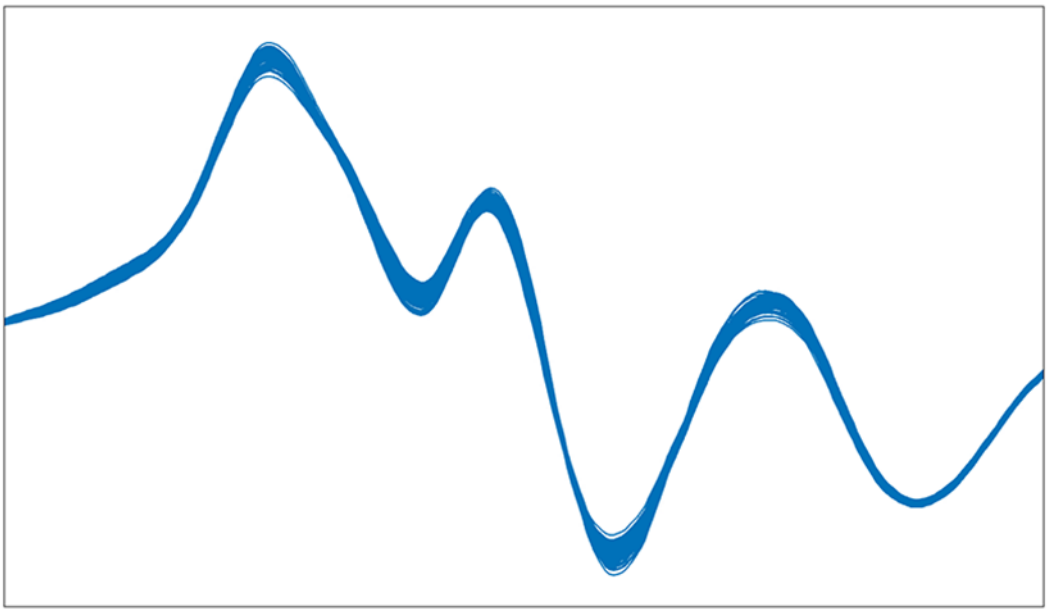
Teoría de OMNIS NIR



Manual

8.0600.8101ES / v9 / 2025-10-10

Metrohm

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

Teoría de OMNIS NIR

Manual

8.0600.8101ES / v9 /
2025-10-10

Technical Communication
Metrohm AG
CH-9100 Herisau

Esta documentación constituye un documento original.

Esta documentación se ha elaborado con la mayor precisión. No obstante puede que haya algún error. Le rogamos nos informe de eventuales errores a la dirección arriba indicada.

Exención de responsabilidad

La garantía no incluye deficiencias que surjan por circunstancias que no sean responsabilidad de Metrohm, tales como un almacenamiento inadecuado, uso inapropiado, etc. Las modificaciones no autorizadas en el producto (por ejemplo, conversiones o accesorios) excluyen cualquier responsabilidad del fabricante por los daños resultantes y sus consecuencias. Deben seguirse estrictamente las instrucciones y notas de la documentación del producto de Metrohm. En caso contrario, queda excluida la responsabilidad de Metrohm.

Índice

1 Información general

1.1 Introducción

La espectroscopía del infrarrojo cercano (espectroscopía NIR) es un método de análisis no destructivo, rápido y sin reactivos, adecuado para una amplia gama de muestras. Puede analizar varios parámetros simultáneamente y determinar las propiedades físicas y químicas de un material. Entre otras: Las concentraciones de analitos, la densidad, el tamaño de partículas y la viscosidad intrínseca.

La espectroscopía NIR también permite identificar muestras desconocidas (a partir de OMNIS Software 4.0) y la verificación de muestras (a partir de OMNIS Software 4.2).

La capacidad de medir muestras a distancia y de forma no destructiva es fundamental en el control de calidad y la supervisión de procesos.

En este manual se describen las técnicas y los algoritmos para la adquisición, el procesamiento y el análisis de los espectros del infrarrojo cercano, tal y como se han implementado en OMNIS Software. En el capítulo 2 se explica brevemente cómo se convierten las señales de medida en espectros de absorción. El capítulo 3 se refiere a la calibración del aparato. El capítulo 4 describe el desarrollo de modelos que pueden predecir el parámetro de interés (cuantificación) o la pertenencia al producto (identificación). El capítulo 5 se ocupa de la predicción de incógnitas. El capítulo 6 constituye el apéndice con explicaciones de diversos algoritmos.

1.2 Marco conceptual

Los procesos presentados se integran en el siguiente marco:

1. **Calibración, estandarización y pruebas de rendimiento**
   Garantizar la transferibilidad y fiabilidad de los espectros de absorción registrados con el aparato.
2. **Desarrollo del modelo**
   Se desarrolla un modelo para la predicción de un parámetro cuantitativo o para la identificación de muestras.
   El desarrollo se basa en muestras con un parámetro de interés conocido o una pertenencia al producto conocida.
3. **Análisis de muestras**
   Se registra un espectro de la muestra analizada. A partir del espectro, un modelo de cuantificación suministra una predicción cuantitativa o un modelo de identificación identifica o verifica la muestra.

4. **Monitorización**
   La monitorización del modelo y del aparato confirma que el sistema es adecuado para otros análisis.

1.3 Acerca de la documentación

Convenciones gráficas

Las letras mayúsculas en negrita representan matrices, y las letras minúsculas en negrita representan vectores. La transpuesta de una matriz o un vector se indica mediante un superíndice t, por ejemplo $\mathbf{x}^{\mathbf{t}}$.

Los escalares se representan con letras minúsculas. El circunflejo (^) indica los valores estimados (calculados), por ejemplo, $\hat{y}$. Una barra superior indica un valor medio, por ejemplo, $\overline{y}$.

Un escalar puede representar una variable que depende de la longitud de onda, por ejemplo, la absorbancia A. La notación escalar se puede intercambiar con la notación vectorial.

1.4 Información adicional

En las siguientes páginas se encuentra información adicional sobre el producto:

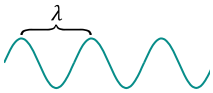
- Sitio web de Metrohm *https://www.metrohm.com* – Visión conjunta de la familia de productos, documentos en formato PDF, datos de los accesorios e información sobre aplicaciones.
- Ayuda de OMNIS Software *https://guide.metrohm.com* – Información filtrada por tema sobre OMNIS Software.

2 Luz de infrarrojo cercano y espectros

2.1 La luz y su interacción con la materia

Un espectrómetro mide cómo interactúa una muestra con la luz. La luz puede ser absorbida o dispersada en diferentes grados. La interacción depende de las propiedades de la luz, especialmente su longitud de onda, y de las propiedades del material, en particular, de su estructura molecular.

Longitud de onda



La luz es radiación electromagnética. Se mueve como una onda con campos eléctricos y magnéticos oscilantes a través del espacio. Las ondas se propagan en el espacio y el tiempo. Por lo tanto, una onda se caracteriza por su longitud de onda λ (por ejemplo, en nanómetros = 10^{-9} metros) y su frecuencia (Hz).

La longitud de onda es inversamente proporcional a la frecuencia de la onda. Las ondas con frecuencias más altas (más oscilaciones por segundo) tienen longitudes de onda más cortas y a la inversa. Debido a esta relación, la onda se puede describir ya sea por su longitud de onda (nm) o por su frecuencia (Hz).

La luz puede intercambiar energía en unidades cuánticas discretas llamadas fotones. La energía de un solo fotón, *E*, depende de su frecuencia *f* o bien de su longitud de onda λ:

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

Donde *h* es la constante de Planck y *c* es la velocidad de la luz.

La *figura 1* muestra diferentes gamas de radiación electromagnética.

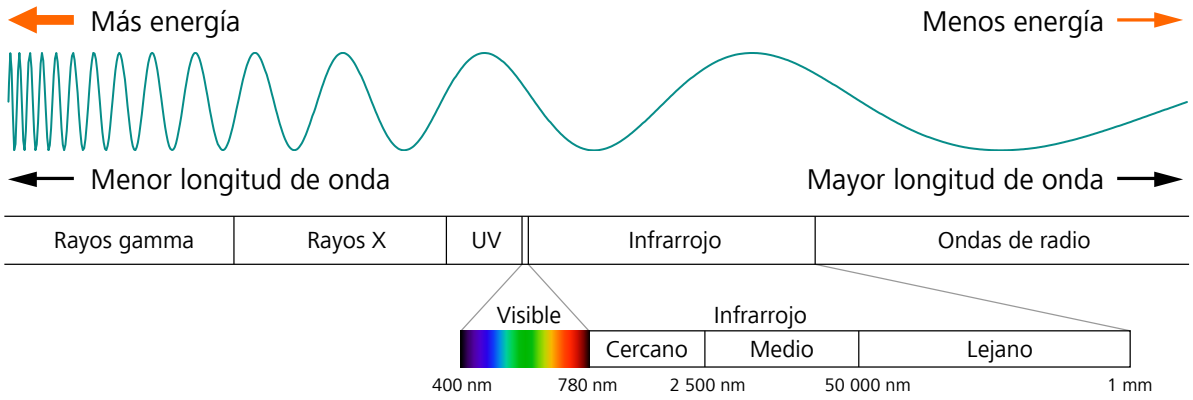


Figura 1 Gamas de radiación electromagnética. La región del infrarrojo cercano (NIR) se encuentra adyacente al rango de la luz visible. La NIR abarca longitudes de onda de 780 nm a 2500 nm.

Fuentes de radiación

Las diferentes gamas de radiación electromagnética tienen diferentes fuentes de radiación. En la gama NIR, la fuente es la **radiación térmica**. Por ejemplo, una cámara de infrarrojos reconoce el cuerpo humano porque tiene una temperatura más alta que su entorno.

Intensidad

La amplitud de la onda electromagnética determina la intensidad de la luz. A mayor amplitud, mayor intensidad. En el caso de la luz visible, la intensidad se percibe como brillo.

Interacción de luz y materia

A continuación, se describe el proceso de absorción de la luz por las moléculas.

La forma en que la luz interactúa con la materia depende del rango de radiación electromagnética. Por ejemplo, cuando la energía de la luz visible se transfiere a las moléculas, los electrones en estas moléculas pasan de un nivel de energía más bajo a uno más alto (transición electrónica). En el rango del infrarrojo, se producen **transiciones vibracionales**. Los enlaces químicos, los grupos funcionales y las moléculas pueden vibrar de diferentes maneras, por ejemplo, a través de vibración de estiramiento, vibración de deformación o vibración de torsión.

Las moléculas solo pueden ocupar estados vibracionales discretos. A temperatura ambiente, la mayor parte de las moléculas están vibrando en el estado fundamental (nivel 0). Las transiciones del estado de vibración fundamental a los estados excitados se denominan según el siguiente esquema:

Transición vibracional *i* ➔ *j*	Nombre
0 ➔ 1	Transición fundamental
0 ➔ 2	Primer sobretono
0 ➔ 3	Segundo sobretono

Un estado vibracional más alto corresponde a un nivel de energía mayor. Para la transición del estado *i* al estado excitado *j* la molécula debe absorber una determinada cantidad de energía de transición ΔE_{ij}.

La luz puede intercambiar energía en partes de $E = hf$, en donde f es la frecuencia de la luz. La absorción de la luz tiene lugar si la energía del fotón hf es igual a la energía de transición ΔE_{ij}.

Los estados vibracionales permitidos dependen, entre otras cosas, de la fuerza de los enlaces y de la masa de los átomos implicados. Por lo tanto, un tipo específico de enlace puede asociarse con energías de transición características o longitudes de onda absorbidas.

Para que se absorba la luz, deben cumplirse otras condiciones. La transición vibracional debe desplazar la distribución de carga, de manera que cambie el momento dipolar eléctrico de la molécula. La probabilidad de absorción de energía depende de la magnitud del cambio en el momento dipolar a lo largo del enlace químico afectado.

Una transición vibracional puede provocar un cambio del momento dipolar tanto en moléculas polares como apolares y grupos funcionales. Las moléculas diatómicas homonucleares como N_2 no absorben luz infrarroja.

La duración del estado vibracional excitado es limitada. Cuando la molécula vuelve a un estado vibracional inferior, la energía se convierte en calor.

La gama espectral NIR

Las longitudes de onda correspondientes a las transiciones fundamentales se encuentran en la región del infrarrojo medio. La región del infrarrojo cercano comprende los sobretonos y las bandas de combinación. La *figura 2* muestra las bandas de longitudes de onda absorbidas por diferentes moléculas y grupos funcionales.

