de préparation suivantes sont nécessaires :

1. **++++++Prétraitement des données** : le logiciel OMNIS applique le prétraitement des données défini sur les spectres *(voir "Prétraitement des données", Chapitre 4.3.1, page 40)*.
2. **Gamme de longueurs d'onde** : le logiciel OMNIS applique la sélection de longueurs d'onde définie sur les spectres *(voir "Gammes de longueurs d'onde", Chapitre 4.3.2, page 49)*.
3. **Centrage de la moyenne** : pour chaque longueur d'onde on calcule la valeur d'absorbance moyenne que l'on soustrait de la valeur respective dans chaque spectre.

Les premiers composants principaux

Après les étapes de préparation, l'ACP réorganise les informations dans les données spectrales et sépare les données pertinentes du bruit de fond. À cette fin, l'ACP transforme les variables de longueur d'onde dans un nouvel espace de variables appelées **composantes principales** (**PC**, en anglais Principal Components).

L'ACP convertit les informations pertinentes issues d'une multitude de variables de longueur d'onde en quelques composantes principales. Pour représenter le concept à l'aide d'un exemple simple, on suppose qu'au lieu de milliers, il n'y a que 2 variables de longueur d'onde et que ces 2 variables sont réduites à 1 composante principale.

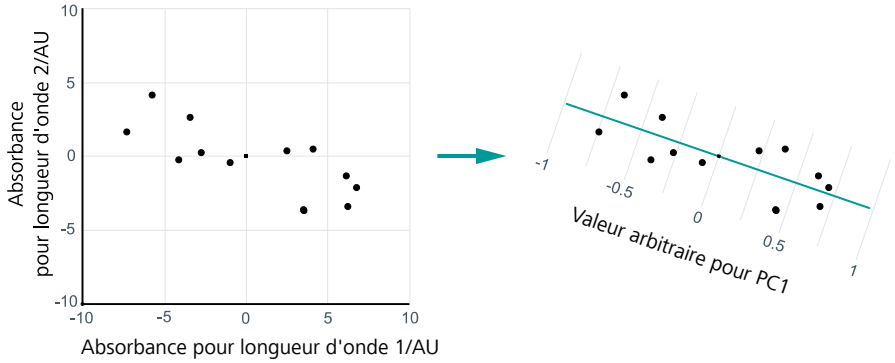


Figure 17 Points représentant les spectres dans un espace de longueur d'onde à 2 dimensions (à gauche). Les mêmes points dans un espace de composantes principales à 1 dimension (à droite).

Sur la *figure 17* à gauche, l'axe horizontal et l'axe vertical forment l'espace initial des longueurs d'onde avec 2 variables. Chaque point représente ainsi un spectre avec seulement 2 longueurs d'onde. La valeur moyenne de toutes les valeurs de longueur d'onde constitue le point zéro.

À droite, la composante principale PC1 est représentée par la direction à travers les données qui explicite la variance maximale. Dans cet exemple, PC1 est la seule variable dans l'espace des composantes principales. En conséquence, les 2 variables initiales sont réduites à 1.

Scores et résidus

La *figure 18* indique les grandeurs qui caractérisent un spectre *i* :

- La distance s_i par rapport au centre, mesurée dans l'espace des composantes principales. Dans l'exemple avec 1 seule composante principale, s_i est mesurée dans la direction de PC1. La distance s_i est désignée comme le **score** du spectre *i*.
- Le décalage e_i entre l'espace de la composante principale et le spectre. La distance s_i est désignée comme le **résidu** du spectre *i*.

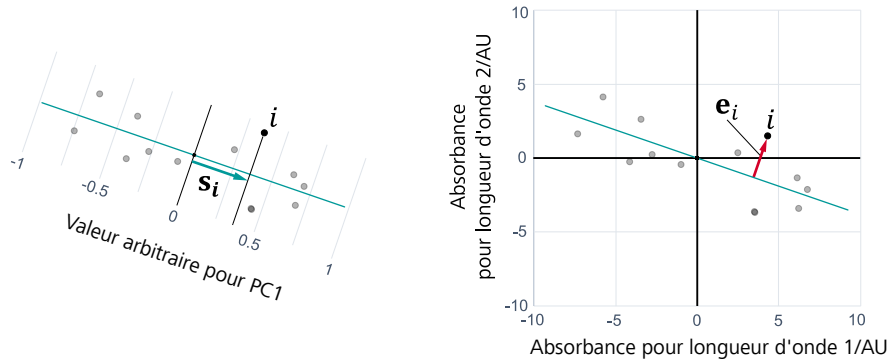


Figure 18 Spectre i avec score (gauche) et résidu (droite).

Le score s_i est mesuré dans l'espace de la composante principale. Le résidu e_i est mesuré dans l'espace original de la longueur d'onde.

Conversion en plusieurs composantes principales

Généralement, une description raisonnable des données spectroscopiques nécessite plus d'une composante principale.

Sur la *figure 19* sont représentées 3 variables originales x_1 x_2 et x_3. Chaque point représente un spectre avec 3 longueurs d'onde.

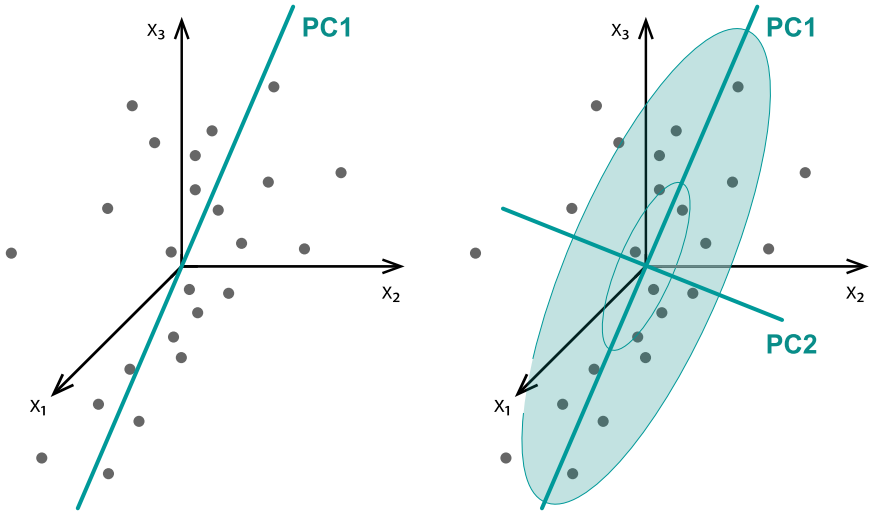


Figure 19 3 variables originales sont réduites à 1 composante principale (gauche) ou 2 composantes principales (droite). PC1 et PC2 constituent un espace de composantes principales à 2 dimensions.

La première composante principale PC1 est ici encore la direction à travers les données qui explicite la variance maximale.

La deuxième composante principale PC2 est la direction à travers les données qui explicite la variance résiduelle maximale. Il en va de même pour toutes les composantes principales suivantes, qui décrivent chacune la variance résiduelle maximale restante. Les quelques premières composantes principales représentent donc la plus grande partie de la variance dans les données, tandis que les autres contiennent principalement du bruit de fond et peuvent être rejetés. De cette manière, il est possible de réduire le nombre de variables.

Une caractéristique essentielle de l'ACP est que tous les composantes principales sont **orthogonales** (à angle droit) les unes par rapport aux autres. Les scores ne sont pas corrélés.

Distance de Mahalanobis

Comme illustré ci-dessus, le score d'un spectre *i* est mesuré dans l'espace des composantes principales, tandis que le résidu est mesuré dans l'espace des longueurs d'onde initial.

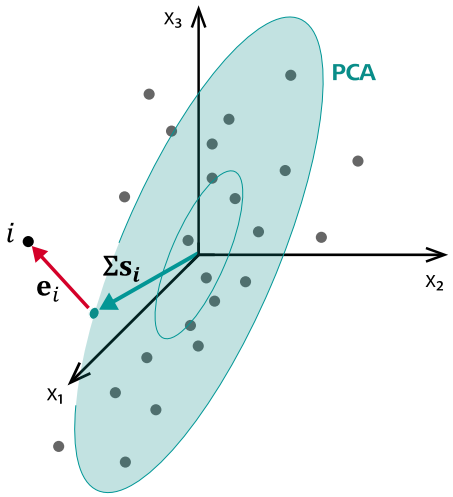
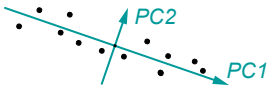


Figure 20 Score et résidu du spectre i. Le point vert est la projection orthogonale du point i (qui représente le spectre i) sur l'espace de composantes principales.

Dans la *figure 20*, le vecteur-score $\mathbf{\Sigma s}_i$ représente la distance absolue (distance euclidienne) du point central du modèle ACP à la projection orthogonale du spectre sur l'espace des composantes principales.



Dans l'exemple, les distances euclidiennes des spectres sont plus éloignées dans la direction PC1 que dans la direction PC2. La dispersion peut être mesurée comme **variance**. La variance dans PC1 est supérieure à celle dans PC2.

Le vecteur-score normalisé $\mathbf{s}_i$ représente une distance normalisée appelée la **distance de Mahalanobis**. La distance de Mahalanobis tient compte de la variance différente dans les différentes directions des composantes principales. Chaque direction a la même pondération. Par conséquent, une petite distance euclidienne dans une direction à faible variance peut compter autant qu'une grande distance euclidienne dans une direction à variance plus élevée.

Conversion de spectres à différentes longueurs d'onde

Les mêmes concepts s'appliquent à la conversion de spectres avec une variété de variables de longueurs d'onde dans les composantes principales. Sur la *figure 21*, chaque spectre est représenté par une courbe (à gauche) et un point (à droite).

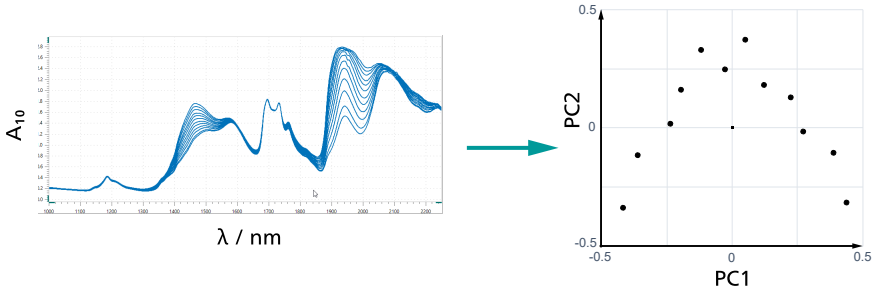


Figure 21 Conversion des données spectrales dans un espace de composantes principales. Les scores à droite sont exprimés en unités arbitraires.

La figure de droite montre les 2 premières composantes principales PC1 et PC2. Les composantes principales suivantes PC3, PC4, etc. peuvent être visualisées de la même manière.

Un modèle ACP utilise un nombre fixe de composantes principales. Plus les composantes principales sont nombreuses, plus le modèle explique les variations spectrales pertinentes. Simultanément, le modèle acquiert aussi des variations spectrales (bruit de fond) moins pertinentes. Un compromis équilibré est nécessaire.

Si le logiciel OMNIS exécute une analyse en composantes principales, leur nombre doit être sélectionné de manière à ce que la variance expliquée représente au moins 95 %.

Algorithme ACP

Il existe plusieurs possibilités de conversion des données originales dans un espace de composantes principales. Le logiciel OMNIS exécute une décomposition en valeurs singulières *(voir "Algorithme ACP", Chapitre 6.2, page 93)*.

4.3 Préparation des données

4.3.1 Prétraitement des données

Les modèles spectroscopiques s'appuient sur la relation entre les valeurs d'absorbance et les paramètres d'intérêt (quantification) ou l'appartenance du produit (identification, vérification). Le **paramétrage** des spectres garantit que les spectres expriment bien cette relation. L'objectif est d'éliminer la variance non pertinente sans perdre les informations utiles. Les artefacts et les non-linéarités sont corrigés. Un paramétrage correctement réalisé augmente l'exactitude et la robustesse du modèle ainsi que la répétabilité et la reproductibilité des prédictions.

Le paramétrage est appliqué à l'ensemble de données de calibrage, à l'ensemble de données validées et à l'ensemble de données atypiques,

ainsi qu'à tous les échantillons futurs inconnus et analysés avec le même modèle.

Le **prétraitement des données** est la première étape du paramétrage. Le prétraitement des données s'effectue dans l'ordre indiqué. Les gammes de longueurs d'onde pertinentes peuvent être définies dans une deuxième étape du paramétrage *(voir "Gammes de longueurs d'onde", Chapitre 4.3.2, page 49)*.

Réduction du bruit de fond

Les spectres peuvent contenir différents types de variations aléatoires autour du signal. Il s'agit, par exemple, du bruit de fond à haute fréquence dû au détecteur et aux circuits électroniques de l'appareil ou du bruit de fond à basse fréquence causé par la dérive de l'appareil pendant les mesures de balayage.

Le spectromètre fournit un spectre dont la moyenne est calculée à partir d'une série de mesures individuelles. Le bruit de fond à haute fréquence est ainsi considérablement réduit. Un filtre de lissage permet également d'obtenir une réduction du bruit de fond supplémentaire. Ces filtres s'appuient sur l'idée que le bruit de fond est à haute fréquence et le signal à basse fréquence. Ils approximent le signal par les valeurs d'absorbance voisines et réduisent le bruit de fond en calculant la moyenne.

Correction de la dispersion

Par dispersion on entend la modification de la direction de la lumière provoquée par l'interaction avec l'échantillon. La lumière diffuse que le détecteur n'atteint pas entraîne des variations de la ligne de base dans les spectres.

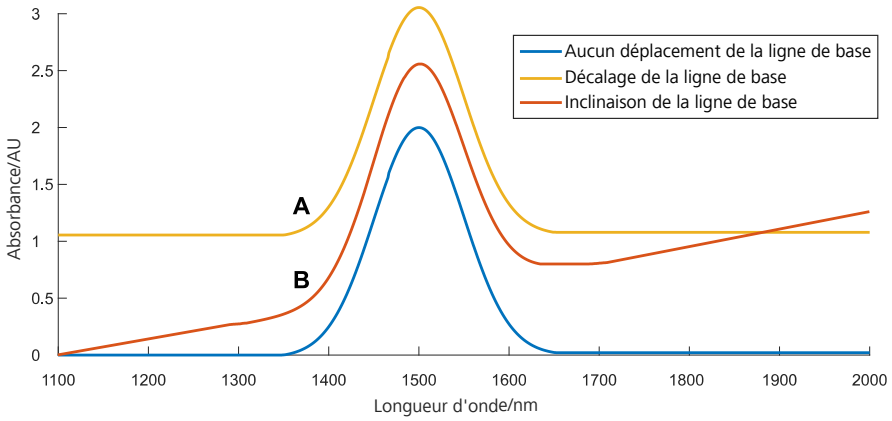


Figure 22 Les principaux types de déplacement de la ligne de base sont le décalage et l'inclinaison de la ligne de base.

On distingue différents types de déplacement de la ligne de base :

- Un facteur additif constant qui conduit à un **décalage de la ligne de base** (spectre **A**).

- Un facteur multiplicatif dépendant de la longueur d'onde qui conduit à une **inclinaison de la ligne de base** (spectre **B**).
- Un facteur multiplicatif dépendant de la longueur d'onde de second ordre qui conduit à une **inclinaison quadratique de la ligne de base**.
  Des inclinaisons de ligne de base d'ordre plus élevé peuvent également se produire.
- Des facteurs multiplicatifs dépendants de l'absorbance qui conduisent à une **amplification**. Des amplifications sont toutefois non pertinentes.

Les échantillons solides produisent une dispersion plus notable. Les déplacements de la ligne de base qui en résultent peuvent être utilisés pour détecter les changements de taille des particules ou d'autres variations physiques.

Si toutefois les variations chimiques ont un intérêt, les déplacements de la ligne de base doivent être réduits par un prétraitement adapté.

Prétraitements des données

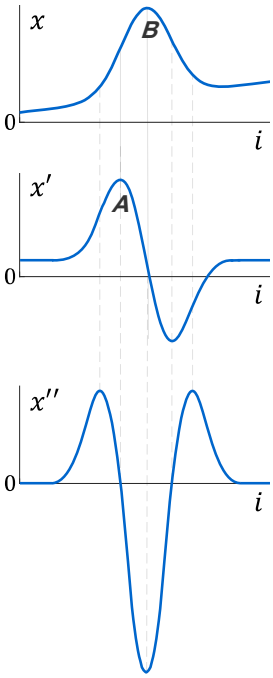
Tous les prétraitements des données suivants sont appliqués sur un seul spectre. Aucun autre spectre n'est pris en compte dans les calculs.

Le prétraitement des données modifie les valeurs du signal. Les chiffres sur l'axe y n'ont plus de signification.

Les prétraitements de données appliqués sont des conversions linéaires. Par conséquent, la loi de Beer-Lambert continue de s'appliquer.

4.3.1.1 Dérivées

Des dérivées peuvent être obtenues avec le filtre Gap-Segment ou le filtre Savitzky-Golay.



La dérivée d'un spectre décrit la pente de la courbe en chaque point. La pente est la vitesse de variation du spectre de sortie.

Dans le spectre, x_i est l'absorbance pour la longueur d'onde *i*. La dérivée de premier ordre x_i' représente la pente du spectre pour la longueur d'onde *i*. Là où la pente du spectre de sortie est la plus élevée, la dérivée de premier ordre présente un maximum (**A**). Au sommet d'un pic du spectre de sortie (**B**), la dérivée de premier ordre est égale à 0.

La dérivée de premier ordre élimine les décalages de la ligne de base et convertit les inclinaisons de la ligne de base en décalages.

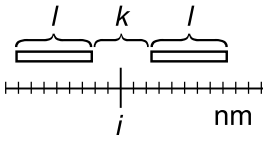
La dérivée de second ordre x_i'' correspond à la pente de la dérivée de premier ordre pour la longueur d'onde *i*. Les pics positifs dans le spectre original (**B**) deviennent des pics négatifs et inversement.

La dérivée de second ordre élimine les décalages et les inclinaisons de la ligne de base issus du spectre original.

Il convient d'être prudent lorsque le niveau de bruit de fond du spectre original est important. Chaque dérivation dégrade considérablement le rapport signal/bruit de fond. C'est pourquoi les dérivées sont associées à une fonction de lissage par le filtre Gap-Segment ou le filtre Savitzky-Golay.

4.3.1.2 Gap-Segment

Le filtre Gap-Segment lisse le spectre. En option, le filtre Gap-Segment exécute une dérivée de premier ou de second ordre. Le calcul dépend de l'utilisation ou non de dérivées :



- **Dérivation d'ordre 0** : pour chaque longueur d'onde *i*, le filtre Gap-Segment calcule la valeur moyenne à partir de 2 segments dont la taille de segment est *l*, p. ex. 10 nm. Les 2 segments sont séparés par une distance de grandeur *k*, p. ex. 5 nm.
- **Dérivation d'ordre 1** : les valeurs moyennes des 2 segments sont calculées séparément pour la dérivée de premier ordre. La différence entre les deux moyennes est ensuite calculée.
- **Dérivation d'ordre 2** : la dérivée de second ordre peut être calculée de la même manière à partir de la dérivée de premier ordre.

Au début et à la fin du spectre, les longueurs d'onde $l + k/2$ sont calculées, en utilisant des valeurs nulles pour les longueurs d'onde des segments situés en dehors du spectre.