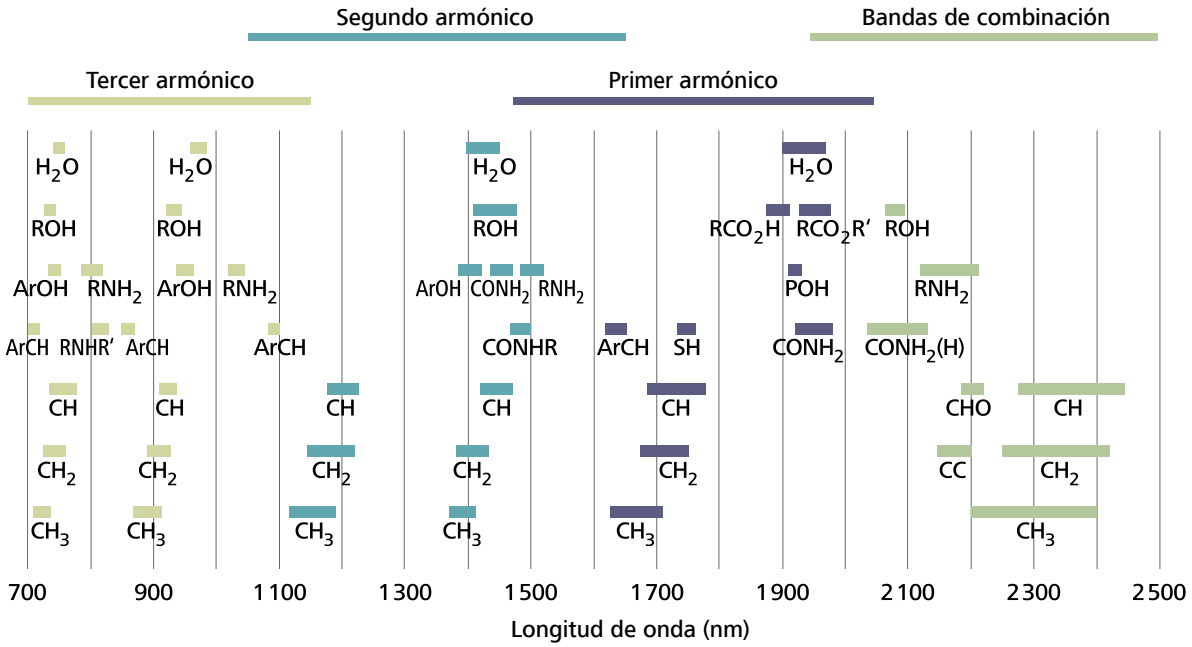


Figura 2 Bandas de absorción del NIR

La transición fundamental es la más probable y la que se produce con mayor frecuencia. Las transiciones de sobretono son menos probables. Por lo tanto, la transición fundamental absorbe más luz que las transiciones de sobretono. En general, la absorción disminuye con cada sobretono. Esto hace que los sobretonos sean adecuados para moléculas fuertemente absorbentes.

Dos o más vibraciones fundamentales pueden excitarse simultáneamente con una frecuencia de luz única correspondiente a las frecuencias combinadas de las vibraciones fundamentales. Las bandas de absorción correspondientes se denominan **bandas de combinación**. Algunas bandas de combinación se encuentran en el rango NIR, concretamente entre 1900 y 2500 nm.

2.2 Fundamentos matemáticos

2.2.1 Ley de Beer-Lambert

La ley de Beer-Lambert describe de qué modo la absorción de la luz por una muestra homogénea depende de las propiedades de una sustancia absorbente en la muestra.

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

Aquí, A corresponde a la absorbancia, ε al coeficiente de extinción molar del absorbedor (L/mol/cm), c a la concentración del absorbedor (mol/L) y l al espesor de capa de la muestra (cm).

El coeficiente de extinción molar ε es una constante que indica cuánto absorbe una sustancia. Este coeficiente es específico para una determinada longitud de onda *i* y una sustancia *j* determinada. La absorbancia total de una mezcla es la suma de la absorbancia de todas las sustancias contenidas en la mezcla:

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

Donde A_i corresponde a la absorbancia a la longitud de onda *i*; *N* es el número de sustancias en la mezcla; ε_{ij} es el coeficiente de extinción molar para la longitud de onda *i* y la sustancia *j* y c_j la concentración de la sustancia *j*.

La ley de Beer-Lambert asume una relación lineal entre la absorbancia y la concentración, así como una relación lineal entre la absorbancia y el coeficiente de extinción molar. Este comportamiento lineal se cumple en muchas situaciones.

Basándose en la ley de Beer-Lambert, las mediciones espectroscópicas de absorción pueden detectar:

- Fluctuaciones en la concentración de un absorbedor.
  Esta es la aplicación más común.
- Fluctuaciones en los factores que afectan al coeficiente de extinción molar.
  La temperatura, la viscosidad, el valor de pH o la constante dieléctrica del disolvente pueden influir en el coeficiente de extinción molar. En algunos casos, esto puede ser aprovechado para las medidas espectroscópicas.

Los efectos de dispersión no están relacionados con la ley de Beer-Lambert. En ocasiones, los efectos de dispersión se pueden usar para detectar, por ejemplo, fluctuaciones en el tamaño de las partículas.

2.2.2 Regresión lineal

Una longitud de onda

En el caso más simple, una mezcla tiene solo un absorbedor. Según la ley de Beer-Lambert, la absorbancia a una longitud de onda específica es lineal con respecto a la concentración del absorbedor.

En la figura 1, cada punto representa una muestra con una concentración conocida del absorbedor (eje x) y la absorbancia medida (eje 'y'). Una regresión lineal produce la recta de regresión **A**.

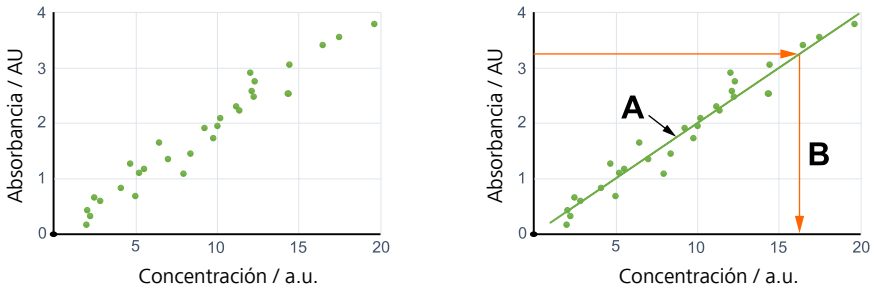


Figura 3 Relación entre valores de absorbancia y concentraciones

Para una muestra con una concentración desconocida del absorbedor, la concentración se puede determinar de la siguiente manera:

1. Medida de la absorbancia a la longitud de onda especificada.
2. Uso de las líneas de regresión para determinar la concentración (**B**). La recta de regresión es un modelo de cuantificación simple para predecir el parámetro de interés, como la concentración del absorbedor.

Sin embargo, este procedimiento falla si la mezcla contiene varios absorbedores con diferentes concentraciones.

Varias longitudes de onda

En lugar de medir la absorción en una sola longitud de onda, se puede medir la absorción en varias longitudes de onda, obteniendo así un espectro. Como en el caso anterior, se puede determinar la relación entre los espectros y el parámetro de interés mediante una regresión lineal. Para múltiples longitudes de onda, se requiere una regresión lineal múltiple (MLR).

Se ha demostrado que la regresión lineal múltiple puede predecir el parámetro de interés incluso cuando la mezcla contiene varios absorbedores con diferentes concentraciones *(véase "Ejemplo de una regresión lineal", capítulo 6.1, página 90)*.

Sin embargo, para una regresión lineal múltiple, se necesitaría un mayor número de muestras que la cantidad de longitudes de onda. Para predicciones espectroscópicas, se pueden usar otros métodos como, por ejemplo, PCA o PLS.

2.3 Cómo se convierte la luz en un espectro

Un espectrómetro (o espectrofotómetro) consta de una fuente de luz y una unidad de detección. La fuente de luz emite luz con un amplio espectro de longitudes de onda, es decir, luz policromática. La luz interactúa con la muestra. A continuación, el espectrómetro captura la luz restante como una función de la longitud de onda.

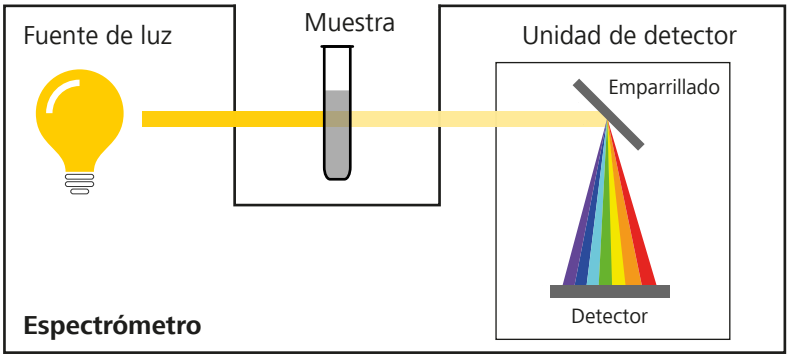


Figura 4 Un espectrómetro con fuente de luz y unidad de detección.

En el espectrómetro, la luz se descompone en longitudes de onda individuales mediante una red de difracción. Un detector mide la luz, y cada longitud de onda incide en un elemento diferente -o píxel- de un sensor lineal.

Un **escaneo** es una medida a través de todos los píxeles. Cada píxel genera una señal fotoeléctrica que es proporcional a la intensidad de luz. Las señales pueden graficarse en función de los píxeles.

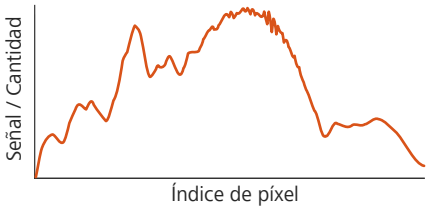


Figura 5 Espectro de la señal del detector como función de los píxeles.

Tiempo de integración

El tiempo de integración es el período durante el cual el detector capta la luz. Un mayor tiempo de integración aumenta la señal.

Sin embargo, tiempos de integración demasiado largos generan la saturación del detector y provocan la pérdida de información. En cambio, tiempos de integración demasiado breves disminuyen la señal y, por lo tanto, la relación señal/ruido.

Un **tiempo de integración automático** garantiza una exposición óptima, es decir, una relación señal/ruido óptima sin saturación. Antes

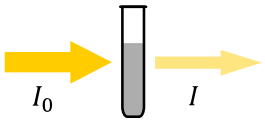
de cada escaneo de la muestra y de cada escaneo de referencia se realizan varias medidas. El tiempo de integración se ajusta para que la señal de mayor intensidad alcance aproximadamente el 90% del rango detectable.

Las diferencias en los tiempos de integración se consideran en los cálculos posteriores.

- **OMNIS NIR Analyzer**
  El tiempo de integración siempre se ajusta automáticamente.
- **2060 The NIR**
  El tiempo de integración se puede ajustar de modo manual o automático.
  Un ajuste manual del tiempo de integración puede reducir la duración de las medidas repetidas y similares. Para evitar la saturación del detector con tiempos de integración demasiado largos, el tiempo de integración se debería ajustar con suficiente margen *(véase "Saturación", página 50)*.

Absorbancia

Las medidas espectroscópicas determinan cuánta luz absorbe o dispersa una muestra. La muestra se expone a un haz de luz. Un detector mide la luz emitida por una fuente de luz y la luz restante después de la interacción con la muestra.



La absorbancia A se define como el logaritmo decimal de la relación entre la intensidad de luz antes de la interacción del haz con la muestra (I_0) y la intensidad de luz después de la interacción del haz con la muestra (I):

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

Una absorbancia de 1 significa que el 10% de la luz atraviesa la muestra, mientras que una absorbancia de 2 significa que solo el 1% la atraviesa.

La absorbancia no tiene unidad.

Escaneo de referencia y escaneo de muestra

De acuerdo con la fórmula anterior, se necesitan dos escaneos para efectuar el cálculo de un espectro de absorción. Los escaneos miden, para cada píxel, una señal fotoeléctrica S:

- Un **escaneo de referencia** mide la señal S_0 antes de la interacción del haz con la muestra.
- Un **escaneo de muestra** mide la señal S después de la interacción del haz con la muestra.