Au début et à la fin du spectre, des valeurs nulles sont utilisées pour les longueurs d'onde des segments situés en dehors du spectre.

Le lissage peut entraîner un léger décalage des pics et un certain biais.

Paramétrages

Un plus grand lissage est obtenu par :

- une dérivée d'ordre inférieur,
- une taille du segment plus grande,
- un espace entre les segments plus grand.

Un lissage excessif entraîne une perte de variance pertinente, ce qui réduit la capacité prédictive du modèle.

4.3.1.3 Savitzky-Golay

Comme le filtre Gap-Segment, le filtre Savitzky-Golay lisse le spectre et effectue éventuellement une dérivée de premier ou de second ordre. Le filtre Savitzky-Golay utilise toutefois une autre méthode de lissage.

Pour la longueur d'onde *i*, le filtre Savitzky-Golay ajuste un polynôme d'ordre inférieur dans la gamme de longueur d'onde concernée. La valeur que prend le polynôme à la longueur d'onde *i*, est la valeur lissée. Si on effectue la dérivation, c'est la valeur de la dérivée qui est utilisée.

Une somme pondérée de valeurs voisines calcule tout en une fois :

$$x_i = \sum_{j=-k/2}^{k/2} c_j x_{i+j}$$

où *k* est la bande passante, les c_j sont des coefficients de déconvolution qui, consultables dans des tableaux, dépendent de l'ordre de dérivation, du degré du polynôme et de la bande passante, et les x_{i+j} sont les valeurs d'absorbance du spectre de sortie pour la longueur d'onde *i+j*.

Au début et à la fin du spectre, des valeurs extrapolées sont utilisées pour les longueurs d'onde du filtre situées en dehors du spectre (extrapolation horizontale).

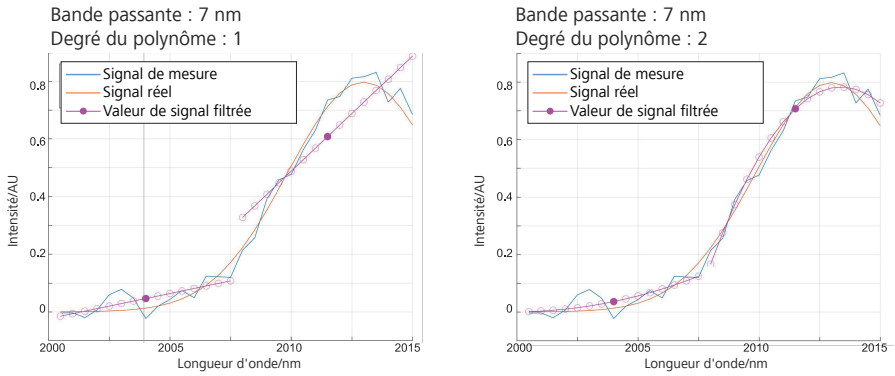


Figure 23 Filtrage Savitzky-Golay à différents degrés du polynôme

La *figure 23* représente le filtrage Savitzky-Golay. La bande passante est de 7 nm. Les polynômes sont représentés pour les longueurs d'onde 2 004 nm et 2 011,5 nm. Les nouvelles valeurs sont les points remplis ou, si l'on utilise des dérivées, leur dérivée. Toutes les autres longueurs d'onde sont traitées de la même manière.

La bande passante définit la gamme de longueurs d'onde dans laquelle chaque polynôme est ajusté. La convolution est pondérée de façon à ce que l'influence des valeurs d'absorbance diminue de part et d'autre de la longueur d'onde correspondante.

Paramétrages

Un plus grand lissage est obtenu par :

- une dérivée d'ordre inférieur,
- une plus grande bande passante,
- un degré du polynôme inférieur.

Un lissage excessif entraîne une perte de variance pertinente, ce qui réduit la capacité prédictive du modèle.

4.3.1.4 SNV – Standard Normal Variate

La SNV normalise un spectre unique à la variance 1 et la valeur moyenne 0. La SNV normalise les valeurs d'absorbance x_i pour chaque longueur d'onde i dans une gamme de longueurs d'onde spécifiée comme suit :

$$x_i = \frac{x_i - m}{s}$$

où m correspond à la valeur moyenne et s à l'écart-type de toutes les valeurs d'absorbance dans la gamme de longueurs d'onde définie.

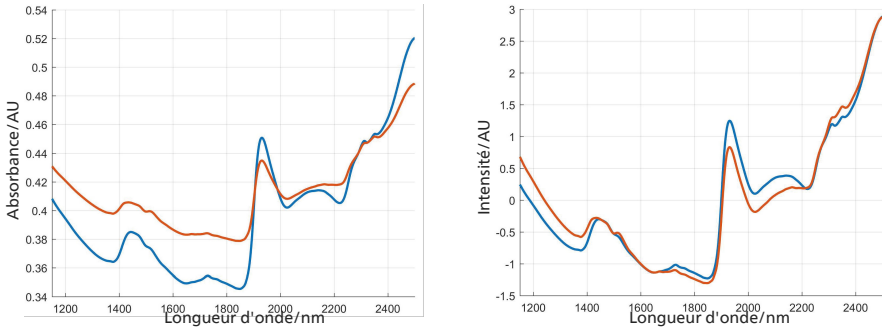


Figure 24 Spectres d'absorption (gauche) et spectres traités avec la SNV (droite).

La normalisation permet d'éliminer la variance entre les spectres. Cela est utile si la variance résulte de propriétés qui ne présentent pas d'intérêt, par exemple des épaisseurs de couche différentes dans des échantillons granuleux ou poudreux ou dans des milieux troubles.

Remarque : si une dérivée est appliquée après la SNV, la variance éliminée peut réapparaître partiellement. Les dérivées doivent donc être appliquées avant la SNV. Si, exceptionnellement, la SNV doit être effectuée avant la dérivée, l'ordre suivant peut être envisagé : detrend (tendance), SNV, dérivée. L'ordre courant est toutefois : dérivée, SNV, detrend *(voir "Ordre en*

cas de plusieurs étapes de prétraitement des données", Chapitre 4.3.1.7, page 48).

Paramètres

- **Gammes de longueurs d'onde**
  Si des artefacts compromettent certaines gammes de longueurs d'onde (p. ex. par saturation ou fort bruit de fond), elles peuvent être exclues. Seules les gammes de longueurs d'onde définies sont prises en compte dans les calculs de la valeur moyenne et de l'écart-type. La normalisation qui suit est exécutée à la fois sur les gammes de longueurs d'onde définies et dans les gammes intermédiaires. Pour les longueurs d'ondes exclues au début du spectre, la valeur normalisée de la longueur d'onde initiale voisine est appliquée. Pour les longueurs d'ondes exclues à la fin du spectre, la valeur normalisée de la longueur d'onde finale voisine est appliquée.
  Au besoin, les longueurs d'onde exclues peuvent également être exclues pour le calcul du modèle *(voir "Gammes de longueurs d'onde", Chapitre 4.3.2, page 49)*.
  Remarque : à partir de la version logicielle d'OMNIS 4.6, plusieurs gammes de longueurs d'onde peuvent être définies.

4.3.1.5 Detrend

Detrend utilise la méthode des moindres carrés pour ajuster un polynôme de second ordre à l'ensemble du spectre. Detrend soustrait ensuite le polynôme du spectre.

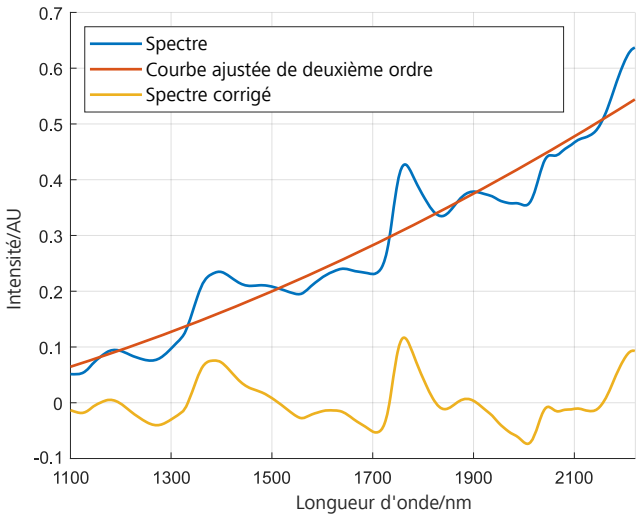


Figure 25 Detrend convertit le spectre bleu en spectre jaune.

Detrend réduit les effets de dispersion liés la longueur jusqu'aux inclinaisons quadratiques de la ligne de base.

La figure ci-dessus présente un spectre (bleu) dans lequel la tendance domine. Si cette tendance dominante est similaire pour tous les spectres,

detrend peut fonctionner correctement. Dans les autres cas, detrend tend à supprimer les variations utiles. Pour ces cas, les dérivées restent vraisemblablement le meilleur choix.

Un polynôme distinct étant adapté à chaque spectre, une variance perturbatrice supplémentaire peut apparaître. Normalement, on applique la SNV avant detrend. Cela permet d'obtenir des valeurs estimées des coefficients du polynôme plus robustes.

Paramètres

- **Gammes de longueurs d'onde**
  Si des artefacts compromettent certaines gammes de longueurs d'onde (p. ex. par saturation ou fort bruit de fond), elles peuvent être exclues. Un polynôme est ajusté à toutes les valeurs d'intensité des gammes de longueurs d'onde définies. Ensuite, dans toutes les gammes de longueurs d'onde définies, le polynôme est soustrait du spectre. La valeur d'intensité est mise à zéro pour toutes les longueurs d'onde exclues.
  Au besoin, les longueurs d'onde exclues peuvent également être exclues pour le calcul du modèle *(voir "Gammes de longueurs d'onde", Chapitre 4.3.2, page 49)*.
  Remarque : à partir de la version logicielle d'OMNIS 4.6, plusieurs gammes de longueurs d'onde peuvent être définies.

4.3.1.6 Aperçu du prétraitement des données

Prétraitement	Destination	Effets positifs	Effets négatifs
Gap-Segment	Lissage Un plus grand lissage est obtenu avec un ordre de dérivation inférieur, une taille de segment plus grande ou un espace entre les segments plus grand.	▪ Réduction du bruit de fond à haute fréquence.	▪ Un lissage excessif entraîne une perte de variance pertinente.
Dérivées avec Gap-Segment	Correction de la ligne de base	▪ Dérivée de premier ordre : élimine les décalages de la ligne de base. ▪ Dérivée de second ordre : élimine les décalages et les inclinaisons de la ligne de base.	▪ Renforce le bruit de fond. ▪ Modifie l'apparence du spectre.

Prétraitement	Destination	Effets positifs	Effets négatifs
Savitzky-Golay	Lissage Un plus grand lissage est obtenu avec un ordre de dérivation inférieur, une bande passante plus grande ou un degré du polynôme inférieur.	▪ Réduction du bruit de fond à haute fréquence.	▪ Un lissage excessif entraîne une perte de variance pertinente.
Dérivées dans Savitzky-Golay	Correction de la ligne de base	▪ Dérivée de premier ordre : élimine les décalages de la ligne de base. ▪ Dérivée de second ordre : élimine les décalages et les inclinaisons de la ligne de base.	▪ Renforce le bruit de fond. ▪ Modifie l'apparence du spectre.
SNV – Standard Normal Variate	Correction de la dispersion *	▪ Élimination des décalages de la ligne de base.	▪ La variance pertinente peut le cas échéant être supprimée.
Detrend	Correction de la dispersion *	▪ Élimination des décalages de la ligne de base. ▪ Élimine les inclinaisons et les inclinaisons quadratiques de la ligne de base.	▪ La variance pertinente peut le cas échéant être supprimée. ▪ Cela peut entraîner des variances non pertinentes.

* Remarque : les gammes de longueurs d'onde avec artefacts (p. ex. saturation ou fort bruit de fond), devraient être exclues.

4.3.1.7 Ordre en cas de plusieurs étapes de prétraitement des données

Si on utilise plusieurs étapes de prétraitement des données, l'ordre peut être déterminant. La règle de base est la suivante.

Utiliser de préférence Gap-Segment ou Savitzky-Golay avant SNV et SNV avant detrend.

Exemple d'une dérivée de premier ordre et SNV : la dérivée de premier ordre transforme les pentes de la ligne de base en décalages. Une SNV subséquente élimine ces décalages. Si l'ordre est inversé, la SNV ne modifie pas les pentes de la ligne de base. La dérivée de premier ordre qui en résulte les transforme en décalages de la ligne de base. Les décalages resteraient.

Exemple d'une dérivée de second ordre et SNV : les décalages et pentes de la ligne de base sont éliminés dans tous les cas. Néanmoins, l'application de la dérivée de second ordre et de la SNV dans le bon ordre permet d'éliminer les pentes quadratiques de la ligne de base. La dérivée de second ordre transforme les pentes quadratiques de la ligne de base en décalages. Une SNV subséquente élimine ces décalages. Si l'ordre est inversé, la SNV ne modifie pas les pentes quadratiques de la ligne de base. La dérivée de second ordre qui en résulte les transforme en décalages de la ligne de base. Les décalages resteraient.

4.3.2 Gammes de longueurs d'onde

Après le prétraitement des données *(voir "Prétraitement des données", Chapitre 4.3.1, page 40)*, la deuxième étape concerne le paramétrage. La sélection de gammes de longueurs d'onde permet d'exclure les gammes qui ne conviennent pas à la finalité. Des gammes de longueurs d'onde bruyantes ou saturées notamment peuvent compromettre les calculs suivants et doivent être exclues.

Bruit de fond

Un bruit de fond se produit en présence de valeurs d'absorbance élevées, lorsque seule une faible quantité de lumière atteint le détecteur. La figure suivante présente des gammes bruyantes.

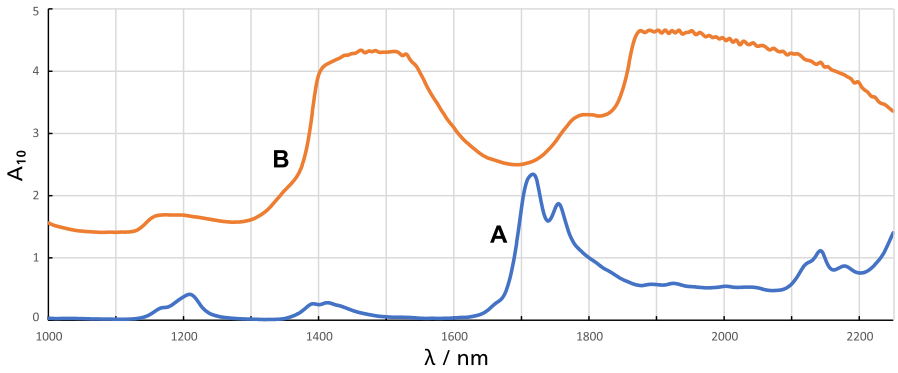


Figure 26 Exemple de gammes de longueurs d'onde bruyantes.

Le spectre **A** comporte des formes de pic normales. Le spectre **B** a 2 gammes bruyantes, une de 1 400 à 1 550 nm et une supérieure à 1 870 nm. Ces gammes sont extrêmement bruyantes et ne ressemblent pas à une courbe en cloche ni à une combinaison des deux.

Saturation

Une saturation du détecteur peut se produire si une grande quantité de lumière atteint le détecteur, c'est-à-dire lorsque les valeurs d'absorbance sont basses.

- **OMNIS NIR Analyzer**
  Le temps d'intégration est toujours réglé automatiquement. Cela évite la saturation et minimise le bruit de fond *(voir « Temps d'intégration », page 9)*.
- **2060 The NIR**
  Si le temps d'intégration automatique est activé, il n'y a pas de saturation.
  Si le temps d'intégration manuel est activé, des temps d'intégration trop longs peuvent entraîner la saturation du détecteur. Des gammes saturées se produisent lorsque les valeurs d'absorbance sont basses, mais elle ne sont pas toujours faciles à détecter visuellement. Le temps d'intégration manuel doit, par conséquent, être réglé avec une marge suffisante *(voir « Temps d'intégration », page 9)*.

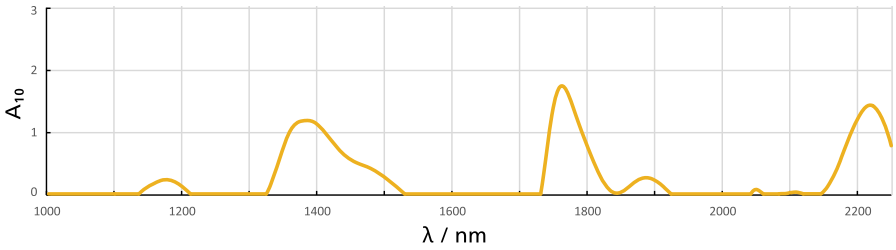


Figure 27 Exemple de gammes de longueurs d'onde saturées.

Autres raisons

Il existe d'autres raisons d'inclure ou d'exclure des gammes de longueurs d'onde. La sélection peut s'appuyer sur la connaissance des paramètres d'intérêt et de leurs bandes d'absorption respectives *(voir "La lumière et son interaction avec les matières", Chapitre 2.1, page 3)*. Dans ce cas, il faut toutefois tenir compte du fait que, selon le type de prétraitement, les informations pertinentes peuvent éventuellement être déplacées vers d'autres gammes de longueurs d'onde.

Si des anomalies ont été introduites au début et à la fin des spectres lors du prétraitement des données, les longueurs d'onde correspondantes peuvent être exclues.

Les variations des composants chimiques ou des conditions ambiantes peuvent influencer certaines gammes de longueurs d'onde. L'exclusion de celles-ci peut améliorer la robustesse du modèle.

Il convient d'être prudent lors de l'exclusion de gammes de longueurs d'onde bien formées. Des gammes ne contenant apparemment aucune information peuvent en réalité fournir des informations dissimulées et importantes. Elles peuvent contribuer à la détection de valeurs atypiques ou au traitement de bandes d'absorption interférentes. En fait, les bandes d'absorption interférentes sont la principale raison de la réalisation de mesures multivariées *(voir "Exemple d'une régression linéaire", Chapitre 6.1, page 89)*.

4.3.3 Valeur spectrale atypique

Un spectre qui se distingue de la plupart des autres spectres est désigné comme valeur spectrale atypique.

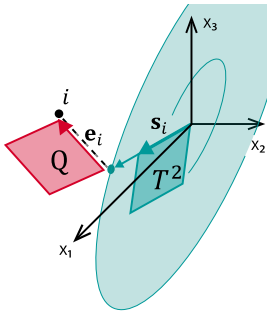
Les valeurs atypiques doivent être vérifiées avec soin. Une valeur atypique peut biaiser le modèle si, par exemple, un échantillon contaminé ou une erreur de mesure en est la cause. Dans ce cas, la valeur atypique ne doit pas être prise en compte lors du calcul du modèle.

D'autre part, une valeur atypique peut représenter des propriétés qui ne sont pas bien couvertes par les autres spectres. Dans ce cas, la valeur atypique peut même améliorer le modèle. Si la valeur atypique semble être un échantillon valide, il convient de vérifier si les échantillons de calibrage sont répartis uniformément sur l'intervalle de variation.

Les T^2 de Hotelling et résidus Q sont des mesures appropriées pour la détection des mesures atypiques.

T² de Hotelling et résidus Q

Lors de la conversion de données spectroscopiques dans un espace de composants principaux, les spectres peuvent être caractérisés par leurs scores et résidus *(voir "Analyse en composantes principales (ACP)", Chapitre 4.2, page 36)*. Cela s'applique également à la conversion dans un espace de variables latentes *(voir "Régression MCP", Chapitre 4.4.1, page 61)*.