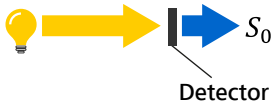
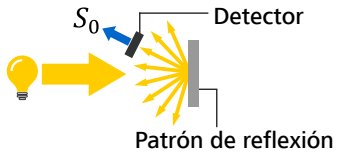
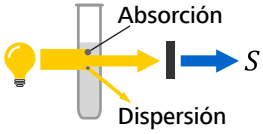
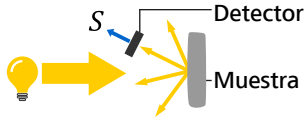
La señal fotoeléctrica medida de un píxel es proporcional al valor medio de la intensidad de luz sobre el área del píxel. Por lo tanto, $S_0/S = I_0/I$. De este modo, las señales fotoeléctricas se pueden usar para calcular la absorbancia:

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

Transmisión y reflexión

En el **modo de transmisión** se mide la luz que atraviesa la muestra. S_0 se mide en ausencia de la muestra. S se mide con la luz que ha pasado a través de la muestra.

En el **modo de reflexión** se mide la luz reflejada por la muestra. Como referencia, en lugar de la muestra, se usa un patrón de reflexión. El patrón de reflexión refleja idealmente el 100% de la luz. Una parte de la luz reflejada se dirige al detector y proporciona la señal S_0. La señal S se mide de la misma manera, pero con la muestra que refleja la luz.

	Transmisión	Reflexión difusa
Sin muestra		
Con muestra		

La absorbancia calculada A representa toda la luz que no llega al detector. Por lo tanto, A incluye no solo la luz absorbida por la muestra, sino también:

- La luz que no llega al detector porque se dispersa lejos de él.
- La luz que se dispersa incorrectamente hacia el detector.

Espectro de absorción

El espectro de absorción se calcula usando el escaneo de referencia (señal S_0) y el escaneo de muestra (señal S) según la fórmula mencionada anteriormente.

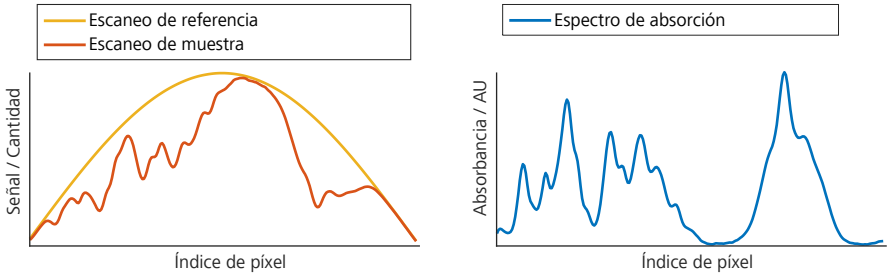


Figura 6 Escaneo de referencia y escaneo de muestra (a la izquierda), así como el espectro de absorción calculado (a la derecha) como función del índice de píxeles.

Para el cálculo anterior se asume que el escaneo de referencia y el escaneo de muestra utilizan los mismos caminos ópticos o caminos ópticos con propiedades ópticas similares. En entornos de procesos se usa un enfoque de referencia en múltiples etapas *(véase "2060 The NIR", capítulo 3.2.2, página 17)*.

De píxeles a longitudes de onda

La escala de píxeles se convierte en la escala de longitudes de onda. El aparato asigna una longitud de onda precisa a cada píxel, por ejemplo:

Pixel 6 ➔ longitud de onda 1009,4 nm

La longitud de onda exacta para cada píxel se determina mediante la calibración de las longitudes de onda *(véase "Calibración de las longitudes de onda", capítulo 3.1, página 14)*.

Conversión de la escala de longitudes de onda

El espectro se convierte a la escala de longitudes de onda estándar mediante interpolación:

1000,0 nm, 1000,5 nm, 1001,0 nm, …

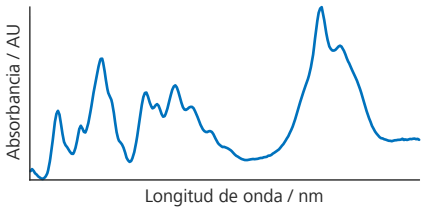


Figura 7 El espectro de absorción en la escala de longitudes de onda.

3 Configuración del aparato

Los siguientes pasos garantizan que se obtengan espectros idénticos para una muestra determinada, con una configuración de medición de muestra específica, independientemente de que se hayan registrado en momentos diferentes y con el mismo dispositivo o con un dispositivo diferente de la familia de productos **OMNIS NIR Analyzer** o con otro aparato del tipo **2060 The NIR**.

Deben tenerse en cuenta tanto el eje x como el eje 'y' de los espectros:

- **Eje x**: calibración de las longitudes de onda *(véase "Calibración de las longitudes de onda", capítulo 3.1, página 14)*
- **Eje 'y'**: normalización de referencia *(véase "Normalización de referencia", capítulo 3.2, página 15)*

Además, se debe garantizar que el rendimiento del aparato cumpla con los requisitos, es decir:

- Las **Pruebas de rendimiento del aparato** deben realizarse correctamente primero para que se puedan registrar espectros con el aparato *(véase "Pruebas de rendimiento del aparato", capítulo 3.3, página 26)*.

Patrones

Para la calibración de las longitudes de onda y la Pruebas de rendimiento del aparato el aparato usa un **patrón de longitud de onda** interno, con trazabilidad metrológica.

Al usar el modo de reflexión, también se necesita un **patrón de reflexión** para la normalización de referencia y las Pruebas de rendimiento del aparato, que varía según el tipo de aparato:

- **OMNIS NIR Analyzer**: patrón de reflexión interno
- **2060 The NIR**: patrón de reflexión externo

3.1 Calibración de las longitudes de onda

La calibración de las longitudes de onda estandariza los valores de las longitudes de onda, es decir, el eje x de los espectros. Asigna cada píxel de la matriz de fotosensores a una longitud de onda.

La calibración de las longitudes de onda utiliza un patrón de longitud de onda interno, con trazabilidad metrológica. El patrón de longitud de onda tiene un espectro de absorción con picos definidos y posiciones de los picos conocidas.

La instrucción **CAL WL** lleva a cabo los siguientes pasos:

1. Se registra el espectro de absorción del patrón de longitud de onda utilizando un camino de referencia interno en la escala de píxeles.
2. Se identifican las posiciones de los picos en el espectro registrado con precisión sub-píxel.
3. Se realiza una regresión polinomial usando las posiciones de los picos medidas en la escala de píxeles y las posiciones de los picos nominales del patrón de longitud de onda.
4. El polinomio de regresión asigna cada píxel a su respectiva longitud de onda.

Los coeficientes de regresión se almacenan en el aparato:

- **OMNIS NIR Analyzer**: para cada presentación de muestras, se guarda un conjunto de coeficientes de regresión en el aparato. Esto significa que para el OMNIS NIR Analyzer Liquid/Solid se debe realizar una calibración de las longitudes de onda y una validación en ambas unidades funcionales.
- **2060 The NIR**: los coeficientes de regresión son específicos de cada aparato. Para todos los canales se usa el mismo conjunto.

Validación de la calibración de las longitudes de onda

Después de realizar una calibración de las longitudes de onda, esta se debe validar. La instrucción **VAL WL** lleva a cabo los siguientes pasos:

1. Se registra el espectro de absorción del patrón de longitud de onda.
2. Validación de las longitudes de onda:
    a. Se identifican las posiciones de los picos en la escala de longitudes de onda en el espectro registrado.
    b. Se calculan los residuos de las longitudes de onda entre las posiciones de los picos medidas y las posiciones de los picos conocidas.
    c. Para cada pico, el residuo de la longitud de onda debe estar dentro de la tolerancia para pasar la prueba.

3. Validación de los anchos de banda:
   a. En el espectro registrado, se determina el ancho de los picos.
   b. Se calculan los residuos de los anchos de banda entre los anchos de los picos medidos y los anchos de los picos conocidos.
   c. Para cada pico, el residuo del ancho de banda debe estar dentro de la tolerancia para pasar la prueba.
4. El estado general de la validación es correcto si todos los residuos mencionados están dentro de la tolerancia.

La validación debe realizarse correctamente primero para poder registrar espectros con el aparato.

3.2 Normalización de referencia

La normalización de referencia normaliza los valores de absorbancia, es decir, el eje 'y' de los espectros.

Determinación de la absorbancia

Para calcular la absorbancia A de una muestra, se necesitan las señales S_0 (escaneo de referencia) y S (escaneo de muestra) *(véase "Cómo se convierte la luz en un espectro", capítulo 2.3, página 9)*:

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

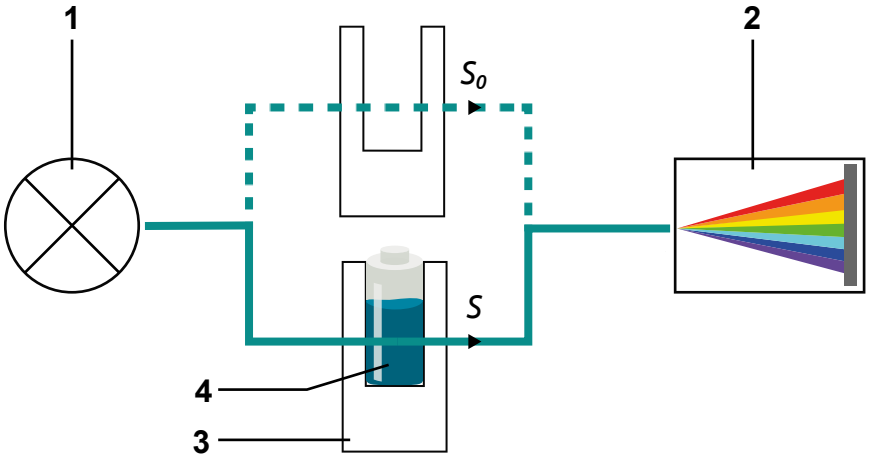


Figura 8 Camino óptico en el modo de transmisión (por ejemplo, con una presentación de muestras líquidas).

En la *figura 8* la luz pasa desde la fuente de luz (**1**) a través de un soporte de muestras (**3**) hasta el detector (**2**).

La señal de referencia S_0 se mide sin la muestra, mientras que la señal S se mide con la muestra (**4**). En caso contrario, las propiedades ópticas para ambos caminos ópticos son idénticas y generan el mismo porcentaje de

atenuación para ambas señales. Esto no altera en nada el resultado de la fórmula anterior.

La ecuación relaciona S con S_0, la referencia. Ambas señales son igualmente importantes. Una desviación en cualquiera de las dos señales conduce a un valor de absorbancia diferente y, en última instancia, a un espectro diferente.

Tanto S como S_0 se ven afectados por las fluctuaciones del aparato y por las condiciones ambientales. Para garantizar que estas influencias se contrarresten mutuamente, ambas señales se deberían medir en un período de tiempo breve.

La implementación de este principio depende del tipo de aparato:

- **OMNIS NIR Analyzer**
  El espectro de absorción de la muestra se calcula con las señales S y S_0*(véase "OMNIS NIR Analyzer", capítulo 3.2.1, página 16)*.
- **2060 The NIR**
  En entornos de proceso, no resulta práctico usar el mismo camino óptico para las medidas de S y S_0. Por lo tanto, deberán tomarse precauciones adicionales *(véase "2060 The NIR", capítulo 3.2.2, página 17)*.

3.2.1 OMNIS NIR Analyzer

La normalización de referencia se realiza midiendo las señales S_0 y S, así como calculando la absorbancia A.

Registrar el espectro de una muestra

Para poder registrar espectros con una unidad funcional, deben realizarse correctamente primero las Pruebas de rendimiento del aparato para esa unidad funcional *(véase "Pruebas de rendimiento del aparato", capítulo 3.3, página 26)*.

1. La muestra debe estar lista en la presentación de muestras.
2. El espectro de absorción se calcula usando el último espectro de referencia registrado S_0. Para obtener un valor actual para S_0, se puede ejecutar la instrucción **MEAS REF SPEC**.
   Al usar la presentación de muestras de materia sólida, el aparato inserta automáticamente un patrón de reflexión en el camino óptico. Este patrón de reflexión no necesita corregir la señal S_0.
3. Con la instrucción **MEAS SPEC** se mide la muestra. Esto da como resultado la señal S.
4. El software calcula A, la absorbancia de la muestra:

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

   En este caso, S_0 corresponde a la señal medida en el camino de referencia y S corresponde a la señal medida a través de la muestra.

3.2.2 2060 The NIR

Los aparatos del tipo **2060 The NIR** requieren una normalización de referencia externa.

Normalización de referencia externa

La medida repetida de la señal S (con la muestra) y de la señal S_0 (sin la muestra) a través de caminos ópticos con propiedades ópticas idénticas necesita mucho tiempo y es susceptible de errores.

Por este motivo, se introducen dos caminos adicionales *(véase figura 9, página 17)*:

- Una **referencia interna** en el aparato. El camino de referencia interna proporciona la señal S_{ref}, que se puede medir fácilmente.
- Otro camino óptico externo en el que la fibra óptica está conectada a un **dispositivo de calibración**. Este camino óptico proporciona la señal S_{fiber}.

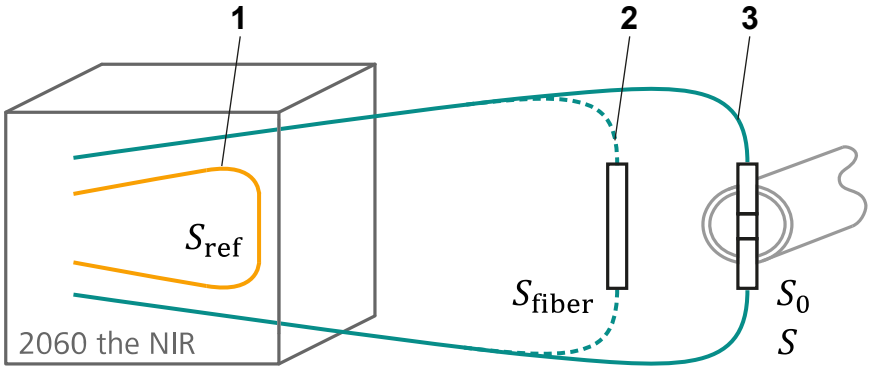


Figura 9 Caminos ópticos como ejemplo en el modo de transmisión: camino de referencia interna (1), fibra óptica externa conectada a un dispositivo de calibración (2) y fibra óptica externa conectada a la sonda con o sin muestra (3). Los caminos ópticos 2 y 3 representan la misma fibra óptica, que solo está conectada de forma diferente.

El dispositivo de calibración fija la fibra óptica y genera así el camino de referencia (**2**). En el modo de transmisión, el aire sirve de referencia y transmite el 100% de la luz. En el modo de reflexión, el dispositivo de calibración también registra el patrón de reflexión. En un primer momento, se parte de un patrón de reflexión ideal, que refleja el 100% de la luz.

A partir de las señales AS_0 y S se calcula la absorbancia de la muestra. Si se añaden las dos señales adicionales S_{ref} y S_{fiber} al numerador y al denominador, el resultado no se modifica:

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

Esta ecuación puede convertirse en:

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

Los 3 términos representan valores de absorbancia y pueden etiquetarse del siguiente modo:

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

La *figura 10* ilustra de qué modo se miden las señales S_{ref}, S_{fiber}, S_0 y S.

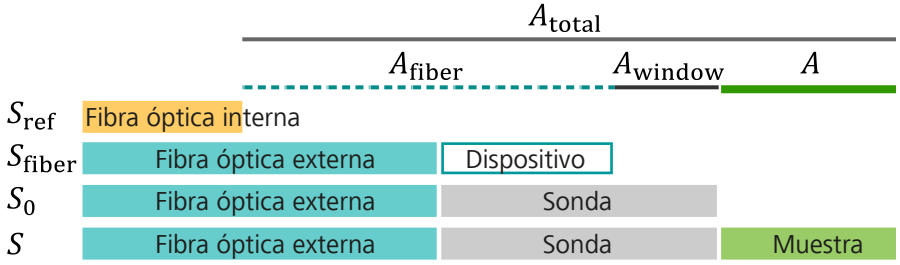


Figura 10 Normalización de referencia externa

A_{total} es la absorbancia de la fibra óptica externa, la sonda y la muestra, en relación con la fibra óptica interna.

A_{fiber} es la absorbancia de la fibra óptica externa más el dispositivo de calibración, en relación con la fibra óptica interna.

A_{window} es la absorbancia de la sonda menos el dispositivo de calibración.

Eliminación de las fluctuaciones ambientales

Para determinar la absorbancia A de la muestra, se miden 3 valores de absorbancia. A partir de los 3 valores se calcula A de acuerdo con la ecuación anterior:

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

Esto permite determinar los 3 valores de absorbancia en diferentes momentos. De este modo, se pueden eliminar fácilmente las fluctuaciones en el aparato o las fluctuaciones de las condiciones ambientales para cada una de las 3 determinaciones:

- A_{total} se determina con cada medida de muestra. Para ello, S_{ref} y S deberían medirse en un intervalo de tiempo breve para eliminar las fluctuaciones.
- La **corrección de la fibra de vidrio** A_{fiber} se puede determinar con menos frecuencia. Para ello, S_{ref} y S_{fiber} deberían medirse en un intervalo de tiempo breve para eliminar las fluctuaciones.

- La **corrección de la ventana** A_{window} puede determinarse con menos frecuencia, generalmente solo una vez después de la instalación. Para ello, S_{fiber} y S_0 deberían medirse en un intervalo de tiempo breve para eliminar las fluctuaciones.

¿Cuándo se necesita una corrección de la ventana?

Si el dispositivo de calibración reproduce completamente las propiedades ópticas de la sonda, A_{window} es igual a 0. En este caso, se puede omitir la corrección de la ventana.

Por lo general, en el modo de transmisión, se necesita una corrección de la ventana. Para el modo de reflexión, no se requiere una corrección de la ventana. Sin embargo, hay algunas excepciones, que se enumeran en la tabla siguiente:

Modo de medida	Sonda	Normalización de referencia
Transmisión	Par de transmisión Sonda de transmisión Sonda de transflexión con fibras individuales	Fibra de vidrio + ventana
Reflexión	Sonda de reflexión	Fibra de vidrio
	Sonda de transflexión con MicroBundle	Fibra de vidrio + ventana

Para decidir si se necesita una corrección de la ventana, se debe examinar cada combinación del dispositivo de calibración y sonda por separado. Si el dispositivo de calibración *no* reproduce completamente las propiedades ópticas de la sonda, se necesita una corrección de la ventana.

Canales

Un aparato del tipo **2060 The NIR** ofrece múltiples canales. Cada canal puede estar conectado a una configuración diferente de fibra óptica y sonda. Por lo tanto, la normalización de referencia debe realizarse por separado para cada canal.

Todos los canales usan el mismo camino de referencia interno, es decir, la misma señal S_{ref}. Un múltiplex cambia entre la referencia interna y los diferentes canales de medida.

Realización de una corrección de la fibra

Después de la puesta en marcha o si cambia la configuración de la fibra óptica de un canal, debe realizarse una corrección de la fibra. Además, si se efectúa un cambio de lámpara o una modificación extrema de las condiciones ambientales puede ser aconsejable realizar una nueva normalización.

En este procedimiento se utiliza un material de referencia:

- En el modo de reflexión, el material de referencia es un patrón de reflexión. Se parte de un patrón de reflexión no ideal (por ejemplo, 99%). El patrón de reflexión tiene un espectro de absorción nominal conocido $A_{nominal}$.
- En el modo de transmisión, se usa el aire como referencia. El espectro de absorción nominal es una línea cero ($A_{nominal} = 0$), ya que se supone que el aire no absorbe ninguna luz.

La *figura 11* ilustra el siguiente procedimiento:

1. Las fibras ópticas externas deben conectarse al dispositivo de calibración.
   En el modo de reflexión, el dispositivo de calibración se combina con el patrón de reflexión.
2. La instrucción **REF STD** con la interfaz **Fibra de vidrio** realiza los siguientes escaneos:
   a. Un escaneo de referencia interno proporciona un valor para S_{ref}.
   b. Un escaneo externo mide las fibras ópticas externas, el dispositivo de calibración y el material de referencia. Esto da como resultado la señal S_{raw}.
3. El software calcula A_{raw} (**Espectro primario medido**):
   $$A_{raw} = \log_{10} \frac{S_{ref}}{S_{raw}}$$
   Aquí, A_{raw} corresponde a la absorbancia de las fibras ópticas externas, al dispositivo de calibración y al material de referencia, en relación con el camino óptico interno.
4. El espectro de absorción nominal del material de referencia, $A_{nominal}$, se muestra en el software como **Espectro de referencia**. El espectro de referencia debe restarse de A_{raw} para obtener A_{fiber}:
   $$A_{fiber} = A_{raw} - A_{nominal}$$
   Aquí, A_{fiber} corresponde a la absorbancia de las fibras ópticas externas y al dispositivo de calibración, en relación con el camino óptico interno.
   Nota: En el modo de transmisión, es $A_{nominal} = 0$ y $A_{fiber} = A_{raw}$.
   A_{fiber} representa el **Espectro de corrección** de fibra de vidrio.
5. A_{fiber} permanece sin cambios hasta que se realiza una nueva corrección de la fibra para el canal correspondiente.

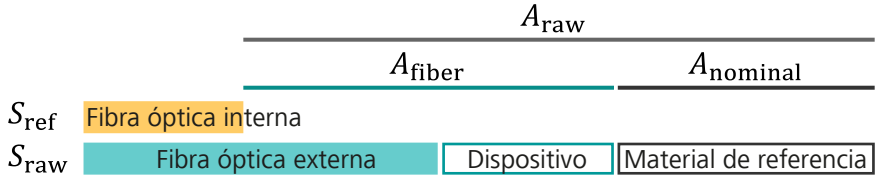


Figura 11 Corrección de la fibra

Validar la corrección de la fibra

La corrección de la fibra debe validarse con los mismos parámetros de medida y el mismo dispositivo de calibración.

En este procedimiento se utiliza un material de referencia:

- En el modo de reflexión, el material de referencia es un patrón de reflexión. Se parte de un patrón de reflexión no ideal (por ejemplo, 99%). El patrón de reflexión tiene un espectro de absorción nominal conocido $A_{nominal}$.
- En el modo de transmisión, se usa el aire como referencia. El espectro de absorción nominal es una línea cero ($A_{nominal} = 0$), ya que se supone que el aire no absorbe ninguna luz.

La *figura 12* representa el modo en que se determinan los residuos de la validación:

1. Las fibras ópticas externas deben conectarse al dispositivo de calibración.
   En el modo de reflexión, el dispositivo de calibración se combina con el patrón de reflexión.
2. La instrucción **VAL REF STD** con la interfaz **Fibra de vidrio** realiza los siguientes escaneos:
   a. Un escaneo de referencia interno proporciona un valor para S_{ref}.
   b. Un escaneo externo en el canal correspondiente mide las fibras ópticas externas, el dispositivo de calibración y el material de referencia. Esto da como resultado la señal S_{raw}.
3. El software calcula A_{raw} (**Espectro primario medido**):
   $$A_{\mathrm{raw}} = \log_{10} \frac{S_{\mathrm{ref}}}{S_{\mathrm{raw}}}$$
4. A_{raw} se corrige mediante el espectro de corrección de la fibra de vidrio para eliminar la absorbancia de las fibras ópticas y del dispositivo de calibración:
   $$A_{\mathrm{corrected}} = A_{\mathrm{raw}} - A_{\mathrm{fiber}}$$
   $A_{corrected}$ se muestra en el software como **Espectro corregido medido**.
5. Idealmente, $A_{corrected}$ debería coincidir con **Espectro de referencia** $A_{nominal}$. Las diferencias entre ambos se calculan como **Residuos de la validación**:
   $$A_{\mathrm{residual}} = A_{\mathrm{corrected}} - A_{\mathrm{nominal}}$$
   Nota: En el modo de transmisión, es $A_{nominal} = 0$ y $A_{residual} = A_{corrected}$.

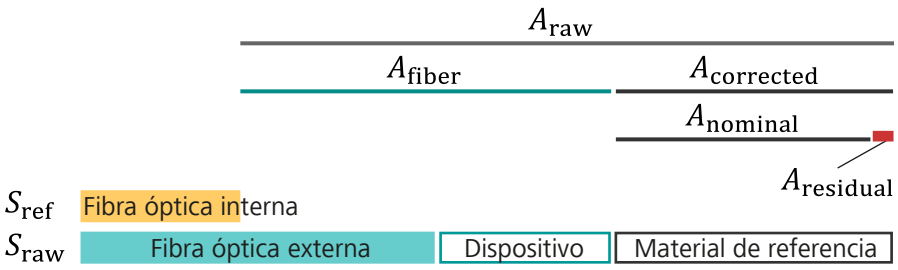


Figura 12 Residuos para la validación de la corrección de la fibra

Para analizar los residuos de la validación, la gama de longitudes de onda se divide en varios segmentos. Para cada segmento, el valor medio cuadrático de los residuos a través de las longitudes de onda da el **residuo RMS** (unidad: mAU):

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

Aquí, f corresponde a la cantidad de longitudes de onda en el segmento y $A_{residual\ i}$ al residuo de la longitud de onda i.

Cada segmento debe cumplir con una tolerancia predefinida para A_{RMS}. Si todos los segmentos cumplen con la tolerancia, la validación total se ha realizado correctamente.