Exemple : un espace de longueur d'onde en 3 dimensions (x_1, x_2, x_3) est converti en espace en 2 dimensions (vert). Le point *i* représente le spectre *i* et est projeté d'un espace tridimensionnel dans un espace bidimensionnel. Il en découle ce qui suit :

- un vecteur de score $\Sigma \mathbf{s}_i$ dans l'espace bidimensionnel ou son vecteur de score normalisé $\mathbf{s}_i$, qui représente la distance de Mahalanobis ;
- un résidu $\mathbf{e}_i$ dans l'espace tridimensionnel.

Les quantités suivantes peuvent être dérivées de $\mathbf{s}_i$ et $\mathbf{e}_i$ *(voir "T² de Hotelling et résidus Q", Chapitre 6.4, page 96)* :

- Le T^2 de **Hotelling** ou simplement $\boldsymbol{T^2}$ est la distance de Mahalanobis au carré, c'est-à-dire le carré de la distance normalisée entre le centre du modèle et la projection orthogonale du spectre sur l'espace des composants principaux ou l'espace des variables latentes.
  Si tous les scores d'un spectre correspondent à la valeur moyenne, T^2 est égal à 0 et le spectre se situe au point central du modèle. C'est près du point central que le modèle s'adapte le mieux.
  Loin du point central, le modèle risque de mal s'adapter. Les valeurs T^2 sont élevées. Une valeur T^2 élevée indique un spectre extrême qui, par exemple, représente un échantillon avec une composition extrême de composants chimiques.

- Le **résidu Q** est le résidu au carré, c'est-à-dire la distance orthogonale au carré entre le spectre et l'espace des composants principaux ou l'espace des variables latentes.
  Les résidus Q indiquent les variations qui ne sont pas expliquées par le modèle. Un résidu Q élevé indique que le spectre risque de ne pas correspondre au modèle, par exemple, si l'échantillon mesuré contient une autre substance.

Détection de valeurs spectrales atypiques

La détection de valeurs spectrales atypiques identifie des spectres qui s'écartent de l'intégralité.

1. Le paramétrage est considéré comme suit :
   a. À partir de la version logicielle d'OMNIS 4.2 : l'utilisateur décide si le paramétrage (prétraitement des données et sélection des longueurs d'onde) doit être appliqué ou non.
   Les modifications ultérieures du paramétrage n'ont aucune incidence sur la répartition des ensembles de données.
   b. À partir de la version logicielle d'OMNIS 3.3 jusqu'à la version 4.1 : l'utilisateur décide si le prétraitement des données doit être pris en compte ou non. La sélection des longueurs d'onde et les modifications ultérieures du prétraitement des données n'ont aucune incidence sur la répartition des ensembles de données.
   c. Jusqu'à la version logicielle d'OMNIS 3.2 : le prétraitement des données est pris en compte tel qu'il a été défini au moment de la détection des valeurs atypiques. La sélection des longueurs d'onde et les modifications ultérieures du prétraitement des données n'ont aucune incidence sur la répartition des ensembles de données.
2. La détection des valeurs spectrales atypiques s'appuie sur le modèle ACP de tous les spectres centrés sur les valeurs moyennes *(voir "Analyse en composantes principales (ACP)", Chapitre 4.2, page 36)*. Le nombre de composants principaux est choisi de manière à ce que la variance expliquée soit d'au moins 95 %.
3. Pour détecter les valeurs spectrales atypiques, on utilise les valeurs de T^2 de Hotelling et les résidus Q des spectres. L'algorithme évalue si le T^2 de Hotelling ou le résidu Q du spectre testé est le résultat d'une variation aléatoire ou systématique.
   Une description de l'algorithme est disponible en annexe *(voir "Valeur spectrale atypique - algorithme", Chapitre 6.5, page 98)*.

4.3.3.1 Graphique d'influence

Le graphique d'influence indique les propriétés de base des spectres et aide à l'analyse des valeurs spectrales atypiques.

Le fondement du graphique d'influence est un modèle ACP *(voir "Analyse en composantes principales (ACP)", Chapitre 4.2, page 36)* ou un modèle MCP :

- **Quantification** : le graphique d'influence s'appuie au choix sur **ACP** ou **MCP**.
  La régression MCP réduit comme l'ACP les données spectroscopiques à un nombre plus restreint de variables. Les MCP prennent également en compte les valeurs de référence.
  Les composants principaux de l'ACP sont désignés comme **variables latentes** pour les MCP.
  (voir "Régression MCP", Chapitre 4.4.1, page 61)
- **Identification** : un graphique d'influence basé sur **ACP** est disponible (à partir de la version logicielle 4.3 d'OMNIS).

Types de valeurs spectrales atypiques

Le graphique d'influence fait apparaître pour chaque spectre les valeurs T^2 de Hotelling et le résidu Q *(voir « T^2 de Hotelling et résidus Q », page 51)*. Les T^2 de Hotelling et les résidus Q permettent d'identifier différents types de valeurs spectrales atypiques :

- **Valeur atypique T^2 de Hotelling**, également appelée valeur atypique par effet de levier (en anglais Leverage outlier) : un T^2 élevé signifie que la projection du spectre sur l'espace des composantes principales (ACP) ou l'espace des variables latentes (MCP) est éloignée du centre du modèle.
- **Valeurs atypiques résidus Q** : un résidu Q élevé signifie que le spectre du modèle est mal décrit.

La *figure 28* présente plusieurs spectres sous différentes vues :

- Graphique d'influence à gauche : les résidus Q décrivent les variations non expliquées par le modèle tandis que le T^2 de Hotelling prend en compte les variations au sein du modèle lui-même.
  Les lignes en pointillés indiquent les **valeurs critiques** ou les **valeurs limites** pour la pertinence définie *(voir "Valeur spectrale atypique - algorithme", Chapitre 6.5, page 98)*.
  Plus la pertinence est élevée, plus les valeurs limites sont basses, de sorte que davantage de points peuvent se situer en dehors des valeurs limites.

- À droite : exemple d'espace initial avec 3 variables x_1, x_2, x_3, transformé en un espace à 2 dimensions avec des variables latentes.
  Pour les points A à D, la distance orthogonale au plan (lignes pointillées) ainsi que le point modélisé dans l'espace des variables latentes (points verts) sont représentés.

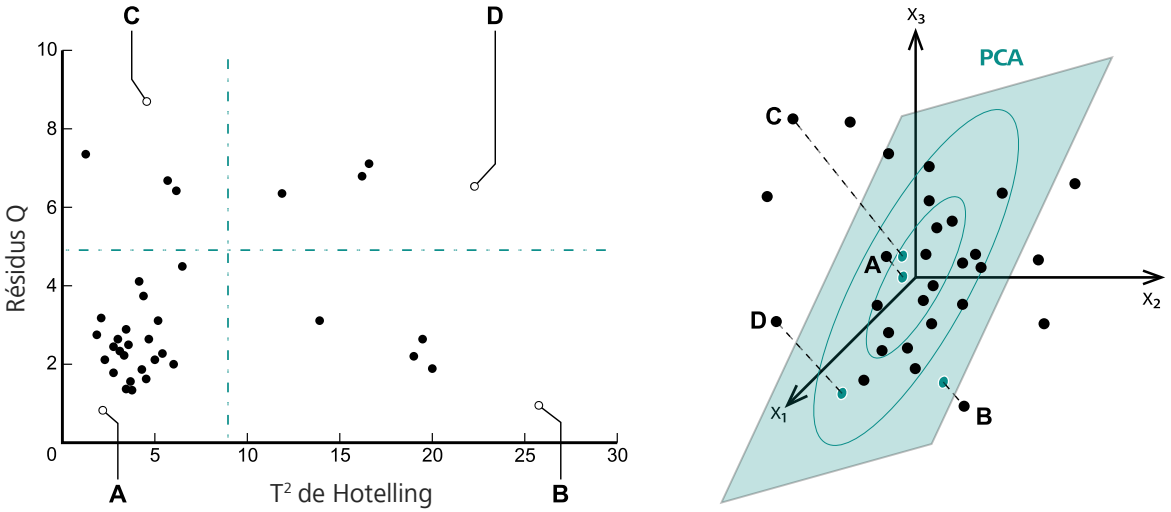


Figure 28 Graphique d'influence (gauche), espace initial et espace de variables latentes (droite). Chaque spectre est représenté par un point sur la figure de gauche et un point sur celle de droite.

Dans les deux vues, 4 points avec des caractéristiques différentes sont mis en évidence :

- Le spectre A a des scores et des résidus faibles. Il est proche du centre du modèle et est bien expliqué par le modèle.
- Le spectre B est une valeur atypique *T*² de Hotelling. Il est éloigné du point central, mais il est bien expliqué par le modèle.
- Le spectre C est une valeur atypique de résidus Q. Il est éloigné du point central et est mal expliqué par le modèle.
- Le spectre D est aussi bien une valeur atypique *T*² de Hotelling qu'une valeur atypique de résidus Q. Il est éloigné du point central et est expliqué partiellement par le modèle.

Le graphique d'influence montre comment plusieurs spectres influencent le modèle. Comme toutes les variables latentes passent par le point central, les spectres proches de celui-ci (par exemple le spectre A) ont peu de chance de changer la direction des variables latentes. Il n'ont pas d'effet de levier. Plus la distance par rapport au centre est grande, plus l'effet de levier et le potentiel d'influence sur le modèle sont importants. Certains spectres parviennent effectivement à attirer le modèle dans leur direction (spectre B), alors que pour d'autres, ce n'est que dans une certaine mesure (spectre D) ou pas du tout (spectre C).

Comparé à un modèle basé sur tous les spectres, le calcul d'un modèle sans spectre B modifiera probablement davantage le modèle qu'un modèle sans spectre D et encore plus qu'un modèle sans spectre C. Le spectre B influence probablement fortement le modèle de quantification - en bien ou en mal. La décision de supprimer ou non une valeur atypique potentielle dans le quadrant inférieur droit du graphique d'influence nécessite une attention particulière.

Dans le cas idéal, le modèle devrait acquérir la variance d'un grand nombre de spectres. Il n'est pas souhaitable que le modèle ne soit influencé que par quelques spectres. Dans la figure ci-dessus, quelques spectres sont très éloignés du centre et également de la plupart des autres spectres. Ce qui est suspect. Le modèle n'est influencé que par quelques spectres. Il s'agit donc de valeurs atypiques potentielles nécessitant un examen. En outre, il faut s'assurer que les échantillons sont répartis uniformément sur l'étendue de la variation.

Graphique d'influence ACP et MCP

Le graphique d'influence ACP dépend uniquement des spectres. Le graphique d'influence MCP dépend des spectres et des valeurs de référence.

Le tableau suivant indique l'effet de différents réglages sur les graphiques d'influence ACP et MCP.

	Graphique d'influence ACP	Graphique d'influence MCP
Spectres	Le modèle ACP sous-jacent est basé sur tous les spectres de l'ensemble de données de calibrage, de l'ensemble de données validées et de l'ensemble de données atypiques.	Le modèle MCP sous-jacent est basé sur tous les spectres de l'ensemble de données de calibrage. Sur la base de ce modèle MCP, les valeurs T^2 et de résidu Q des spectres sont calculées dans les 3 ensembles de données et représentées sur le graphique.
Paramétrage	Prend en compte les prétraitements de données et les gammes de longueurs d'onde sélectionnés. Remarque : la détection des valeurs atypiques s'appuie sur l'ACP et prend en compte le paramétrage selon les paramètres utilisateur et la version logicielle d'OMNIS *(voir « Détection de valeurs spectrales atypiques », page 52)*.	Prend en compte les prétraitements de données et les gammes de longueurs d'onde sélectionnés. Remarque : l'évaluation des valeurs atypiques lors de la prédiction est basée sur les MCP et prend en compte les prétraitements des données et les gammes de longueur d'onde.

	Graphique d'influence ACP	Graphique d'influence MCP
Nombre de variables	Utilise le nombre de composants principaux qui atteint une variance expliquée d'au moins **95 %**.	Utilise le nombre de variables latentes actuellement sélectionnées.
Pertinence et valeurs critiques	Utilise la pertinence actuellement sélectionnée pour calculer et visualiser les valeurs critiques (lignes pointillées). Identification : si la dernière répartition de l'ensemble de données a été réalisée sans détermination de valeurs atypiques, le graphique d'influence utilise une pertinence de 5 %. Remarque : une augmentation de la pertinence entraîne des valeurs critiques plus faibles et donc davantage de valeurs atypiques lors du développement des modèles.	Utilise la pertinence actuellement sélectionnée pour calculer et visualiser les valeurs critiques (lignes pointillées). Remarque : une augmentation de la pertinence entraîne des valeurs critiques plus faibles et donc davantage de valeurs atypiques lors de la prédiction.
Valeurs de référence (quantification)	Les valeurs de référence n'ont aucune influence sur le modèle ACP. Cependant, chaque spectre peut avoir une valeur de référence atypique associée et être ainsi marqué comme atypique.	Les valeurs de référence influencent le modèle MCP et donc le graphique d'influence MCP. En outre, chaque spectre peut avoir une valeur de référence atypique associée et être ainsi marqué comme atypique.

Analyse des valeurs atypiques

Les facteurs suivants doivent être considérés lors de l'analyse de valeurs atypiques potentielles.

- Une valeur atypique T^2 de Hotelling indique un échantillon dont la composition des composants chimiques est extrême par rapport aux autres échantillons.
- Une valeur atypique de résidus Q peut, par exemple, indiquer un échantillon contaminé ou une erreur lors de l'enregistrement de spectre.

Les valeurs atypiques potentielles devraient être examinées avec soin. Les véritables valeurs atypiques doivent être supprimées de la liste des spectres. Les échantillons valides doivent être conservés. Si l'ensemble de données est ensuite à nouveau réparti, la détection des valeurs atypiques trouve, le cas échéant, les valeurs atypiques potentielles qui n'ont pas été trouvées lors du premier passage. Une possibilité est que le nouveau modèle ACP nécessite moins de composantes principales pour atteindre la

variance expliquée de 95 %. Si les nouvelles valeurs atypiques détectées s'avèrent être des échantillons valides, elles doivent être conservées. Dans ce cas, la répartition automatique peut être répétée sans détection de valeurs atypiques.

4.3.3.2 Graphique des scores

Le fondement du graphique des scores est un modèle ACP ou un modèle MCP :

- **Quantification** : le graphique des scores (à partir de la version logicielle 3.0 d'OMNIS) s'appuie sur **MCP** *(voir "Régression MCP", Chapitre 4.4.1, page 61)*.
- **Identification** : le graphique des scores (à partir de la version logicielle 4.3 d'OMNIS) s'appuie sur **ACP** *(voir "Analyse en composantes principales (ACP)", Chapitre 4.2, page 36)*.

Chaque spectre possède une valeur de scores pour chaque composant principal ou variable latente. Chaque spectre est représenté par un point dans le graphique des scores. L'axe x représente, par ex., le score pour la première variable latente, l'axe y celui pour la deuxième variable latente. De la même manière, toute paire de variables latentes peut être affichée.

Comme les valeurs d'absorbance ont été centrées sur la moyenne pour chaque variable de longueur d'onde, les scores de chaque variable latente sont également centrés sur la moyenne. Un point proche du centre du graphique des scores (0/0) représente un spectre moyen par rapport aux deux variables latentes affichées. Les points proches les uns des autres représentent des spectres similaires, les points plus éloignés les uns des autres représentent des spectres dissemblables par rapport aux deux variables latentes affichées.

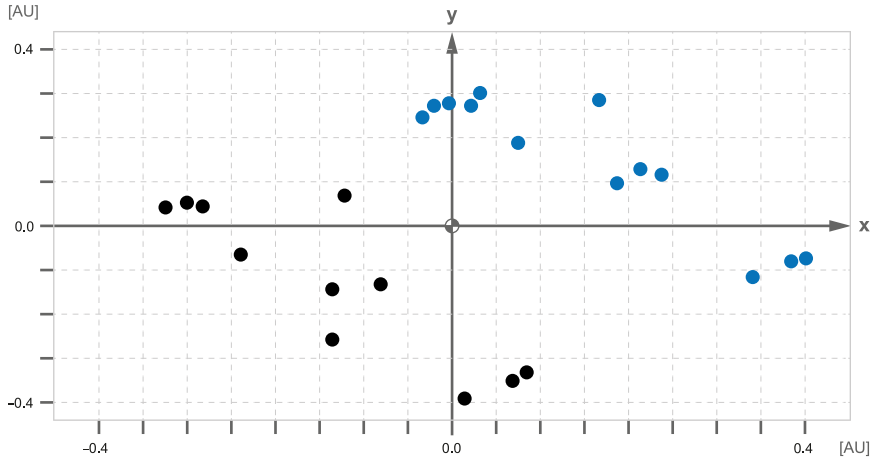


Figure 29 Graphique des scores pour la variable latente 1 (axe x) et la variable latente 2 (axe y). AU = unité arbitraire.

La *figure 29* présente un exemple de 2 jeux de données mesurées dans des conditions différentes. Les scores sont normalisés, chaque variable latente a la même pondération.

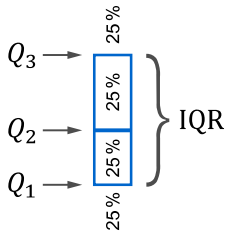
Les scores de tous les composants principaux ou de toutes les variables latentes d'un spectre peuvent être combinés en une valeur unique (T^2 de Hotelling), qui est affichée sur l'axe x du graphique d'influence MCP.

4.3.4 Valeur de référence atypique (quantification)

Dans les modèles de quantification, on détermine les valeurs spectrales atypiques et les valeurs de référence atypiques. Les valeurs de référence atypiques indiquent des anomalies dans la valeur de référence.

Généralement, les valeurs de référence atypiques sont des chiffres mal transmis, par exemple 143 au lieu de 14,3 ou 15,9 au lieu de 51,9. La détection des valeurs atypiques identifie ces erreurs de transfert ou de transcription en s'appuyant sur une approche empirique. Seules les erreurs évidentes sont identifiées pour un examen plus approfondi.

Box plots



Les valeurs de référence atypiques sont identifiées à l'aide d'une méthode basée sur les box plots. Un **box plot** trie les valeurs de référence dans l'ordre croissant. Les quartiles divisent le jeu de données en 4 parties. Chaque partie contient 25 % de valeurs de référence.

Le premier quartile Q_1 sépare les 25 % de valeurs les plus basses du reste. Q_2 est la médiane et sépare les 50 % de valeurs les plus basses du reste. Le troisième quartile Q_3 sépare les 75 % de valeurs les plus basses du reste. Un rectangle vertical représente les 50 % centraux des données, l'intervalle interquartile (IQR, en anglais interquartile range).

Les données se situant hors du box IQR d'un certain montant sont considérées comme des valeurs potentiellement atypiques et peuvent être représentées par de petits cercles. Les valeurs limites inférieure et supérieure des valeurs atypiques sont souvent définies comme 1,5 fois l'IQR :

$$[Q_1 - 1.5\ \mathrm{IQR}\,;\, Q_3 + 1.5\ \mathrm{IQR}]$$

Dans ce cas, Q_1 représente le premier quartile, Q_3 le troisième et IQR l'intervalle interquartile (Q_3 - Q_1).

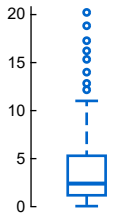
Pour compléter le box plot, les moustaches au-dessus et en dessous du box s'étendent jusqu'aux points les plus éloignés qui ne sont pas identifiés comme des valeurs atypiques potentielles.