La validación debe realizarse correctamente primero para que se puedan registrar espectros en el canal correspondiente con el aparato.

Realización de una corrección de la ventana

En caso de que sea necesario realizar una corrección de la ventana, esta debe efectuarse después de la puesta en marcha o cada vez que se cambie la configuración de la sonda o de la fibra óptica de un canal. Además, si se efectúan cambios en la sonda como, por ejemplo, por suciedad, puede ser aconsejable realizar una nueva normalización.

En este procedimiento se utiliza un material de referencia:

- En el modo de reflexión, el material de referencia es un patrón de reflexión. Se parte de un patrón de reflexión no ideal (por ejemplo, 99%). El patrón de reflexión tiene un espectro de absorción nominal conocido $A_{nominal}$.
- En el modo de transmisión, se usa el aire como referencia. El espectro de absorción nominal es una línea cero ($A_{nominal} = 0$), ya que se supone que el aire no absorbe ninguna luz.

La *figura 13* ilustra el siguiente procedimiento:

1. Para obtener un valor actual de A_{fiber}, se debería realizar una corrección de la fibra como se describió anteriormente.
   Importante: La corrección de la fibra debe realizarse antes de la corrección de la ventana.
2. Las fibras ópticas externas se deben conectar a la sonda sin ninguna muestra presente. Si es necesario, un patrón de reflexión ocupa el lugar de la muestra.
3. La instrucción **REF STD** con la interfaz **Ventana** realiza los siguientes escaneos:
   a. Un escaneo de referencia interno proporciona un valor para S_{ref}.
   b. Un escaneo externo en el canal correspondiente mide las fibras ópticas externas, la sonda y el material de referencia. Esto da como resultado la señal S_{probe}.
4. La absorbancia A_{probe} con respecto al camino óptico interno es:
   $$A_{\mathrm{probe}} = \log_{10} \frac{S_{\mathrm{ref}}}{S_{\mathrm{probe}}}$$
5. El software calcula A_{raw} (**Espectro primario medido**):
   $$A_{\mathrm{raw}} = A_{\mathrm{probe}} - A_{\mathrm{fiber}}$$
   Aquí, A_{raw} corresponde a la absorbancia de la sonda y del material de referencia, con respecto al dispositivo de calibración.
6. El espectro de absorción nominal del material de referencia, $A_{nominal}$, se muestra en el software como **Espectro de referencia**. El espectro de referencia se debe restar de A_{raw} para obtener A_{window}:
   $$A_{\mathrm{window}} = A_{\mathrm{raw}} - A_{\mathrm{nominal}}$$
   Aquí, A_{window} corresponde a la absorbancia de la sonda, con respecto al dispositivo de calibración.
   Nota: En el modo de transmisión, es $A_{nominal} = 0$ y $A_{window} = A_{raw}$. A_{window} representa el **Espectro de corrección** de la ventana.
7. A_{window} permanece sin cambios hasta que se realiza una nueva corrección de la ventana para el canal correspondiente.

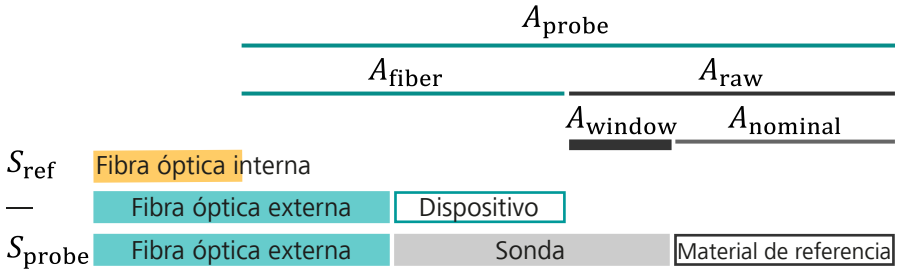


Figura 13 Corrección de la ventana

Validación de la corrección de la ventana

La corrección de la ventana debe validarse con los mismos parámetros de medida y el mismo dispositivo de calibración.

En este procedimiento se utiliza un material de referencia:

- En el modo de reflexión, el material de referencia es un patrón de reflexión. Se parte de un patrón de reflexión no ideal (por ejemplo, 99%). El patrón de reflexión tiene un espectro de absorción nominal conocido $A_{nominal}$.
- En el modo de transmisión, se usa el aire como referencia. El espectro de absorción nominal es una línea cero ($A_{nominal} = 0$), ya que se supone que el aire no absorbe ninguna luz.

La *figura 14* representa el modo en que se determinan los residuos de la validación:

1. Las fibras ópticas externas deben conectarse a la sonda sin ninguna muestra presente. Si es necesario, un patrón de reflexión ocupa el lugar de la muestra.
2. La instrucción **VAL REF STD** con la interfaz **Ventana** realiza los siguientes escaneos:
   a. Un escaneo de referencia interno proporciona un valor para S_{ref}.
   b. Un escaneo externo en el canal correspondiente mide las fibras ópticas externas, la sonda y el material de referencia. Esto da como resultado la señal S_{probe}.
3. La absorbancia A_{probe} con respecto al camino óptico interno es:

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

4. El software calcula A_{raw} (**Espectro primario medido**):

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

5. Al sustraer el espectro de corrección de la ventana de A_{raw}, se eliminan las diferencias de absorbancia entre el dispositivo de calibración y la sonda:

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

   $A_{corrected}$ se muestra en el software como **Espectro corregido medido**.
6. Idealmente, $A_{corrected}$ debería coincidir con **Espectro de referencia** $A_{nominal}$. Las diferencias entre ambos se calculan como **Residuos de la validación**:

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

   Nota: En el modo de transmisión, es $A_{nominal} = 0$ y $A_{residual} = A_{corrected}$.

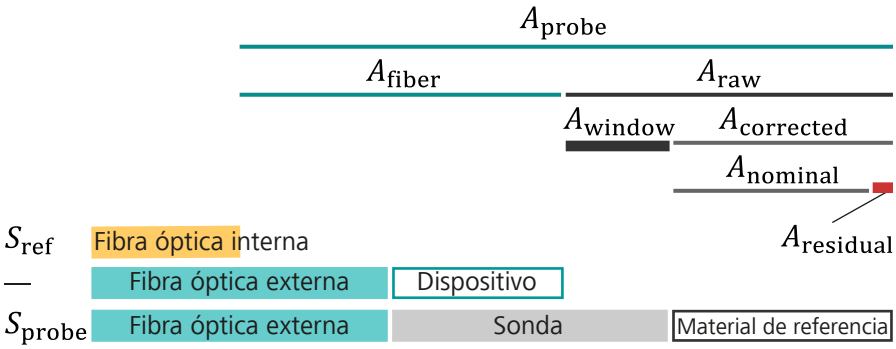


Figura 14 Residuos para la validación de la corrección de la ventana

Para analizar los residuos de la validación, la gama de longitudes de onda se divide en varios segmentos. Para cada segmento, el valor medio cuadrático de los residuos a través de las longitudes de onda da el **ruido RMS** (unidad: mAU):

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

Aquí, *f* corresponde a la cantidad de longitudes de onda en el segmento y $A_{residual\ i}$ al residuo de la longitud de onda *i*.

Cada segmento debe cumplir con una tolerancia predefinida para A_{RMS}. Si todos los segmentos cumplen con la tolerancia, la validación total se ha realizado correctamente.

Registrar el espectro de una muestra

Para poder registrar espectros con el aparato, deben efectuarse correctamente primero las Pruebas de rendimiento del aparato en el canal correspondiente *(véase "Pruebas de rendimiento del aparato", capítulo 3.3, página 26)*.

El procedimiento para la adquisición del espectro de una muestra se ilustra en la *figura 15*:

1. Las fibras ópticas externas deben estar conectadas a la sonda. Debe haber una muestra presente.
2. El espectro de absorción se calcula usando el último espectro de referencia registrado S_{ref}. Para obtener un valor actual para S_{ref}, se puede ejecutar la instrucción **MEAS REF SPEC**.
3. La instrucción **MEAS SPEC** mide la muestra, incluida la sonda y las fibras ópticas. Esto da como resultado la señal S.
4. El software calcula A_{total}, la absorbancia de la muestra, incluida la sonda y las fibras ópticas, con referencia al camino óptico interno:

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

5. A continuación, se calcula la absorbancia de la muestra como se describió anteriormente usando el espectro de corrección de la fibra óptica A_{fiber} y el espectro de corrección de la ventana A_{window} del canal correspondiente:

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

   A indica el espectro de la muestra.

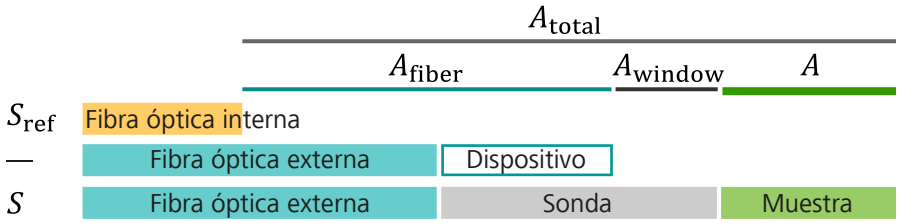


Figura 15 Registrar el espectro de una muestra

3.3 Pruebas de rendimiento del aparato

Las Pruebas de rendimiento del aparato pueden realizarse a través de caminos ópticos internos y externos.

- **OMNIS NIR Analyzer**
  - **Internas Pruebas de rendimiento del aparato** (obligatorias): las pruebas internas utilizan el camino de referencia de la respectiva presentación de muestras. Mediante las pruebas se verifican las longitudes de onda y el ruido de la señal.
    Antes de las pruebas, la calibración de las longitudes de onda para la respectiva presentación de muestras debe haberse realizado y validado correctamente.
    Para poder registrar espectros con el aparato en la respectiva presentación de muestras, primero deben efectuarse correctamente las pruebas internas.
  - **Externas Pruebas de rendimiento del aparato** (opcionales): las pruebas externas respaldan la validación efectuada según las farmacopeas como USP <856>, Ph.Eur 2.2.40 y JP 2.27. Se comprueban las longitudes de onda, el ruido de la señal y la linealidad fotométrica *(véase "Pruebas de rendimiento del aparato (OMNIS NIR Analyzer) externas", capítulo 3.3.1, página 30)*.

- **2060 The NIR**
  Las pruebas pueden usar el camino óptico interno o uno de los caminos ópticos externos. Mediante las pruebas se verifican las longitudes de onda y el ruido de la señal.
  Antes de las pruebas, se deben haber efectuado correctamente y validado la calibración de las longitudes de onda y la normalización de referencia externa en el canal correspondiente.
  Las Pruebas de rendimiento del aparato deben completarse correctamente primero para que se puedan registrar espectros con el aparato en el canal correspondiente. Se deben cumplir las tolerancias predefinidas. Las tolerancias permitidas dependen de la configuración de fibra óptica del canal correspondiente, que se indica en los datos específicos del aparato (modo de medida, tipo de fibra y longitud de la fibra).

Prueba de longitud de onda

En la prueba de longitud de onda se examinan la exactitud y precisión de la longitud de onda. Para ello se utiliza un patrón de longitud de onda que tiene un espectro de absorción con picos definidos y posiciones de los picos conocidas:

- **Interna**: el espectro de absorción del patrón de longitud de onda interno, con trazabilidad metrológica, se determina a través del camino óptico interno:

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

  En este caso. A_{WL} corresponde a la absorbancia del patrón de longitud de onda interno; S_{ref} corresponde a la señal medida en el camino de referencia interno y $S_{\mathrm{ref,WL}}$ corresponde a la señal medida en el camino de referencia interno con el patrón de longitud de onda interno.
- **Externa** para aparatos de la familia de productos **OMNIS NIR Analyzer**: la prueba de longitud de onda externa es opcional *(véase "Prueba de longitud de onda externa", página 30)*.

- **Externa** para aparatos del tipo **2060 The NIR**:
  Las fibras ópticas externas se deben conectar al dispositivo de calibración en el caso de fibras individuales y al patrón de reflexión en el caso de MicroBundle.
  El espectro de absorción del patrón de longitud de onda interno, con trazabilidad metrológica, se determina a través de un camino óptico externo:

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

  En este caso, A_{WL} corresponde a la absorbancia del patrón de longitud de onda interno; S_{ref} corresponde a la señal medida en el camino de referencia interno; $S_{\mathrm{fiber,WL}}$ corresponde a la señal medida en el camino óptico del respectivo canal, a cuyo efecto las fibras están conectadas al dispositivo de calibración o al patrón de reflexión y el patrón de longitud de onda interno está introducido en el camino óptico y A_{fiber} corresponde al espectro de corrección de la fibra de la normalización de referencia.