Ajustement pour les distributions asymétriques

Le box plot courant suppose que la distribution des données est à peu près symétrique. Généralement pour les distributions biaisées beaucoup

de valeurs de référence régulières sont indiquées valeurs atypiques potentielles. C'est pourquoi le logiciel OMNIS utilise une version adaptée du box plot qui prend en compte l'asymétrie de la distribution.

Dans l'exemple suivant, la distribution est décalée vers les valeurs de référence les plus élevées. Il en résulte 8 valeurs atypiques élevées. Après ajustement des valeurs atypiques limites, il n'en reste plus que 2. De l'autre côté, une nouvelle valeur atypique faible apparaît.

Distribution asymétrique	Valeurs atypiques limites ajustées
	

Une description de l'ajustement est disponible en annexe *(voir "Valeur de référence atypique - algorithme", Chapitre 6.6, page 100)*.

4.3.5 Répartition des ensembles de données

L'ensemble de données consiste en spectres et valeurs de référence (quantification) ou en spectres et noms de produit correspondants (identification, vérification). L'ensemble de données permet de développer et de valider le modèle. L'ensemble de données doit donc être réparti dans un ensemble de données de calibrage, un ensemble de données validées et un ensemble de données atypiques. La répartition peut être effectuée manuellement ou automatiquement.

En supposant qu'il y ait suffisamment de spectres disponibles, on peut par exemple utiliser entre 20 et 30 % des données totales pour l'ensemble de données validées.

Algorithme de répartition automatique

La répartition automatique de l'ensemble de données garantit que l'ensemble de données de calibrage et l'ensemble de données validées sont représentatifs de la population et indépendants l'un de l'autre. L'objectif est de répartir les données en deux ensembles qui couvrent approximativement la même gamme dans l'espace ACP et présentent des propriétés statistiques similaires. L'algorithme utilisé est un algorithme duplex légèrement modifié tiré de R. D. Snee, *Validation of Regression Models: Methods and Examples*, Technometrics Vol. 19, N° 4 (Nov. 1977), p. 415-428.

Réplicats

Il ne doit y avoir aucun réplicat dans l'ensemble de données. L'algorithme ne supprime pas les réplicats ou les pseudo-réplicats et ne force pas l'ajout de réplicats d'un échantillon donné au même ensemble de données.

1. Le paramétrage est considéré comme suit :
   a. À partir de la version logicielle d'OMNIS 4.2 : l'utilisateur décide si le paramétrage (prétraitement des données et sélection des longueurs d'onde) doit être appliqué ou non. Les modifications ultérieures du paramétrage n'ont aucune incidence sur la répartition des ensembles de données.
   b. À partir de la version logicielle d'OMNIS 3.3 jusqu'à la version 4.1 : l'utilisateur décide si le prétraitement des données doit être pris en compte ou non. La sélection des longueurs d'onde et les modifications ultérieures du prétraitement des données n'ont aucune incidence sur la répartition des ensembles de données.
   c. Jusqu'à la version logicielle d'OMNIS 3.2 : le prétraitement des données est pris en compte tel qu'il a été défini au moment de la détection des valeurs atypiques. La sélection des longueurs d'onde et les modifications ultérieures du prétraitement des données n'ont aucune incidence sur la répartition des ensembles de données.
2. La répartition s'appuie sur le modèle ACP de tous les spectres centrés sur les valeurs moyennes. Le nombre de composants principaux est choisi de manière à ce que la variance expliquée soit d'au moins 95 %.
3. Les distances entre toutes les paires de spectres possibles sont calculées à partir des scores ACP.
4. Les 2 spectres les plus éloignés l'un de l'autre sont attribués à l'ensemble de données de calibrage.
5. Parmi les spectres restants, les 2 plus éloignés sont attribués à l'ensemble de données validées.
6. Parmi les spectres restants, celui qui est le plus éloigné des spectres déjà contenus dans l'ensemble de données de calibrage est attribué à l'ensemble de données de calibrage.
7. Parmi les spectres restants, celui qui est le plus éloigné des spectres déjà contenus dans l'ensemble de données validées est attribué à l'ensemble de données validées.
8. Le changement se poursuit jusqu'à ce qu'un des ensembles de données ait atteint sa taille prévue. Les spectres restants sont affectés à l'autre ensemble de données.

4.4 Quantification

4.4.1 Régression MCP

Le logiciel OMNIS utilise la régression MCP pour calculer les modèles de quantification.

Comme l'ACP, la **régression des moindres carrés partiels** (**Régression MCP**, en anglais *partial least squares regression*) réduit les données spectroscopiques à un nombre plus restreint de variables. Cependant :

- Les composants principaux de l'ACP sont désignés comme **variables latentes** pour les MCP. Les directions des variables latentes et des composantes principale sont généralement similaires, mais pas identiques.
- Outre les spectres, la régression MCP prend en compte également les **valeurs de référence**. Par conséquent, moins de variables latentes sont nécessaires pour obtenir une corrélation suffisante avec les valeurs de référence, ce qui entraîne moins de bruit de fond.

À partir des données spectroscopiques volumineuses et hautement redondantes, la MCP extrait les **variables latentes** sous-jacentes. Les variables latentes sont également désignées comme variables cachées, car elles ne sont pas mesurées directement. Les variables latentes expliquent la plus grande partie possible de la variation des données tout en modélisant bien les valeurs de référence.

Étape de préparation

Les étapes de préparation pour la régression MCP sont similaires à celle de l'ACP.

1. **Paramétrage** : le logiciel OMNIS applique sur les spectres le prétraitement des données et la sélection de longueurs d'onde définis.
2. **Centrage de la moyenne** : pour chaque longueur d'onde on calcule la valeur d'absorbance moyenne que l'on soustrait de la valeur respective dans chaque spectre.
   Les valeurs de référence sont également centrées sur la moyenne.

Conversion en variables latentes

Après les étapes de préparation, la régression MCP convertit les données spectrales en un espace de variables latentes en tenant compte des valeurs de référence. La *figure 30* indique les 2 premières variables latentes LV1 et LV2. Les variables latentes LV3, LV4, etc. peuvent être visualisées de la même manière.

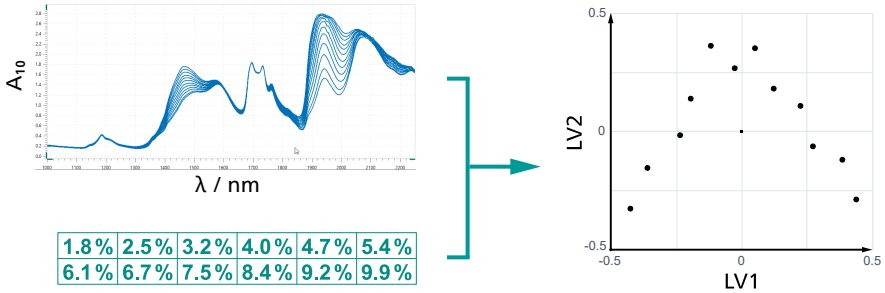
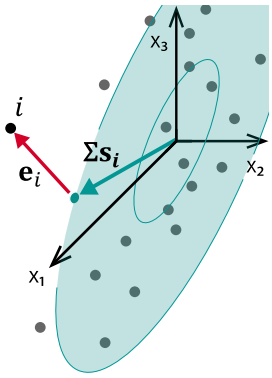


Figure 30 Conversion de spectres et de valeurs de référence en un espace de variables latentes. Les scores à droite sont exprimés en unités arbitraires.

La première variable latente LV1 explique le mieux la variance des données spectrales et présente également la corrélation la plus élevée possible avec les valeurs de référence. Toutes les variables latentes suivantes LV2, LV3, etc. expliquent la variance restante et présentent également la corrélation la plus élevée possible avec les valeurs de référence. Ainsi, les quelques premières variables latentes expliquent la majeure partie de la variance et maximisent la corrélation, tandis que les autres contiennent principalement du bruit de fond et peuvent être rejetées.

Scores et résidus



La MCP des grandeurs similaires à celles de l'ACP :

- **Scores** : les scores sont mesurés dans l'espace des variables latentes. La projection orthogonale de l'échantillon *i* sur chaque direction des variables latentes donne le vecteur de score $\Sigma \mathbf{s}_i$ qui représente la distance euclidienne du point central.
  La **distance de Mahalanobis** $\mathbf{s}_i$ est le vecteur de score normalisé qui apporte la même pondération à chaque direction.
- **Résidus** : le vecteur de résidu $\mathbf{e}_i$ est le décalage entre l'échantillon *i* et l'espace des variables latentes, mesuré dans l'espace de longueur d'onde initial.

Algorithme MCP

L'algorithme MCP maximise la covariance entre les spectres et les valeurs de référence *(voir "Algorithme MCP", Chapitre 6.3, page 95)*.

4.4.1.1 Nombre de variables latentes

Le choix du nombre de variables latentes dans le modèle de quantification est essentiel pour la capacité prédictive du modèle. Avec trop peu de variables latentes, des variations spectrales pertinentes ne seront pas détectées. C'est ce que l'on appelle le **sous-ajustement** qui entraîne des prédictions de faible précision.

Avec trop de variables latentes, les échantillons de calibrage seront trop bien modélisés. Le modèle acquiert des variations spectrales non perti-

nentes (bruit de fond). C'est ce que l'on appelle le **surajustement** qui entraîne des prédictions fluctuantes, instables et moins précises d'échantillons inconnus.

Pour trouver le nombre optimal de variables latentes, il faut obtenir un équilibre entre les objectifs suivants :

- La SEP est suffisamment proche de sa valeur minimale.
  Sans jeu de données validées, la SECV est utilisée.
- Le modèle de quantification doit utiliser le moins de variables latentes possibles. En cas de doute, le nombre le plus faible devrait être utilisé.
- Le diagramme de corrélation pour le jeu de données validées est proche de son optimum. Dans le cas idéal, la pente est proche de 1, l'ordonnée à l'origine y est proche de 0 et les points de données indiquent une dispersion minimale.
  Sans jeu de données validées, on utilise les valeurs de la validation croisée *(voir « Validation croisée », page 64)*.

Il convient de noter que pour les figures de mérite il s'agit uniquement des valeurs estimées sur la base des échantillons de calibrage et des échantillons de validation disponibles. En général, un modèle de quantification basé sur moins de variables latentes est plus robuste.

4.4.1.2 Graphique des poids factoriels

Le graphique des poids factoriels dans le logiciel OMNIS repose sur le modèle MCP *(voir "Algorithme MCP", Chapitre 6.3, page 95)*.

Les poids factoriels MCP montrent comment les variables de longueur d'onde initiales (y compris le paramétrage) contribuent à la formation de chaque variable latente. Que les poids factoriels soient positifs ou négatifs ne joue aucun rôle.

Les poids factoriels sont calculés de façon à ce que la première variable latente acquière la variance la plus pertinente pour le paramètre de référence. Toutes les variables latentes suivantes acquièrent la variance restante la plus pertinente pour le paramètre de référence. Des poids factoriels MCP très différents de 0 indiquent que les longueurs d'onde respectives sont bien adaptées à la modélisation du paramètre de référence.

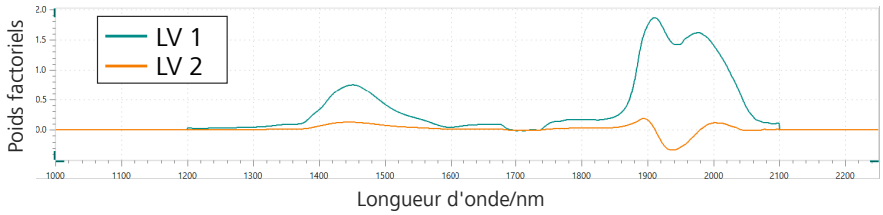


Figure 31 Graphique des poids factoriels pour les variables latentes LV1 et LV2.

Dans la *figure 31*, la sélection des longueurs d'onde a été limitée à la gamme de 1 200 nm à 2 100 nm. C'est pourquoi il n'y a aucun poids factoriel hors de cette gamme.

4.4.2 Validation des modèles de quantification

La validation consiste à vérifier que le modèle répond aux exigences de performance et de robustesse. Pour ce faire, l'erreur attendue de la prédiction doit être estimée de la manière la plus réaliste possible.

Un modèle de quantification est en général validé comme suit.

1. En partant d'un nombre d'échantillons limité et en l'absence d'un ensemble de données validées, un modèle est développé et testé par validation croisée (voir ci-dessous).
2. Avec un nombre d'échantillons plus important, l'ensemble de données est réparti dans un ensemble de données de calibrage pour développer un modèle et un ensemble de données validées pour valider le modèle.
3. Enfin, des échantillons sont recueillis et mesurés pour l'ensemble de données validées un autre jour, le cas échéant par une autre personne et avec un autre appareil.

Validation croisée

La validation croisée repose exclusivement sur l'ensemble de données de calibrage. Elle donne une valeur estimée pour chaque échantillon de calibrage en utilisant un modèle temporaire créé sans l'échantillon de calibrage en question.

La validation croisée consiste à utiliser une procédure à plusieurs tours de l'une des manières suivantes :

- **Leave-One-Out**
  À chaque tour, la validation croisée Leave-One-Out (LOO-CV) met de côté 1 échantillon tandis qu'un modèle est créé à partir des autres échantillons. Ce modèle prédit ensuite le paramètre d'intérêt pour l'échantillon mis de côté. Cette prédiction sert de valeur estimée pour l'échantillon.
  Le cycle se poursuit jusqu'à ce que chaque échantillon ait été retenu une fois.

- **K-fold**
  Pour la validation croisée K-fold, l'ensemble de données de calibrage est subdivisé en blocs *k* de taille égale si possible.
  À chaque tour, un bloc est retenu tandis qu'un modèle est créé à partir des blocs restants. Ce modèle prédit ensuite le paramètre d'intérêt pour les échantillons mis de côté.
  Le cycle se poursuit jusqu'à ce que chaque bloc ait été retenu une fois.
  Les blocs peuvent être sélectionnés de différentes manières :
  - **Fixed Blocks (DUPLEX)** (à partir de la version logicielle d'OMNIS 3.2) : les blocs sont sélectionnés reproductibles sur la base d'un algorithme duplex. Chaque prédiction sert directement de valeur estimée pour l'échantillon respectif.
  - **Random** : les blocs sont sélectionnés aléatoirement. La procédure décrite ci-dessus est répétée plusieurs fois. L'ensemble de données de calibrage est chaque fois divisé en blocs *k* différemment. Enfin, il existe plusieurs valeurs estimées pour chaque échantillon. La valeur moyenne sert de valeur estimée pour l'échantillon correspondant.

En général la méthode Leave-One-Out est privilégiée. Pour les grands ensembles de données de calibrage, la validation croisée K-fold peut être utilisée pour économiser du temps de calcul. 5 est une valeur type pour *k*.

4.4.2.1 Diagramme de corrélation

Le diagramme de corrélation visualise la corrélation entre les valeurs de référence et les valeurs calculées. Il permet d'évaluer le modèle de quantification en un coup d'œil.

Les valeurs calculées sont déterminées comme suit.

- Jeu de données validées et jeu de données atypiques : prédiction via le modèle de quantification
- Jeu de données de calibrage : valeurs estimées de la validation croisée

Dans le diagramme de corrélation, chaque échantillon est représenté par un point. La valeur calculée de l'échantillon se situe sur l'axe x et la valeur de référence sur l'axe y. Une droite de régression montre la relation systématique entre les variables. Dans le cas idéal, la droite de régression a une pente de 1 et une ordonnée à l'origine y de 0 et tous les points sont situés sur la droite. Cela signifie que pour chaque échantillon la valeur calculée correspond à la valeur de référence.

Les écarts par rapport au cas idéal permettent de distinguer les erreurs systématiques des erreurs aléatoires. La position de la droite de régression indique les erreurs systématiques. Les distances entre les points de la droite de régression représentent les erreurs aléatoires.

Le diagramme de corrélation suivant **A** montre une bonne corrélation. Les autres diagrammes montrent différents types d'erreurs qui seront expliquées ultérieurement.

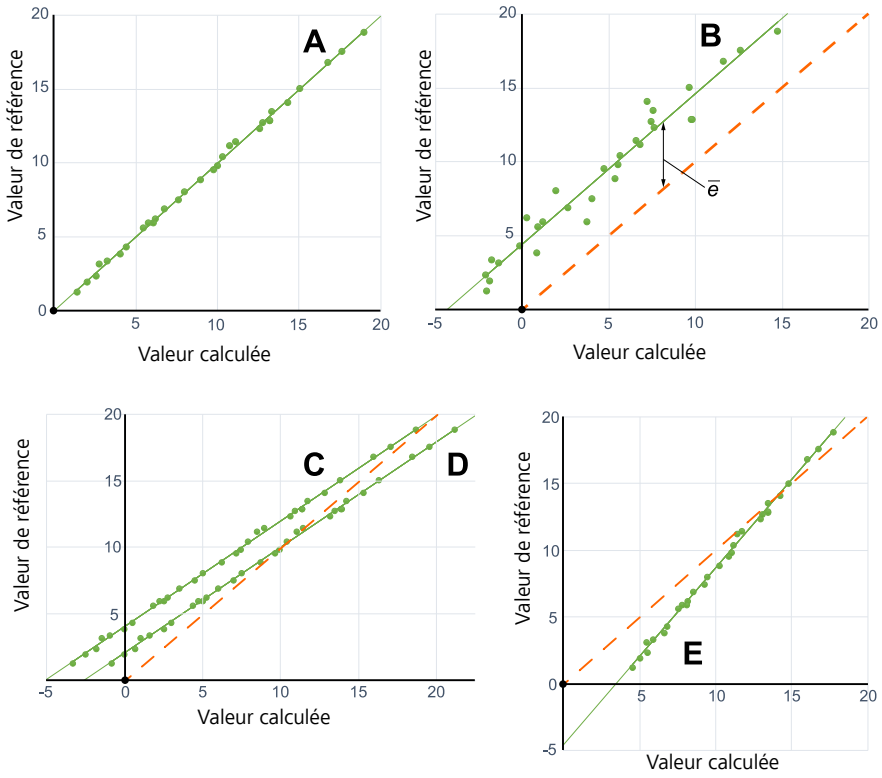


Figure 32 Diagrammes de corrélation Chaque point représente un échantillon et indique sa valeur de référence et sa valeur calculée. La ligne en pointillés représente la droite idéale à 45°.

Erreur systématique

Les erreurs systématiques sont des erreurs qui se produisent toujours et peuvent être répétées pour une application déterminée. Les erreurs systématiques peuvent être corrigées. Elles sont quantifiées par le biais $\bar{e}$ et la pente *b* des droites de régression :

$$y = b\hat{y} + \bar{e}$$

Si la pente est égale à 1 et le biais à 0, il n'y a pas d'erreur systématique.

Le **biais** est l'erreur moyenne entre les valeurs de référence et les valeurs calculées :

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \bar{y} - \bar{\hat{y}}$$

Dans ce cas, *n* représente le nombre d'échantillons, e_i l'erreur du *i*ème échantillon; y_i la valeur de référence du *i*ème échantillon, $\hat{y}_i$ la valeur calculée du *i*ème échantillon, $\bar{y}$ la valeur moyenne des valeurs de référence et $\bar{\hat{y}}$ la valeur moyenne des valeurs calculée.

Les droites **B** et **C** ont un biais positif, la droite **E** un biais négatif.

Les erreurs de signes opposés s'annulent mutuellement. Par conséquent, le biais de la droite **D** est proche de 0.