A_{WL} contiene también la absorbancia del patrón de reflexión, que es irrelevante para los cálculos siguientes. En el escenario ideal, las posiciones de los picos de A_{WL} son idénticas a las posiciones de los picos nominales del patrón de longitud de onda.

La exactitud y la precisión de longitud de onda se comprueban de la siguiente manera:

1. Se registra una serie de espectros de absorción del patrón de longitud de onda como se describió anteriormente (A_{WL}).
2. En los espectros registrados se identifican las posiciones de los picos.
3. Para cada posición del pico, se calculan las siguientes estadísticas sobre los espectros registrados:
   a. Valor medio (unidad: nm)
   b. Desviación estándar (unidad: nm)
4. **Exactitud**: en cada pico, la diferencia entre la posición del pico media y la posición del pico nominal debe estar dentro de la tolerancia predefinida.
   ➨ **Nota**: Las posiciones nominales de los picos pueden variar ligeramente de una prueba a otra. Esto se debe a que las posiciones de los picos se corrigen con la temperatura. Se realiza una corrección de la temperatura similar durante la calibración de las longitudes de onda. Esto permite asegurarse de que las medidas a una determinada temperatura proporcionan resultados comparables en todos los aparatos.
5. **Precisión**: en cada pico, la desviación estándar debe estar dentro de la tolerancia predefinida.
6. Si se cumplen las tolerancias para todos los picos, el estado general de la prueba de longitud de onda es correcto.

Prueba de ruido

El ruido de la señal puede comprobarse interna o externamente:

- **Interna**: el ruido se determina como la absorbancia del camino óptico interno, en relación con la absorbancia de otra medida en el mismo camino óptico:

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

  En este caso, $S_{\text{ref},1}$ y $S_{\text{ref},2}$ corresponden a las señales medidas en el camino de referencia interno.
- **Externa** para aparatos de la familia de productos **OMNIS NIR Analyzer**: la prueba de ruido externa es opcional *(véase "Pruebas de ruido externas", página 30)*.
- **Externa** para aparatos del tipo **2060 The NIR**:
  Las fibras ópticas externas se deben conectar al dispositivo de calibración en el caso de fibras individuales y al patrón de reflexión en el caso de MicroBundle.
  El ruido se determina como la diferencia entre el espectro de absorción medido y el espectro de absorción nominal:

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

  En este caso, S_{ref} corresponde a la señal medida en el camino de referencia interno; S_{fiber} corresponde a la señal medida en el camino óptico externo, a cuyo efecto las fibras están conectadas al dispositivo de calibración o al patrón de reflexión; A_{fiber} corresponde al espectro de corrección de la fibra de la normalización de referencia y A_{nominal} corresponde al espectro de absorción nominal del patrón de reflexión.
  Nota: En el modo de transmisión, A_{nominal} es = 0.

En el escenario ideal, A_{noise} es = 0.

En la prueba de ruido se ejecutan los siguientes pasos:

1. Se registra una serie de espectros de ruido como se describió anteriormente (A_{noise}).
2. Los espectros de ruido se dividen en diferentes segmentos de longitud de onda.
3. Para cada espectro de ruido y cada segmento se calculan 3 magnitudes:
   a. Ruido fotométrico (unidad: mAU)
   b. Ruido de pico a pico (unidad: mAU)
   c. Desviación de la línea base del ruido (unidad: mAU)
4. Para cada una de las 3 magnitudes en cada segmento se calcula el valor medio de los espectros de ruido registrados.
5. Si todos los valores medios están dentro de las tolerancias predefinidas, el estado general de la prueba de ruido es correcto.

3.3.1 Pruebas de rendimiento del aparato (OMNIS NIR Analyzer) externas

Los aparatos de la familia de productos **OMNIS NIR Analyzer** se pueden validar según farmacopeas como USP <856>, Ph.Eur 2.2.40 y JP 2.27 (a partir de la versión 4.4. de OMNIS Software). Estas pruebas requieren patrones de referencia externos, con trazabilidad metrológica. Los patrones de referencia tienen espectros de absorción nominales individuales que se han medido con un aparato de referencia a temperatura ambiente.

Prueba de longitud de onda externa

La exactitud y la precisión de longitud de onda se comprueban de la siguiente manera:

1. El patrón de longitud de onda externo (transmisión o reflexión) debe colocarse en posición.
2. Se registra una serie de espectros de absorción del patrón de longitud de onda externo:

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

   En este caso, A_{WL} corresponde a la absorbancia del patrón de longitud de onda externo; S_{ref} corresponde a la señal medida en el camino de referencia interno y S_{WL} corresponde a la señal medida a través del patrón de longitud de onda externo.
3. En los espectros registrados se identifican las posiciones de los picos.
4. Para cada posición del pico se calculan las siguientes estadísticas sobre los espectros registrados:
   a. Valor medio (unidad: nm)
   b. Desviación estándar (unidad: nm)
5. **Exactitud**: en cada pico, la diferencia entre la posición del pico media y la posición del pico nominal debe estar dentro de la tolerancia predefinida.
6. **Precisión**: en cada pico, la desviación estándar debe estar dentro de la tolerancia predefinida.
7. Si se cumplen las tolerancias para todos los picos, el estado general de la prueba de longitud de onda es correcto.

Pruebas de ruido externas

El ruido de la señal se comprueba una vez con una corriente de fotones baja (bajo flujo) y una vez con una corriente de fotones alta (alto flujo):

1. El patrón de referencia externo (transmisión o reflexión) para el test de bajo flujo o el test de alto flujo debe colocarse en posición.

2. Se registra una serie de espectros de ruido como la diferencia entre los espectros de absorción medidos y el espectro de absorción nominal del patrón de referencia:

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

   En este caso, S_{ref} corresponde a la señal medida en el camino de referencia interno; S_{ND} corresponde a la señal medida a través del patrón de referencia externo y A_{nominal} corresponde al espectro nominal del patrón de referencia.
3. Los espectros de ruido se dividen en diferentes segmentos de longitud de onda.
4. Para cada espectro de ruido y cada segmento se calculan 3 magnitudes:
   a. Ruido fotométrico (unidad: mAU)
   b. Ruido de pico a pico (unidad: mAU)
   c. Desviación de la línea base del ruido (unidad: mAU)
5. Para cada una de las 3 magnitudes en cada segmento se calcula el valor medio de los espectros de ruido registrados.
6. Si todos los valores medios están dentro de las tolerancias predefinidas, el estado general de la prueba de ruido es correcto.

Linealidad fotométrica

El objetivo de esta prueba consiste en demostrar, en toda la gama de longitudes de onda, una relación lineal entre la reflectancia (o la transmitancia) y la absorbancia medida:

1. Se registran espectros de absorción de 5 patrones de referencia con diferente reflectancia (o transmitancia).
2. Se comprueba la relación lineal entre la reflectancia (o transmitancia) y la absorbancia medida mediante una regresión lineal en varias longitudes de onda.
3. Si las pendientes y los sectores de eje de coordenadas 'y' de todas las rectas de regresión están dentro de las tolerancias predefinidas, el estado general de la prueba es correcto.

4 Desarrollo del modelo

Se distinguen los siguientes tipos de modelo:

- Un **Modelo de cuantificación** describe la dependencia de un parámetro de interés (por ejemplo, el contenido de agua) respecto a los espectros registrados de las muestras.
- Un **Modelo de identificación** (a partir de la versión de OMNIS Software 4.0) clasifica las muestras de acuerdo a los espectros registrados en diferentes productos (por ejemplo, diferentes variedades de granos de café).
  Un producto se refiere a una sustancia química específica o a una sustancia química con propiedades físicas específicas (por ejemplo, tamaño de partículas).

Para analizar una muestra desconocida, se registra un espectro de la muestra. Según la aplicación, el espectro se utiliza de la siguiente manera:

- Cuantificación: A partir del espectro, un modelo de cuantificación realiza una predicción, por ejemplo, del contenido de agua de la muestra.
- Identificación: A partir del espectro, un modelo de identificación identifica la muestra, por ejemplo, un café arábica.
- Verificación (a partir de la versión de OMNIS Software 4.2): A partir del espectro, un modelo de identificación verifica, por ejemplo, si la muestra es de café arábica o no.

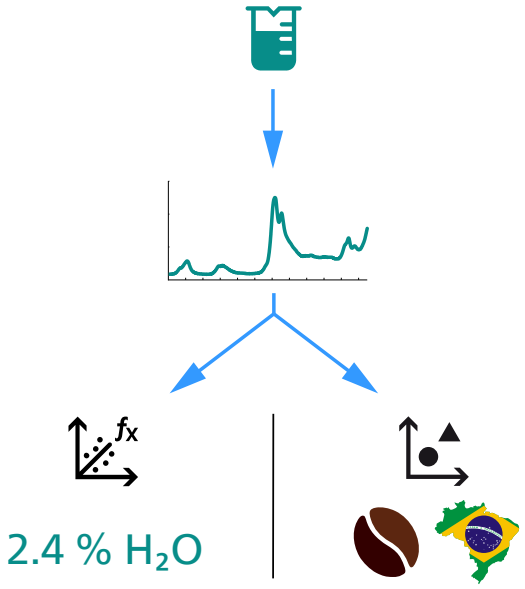


Figura 16 Cuantificación (abajo a la izquierda) e identificación (abajo a la derecha).

El desarrollo de un modelo y el análisis de muestras comprenden los siguientes pasos:

1. **Muestreo**
   Se registran y procesan muestras físicas:
   a. Para cada muestra, se adquiere un espectro.
   b. Para la cuantificación, se mide un valor de referencia para el parámetro de interés (por ejemplo, contenido de agua) con un método de referencia (por ejemplo, titulación). La medida de referencia debe ser precisa y exacta.
   c. Para la identificación, debe conocerse la pertenencia de la muestra al producto.
2. **Desarrollo del modelo**
   El desarrollo del modelo se lleva a cabo en un proceso iterativo que comprende los siguientes pasos:
   a. División del conjunto de datos en un conjunto de calibración, un conjunto de validación y un conjunto de valores discrepantes.
   b. Aplicación de un pretratamiento de datos adecuado y de gamas de longitudes de onda a los espectros.
   c. Cálculo de un modelo basado en el conjunto de calibración.
   d. La validación del modelo asegura que el modelo cumple con los requisitos. La validación se basa principalmente en el conjunto de validación que no se usó durante el desarrollo del modelo.
      En la cuantificación, el modelo predice el parámetro de interés para los espectros en el conjunto de validación. Luego, los valores calculados se comparan con los valores de referencia conocidos.
      Para la identificación, el modelo asigna los espectros a productos diferentes. Las pertenencias a productos predichas se comparan con las pertenencias a productos reales.
3. **Monitorización**
   La monitorización del modelo asegura que la capacidad predictiva no disminuya con el tiempo. Cualquier modificación en el proceso o en las muestras requiere una revalidación.

4.1 Muestras físicas

El primer paso es la recolección y análisis de muestras físicas. Un muestreo correcto es un requisito para el desarrollo de un modelo robusto. Aquí se deben considerar varios aspectos.

Rango de variación

Las muestras deberían incluir las variaciones típicas en las muestras que cabe esperar en el futuro. Las concentraciones de todos los componentes químicos y tamaños de partículas deberían cubrir al menos el rango de variación esperado.

Las muestras deberían abarcar una variación adecuada de las condiciones y un período de tiempo adecuado. Se deberían considerar todas las variaciones como, por ejemplo, las fluctuaciones del proceso, las fluctuaciones estacionales o las fluctuaciones de las condiciones ambientales.

Las muestras se deberían distribuir uniformemente a lo largo del rango de variación. En la cuantificación, esto comprende el rango de los valores de referencia. Por ejemplo, si el rango de valores de referencia va del 1% al 10%, las muestras deberían estar distribuidas de forma uniforme entre el 1% y el 10%.

Muestras de calibración y validación

Por lo general, se utilizan 2 conjuntos de muestras:

- Conjunto de calibración: muestras usadas para el desarrollo del modelo.
- Registro de validación: muestras usadas para la validación del modelo.

Ambos conjuntos deberían cubrir las variaciones esperadas. Para el registro de validación, es preferible recolectar muestras independientes para verificar la robustez del modelo. Se deberían considerar escenarios como operadores diferentes, proveedores diferentes o, si es necesario, aparatos diferentes.

Si no resulta factible recoger muestras independientes para la calibración y la validación, de forma alternativa se pueden dividir los espectros de las muestras disponibles en un conjunto de calibración y otro de validación. Para garantizar la independencia de los conjuntos, se debería utilizar el algoritmo de distribución automática.

Una vez desarrollado un modelo, se puede considerar el uso de muestras explícitas de valores discrepantes para verificar la detección de valores discrepantes.

Todas las muestras deberían tratarse de la misma manera. Para la adquisición de los espectros, se debería usar el mismo método con la misma configuración de hardware y los mismos parámetros de medida. Para la

cuantificación, se debería usar el mismo método de referencia con los mismos parámetros de medida.

Número de muestras

Cuantas más variaciones de las condiciones, los componentes químicos o tamaños de partículas haya que cubrir, más muestras serán necesarias.

Cuantificación: Para que el análisis estadístico funcione correctamente, se requiere un número mínimo de aproximadamente 50 muestras, por lo que el registro de calibración y el registro de validación deben contener al menos entre 20 y 25 muestras cada uno.

Identificación y verificación: Para cada producto, las muestras deben cubrir las variaciones esperadas. Los productos pueden tener un número de muestras diferente, el número mínimo es 3.

Con al menos 10 a 20 muestras (dependiendo del número de variaciones) se puede desarrollar un primer modelo sin registro de validación. Si la validación cruzada (para modelos de cuantificación) o la validación interna (para modelos de identificación) indican que se puede crear un modelo adecuado, se deben recolectar más muestras de calibración y validación para el desarrollo del modelo final.

Réplicas

En ocasiones, solo hay muy pocas muestras en un determinado rango de condiciones o valores de referencia. Para compensar esto, se podría intentar replicar esas muestras. Sin embargo, esto puede resultar problemático. Si los duplicados de una muestra se encuentran tanto en el conjunto de calibración como en el registro de validación, las figuras de mérito serán engañosas (demasiado optimistas). También se deben evitar duplicados en el mismo conjunto.

Método de referencia (para cuantificación)

En la cuantificación, se usa un método de referencia para medir los valores de referencia. El **error estándar del laboratorio (SEL)** para el método de referencia usado juega un papel importante en el desarrollo de un modelo de cuantificación. El SEL es la desviación estándar de las diferencias entre las medidas de muestras duplicadas.

Con frecuencia, el SEL es el mayor error que contribuye al error estándar de la predicción (SEP) en el método NIR *(véase "SEP - Error estándar de la predicción", página 72)*. El SEL no debería superar 0,7 veces, de preferencia 0,5 veces, el SEP requerido. El rango de los valores de referencia debería ser al menos de 3 veces el SEL, de preferencia 5 veces.

Una manera de reducir el SEL es realizar medidas de referencia repetidas en cada muestra. La media de los valores medidos debería establecerse como el valor de referencia de la muestra. Para cada muestra se debería realizar el mismo número de medidas de referencia. Las figuras de mérito

se expresan en relación con un determinado número de medidas de referencia repetidas. Un número diferente de medidas de referencia repetidas daría lugar a valores estimados incorrectos de las figuras de mérito y debería evitarse.

Temperatura de muestras

La temperatura de muestras influye notablemente en los espectros de los líquidos que contienen agua u otros enlaces de hidrógeno. Los espectros de otros líquidos polares pueden verse afectados del mismo modo que los espectros de los sólidos que contienen agua, humedad o disolventes. Tales muestras deberían medirse a una temperatura definida.

Valor discrepante

Algunas muestras se pueden identificar posteriormente como valores discrepantes. Un valor discrepante es una muestra que, por alguna razón, difiere de la mayoría de las muestras. Para evitar un impacto negativo en el modelo, las muestras identificadas como valores discrepantes no se incluyen en el cálculo del modelo.

Hay diferentes tipos de valores discrepantes:

- **Valor discrepante espectral**
  Si el espectro detectado de una muestra difiere de la mayoría de los demás espectros, la muestra puede ser identificada como un valor discrepante espectral *(véase "Valores discrepantes espectrales", capítulo 4.3.3, página 52)*.
- **Valor de referencia discrepante (para cuantificación)**
  En la cuantificación, algunos valores de referencia pueden mostrar anomalías y ser identificados posteriormente como valores de referencia discrepantes *(véase "Valor de referencia discrepante (para cuantificación)", capítulo 4.3.4, página 59)*.
  El OMNIS Model Developer (OMD) identifica las muestras correspondientes como valores discrepantes basándose en la detección de valores discrepantes según ASTM D8321-22 *(véase "OMNIS Model Developer (OMD)", capítulo 4.4.3, página 73)*.

4.2 Análisis de componentes principales (PCA)

Los datos espectroscópicos de las muestras de calibración contienen un gran número de variables (longitudes de onda). Las variables están fuertemente correlacionadas entre sí. Los datos son, por lo tanto, muy redundantes. Para tratar este tipo de datos se usan modelos de variables latentes como PCA y PLS.

El **análisis de componentes principales** (**PCA**, por sus siglas en inglés: *principal component analysis*) se centra en los espectros, sin considerar los valores de referencia.

OMNIS Software usa PCA durante el desarrollo del modelo para la división de conjunto de datos automática y la detección de valores discrepantes espectrales.

Pasos preparatorios

Es necesario realizar los siguientes pasos preparatorios:

1. **Pretratamiento de datos**: OMNIS Software aplica el pretratamiento de datos especificado a los espectros *(véase "Pretratamiento de datos", capítulo 4.3.1, página 41)*.
2. **Gama de longitudes de onda**: OMNIS Software aplica la selección de longitudes de onda especificada a los espectros *(véase "Gama de longitudes de onda", capítulo 4.3.2, página 50)*.
3. **Centrado en la media**: para cada longitud de onda se calcula el valor de absorbancia medio y se resta del valor respectivo en cada espectro.

Los primeros componentes principales

Después de los pasos preparatorios, el PCA reorganiza la información en los datos espectrales y separa los datos relevantes del ruido. Para ello, el PCA transforma las variables de longitudes de onda en un nuevo espacio de variables, denominadas **componentes principales** (**PC** por sus siglas en inglés: Principal Components).

El PCA convierte la información relevante de un gran número de variables de longitudes de onda en solo unos pocos componentes principales. Para ilustrar el concepto con un ejemplo sencillo, supongamos que solo hay 2 variables de longitudes de onda en lugar de miles, y que estas 2 variables se reducen a 1 componente principal.

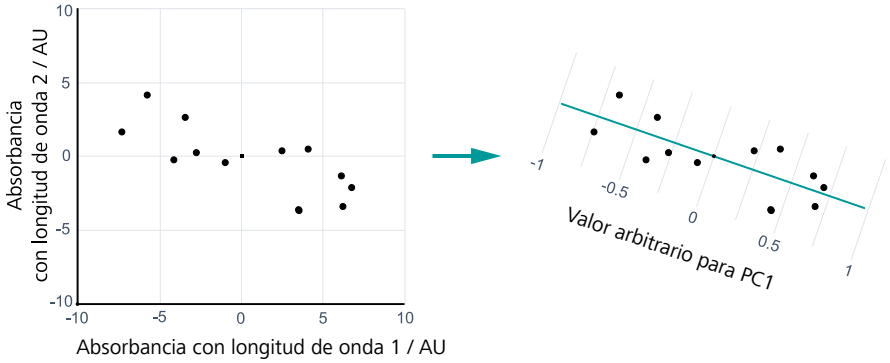


Figura 17 Puntos que representan espectros en un espacio bidimensional de longitudes de onda (a la izquierda). Los mismos puntos en un espacio unidimensional de componentes principales (a la derecha).

En la *figura 17* de la izquierda, los ejes horizontal y vertical representan el espacio de longitudes de onda original con 2 variables. De este modo, cada punto representa un espectro con solo 2 longitudes de onda. El valor medio de todos los valores de las longitudes de onda constituye el punto cero.

A la derecha, la dirección a través de los datos que explica la varianza máxima es el componente principal PC1. En este ejemplo, PC1 es la única variable en el espacio de componentes principales. Como resultado, las 2 variables originales se reducen a 1.

Distribuciones y residuos

La *figura 18* muestra las magnitudes que caracterizan un espectro *i*:

- La distancia s_i desde el centro, medida en el espacio de componentes principales. En el ejemplo con solo 1 componente principal, s_i se mide en la dirección de PC1. La distancia s_i se denomina **distribución** del espectro *i*.
- El desplazamiento e_i del espacio de componentes principales hasta el espectro. La distancia e_i se denomina **residuo** del espectro *i*.

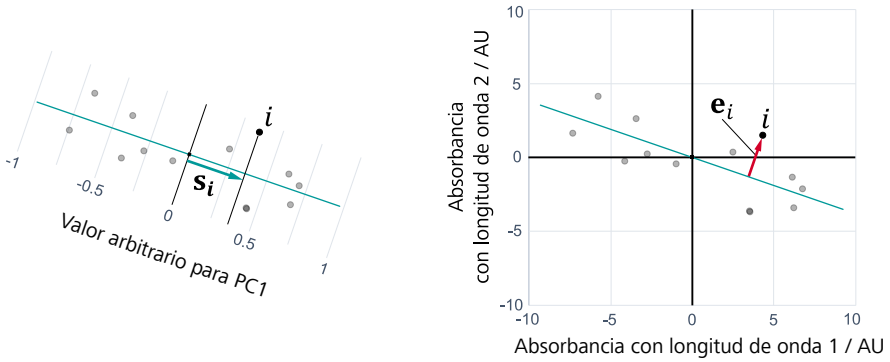


Figura 18 Espectro i con distribución (a la izquierda) y residuo (a la derecha).

La distribución s_i se mide en el espacio de componentes principales. El residuo e_i se mide en el espacio original de longitudes de onda.

Conversión en varios componentes principales

Por lo general, se necesita más de un componente principal para una descripción adecuada de los datos espectroscópicos.

En la *figura 19* hay 3 variables originales x_1, x_2, x_3. Cada punto representa un espectro con 3 longitudes de onda.

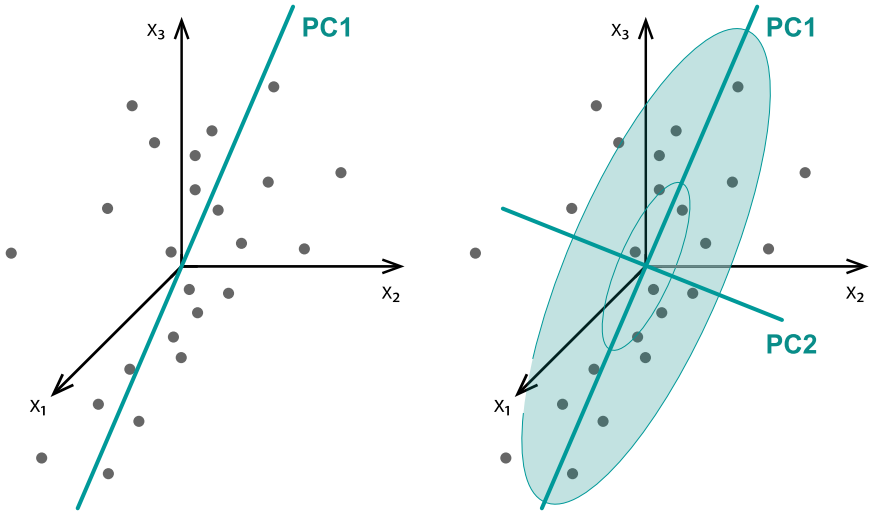


Figura 19 3 variables originales se reducen a 1 componente principal (a la izquierda) o a 2 componentes principales (a la derecha). PC1 y PC2 forman un espacio bidimensional de componentes principales.

De nuevo, el primer componente principal PC1 es la dirección a través de los datos que explica la varianza máxima.

El segundo componente principal PC2 es la dirección a través de los datos que explica la varianza máxima restante. Lo mismo puede decirse de todos los componentes principales siguientes, cada uno de los cuales describe la máxima varianza restante. Por lo tanto, los primeros componentes principales explican la mayor parte de la varianza de los datos, mientras que los demás contienen principalmente ruido y se pueden descartar. De este modo, se puede reducir el número de variables.

Una característica fundamental del PCA es que todos los componentes principales son **ortogonales** (en ángulo recto) entre sí. Por lo tanto, las distribuciones no están correlacionadas.

Distancia de Mahalanobis

Como se mencionó anteriormente, la distribución de un espectro *i* se mide en el espacio de componentes principales, mientras que el residuo se mide en el espacio original de longitudes de onda.