La **pente** des droites de régression est :

$$b = \frac{s_{\hat{y}y}}{s_{\hat{y}}^2}$$

Ici, $s_{\hat{y}y}$ représente la covariance entre valeurs de référence et valeurs calculées et $s_{\hat{y}}^2$ la variance des valeurs calculées.

La pente peut être considérée comme une erreur dépendant de la propriété :

- $b > 1$ (droite **E**) : plus la valeur calculée est élevée, plus l'erreur contribuant au biais est importante (positive).
- $b < 1$ (droites **C** et **D**) : plus la valeur calculée est élevée, plus l'erreur contribuant au biais est faible (négative).
- $b = 1$ (droites **A** et **B**) : l'erreur contribuant au biais est constante.

L'**ordonnée à l'origine y** de la droite de régression avec l'axe y est $\bar{y} - b\bar{\hat{y}}$.

La pente et l'ordonnée à l'origine y sont calculées en utilisant les valeurs de référence comme variable dépendante (axe y) et les valeurs calculées comme variable indépendante (axe x).

Erreur aléatoire

Si tous les points se situent directement sur la droite de régression, il n'y a pas d'erreur aléatoire. Plus les points sont dispersés, plus les erreurs aléatoires sont élevées.

Dans le diagramme de corrélation **B** les erreurs aléatoires sont plus grandes que celles des autres diagrammes de corrélation.

Visualisation des types d'erreurs

Les droites des figures ci-dessus montrent les types d'erreurs suivants :

	Erreur systématique			Erreur aléatoire
Ligne droite	**Biais**	**Pente**	**Ordonnée à l'origine**	
A	~ 0	~ 1	~ 0	petit
B	> 0	~ 1	> 0	grand
C	> 0	< 1	> 0	petit
D	~ 0	< 1	> 0	petit
E	< 0	> 1	< 0	petit

4.4.2.2 Figures de mérite

Les figures de mérite expriment en chiffres la concordance entre les valeurs de référence et les valeurs calculées. Les valeurs calculées sont déterminées par le modèle de quantification.

Coefficient de détermination R²

Le **coefficient de détermination R²** (en anglais coefficient of determination) mesure la qualité de l'ajustement du modèle de quantification. Pour un jeu de données donné, il s'agit de la part de la variation de la valeur de référence expliquée par le modèle de quantification :

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

où SS_{reg} correspond à la somme des carrés de la régression (variance des valeurs calculées, c'est-à-dire variance expliquée), SS_{tot} à la somme quadratique totale (variance de la valeur de référence), SS_{res} à la somme des carrés des résidus (variance résiduelle, c'est-à-dire la variance inexpliquée), y_i à la valeur de référence du *i*ème échantillon, $\hat{y}_i$ à la valeur calculée du *i*ème échantillon, et $\bar{y}$ à la moyenne des valeurs de référence.

La valeur R² est une fraction de 1. Un R² de 1 signifie que les valeurs calculées correspondent parfaitement aux valeurs de référence. Un R² de 0,9 signifie que 90 % de la variance de la valeur de référence sont expliqués par les valeurs calculées et 10 % ne le sont pas.

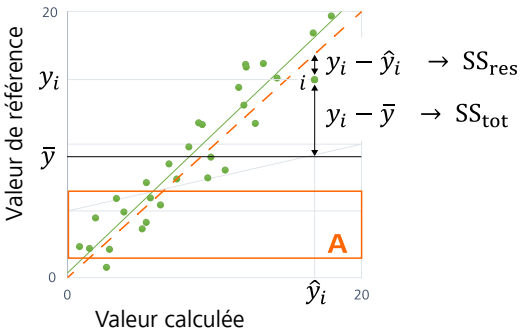


Figure 33 Les composants pour le calcul de R². La ligne en pointillés est la ligne droite à 45°.

La figure ci-dessus présente un diagramme de corrélation avec un échantillon *i* et son résidu issu de la régression MCP. La variation résiduelle est incluse dans SS_{res}, la variance de la valeur de référence dans SS_{tot}.

Une valeur R² élevée ne garantit pas un modèle de quantification utilisable ou des prédictions précises. La taille de R² dépend directement de la variation de la valeur de référence.

Une régression avec une gamme de valeurs de référence plus petite (gamme **A**) a approximativement la même variation résiduelle, mais la variation de la valeur de référence est plus petite. La valeur R² qui en résulte est plus basse.

La raison d'un R² élevé pourrait donc être un intervalle de valeurs de référence d'une grandeur irréaliste. D'un autre côté, les données provenant d'un process de fabrication, par exemple, peuvent avoir une plage de valeurs limitées, ce qui entraîne une valeur R² plus faible. Les erreurs standard doivent être utilisées pour évaluer la capacité prédictive.

La valeur absolue R² doit être considérée avec précaution. Le degré de changement est plus significatif avec chaque variable latente supplémentaire *(voir "Nombre de variables latentes", Chapitre 4.4.1.1, page 62)*.

Selon les valeurs utilisées pour le calcul, on obtient des valeurs R² différentes :

- R²C (non affiché dans le logiciel OMNIS) : calculée avec les valeurs calculées des spectres dans le jeu de données de calibrage.
- **R²CV** : calculée avec les valeurs estimées de la validation croisée des spectres dans le jeu de données de calibrage *(voir « Validation croisée », page 64)*.
- **R²P** : calculée avec les valeurs calculées des spectres dans le jeu de données validées.

Pour le calcul, le logiciel OMNIS utilise le carré du coefficient de corrélation d'échantillon de Pearson $r_{y,\hat{y}}$.

Coefficient de détermination de la validation croisée :

$$\mathrm{R^2CV} = r^2_{y,\hat{y}_{\mathrm{cv}}} = \frac{(\sum_{i=1}^{n}(y_i - \bar{y})\,(\hat{y}_{\mathrm{cv}_i} - \bar{\hat{y}}_{\mathrm{cv}}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(\hat{y}_{\mathrm{cv}_i} - \bar{\hat{y}}_{\mathrm{cv}})^2}$$

où y_i représente la valeur de référence du *i*ème échantillon, $\bar{y}$ la valeur moyenne des valeurs de référence, $\hat{y}_{\mathrm{cv}_i}$ la valeur estimée de la validation croisée du *i*ème échantillon, $\bar{\hat{y}}_{\mathrm{cv}}$ la valeur moyenne des valeurs estimées de la validation croisée et *n* le nombre d'échantillons dans le jeu de données de calibrage. Il convient de noter que chaque échantillon du jeu de données de calibrage a exactement une valeur estimée de la validation croisée.

Coefficient de détermination de la prédiction :

$$R^2P = r^2_{v,\hat{v}} = \frac{(\sum_{i=1}^{v}(v_i - \overline{v})\,(\hat{v}_i - \overline{\hat{v}}))^2}{\sum_{i=1}^{v}(v_i - \overline{v})^2 \cdot \sum_{i=1}^{v}(\hat{v}_i - \overline{\hat{v}})^2}$$

où v_i représente la valeur de référence du *i*ème échantillon de validation, $\overline{v}$ la valeur moyenne des valeurs de référence , $\hat{v}_i$ la valeur calculée du *i*ème échantillon de validation, $\overline{\hat{v}}$ la valeur moyenne des valeurs calculées et v le nombre d'échantillons de validation.

SEC – erreur standard du calibrage

L'**erreur standard du calibrage (SEC)** est liée au jeu de données de calibrage. La SEC peut être considérée comme une valeur estimée de la meilleure précision prédictive théorique. La SEC est l'écart-type des résidus de la régression des moindres carrés partiels (MCP) :

$$SEC = \sqrt{\frac{\mathbf{e}^t\mathbf{e}}{n-k-1}}$$

où **e** correspond au vecteur des résidus , qui contient toutes les variations de valeurs de référence dans le jeu de données de calibrage qui ne sont pas décrites par le modèle, *n* au nombre d'échantillons de calibrage et *k* au nombre de variables latentes. Le dénominateur *n*-*k*–1 est le nombre de degrés de liberté du vecteur des résidus **e**.

En d'autres termes, la SEC est l'écart-type des différences entre les valeurs de référence et les valeurs calculées pour les échantillons du jeu de données de calibrage :

$$SEC = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$$

où y_i représente la valeur de référence du *i*ème échantillon de calibrage, $\hat{y}_i$ la valeur calculée du *i*ème échantillon de calibrage, *n* le nombre d'échantillons de calibrage et *k* le nombre de variables latentes.

La SEC est parfois appelée RMSEC. La SEC comprend les erreurs aléatoires et les erreurs systématiques (pente et biais).

SECV – erreur standard de la validation croisée

L'**erreur standard de la validation croisée (SECV)** est liée au jeu de données de calibrage. La SECV estime la précision prédictive sur la base du jeu de données de calibrage et d'une méthode de validation croisée *(voir « Validation croisée », page 64)*. La SECV peut être utilisée pour une première évaluation du modèle ou pour déterminer le nombre optimal de variables latentes.

La SECV est l'écart-type des différences entre les valeurs de référence et les valeurs calculées de la validation croisée pour les échantillons du jeu de données de calibrage.

$$\mathrm{SECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_{\mathrm{cv}_i})^2}{n}}$$

où y_i représente la valeur de référence du *i*ème échantillon, $\hat{y}_{\mathrm{cv}_i}$ la valeur estimée de la validation croisée du *i*ème échantillon et *n* le nombre d'échantillons du jeu de données de calibrage.

La SECV comprend toutes les erreurs aléatoires et les erreurs systématiques (pente et biais).

D'autres approches utilisent des valeurs séparées pour une SECV corrigée des biais et pour la SECV non corrigée, qui est alors appelée RMSECV. Ces valeurs sont similaires si le biais est faible.

SEP – erreur standard de prédiction

L'**erreur standard de prédiction (SEP)** est liée au jeu de données validées. La SEP fournit donc la valeur estimée la plus réaliste de la précision prédictive.

La SEP est l'écart-type des différences entre les valeurs de référence et les valeurs calculées pour les échantillons du jeu de données validées :

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{v}(v_i - \hat{v}_i)^2}{v}}$$

où v_i représente la valeur de référence du *i*ème échantillon de validation, $\hat{v}_i$ la valeur calculée du *i*ème échantillon de validation et v le nombre d'échantillons de validation.

La SEP comprend toutes les erreurs : aléatoires et systématiques (pente et biais).

D'autres approches utilisent des valeurs séparées pour une SEP corrigée des biais et pour la SEP non corrigée, qui est alors appelée RMSEP. Ces valeurs sont similaires si le biais est faible.

Interprétation des figures de mérite

Les figures de mérite sont des valeurs estimées. Exemple : la valeur SEP est une *valeur estimée* d'un écart-type – sur la base des échantillons présents – et possède également son propre-écart-type. Plus le nombre d'échantillons de validation est élevé, plus la valeur estimée est fiable.

La SEP devrait être comparable à la SECV et à la SEC. Des différences trop importantes peuvent indiquer un surajustement *(voir "Nombre de variables latentes", Chapitre 4.4.1.1, page 62)*. En règle générale, les différences ne devraient pas dépasser 20 %.

Les erreurs standard doivent également être considérées par rapport à l'erreur standard du laboratoire (SEL) pour la méthode de référence. La

précision de la méthode NIR est acceptable si la SEP est 1,4 à 2,0 fois supérieure à la SEL. Une SEP plus grande peut néanmoins être acceptée si elle répond aux exigences.

Une SEC ou SECV inférieure à la SEL pour la méthode de référence suggère un surajustement.

Les relations avec la SEL mentionnées ci-dessus supposent que la SEL est basée sur le nombre correct de mesures de référence répétées pour chaque échantillon *(voir « Méthode de référence (quantification) », page 34)*.

4.4.3 OMNIS Model Developer (OMD)

Le développement d'un modèle de quantification est exigeant, chronophage et nécessite une certaine expertise. L'OMNIS Model Developer (OMD, à partir de la version logicielle d'OMNIS 4.0) automatise le développement et fournit des modèles de quantification bien optimisés.

Fonctionnement

Un échantillonnage correct est la condition préalable *(voir "Échantillons physiques", Chapitre 4.1, page 33)*. Un ensemble de données existant constitué de spectres et des valeurs de référence correspondantes sert d'entrée pour l'OMD.

L'OMD détermine les valeurs spectrales atypiques avec une pertinence de 5 % et l'algorithme présenté en annexe, sans tenir compte du prétraitement des données *(voir « Détection de valeurs spectrales atypiques pour le développement de modèles », page 98)*.

La division de l'ensemble de données dépend du nombre de spectre qui restent après la détection des valeurs atypiques.

Nombre de spectres	Méthode de validation croisée	Ensemble de données validées
> 99	K-fold (5 blocs, DUPLEX)	25 % des spectres
30–99	K-fold (5 blocs, DUPLEX)	—
< 30	Leave-One-Out	—

Après division, des valeurs atypiques supplémentaires sont déterminées dans l'ensemble de données de calibrage selon ASTM D8321-22.

L'évaluation et le classement des modèles se font à l'aide de divers indicateurs. L'OMD optimise le prétraitement des données, la sélection des longueurs d'onde et le nombre de variables latentes en recherchant un équilibre entre le risque de surajustement et le risque de sous-ajustement.

Résultat

Le résultat de l'OMD est une liste de modèles organisés selon leur capacité prédictive. La **capacité prédictive** est calculée sur la base de la complexité du modèle, des figures de mérite et de la taille de l'ensemble de données.

La liste est codée par couleur afin de faciliter le choix du modèle privilégié :

- Vert : bonne capacité prédictive.
  Si le nombre d'échantillons est suffisamment important, le modèle fonctionnera correctement pour tous les échantillons inconnus de même type. Les figures de mérite permettent d'obtenir une valeur estimée fiable des erreurs futures.
- Jaune : capacité prédictive moyenne.
  Si le nombre d'échantillons est suffisamment important, le modèle devrait fonctionner correctement. Les figures de mérite pourraient être trop optimistes pour les futurs échantillons. Une validation séparée est recommandée.
- Rouge : capacité prédictive insuffisante.
  Le modèle présente de sérieux inconvénients. Il ne devrait pas être utilisé.

Les modèles de même couleur sont triés selon un critère d'information qui privilégie un équilibre entre une faible erreur de prédiction estimée et un faible nombre de variables latentes.

Optimisation du paramétrage

Au lieu de créer automatiquement l'ensemble du modèle, il est possible d'optimiser uniquement le paramétrage. Les paramètres actuels (comme la répartition des ensembles de données ou la méthode de validation croisée) restent inchangés et n'ont cependant aucune influence sur l'optimisation.

4.4.4 Correction de pente/d'ordonnée à l'origine

La correction de pente/d'ordonnée à l'origine permet de corriger les erreurs systématiques (biais, pente) lors de l'application d'un modèle de quantification.

Les causes possibles d'erreurs systématiques dans l'ensemble de données de calibrage sont :

- Erreur systématique dans un modèle de quantification. Par exemple valeurs atypiques inconnues ou nombre d'échantillons insuffisant.
- Erreurs systématiques dans la méthode de mesure spectroscopique.
- Erreurs systématiques dans la méthode de mesure de référence.

Si des erreurs systématiques se produisent dans l'ensemble de données validées, il existe d'autres causes :

- Modifications de la méthode de mesure spectroscopique, p. ex. de l'appareil.
- Modifications dans la méthode de mesure de référence, p. ex. nouveau laboratoire, nouvel équipement.
- Modifications des échantillons, p. ex. lors de la manipulation, du stockage ou du transport.

La correction de biais et surtout la correction de pente/d'ordonnée à l'origine doivent être utilisées avec précaution.

Si les erreurs systématiques ne sont pas significatives, aucune correction ne doit être appliquée. Si les erreurs sont significatives, elles doivent faire l'objet d'un examen approfondi. La cause des erreurs doit être éliminée dans toute la mesure du possible. Si les erreurs ne peuvent pas être corrigées pour une raison justifiée, une correction de biais ou une correction de pente/d'ordonnée à l'origine peut être appliquée.

Pour obtenir une valeur estimée fiable du biais, il faut au moins 20 échantillons. Pour obtenir une valeur estimée fiable de la pente, il faut au moins 30 échantillons.

Correction de biais

Le diagramme de corrélation suivant présente une correction de biais. La pente des droites de régression initiales **F** reste inchangée. Après correction (droite de régression **G**), les erreurs positives et négatives s'annulent mutuellement.

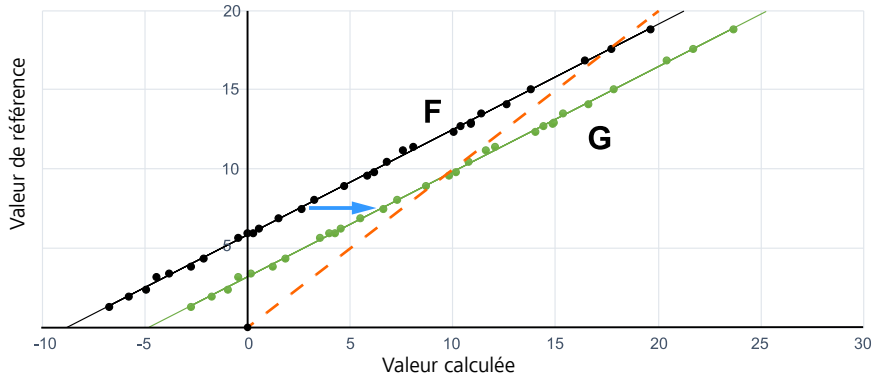


Figure 34 Correction de biais

Correction de pente/d'ordonnée à l'origine

Le diagramme de corrélation suivant présente une correction de pente/d'ordonnée à l'origine. La droite de régression initiale **H** est corrigée en fonction de la pente et de l'ordonnée à l'origine. En conséquence, la pente et le biais sont tous deux corrigés (droite de régression **K**).

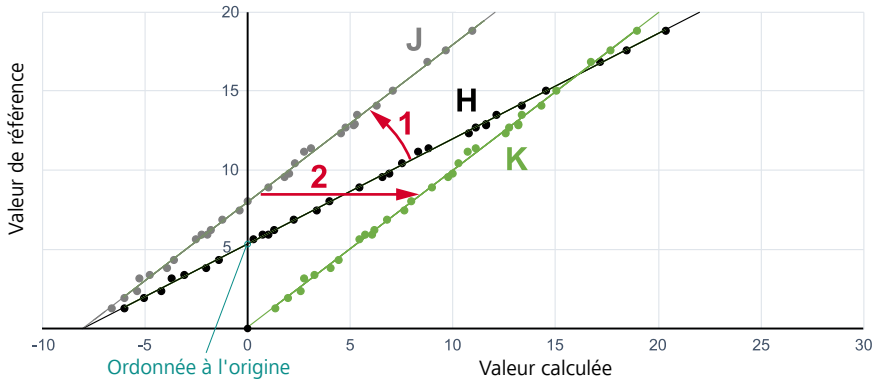


Figure 35 Correction de pente/d'ordonnée à l'origine

SEP

Les échantillons de l'ensemble de données de correction constituent la base de la correction de pente/d'ordonnée à l'origine. Le logiciel OMNIS calcule les **erreurs standard de prédiction (SEP)** suivantes sur la base de ces échantillons. Les dénominateurs tiennent compte respectivement des degrés de liberté :

Type de correction	SEP
Non corrigée	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n}}$
Correction de biais	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-1}}$
Correction de pente/d'ordonnée à l'origine	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-2}}$

où v_i correspond à la valeur de référence du *i*ème échantillon dans l'ensemble de données de correction, $\hat{v}_i$ à la valeur prédite du *i*ème échantillon dans l'ensemble de données de correction et *n* au nombre d'échantillons dans l'ensemble de données de correction.

La SEP comprend toutes les erreurs : aléatoires et systématiques (pente et biais).

4.5 Identification et vérification

4.5.1 Machine à vecteurs de support (SVM)

Les modèles d'identification (à partir de la version logicielle d'OMNIS 4.0) utilisent une machine à vecteurs de support (SVM) pour la classification entre différents produits. Une machine à vecteurs de support est un algorithme d'apprentissage automatique supervisé. Sur la base des échantillons de calibrage, il apprend à attribuer de nouveaux échantillons à un produit.

Pour simplifier, la classification entre 2 produits est décrite ci-dessous. Le concept peut être étendu, le modèle final classifie entre n'importe quel nombre de produits.

Classification linéaire

La *figure 36 (à gauche)* présente des données d'entrée à 2 variables.

Pour simplifier, les spectres paramétrés sont représentés dans un espace de variables à 2 dimensions. Chaque point représente un spectre, la couleur indique l'appartenance du produit.

Les produits sont séparables linéairement. L'algorithme SVM crée un hyperplan entre les produits (figure de droite).

Les hyperplans sont une généralisation des plans de l'espace à trois dimensions aux espaces de n'importe quelle dimension. La dimension d'un hyperplan est inférieure d'une unité à la dimension de l'espace qui l'entoure. Un hyperplan dans un espace à 2 dimensions est une ligne.

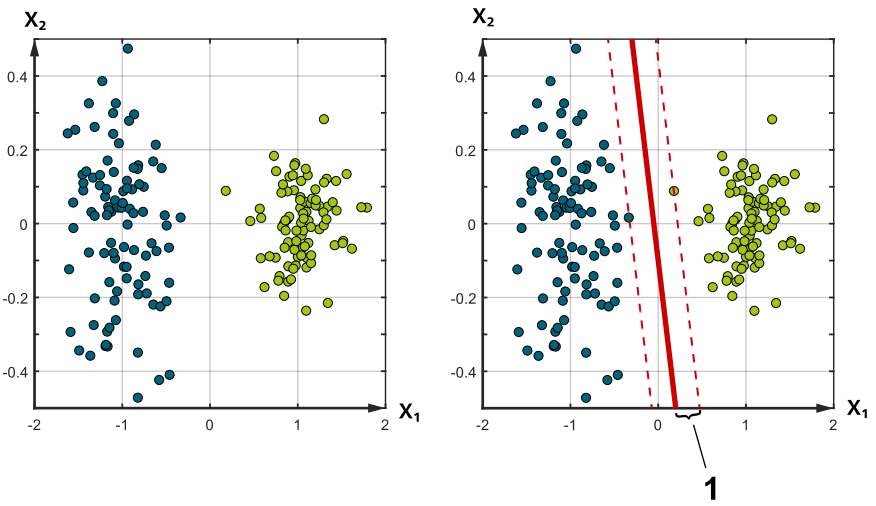


Figure 36 Données d'entrée (à gauche) et hyperplan créé par le SVM (ligne rouge continue à droite). Les valeurs sont indiquées en unités arbitraires.

L'algorithme SVM maximise la marge (**1**) entre l'hyperplan et le point le plus proche de chaque côté. De nouveaux spectres peuvent être représentés sur le même espace et attribués à un produit selon le côté de l'hyperplan sur lequel ils tombent.

Pour la définition de l'hyperplan, l'algorithme SVM tient compte uniquement des points qui sont les plus proches des points du produit opposé. Ces points ou vecteurs prennent en charge la formation de l'hyperplan et sont désignés comme vecteurs de support.

Si les points ne peuvent pas être séparés linéairement - par exemple en raison d'une valeur atypique - il est néanmoins possible de déterminer un hyperplan à classification linéaire. Dans ce cas, un algorithme d'optimisation trouve un compromis entre l'augmentation de la marge de l'hyperplan aux vecteurs de support de chaque côté et garantit que tous les points se trouvent du bon côté de l'hyperplan. Un paramètre de régularisation contrôle le compromis et ainsi la position finale de l'hyperplan.

Classification non linéaire

Dans la *figure 37 (gauche)*, les produits ne peuvent pas être séparés linéairement. Un classificateur non linéaire est nécessaire pour la séparation des produits.

Une fonction noyau linéaire ou non linéaire transforme les données en un espace de caractéristiques de dimension supérieure. La transformation est effectuée de façon à ce que les données dans l'espace de caractéristiques puissent être séparées linéairement par un hyperplan.

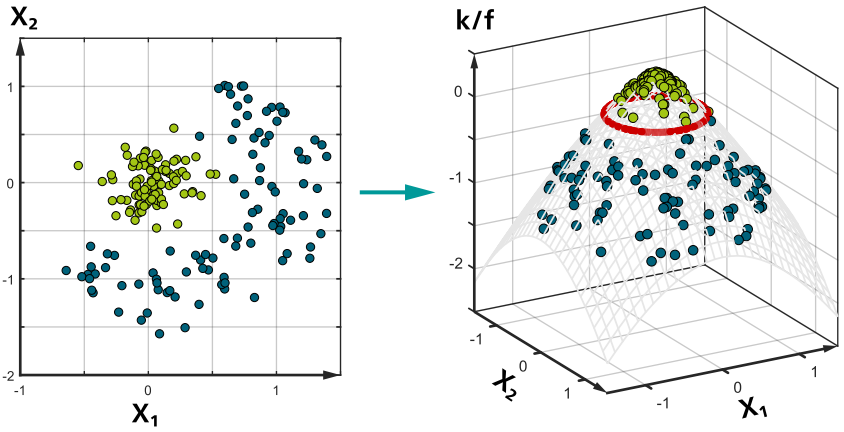


Figure 37 Produits non séparables linéairement (gauche). Une dimension supplémentaire k/f (caractéristique de noyau) facilite la séparation (droite). Les valeurs sont indiquées en unités arbitraires.

Dans la figure, les données sont converties de l'espace à 2 dimensions en espace à 3 dimensions. Les points d'un produit sont élevés au-dessus du niveau d'origine tandis que les points de l'autre produit sont déplacés vers le bas. L'hyperplan linéaire entre les produits est un plan à 2 dimensions qui est très similaire au plan initial. Considérée en 2 dimensions, la limite de décision est une ligne non linéaire, en l'occurrence la ligne circulaire rouge.

Ici également l'hyperplan sert de classificateur. Un nouvel échantillon peut être affecté à un produit selon le côté de l'hyperplan sur lequel son spectre tombe.

Le logiciel OMNIS utilise un noyau avec fonction de base radiale pour transformer les données dans l'espace de caractéristiques. Le noyau utilise un paramètre d'échelle qui ne contrôle pas le degré de non-linéarité.

Sélection des paramètres

Des valeurs adaptées doivent être sélectionnées pour le paramètre de régularisation qui contrôle la position de l'hyperplan et pour le paramètre d'échelle qui contrôle le degré de non-linéarité.

Ces deux paramètres ont un impact sur la capacité de généralisation de la machine à vecteurs de support, c'est-à-dire sur sa capacité à généraliser des spectres de calibrage à des spectres nouveaux et inconnus. Par exemple, si le paramètre d'échelle permet un degré élevé de non-linéarité, il se peut que l'hyperplan soit trop ajusté aux spectres d'étalonnage (suradaptation). Si le paramètre d'échelle permet uniquement un degré faible de non-linéarité, il se peut que l'hyperplan ne soit pas suffisamment ajusté aux spectres de calibrage (sous-ajustement).

Pour une bonne généralisation, on utilise la **recherche par grille**. L'algorithme trouve la meilleure combinaison de paramètres avec le moins d'essais possible :

1. On utilise un jeu de combinaisons de paramètres prédéfinies.
2. Pour certaines combinaisons de paramètres sélectionnées la SVM apprend une règle de classification qui distingue les spectres de calibrage des différents produits avec une exactitude maximale.
3. Pour chaque règle de classification, une validation croisée permet d'estimer le degré de réussite de l'attribution des produits.
4. D'autres combinaisons de paramètres sont sélectionnées sur la base des résultats de la validation croisée.
   La SVM réapprend les règles de classification correspondantes et une méthode de validation croisée estime l'exactitude de la classification. Après quelques répétitions la combinaison de paramètres finale est définie.
5. La SVM évalue la règle de classification finale avec la combinaison de paramètres finale et tous les spectres de calibrage.

Probabilités

Sur la base de la règle de classification, le modèle d'identification calcule la probabilité qu'un échantillon déterminé appartienne ou non à un produit déterminé. La probabilité dépend des distances par rapport à l'ensemble des produits et des paramètres du modèle.

Les probabilités ainsi calculées permettent de contrôler l'attribution des échantillons aux produits *(voir « Attribution d'un échantillon (à partir de la version logicielle d'OMNIS 4.4) », page 79)*.

Pour diminuer les identifications en faux positifs, un modèle de qualification est également formé pour chaque produit à partir de la version logicielle d'OMNIS 4.4 *(voir "Qualification", Chapitre 4.6, page 83)*.

4.5.2 Prédiction de l'appartenance d'un échantillon à un produit

Le modèle d'identification fournit une probabilité par produit pour l'identification d'un échantillon déterminé *(voir « Probabilités », page 79)*.

Chaque probabilité est une valeur individuelle entre 0 % et 100 %. La somme des valeurs de tous les produits ne correspond pas à 100 %.
Les valeurs doivent être considérées comme relatives les unes aux autres, ce qui permet une comparaison des différents produits.

Attribution d'un échantillon (à partir de la version logicielle d'OMNIS 4.4)

L'évaluation s'effectue à l'aide d'un **seuil de probabilité** et avec des qualifications pour certains produits :

1. Une qualification de l'échantillon est effectuée avec le modèle de qualification correspondant pour chaque produit dont la probabilité est supérieure au seuil de probabilité. Si la qualification échoue, la probabilité du produit correspondant est mise à zéro.
2. Évaluation avec les probabilités modifiées de l'étape 1 :
   a. Si aucune probabilité n'est supérieure au seuil de probabilité, l'identification échoue (état de l'identification **Non identifié**).
   b. Si une seule probabilité est supérieure au seuil de probabilité, l'identification de l'échantillon a réussi et il est attribué au produit correspondant (état de l'identification **Identifié**).
   c. Si plusieurs probabilités sont supérieures au seuil de probabilité, la prédiction est équivoque et l'identification échoue (état de l'identification **Équivoque**).

Le tableau 1 présente un exemple d'évaluation avec différents seuils de probabilité. Dans cet exemple, les qualifications des produits A et c ont réussi.

Tableau 1 Exemple d'évaluation avec différents seuils de probabilité (à partir de la version logicielle d'OMNIS 4.4)

Probabilité	Seuil de probabilité	Qualification	Résultat d'identification
Produit A : 87 % Produit B : 71 % Produit C : 68 % Produit D : 30 %	90 %	–	Non identifié
	80 %	Produit A : réussite → 87 %	Produit A
	70 %	Produit A : réussite → 87 % Produit B : échec → 0 %	Produit A
	60 %	Produit A : réussite → 87 % Produit B : échec → 0 % Produit C : réussite → 68 %	Équivoque

Attribution d'un échantillon (version logicielle d'OMNIS 4.0 à 4.3)

Les probabilités de l'échantillon sont évaluées à l'aide d'un **seuil de probabilité** paramétrable :

- Si aucune probabilité n'est supérieure au seuil de probabilité, l'identification échoue (état de l'identification **Non identifié**).
- Si une seule probabilité est supérieure au seuil de probabilité, l'identification de l'échantillon a réussi et il est attribué au produit correspondant (état de l'identification **Identifié**).
- Si plusieurs probabilités sont supérieures au seuil de probabilité, la prédiction est équivoque et l'identification échoue (état de l'identification **Équivoque**).

Le tableau 2 présente un exemple d'évaluation avec différents seuils de probabilité.

Tableau 2 Exemple d'évaluation avec différents seuils de probabilité (version logicielle d'OMNIS 4.0 à 4.3)

Probabilité	Seuil de probabilité	Résultat d'identification
Produit A : 87 %	90 %	Non identifié
Produit B : 72 %	80 %	Produit A
Produit C : 68 %	70 %	Équivoque

4.5.3 Validation des modèles d'identification

Un modèle d'identification est en général validé comme suit.

1. En partant d'un nombre d'échantillons limité et en l'absence d'un ensemble de données validées, un modèle est développé et testé avec les échantillons dans l'ensemble de données de calibrage.
2. Lorsque le nombre d'échantillons est suffisant, l'ensemble de données peut être réparti en un ensemble de données de calibrage et un ensemble de données validées. Les échantillons de l'ensemble de données validées ne sont pas utilisés pour le développement du modèle.
3. Enfin, des échantillons pour un ensemble de données validées externe sont recueillis et mesurés un autre jour, le cas échéant par une autre personne et avec un autre appareil.

Pour chaque échantillon, l'appartenance du produit prédite est comparée à l'appartenance du produit réelle correspondante. Si elles correspondent, la prédiction est correcte (= réussite), sinon elle est incorrecte (= échec).

Validation

Le logiciel OMNIS affiche les grandeurs suivantes indiquant pour les modèles d'identification dans quelle mesure les modèles fonctionnent. Dans l'idéal, tous les indicateurs sont à 100 %.

Succès % (total) mesure l'*exactitude*, l'exactitude globale d'un modèle. Le pourcentage répond à la question : combien d'échantillons le modèle peut-il identifier correctement ?

$$\textit{Succès \% (total)} = \frac{\textit{classifications correctes}}{\textit{toutes les classifications}}$$

Succès % (total) confère le même poids à chaque échantillon. Par conséquent, les produits contenant plus d'échantillons ont un impact plus important sur le pourcentage que les produits contenant moins d'échantillons.

Des nombres similaires pour *Succès %* sont disponibles pour chaque produit.

Améliorer l'identification

Les actions suivantes peuvent contribuer à améliorer le modèle :

- **Ajuster le seuil de probabilité**
  - Le seuil de probabilité peut être augmenté si de nombreuses prédictions sont équivoques ou si de nombreuses probabilités de 0,0 % apparaissent.
  - Si de nombreux échantillons ne sont pas identifiés parce que le seuil de probabilité n'a pas été atteint, celui-ci peut être abaissé.
- **Ajuster le paramétrage**
  Définir le prétraitement des données et la sélection des longueurs d'onde les plus adaptés.
- **Utilisation des hiérarchies de modèles**
  Une hiérarchie des modèles permet une structuration hiérarchique des modèles d'identification.
  Exemple : un modèle d'identification avec 4 produits différents ne peut pas facilement faire la différence entre les produits similaires que sont le fructose et le glucose. Si le fructose et le glucose sont réunis dans un groupe de produits « sucre », le modèle peut faire la distinction entre le sucre et les deux autres produits. Si un échantillon est identifié comme étant du sucre, un autre modèle prend en charge la distinction entre le fructose et le glucose. Comme ce modèle est spécialisé, il permet de distinguer plus facilement les produits similaires.

Échantillons non identifiés

Si des échantillons ne sont pas identifiés par erreur :

- Vérifier si les échantillons et leur traitement présentent des anomalies.
- Vérifier et si nécessaire abaisser le seuil de probabilité.
- Vérifier le prétraitement des données et la sélection des longueurs d'onde.
- Vérifier les spectres des échantillons non identifiés dans le graphique des scores :
  - Si le modèle ne contient pas encore les spectres : ajouter les spectres pour l'ensemble de données validées du produit respectif.
  - Dans le graphique des scores, comparer les scores des spectres à vérifier avec les scores des spectres dans l'ensemble de données de calibrage.

  Si les échantillons ne sont pas des valeurs aberrantes, mais que leurs variations sont sous-représentées dans l'ensemble de données de calibrage, cet ensemble de données de calibrage doit être étendu en conséquence.

4.6 Qualification

Les modèles de qualification (à partir de la version logicielle d'OMNIS 4.4) différencient un groupe d'échantillons des autres échantillons. Ces modèles permettent p. ex. de distinguer des échantillons utilisables (échantillons positifs) des échantillons non utilisables (échantillons négatifs).

4.6.1 Calcul de modèles de qualification

Le calcul d'un modèle de qualification s'effectue de manière similaire à celui d'un modèle d'identification *(voir "Machine à vecteurs de support (SVM)", Chapitre 4.5.1, page 76)*. L'ensemble de données de calibrage pour la qualification contient toutefois un seul type d'échantillons (échantillons positifs).

Une machine à vecteurs de support (SVM) transforme les données d'entrée en un espace tridimensionnel. Un paramètre de régularisation définit la position de l'hyperplan, tandis qu'un paramètre d'échelle détermine le degré de non-linéarité. Une recherche par grille détermine la projection appropriée. La limite de décision pour cette projection constitue la base du modèle de qualification.

4.6.2 Validation de modèles de qualification

Procédure

En général, un modèle de qualification est développé et validé progressivement. Le nombre d'échantillons est progressivement augmenté :

1. Ensemble de données validées positif :
   a. Si un nombre limité d'échantillons positifs est trouvé, dans un premier temps aucun ensemble de données validées positif n'est créé.
   b. Dès qu'un nombre suffisant d'échantillons positifs est disponible, un ensemble de données validées positif est créé à l'aide de la répartition automatique de l'ensemble de données.

   Ensemble de données validées négatif :
   a. Les échantillons négatifs collectés sont attribués à l'ensemble de données validées positif.
   b. La détermination automatique des spectres négatifs peut être utilisée pour détecter des valeurs spectrales atypiques et les attribuer à l'ensemble de données validées négatif *(voir "Valeur spectrale atypique - algorithme", Chapitre 6.5, page 98)*.
2. Enfin, des échantillons sont recueillis et mesurés pour l'ensemble de données validées positif et négatif un autre jour, le cas échéant par une autre personne et avec un autre appareil.

Les échantillons de l'ensemble de données validées négatif et positif ne sont pas utilisés pour le calcul du modèle.

Validation

Le modèle de qualification détermine un résultat (positif ou négatif) pour chaque échantillon. On attend un résultat positif pour les échantillons dans l'ensemble de données de calibrage et dans celui de données validées positif. Pour les échantillons dans l'ensemble de données validées négatif, on attend un résultat négatif. Si le résultat correspond à ce qui est attendu, la prédiction est correcte (= réussite), sinon elle est incorrecte (= échec).

Le logiciel OMNIS affiche les valeurs suivantes pour les modèles de qualification :

Succès % (total) mesure l'exactitude globale d'un modèle.

$$\textit{Succès \% (total)} = \frac{\textit{prédictions correctes}}{\textit{toutes les prédictions}}$$

Des valeurs similaires pour *réussite %* sont disponibles pour chaque ensemble de données. Dans l'idéal, tous les indicateurs sont à 100 %.

Améliorer une qualification

Un paramétrage adapté (prétraitement des données et sélection des longueurs d'onde) peut améliorer le modèle de qualification.

Échantillons non qualifiés

Si des échantillons ne sont pas qualifiés par erreur :

- Vérifier si les échantillons et leur traitement présentent des anomalies.
- Vérifier le prétraitement des données et la sélection des longueurs d'onde.
- Vérifier les spectres des échantillons non qualifiés dans le graphique des scores :
  - Si le modèle ne contient pas encore les spectres : ajouter les spectres à l'ensemble de données validées positif.
  - Dans le graphique des scores, comparer les scores des spectres à vérifier avec les scores des spectres dans l'ensemble de données de calibrage.

  Si les échantillons ne sont pas des valeurs aberrantes, mais que leurs variations sont sous-représentées dans l'ensemble de données de calibrage, cet ensemble de données de calibrage doit être étendu en conséquence.

5 Prédiction

5.1 Quantification

Pour la prédiction d'un échantillon dont le paramètre d'intérêt est inconnu, on procède comme suit :

1. Le spectre de l'échantillon est enregistré.
2. Le modèle de quantification applique le même prétraitement des données et les mêmes gammes de longueurs d'onde que pour les spectres dans l'ensemble de données de calibrage.
3. Sur la base du spectre obtenu, le modèle prédit le paramètre d'intérêt.

Lors du calcul du modèle de quantification, les spectres de calibrage et les valeurs de référence ont été centrés sur la moyenne. Cela est bien sûr pris en compte dans la procédure décrite ci-dessus.

5.1.1 Valeurs atypiques et contrôle du résultat

Lors de la prédiction d'un paramètre d'intérêt quantitatif, différents types de valeurs atypiques peuvent être détectés, et les résultats contrôlés :

Valeur atypique / contrôle		Cause (exemple)
Valeur spectrale atypique	T^2 de Hotelling	La concentration d'un composant chimique se situe hors de la plage de concentration des échantillons de calibrage.
	Résidus Q	L'échantillon contient des composants qui ne sont pas présents dans les échantillons de calibrage.
Valeur atypique Nearest-Neighbor (en option, à partir de la version logicielle d'OMNIS 4.2)		L'échantillon présente des variations qui ne sont pas représentées dans cette combinaison des échantillons de calibrage.
Contrôle du résultat (en option)		La valeur du résultat du paramètre d'intérêt se situe hors des critères d'acceptation sélectionnés par le client.

Valeur spectrale atypique

Les valeurs spectrales atypiques sont déterminées comme suit :

1. Le logiciel calcule les valeurs T^2 de Hotelling et les résidus Q pour le spectre sur la base du modèle de quantification (modèle MCP).
2. Si T^2 ou la valeur du résidu Q du spectre est supérieure à la valeur critique correspondante calculée par le modèle, l'échantillon est identifié comme atypique par rapport au modèle appliqué *(voir « Évaluation de la valeur atypique pour la prédiction (quantification) », page 99)*.

Les T^2 et les valeurs du résidu Q sont disponibles en tant que variables dans le logiciel OMNIS. Elles peuvent être comparées aux valeurs du graphique d'influence MCP du modèle de quantification. Les lignes en pointillés indiquent les valeurs critiques.

Valeur atypique Nearest-Neighbor

(à partir de la version logicielle d'OMNIS 4.2)

En théorie, les échantillons de calibrage couvrent toutes les combinaisons possibles de variations d'échantillon. Dans la pratique, certaines combinaisons se produisent plus fréquemment, d'autres pas du tout. En conséquence, les échantillons de calibrage sont inégalement répartis dans l'espace des variables latentes. Dans certaines parties, il existe de nombreux échantillons de calibrage entre lesquels il y a des espaces.

Si le spectre d'un échantillon inconnu se situe dans un espace entre les échantillons de calibrage, le résultat de prédiction peut être invalide ou inexact. Pour détecter ces cas, on calcule la distance *D* de l'échantillon inconnu *i* par rapport à chaque échantillon de calibrage *u* :

$$D = \sqrt{(\mathbf{s_i} - \mathbf{s_u})^{\mathbf{t}} (\mathbf{s_i} - \mathbf{s_u})}$$

Ici $\mathbf{s_i}$ correspond aux scores de l'échantillon inconnu *i* et $\mathbf{s_u}$ aux scores de l'échantillon de calibrage *u*. Les scores sont normalisés et orthogonaux.

La plus petite distance est la distance par rapport à l'échantillon de calibrage le plus proche et est désignée comme **Nearest Neighbor Distance** (NND) :

Si la valeur NND dépasse une valeur limite NND définie, l'échantillon inconnu est désigné comme valeur atypique Nearest Neighbor.

La valeur limite NND est définie comme suit :

1. Une valeur NND est définie pour chaque échantillon de calibrage. Cette valeur correspond à la distance au plus proche des échantillons de calibrage restants.
2. La valeur NND maximale de tous les échantillons de calibrage est la valeur limite NND.

La valeur NND de l'échantillon inconnu et la valeur limite NND sont disponibles dans le logiciel OMNIS en tant que variables.

Contrôle du résultat

Le contrôle du résultat peut être utilisé dans le logiciel OMNIS pour définir des limites d'alerte et des limites de contrôle pour la partie des résultats de prédiction. En option, on peut définir des actions qui seront déclenchées si les limites définies ne sont pas respectées.

5.2 Identification et vérification

Lors de l'identification ou de la vérification d'un échantillon, la procédure est la suivante :

1. Le spectre de l'échantillon est enregistré.
2. Le modèle d'identification applique le même prétraitement des données et la même sélection de longueurs d'onde que pour les spectres dans l'ensemble de données de calibrage.
3. Le modèle d'identification évalue le spectre *(voir "Prédiction de l'appartenance d'un échantillon à un produit", Chapitre 4.5.2, page 79)*.
4. Si le modèle d'identification fait partie d'une hiérarchie des modèles et qu'un autre modèle d'identification est lié au produit identifié, ce modèle est exécuté. Si un ou plusieurs modèles de quantification sont liés au produit déterminé, ces modèles sont exécutés.
5. Le résultat d'identification ou de vérification est affiché et, dans le cas d'une hiérarchie de modèles, également les résultats de quantification, le cas échéant.

État d'identification

- Identifié
  L'identification a réussi.
- Équivoque
  Plusieurs produits dépassent le seuil de probabilité. L'identification a échoué.
- Non identifié
  Aucun produit ne dépasse le seuil de probabilité. L'identification a échoué.

État de vérification

- Succès
  L'échantillon a été identifié avec succès et le résultat correspond au produit attendu.
- Échec
  La vérification a échoué.

5.3 Qualification

Pour la qualification d'un échantillon, la procédure est la suivante :

1. Le spectre de l'échantillon est enregistré.
2. Le modèle de qualification applique le même prétraitement des données et la même sélection de longueurs d'onde que pour les spectres dans l'ensemble de données de calibrage.
3. Sur la base du spectre obtenu, le modèle qualifie l'échantillon.
4. Le résultat de qualification s'affiche.

État de la qualification

- Succès
- Échec

6 Annexe

6.1 Exemple d'une régression linéaire

Régression linéaire à une variable

Dans le cas le plus simple, un mélange a un seul absorbeur et le spectre un seul pic. Les échantillons à différentes concentrations de l'absorbeur présentent des pics avec des valeurs d'absorbance différentes *(voir "Loi de Beer-Lambert", Chapitre 2.2.1, page 6)*.

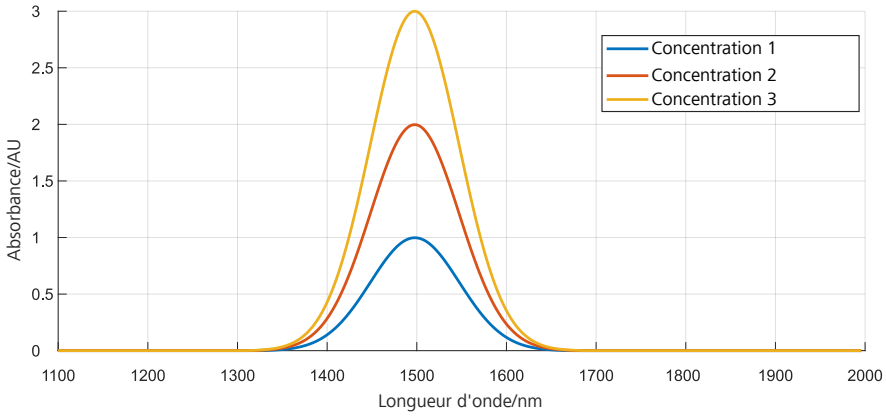


Figure 38 Données modélisées pour 3 échantillons avec un pic à 1 500 nm.

Les 3 valeurs d'absorbance mesurées à 1 500 nm peuvent être appliquées contre la concentration de l'absorbeur.

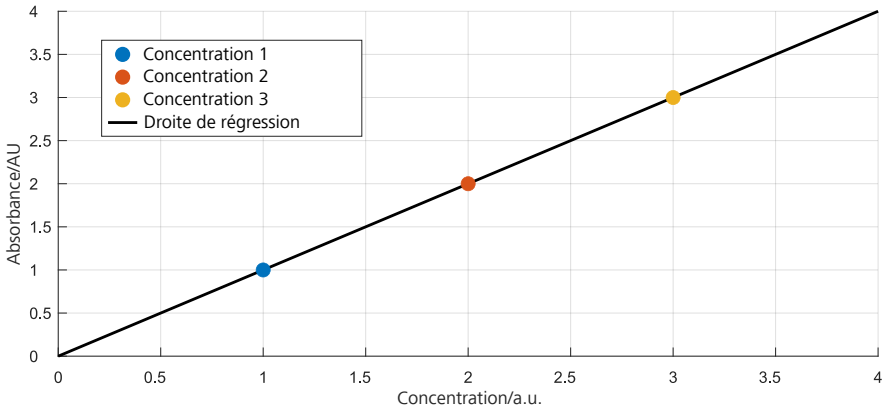


Figure 39 Relation entre valeurs d'absorbance et concentrations

Selon la loi de Beer-Lambert, la relation entre les valeurs d'absorbance et les concentrations est linéaire. Avec le jeu de 3 échantillons, il est possible d'effectuer une régression linéaire qui donne une droite de régression comme illustré.

Comme il n'y a qu'une seule variable (1 longueur d'onde), il s'agit d'une régression linéaire à une variable. Cette régression peut être utilisée comme modèle de quantification. Pour un échantillon avec une concentration inconnue l'absorbance A est mesurée pour 1 500 nm. La concentration correspondante c de l'absorbeur résulte de la droite de régression.

$$c = bA$$

Le coefficient *b* est constant et identique à la pente de la droite de régression.

Il convient de noter que tous les échantillons doivent contenir le même absorbeur avec le même coefficient d'extinction molaire. En outre, toutes les mesures d'absorption doivent être effectuées avec des épaisseurs de couches identiques.

Régression linéaire à plusieurs variables

Les mélanges réels contiennent plus d'un absorbeur. Le spectre enregistré est la somme de tous les spectres d'absorbeurs *(voir "Loi de Beer-Lambert", Chapitre 2.2.1, page 6)*.

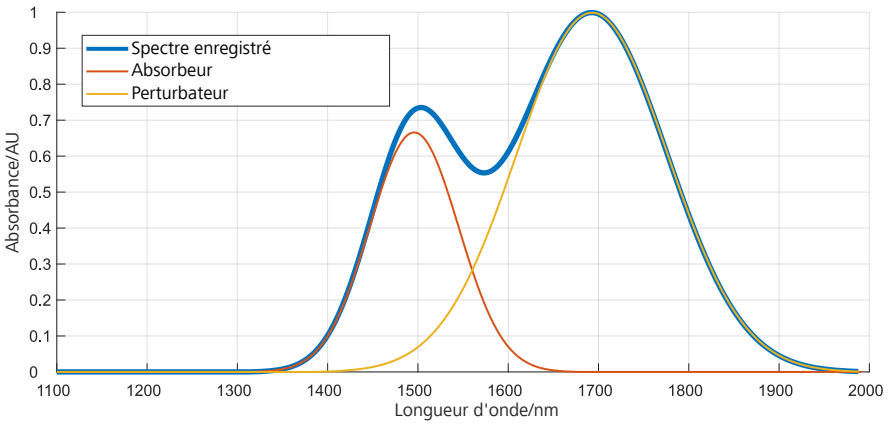


Figure 40 Données modélisées avec 2 composants. L'absorbeur (ligne rouge) doit être quantifié.

Le spectre enregistré (ligne bleue) est la somme du spectre de l'absorbeur pur et d'un spectre de perturbations qui se chevauchent. À 1 500 nm, la valeur d'absorbance mesurée se compose non seulement de l'absorbance de l'absorbeur, mais aussi de celle du perturbateur. Le paramètre d'intérêt ne peut pas être quantifié avec une seule mesure à 1 500 nm. Il est également impossible de savoir si un perturbateur était présent et si la mesure est fiable.

Que se passe-t-il lorsque, par exemple, 2 longueurs d'onde sont mesurées, par exemple, à 1 500 nm et 1 700 nm ? L'absorbance mesurée à la longueur d'onde 1, A_1 est la somme du signal de l'absorbeur pur A_1^a (index a = absorbeur) et du signal perturbateur pur A_1^f (index f = perturbateur). Il en est de même pour l'absorbance mesurée pour la longueur d'onde 2, A_2 :

$$A_1 = A_1^{\mathrm{a}} + A_1^{\mathrm{f}} = \varepsilon_1^{\mathrm{a}} c_{\mathrm{a}} + \varepsilon_1^{\mathrm{f}} c_{\mathrm{f}}$$
$$A_2 = A_2^{\mathrm{a}} + A_2^{\mathrm{f}} = \varepsilon_2^{\mathrm{a}} c_{\mathrm{a}} + \varepsilon_2^{\mathrm{f}} c_{\mathrm{f}}$$

où $\varepsilon_1{}^{\mathrm{a}}$ et $\varepsilon_1{}^{\mathrm{f}}$ représentent les coefficients d'extinction molaires à la longueur d'onde 1 pour l'absorbeur ou le perturbateur et c_{a} et c_{f} les concentrations pour l'absorbeur ou le perturbateur.

Dans les équations ci dessus, l'épaisseur de couche l est exclue de la loi de Beer-Lambert. Cela rend le calcul algébrique un peu plus facile par la suite. L'épaisseur de couche doit naturellement être la même pour tous les échantillons. Les absorbances dans les équations sont par conséquent des absorbances par cm.

Les équations sous forme de matrices peuvent être décrites comme suit :

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} \varepsilon_1^{\mathrm{a}} & \varepsilon_1^{\mathrm{f}} \\ \varepsilon_2^{\mathrm{a}} & \varepsilon_2^{\mathrm{f}} \end{bmatrix} \begin{bmatrix} c_{\mathrm{a}} \\ c_{\mathrm{f}} \end{bmatrix}$$

Par conséquent :

$$\begin{bmatrix} c_{\mathrm{a}} \\ c_{\mathrm{f}} \end{bmatrix} = \begin{bmatrix} \varepsilon_1^{\mathrm{a}} & \varepsilon_1^{\mathrm{f}} \\ \varepsilon_2^{\mathrm{a}} & \varepsilon_2^{\mathrm{f}} \end{bmatrix}^{-1} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

Leur résolution pour la concentration de l'absorbeur donne :

$$c = c_{\mathrm{a}} = \frac{\varepsilon_2^{\mathrm{f}}}{\varepsilon_1^{\mathrm{a}}\varepsilon_2^{\mathrm{f}} - \varepsilon_1^{\mathrm{f}}\varepsilon_2^{\mathrm{a}}} A_1 + \frac{-\varepsilon_1^{\mathrm{f}}}{\varepsilon_1^{\mathrm{a}}\varepsilon_2^{\mathrm{f}} - \varepsilon_1^{\mathrm{f}}\varepsilon_2^{\mathrm{a}}} A_2$$

Il en résulte que la concentration de l'absorbeur peut être calculée même en présence d'un perturbateur en mesurant l'absorbance à deux longueurs d'onde et en multipliant chaque absorbance par une constante.

Les constantes se réfèrent aux coefficients d'extinction molaires et peuvent être consultées dans des tableaux. Mais cela n'arrive jamais dans la réalité. Au lieu de cela, elles sont déterminées par une étape de calibrage et la solution du système d'équations linéaires par une régression linéaire multivariée comme la régression MCP. Les constantes sont donc désignées comme coefficients de régression b_1 et b_2.

$$c = b_1 A_1 + b_2 A_2$$

Plus de 2 absorbeurs

Comme indiqué ci-dessus, il suffit pour 1 absorbeur de déterminer l'absorbance pour 1 longueur d'onde. Pour 2 absorbeurs, il suffit de déterminer l'absorbance pour 2 longueurs d'onde.

Cette opération peut être généralisée. Plus d'absorbeurs demandent plus de valeurs d'absorbance A_i pour différentes longueurs d'onde i. Il s'agit toujours d'une relation linéaire :

$$c = b_1 A_1 + b_2 A_2 + \cdots + b_n A_n$$

Développement d'un modèle de quantification

Avant que l'équation ci-dessus puisse prédire la concentration dans des échantillons inconnus, il faut déterminer les coefficients b_1, b_2, etc. Il faut pour cela une étape de calibrage. Plusieurs échantillons à différentes concentrations du paramètre d'intérêt sont mesurées.

En accord avec la terminologie utilisée ultérieurement pour l'ACP et le MCP, *c* peut être remplacé par *y* et *A* par *x*. L'équation ci-dessus se présente alors comme suit pour chaque échantillon de calibrage :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,f} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,f} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,f} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

où *n* correspond au nombre d'échantillons, *f* au nombre de longueurs d'onde, y_1 à la valeur de référence pour l'échantillon 1, mesuré avec la méthode de référence (titrage), $x_{1,1}$ l'absorbance mesurée de l'échantillon 1 pour la longueur d'onde 1 et $\{x_{1,1} \ldots x_{1,f}\}$ le spectre de l'échantillon 1, mesuré pour *f* longueurs d'onde. $b_1 \ldots b_n$ correspondent aux coefficients de régression et $e_1 \ldots e_n$ aux termes d'erreur, qui montrent à quel point les coefficients de régression modélisent bien les données mesurées.

Sous une forme matricielle plus compacte, cela donne :

$$\mathbf{y} = \mathbf{X^t p} + \mathbf{e}$$

X est défini comme matrice *f* x *n*. $\mathbf{X^t}$ est la matrice transposée **X**, c'est à dire que les lignes et colonnes sont permutées pour recevoir la matrice ci-dessus *n* x *f*. Le vecteur de prédiction **p** correspond aux coefficients de régression ci-dessus **b**.

Le vecteur de prédiction **p** permet de prédire le paramètre d'intérêt d'un nouvel échantillon sur la base de son spectre **x**. La valeur calculée $\hat{y}$ est :

$$\hat{y} = \mathbf{x^t p}$$

La tâche de la régression linéaire multivariée est de déterminer les coefficients de régression qui conduisent aux termes d'erreur minimaux. Une régression linéaire multiple (RLM) nécessiterait toutefois un nombre d'échantillons de calibrage supérieur au nombre de longueurs d'onde. Un autre obstacle est la corrélation élevée entre les variables.

D'autres méthodes peuvent être utilisées dans le cadre des prédictions spectroscopiques. L'ACP peut réduire considérablement le volume de données et éliminer complètement la corrélation. De plus, une régression MCP prend en compte les valeurs de référence des échantillons.

6.2 Algorithme ACP

L'**analyse en composantes principales** (**ACP**, en anglais *principal component analysis*) est utilisée pour la répartition automatique des jeux de données et la détection des valeurs spectrales atypiques *(voir "Analyse en composantes principales (ACP)", Chapitre 4.2, page 36)*.

Pour l'analyse en composantes principales des spectres paramétrés et centrés sur les valeurs moyennes, le logiciel OMNIS effectue une **décomposition en valeurs singulières** (**SVD**, en anglais *singular value decomposition*). Il décompose la matrice de données initiale **X** (les spectres du jeu de données de calibrage) en 3 matrices et calcule n composantes principales :

$$\mathbf{X} = \mathbf{L}\Sigma\mathbf{S}^{\mathbf{t}}$$

où **X** représente les données spectrales initiales (une matrice *f* x *n* avec *f* longueurs d'onde et *n* échantillons), **L** le poids factoriel (une matrice *f* x *n*), **Σ** les valeurs singulières (une matrice *n* x *n*) et **S** les scores (une matrice *n* x *n*). L'équation peut être représentée sous forme graphique *(voir Figure 41, page 94)*.

Les poids factoriels **L** projettent les axes de l'espace des longueurs d'onde sur les axes de l'espace des composantes principales. Les vecteurs de colonnes de **L** sont les axes principaux ou les directions principales.

Les scores **S** sont les projections des spectres initiaux sur les axes principaux. Chaque vecteur de ligne de **S** contient un spectre déterminé.

Σ est une matrice diagonale des valeurs singulières σ_i. Les valeurs singulières décrivent la variance expliquée par chaque composante principale. Par convention elles sont triées de façon à ce que $\sigma_1 > \sigma_2 > \ldots > \sigma_n$.

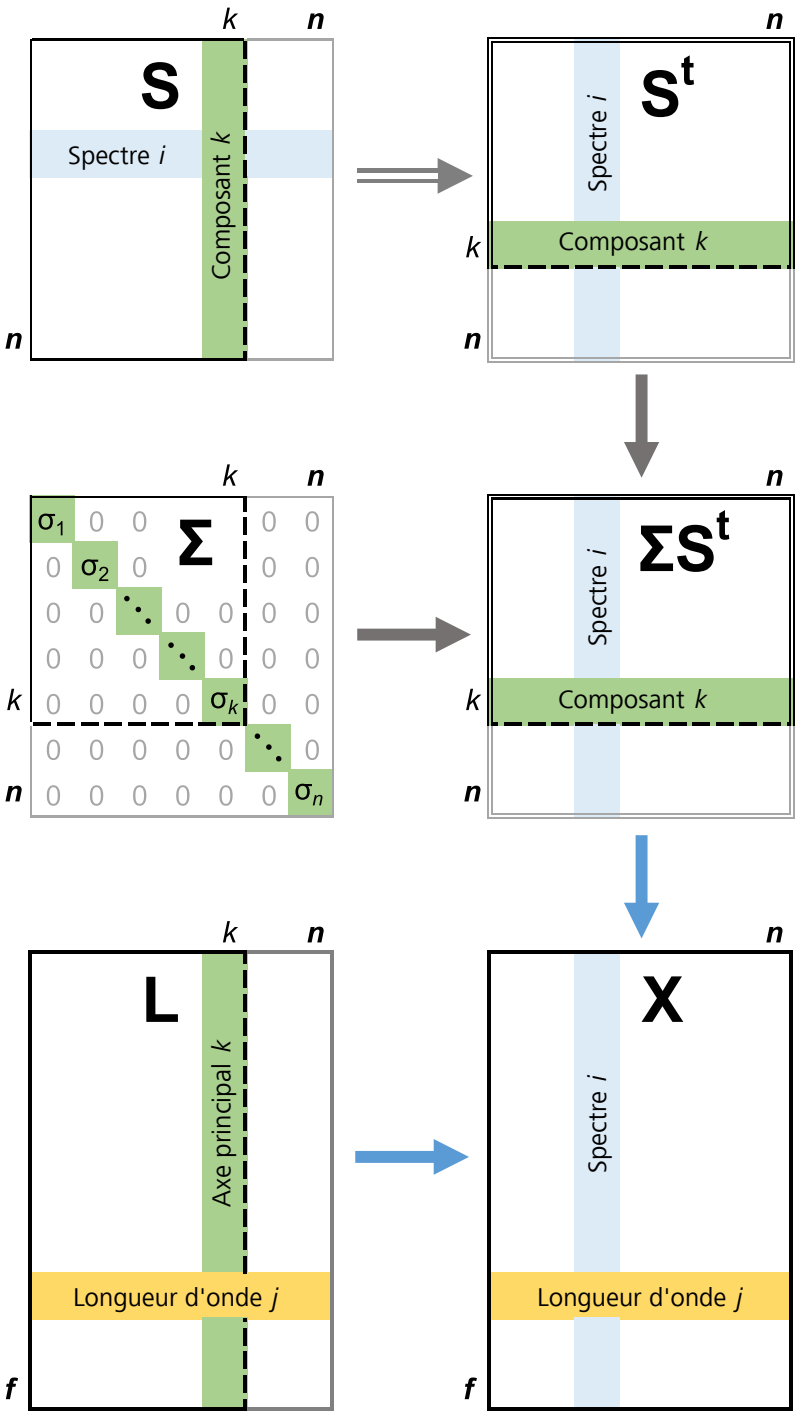


Figure 41 *Équation de la décomposition des valeurs singulières sous forme graphique. Les lignes pointillées indiquent les matrices raccourcies pour un modèle avec k composantes principales. Les informations contenues dans les n-k composantes principales omises sont intégrées dans la matrice résiduelle (une matrice f × n, non représentée).*

Matrice résiduelle

Un modèle ACP utilise uniquement les quelques premières des *n* composantes principales calculées. Sur la *figure 41*, il s'agit des *k* premières composantes principales de *n*. Les données originales **X** peuvent être divisées en données décrites et en données non décrites par le modèle :

$$\mathbf{X} = \mathbf{L_a}\mathbf{\Sigma_a}\mathbf{S_a^t} + \mathbf{E}$$

où **X** correspond aux données spectrales initiales (une matrice *f* x *n*), $\mathbf{L_a}$ (une matrice *f* x *k*) les premières colonnes *k* de **L**, $\mathbf{\Sigma}_a$ (une matrice diagonale *k* x *k*) les premières valeurs singulières *k*, $\mathbf{S_a}$ (une matrice *n* x *k* avec *k* composantes principales) les premières colonnes *k* de **S** et **E** la matrice résiduelle (une matrice *f* x *n*) qui contient toutes les variations spectrales dans **X** qui ne peuvent pas être décrites par le modèle.

En règle générale, *k* est $\ll n \ll f$. Par exemple : $k = 3$ composantes principales, $n = 100$ échantillons et $f = 2\ 500$ longueurs d'onde.

Chaque colonne $\mathbf{e}_i$ de la matrice résiduelle **E** montre la distance orthogonale entre le spectre *i* et l'espace ACP appelé **résidu**. Plus le modèle utilise des composantes principales, plus le résidu est petit.

6.3 Algorithme MCP

La **régression des moindres carrés partiels** (**régression MCP**, en anglais *partial least squares regression*) est utilisée pour calculer le modèle de quantification *(voir "Régression MCP", Chapitre 4.4.1, page 61)*.

La MCP inclut deux blocs de données :

- La matrice paramétrée et centrée sur la valeur moyenne **X** (les spectres).
- Le vecteur **y** centré sur la valeur moyenne (les valeurs de référence).

La MCP décompose la matrice **X** en 2 matrices :

$$\mathbf{X} = \mathbf{L}\mathbf{S}^\mathbf{t} + \mathbf{Z}$$

où **X** (une matrice *f* x *n* avec *f* longueurs d'onde et *n* échantillons) les spectres prétraités et centrés sur la valeur moyenne, **L** les poids factoriels (une matrice *f* x *k* avec *k* variables latentes), **S** les scores (une matrice *n* x *k*) et **Z** la matrice résiduelle (une matrice *f* x *n*), qui contient toutes les variations spectrales dans **X**, qui ne peuvent pas être décrites par le modèle.

Alors que dans l'ACP, la matrice de score **S** explique la variance de **X**, dans la MCP, la matrice de score **S** explique la covariance entre **X** et **y**. La MCP maximise la covariance expliquée par les scores. Cela signifie que les scores expliquent non seulement mieux la variance de **X**, mais qu'ils

présentent également la corrélation la plus élevée possible avec les valeurs de référence.

Afin de maximiser la covariance entre **X** et **y**, l'algorithme MCP échange des données entre **X** et **y**. Par conséquent, **X** et **y** fusionnent en un seul système intégré. La régression des scores **S** se fait alors selon les valeurs de référence **y**, pour obtenir les coefficients de régression **b** :

$$\mathbf{y} = \mathbf{Sb} + \mathbf{e}$$

où **e** est le vecteur résiduel qui contient toutes les variations de la valeur de référence en **y** qui ne peuvent pas être décrites par le modèle.

Prédiction

À partir des coefficients de régression **b**, il est possible de déterminer le vecteur de prédiction **p**. Le vecteur de prédiction **p** et le spectre prétraité et centré sur la moyenne **x** sont utilisés pour prédire le paramètre d'intérêt $\hat{y}$ d'un nouvel échantillon :

$$\hat{y} = \mathbf{x}^{\mathbf{t}}\mathbf{p}$$

Le logiciel OMNIS implémente la MCP avec l'algorithme SIMPLS et un jeu unique de valeurs de référence (PLS-1).

6.4 T² de Hotelling et résidus Q

Les *T²* de Hotelling et les résidus Q caractérisent les spectres dans un modèle ACP ou MCP. Ils aident notamment à identifier des valeurs atypiques potentielles *(voir « T² de Hotelling et résidus Q », page 51)*.

T² de Hotelling

La distance de Mahalanobis est une mesure de la grandeur de l'écart d'un spectre par rapport au centre du modèle. La distance est normalisée. Chaque composante principale ou variable latente a le même poids.

En supposant que les spectres ou les scores sont normalement distribués, les carrés des distances de Mahalanobis, MD², suivent une distribution *T²* de Hotelling :

$$\mathrm{MD}^2 \sim T^2$$

La distance de Mahalanobis au carré pour le spectre *i* est la somme quadratique des scores normalisés pour les *k* premières composantes principales ou variables latentes :

$$\mathrm{MD}_i^2 = \mathbf{s_i}\mathbf{s_i^t} = \sum_{a=1}^{k} s_{i,a}^2$$

où $\mathbf{s_i}$ correspond à la *i*ème ligne de la matrice de scores raccourcie **S**, $s_{i,a}$au score normalisé pour le spectre *i* et la composante principale (ou variable latente) *a* et *k* au nombre de composantes principales ou de variables latentes utilisées.

MD² peut être désigné simplement comme T^2. Un spectre avec $T^2 = 0$ se projette sur le centre du modèle centré sur la moyenne, c'est-à-dire que tous les scores se situent dans la moyenne. Plus la projection d'un spectre sur l'hyperplan s'éloigne du centre, plus la valeur T^2 est élevée. C'est près du point central que le modèle s'adapte le mieux. Loin du point central, le modèle risque de mal s'adapter.

Il est possible de définir une **pertinence** de 5 %, par exemple, pour un test d'hypothèse qui détecte les valeurs atypiques *(voir "Valeur spectrale atypique - algorithme", Chapitre 6.5, page 98)*.

Résidu Q

Le résidu Q pour le spectre *i* est la distance quadratique entre le modèle ACP ou MCP et le spectre :

$$Q_i = \mathbf{e_i^t e_i} = \sum_{j=1}^{f} e_{i,j}^2$$

où $\mathbf{e_i}$ représente la *i*ème colonne de la matrice résiduelle **E**, $e_{i,j}$ le résidu pour le spectre *i*, et *f* le nombre des longueurs d'onde.

Les résidus Q indiquent les variations qui ne sont pas expliquées par le modèle. Un résidu Q élevé indique que le spectre risque de ne pas correspondre au modèle, par ex., si l'échantillon mesuré contient une autre substance.

Il est possible de définir une **pertinence** de 5 %, par exemple, pour un test d'hypothèse qui détecte les valeurs atypiques *(voir "Valeur spectrale atypique - algorithme", Chapitre 6.5, page 98)*.

6.5 Valeur spectrale atypique - algorithme

La détection de valeurs spectrales atypiques identifie des spectres qui s'écartent de l'intégralité *(voir « Détection de valeurs spectrales atypiques », page 52)*. L'algorithme évalue si la valeur du T^2 de Hotelling ou du résidu Q du spectre testé est le résultat d'une variation aléatoire ou systématique.

Détection de valeurs spectrales atypiques pour le développement de modèles

1. Le paramétrage est considéré comme suit :
   a. À partir de la version logicielle d'OMNIS 4.2 : l'utilisateur décide si le paramétrage (prétraitement des données et sélection des longueurs d'onde) doit être appliqué ou non.
   Les modifications ultérieures du paramétrage n'ont aucune incidence sur la répartition des ensembles de données.
   b. À partir de la version logicielle d'OMNIS 3.3 jusqu'à la version 4.1 : l'utilisateur décide si le prétraitement des données doit être pris en compte ou non. La sélection des longueurs d'onde et les modifications ultérieures du prétraitement des données n'ont aucune incidence sur la répartition des ensembles de données.
   c. Jusqu'à la version logicielle d'OMNIS 3.2 : le prétraitement des données est pris en compte tel qu'il a été défini au moment de la détection des valeurs atypiques. La sélection des longueurs d'onde et les modifications ultérieures du prétraitement des données n'ont aucune incidence sur la répartition des ensembles de données.
2. La détection des valeurs spectrales atypiques s'appuie sur le modèle ACP de tous les spectres centrés sur les valeurs moyennes dans la liste des spectres *(voir "Analyse en composantes principales (ACP)", Chapitre 4.2, page 36)*. Le spectre à tester est également inclus dans le modèle ACP. Le nombre de composants principaux est choisi de manière à ce que la variance expliquée soit d'au moins 95 %.

3. **Valeur atypique T² de Hotelling** :
   Sur la base des valeurs *T²* de Hotelling *(voir « T² de Hotelling », page 96)*, un test d'hypothèse est effectué selon H. Hotelling, *The Generalization of Student's Ratio*, The Annals of Mathematical Statistics Vol. 2, N° 3 (août 1931), p. 360-378.
   a. L'hypothèse nulle est que la valeur *T²* du spectre à étudier correspond aux valeurs *T²* du modèle ACP. Si l'hypothèse nulle est vraie, *T²* suit une distribution *T²* de Hotelling. La distribution peut être exprimée en distribution *F* :
   $$T^2 \sim \frac{k(n-1)}{n-k} F_{k,n-k}$$
   où *k* est le nombre de composantes principales, *n* le nombre de spectres et $F_{k,n-k}$ la distribution *F* avec les paramètres *k* et *n-k*.
   b. La pertinence ajustable contrôle la probabilité de rejet de l'hypothèse nulle si elle est vraie (connue sous le nom d'erreur de type I). La valeur par défaut est 5 %.
   c. Une valeur critique est calculée pour *T²* sur la base de cette distribution et de la pertinence.
   d. Si la valeur *T²* pour le spectre est supérieure à la valeur critique, l'hypothèse nulle est rejetée et le spectre est marqué comme potentiellement atypique.
4. **Valeur atypique de résidus Q**
   Sur la base des résidus Q *(voir « Résidu Q », page 97)* un test d'hypothèse est effectué selon E. Jackson et G. S. Mudholkar, *Control Procedures for Residuals Associated With Principal Component Analysis*, Technometrics Vol. 21, N° 3 (août 1979), p. 341–349.
   a. L'hypothèse nulle est que le résidu Q du spectre à étudier correspond aux résidus Q du modèle ACP. Si l'hypothèse nulle est vraie, il est possible d'établir une statistique du test de résidu Q à peu près normalement distribuée.
   b. La pertinence ajustable contrôle la probabilité de rejet de l'hypothèse nulle si elle est vraie (connue sous le nom d'erreur de type I). La valeur par défaut est 5 %.
   c. Une valeur critique est calculée pour Q sur la base de cette statistique de test et de la pertinence.
   d. Si la valeur du résidu Q pour le spectre est supérieure à la valeur critique, l'hypothèse nulle est rejetée et le spectre est marqué comme potentiellement atypique.

Évaluation de la valeur atypique pour la prédiction (quantification)

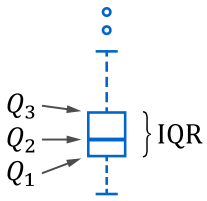
Lors de la quantification, une évaluation des valeurs atypiques est disponible pendant la prédiction :

1. Étape de préparation :
   a. Le modèle de quantification utilise le même algorithme ci-dessus. Toutefois, la base est le modèle de quantification (modèle MCP) avec tous les spectres centrés sur la moyenne de l'ensemble de données de calibrage. Le prétraitement des données, les gammes de longueurs d'onde et le nombre de variables latentes sont pris en compte comme indiqué dans le modèle de quantification.
   b. Sur la base de la pertinence définie, le modèle de quantification calcule les valeurs critiques pour T^2 et Q (lignes en pointillés sur le graphique d'influence MCP). Les valeurs critiques sont enregistrées dans le modèle de quantification.
2. Lors de la prédiction, les valeurs T^2 et Q du spectre sont calculées sur la base du modèle de quantification.
3. Si T^2 ou la valeur Q du spectre est supérieure à la valeur critique correspondante, l'échantillon est considéré comme une valeur atypique potentielle par rapport au modèle de quantification appliqué.

6.6 Valeur de référence atypique - algorithme

Les box plots permettent d'identifier des valeurs atypiques pour les valeurs de référence *(voir "Valeur de référence atypique (quantification)", Chapitre 4.3.4, page 58)*.

Pour tenir compte du caractère asymétrique de la distribution, les valeurs limites atypiques sont ajustées à l'aide des calculs suivants.



Le **Medcouple** (MC) mesure l'asymétrie des valeurs de référence. Le calcul commence par la médiane du box plot, Q_2. Une fonction est calculée avec toutes les paires possibles (en anglais *couples*) de la moitié supérieure et de la moitié inférieure des valeurs de référence. La médiane du résultat est le Medcouple :

$$\mathrm{MC} = \underset{y_i \le Q_2 \le y_j}{\mathrm{med}} \frac{(y_j - Q_2) - (Q_2 - y_i)}{y_j - y_i}$$

où Q_2 correspond au deuxième quartile qui définit la ligne médiane dans le box plot et y_i, y_j une paire de valeurs de références.

Le Medcouple se situe toujours entre -1 et 1. Pour une distribution symétrique MC = 0. Une distribution asymétrique avec MC > 0 est déformée vers les valeurs de référence les plus élevées, avec MC < 0 vers les valeurs de référence les plus basses.

Le calcul des **valeurs limites atypiques ajustées** dépend du côté vers lequel la distribution est décalée :

$$\mathrm{MC} \geq 0\text{: } [Q_1 - 1.5\, e^{-4\mathrm{MC}}\, \mathrm{IQR}\,;\, Q_3 + 1.5\, e^{3\mathrm{MC}}\, \mathrm{IQR}]$$
$$\mathrm{MC} < 0\text{: } [Q_1 - 1.5\, e^{-3\mathrm{MC}}\, \mathrm{IQR}\,;\, Q_3 + 1.5\, e^{4\mathrm{MC}}\, \mathrm{IQR}]$$

Pour une distribution symétrique, (MC = 0), les distances entre les valeurs limites et la boîte sont de 1,5 IQR.

La fonction exponentielle permet une détection précise et robuste des valeurs atypiques pour différentes distributions avec différentes asymétries, comme l'ont démontré empiriquement M. Hubert et E. Vandervieren, *An adjusted boxplot for skewed distributions*, Computational Statistics & Data Analysis Vol. 52, N° 12 (août 2008), p. 5186-5201.

Le pourcentage attendu de valeurs atypiques marquées est d'environ 1 % et est assez similaire au pourcentage du box plot standard pour la distribution normale. Remarque : ce pourcentage ne dépend pas de la pertinence.