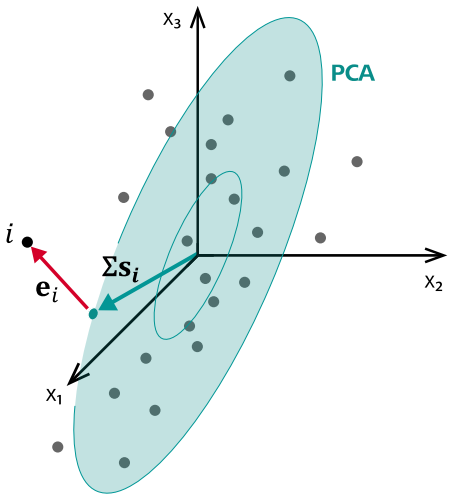
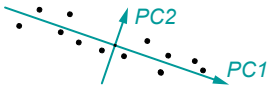


Figura 20 Distribución y residuo del espectro i. El punto verde es la proyección ortogonal del punto i (que representa al espectro i) en el espacio de componentes principales.

En la *figura 20* el vector de distribución $\Sigma \mathbf{s}_i$ representa la distancia absoluta (distancia euclídea) desde el centro del modelo PCA hasta la proyección ortogonal del espectro en el espacio de componentes principales.



En el ejemplo, las distancias euclídeas de los espectros están más separadas en la dirección de PC1 que en la dirección de PC2. La dispersión puede medirse como **varianza**. La varianza en PC1 es mayor que en PC2.

El vector de distribución de variables normalizadas $\mathbf{s}_i$ representa una distancia normalizada, la llamada **distancia de Mahalanobis**. La distancia de Mahalanobis considera la varianza diferente en las distintas direcciones de los componentes principales. Cada dirección recibe la misma ponderación. Por lo tanto, una pequeña distancia euclídea en una dirección con poca varianza puede contar tanto como una gran distancia euclídea en una dirección con mayor varianza.

Conversión de espectros con múltiples longitudes de onda

Los mismos conceptos se aplican para la conversión de espectros con un gran número de variables de longitudes de onda en componentes principales. En la *figura 21*, cada espectro se representa mediante una curva (a la izquierda) y un punto (a la derecha).

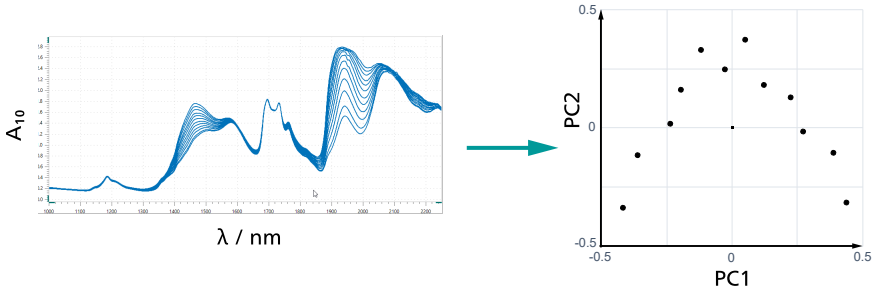


Figura 21 Conversión de datos espectrales en un espacio de componentes principales. Las distribuciones en el lado derecho se expresan en unidades arbitrarias.

La figura de la derecha muestra los 2 primeros componentes principales PC1 y PC2. De la misma manera, también se pueden visualizar los componentes principales siguientes PC3, PC4, etc.

Un modelo PCA usa un número fijo de componentes principales. Cuantos más componentes principales, más variaciones espectrales relevantes explica el modelo. Al mismo tiempo, el modelo también capta más variaciones espectrales irrelevantes (ruido). Se necesita una compensación equilibrada.

Cuando OMNIS Software realiza un análisis de componentes principales, la cantidad de componentes principales se selecciona de manera que la varianza explicada sea al menos del 95%.

Algoritmo PCA

Existen varias formas de convertir los datos originales en un espacio de componentes principales. OMNIS Software realiza una descomposición en valores singulares *(véase "Algoritmo PCA", capítulo 6.2, página 94)*.

4.3 Preparación de datos

4.3.1 Pretratamiento de datos

Los modelos espectroscópicos se basan en la relación entre los valores de absorbancia y el parámetro de interés (cuantificación) o la clase de producto (identificación, verificación). La **parametrización** de los espectros asegura que los espectros expresen bien esta relación. El objetivo es eliminar la varianza irrelevante sin perder la información útil. Se corrigen los artefactos y las faltas de linealidad. Una parametrización realizada correctamente aumenta la precisión y robustez del modelo, así como la repetitividad y reproducibilidad de las predicciones.

La parametrización se aplica al conjunto de calibración, al conjunto de validación y al conjunto de valores discrepantes, así como a todas las muestras desconocidas futuras que se analizarán con el mismo modelo.

El primer paso de la parametrización es el **pretratamiento de datos**. El pretratamiento de datos se lleva a cabo en la secuencia especificada. En un segundo paso de la parametrización, se pueden definir las gamas de longitudes de onda pertinentes *(véase "Gama de longitudes de onda", capítulo 4.3.2, página 50)*.

Reducción del ruido

Los espectros pueden contener distintos tipos de fluctuaciones aleatorias alrededor de la señal. Ejemplos de ello son el ruido de alta frecuencia causado por el detector y los circuitos electrónicos del aparato o el ruido de baja frecuencia causado por la deriva del aparato durante las medidas de escaneado.

El espectrómetro proporciona un espectro promediado a partir de una serie de mediciones únicas. Esto reduce significativamente el ruido de alta frecuencia. Con un filtro de suavizado se puede conseguir una mayor reducción del ruido. Estos filtros se basan en la idea de que el ruido es de alta frecuencia y la señal de baja frecuencia. Aproximan la señal a los valores de absorbancia cercanos y reducen el ruido promediando valores.

Corrección de la dispersión

La dispersión es el cambio de dirección de la luz causado por la interacción con la muestra. La luz dispersa que no llega al detector provoca fluctuaciones de la línea base en los espectros.

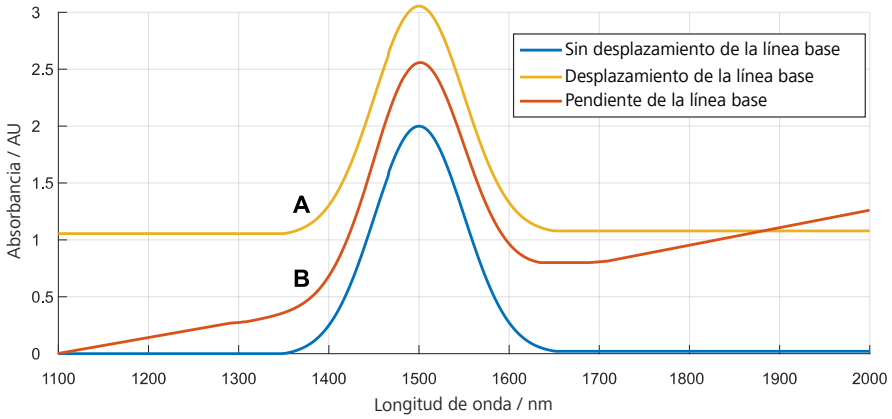


Figura 22 Los tipos más importantes de desviación de la línea base son el desplazamiento de la línea base y la pendiente de la línea base.

Pueden distinguirse diferentes tipos de desplazamientos de la línea base:

- Un factor aditivo constante que provoca un **desplazamiento de la línea base** (espectro **A**).
- Un factor multiplicativo que depende de la longitud de onda que conduce a una **pendiente de la línea base** (espectro **B**).

- Un factor multiplicativo de segundo orden que depende de la longitud de onda que conduce a una **pendiente cuadrática de la línea base**.
  También se pueden producir pendientes de la línea base de orden superior.
- Factores multiplicativos que dependen de la absorbancia que llevan a una **amplificación**. Sin embargo, las amplificaciones son irrelevantes.

La dispersión es más evidente con muestras sólidas. Los desplazamientos de la línea base resultantes se pueden usar para detectar cambios en el tamaño de las partículas u otras variaciones físicas.

Sin embargo, si lo que interesa son las variaciones químicas, se deberían minimizar los desplazamientos de la línea base mediante un pretratamiento adecuado.

Pretratamientos de datos

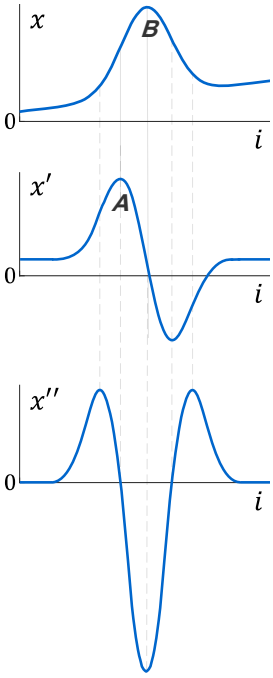
Todos los siguientes pretratamientos de datos se aplican a un único espectro. No se incluye ningún otro espectro en los cálculos.

El pretratamiento de datos modifica los valores de la señal. Los números del eje 'y' dejan de tener significado.

En el caso de los pretratamientos de datos aplicados se trata de transformaciones lineales. Por lo tanto, se sigue aplicando la ley de Beer-Lambert.

4.3.1.1 Derivadas

Las derivadas pueden realizarse usando el filtro Gap-Segment o el filtro Savitzky-Golay.



La derivada de un espectro describe la pendiente de la curva o la inclinación de la curva en cada punto. La pendiente es la tasa de cambio del espectro de salida.

En el espectro, x_i es la absorbancia en la longitud de onda *i*. La derivada de primer orden x_i' representa la pendiente del espectro en la longitud de onda *i*. Donde el espectro de salida es más empinado, la derivada de primer orden tiene un máximo (**A**). Donde el espectro de salida tiene un pico (**B**), la derivada de primer orden es igual a 0.

La derivada de primer orden elimina los desplazamientos de la línea base y convierte las pendientes de la línea base en desplazamientos de la línea base.

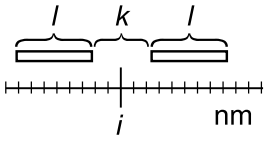
La derivada de segundo orden x_i'' corresponde a la pendiente de la derivada de primer orden en la longitud de onda *i*. Los picos positivos del espectro original (**B**) se convierten en picos negativos y a la inversa.

La derivada de segundo orden elimina los desplazamientos de la línea base y las pendientes de la línea base del espectro original.

Se debe tener cuidado si el espectro original contiene un nivel considerable de ruido. Cada derivada empeora significativamente la relación señal/ruido. Por esta razón, las derivadas se combinan con una función de suavizado en el filtro Gap-Segment o el filtro Savitzky-Golay.

4.3.1.2 Gap-Segment

El filtro Gap-Segment suaviza el espectro. De forma opcional, el filtro Gap-Segment realiza una derivada de primer o segundo orden. El cálculo depende de si se utilizan derivadas:



- **Orden de la derivada 0**: para cada longitud de onda *i*, el filtro Gap-Segment calcula el valor medio de 2 segmentos con el tamaño del segmento *l*, por ejemplo: 10 nm. Los 2 segmentos están separados por una distancia *k*, por ejemplo, 5 nm.
- **Orden de la derivada 1**: para la derivada de primer orden, los valores medios de los 2 segmentos se calculan por separado. A continuación, se calcula la diferencia entre los dos valores medios.
- **Orden de la derivada 2**: la derivada de segundo orden puede calcularse a partir de la derivada de primer orden de la misma manera.

Al principio y al final del espectro, se calculan las longitudes de onda *l* + *k*/2, por lo que se utilizan valores cero para las longitudes de onda del segmento que se encuentran fuera del espectro.

Al principio y al final del espectro, se usan valores cero para las longitudes de onda del segmento que quedan fuera del espectro.

Con el suavizado se puede provocar un ligero desplazamiento de los picos y una cierta desviación.

Ajustes de parámetros

Se puede conseguir un mayor suavizado mediante:

- una derivada de orden inferior,
- un mayor tamaño del segmento,
- una mayor distancia del segmento.

Un suavizado excesivo conduce a una pérdida de varianza relevante, lo que disminuye la capacidad predictiva del modelo.

4.3.1.3 Savitzky-Golay

Al igual que el filtro Gap-Segment, el filtro Savitzky-Golay suaviza el espectro y, opcionalmente, realiza una derivada de primer o segundo orden. Sin embargo, el filtro Savitzky-Golay usa otro método de suavizado.

Para cada longitud de onda *i*, el filtro de Savitzky-Golay ajusta un polinomio de bajo orden en el rango de la longitud de onda correspondiente. El valor del polinomio en la longitud de onda *i* es el valor suavizado. Si se realiza una derivada, se utiliza el valor de la derivada.

Una suma ponderada de valores adyacentes calcula todo de una vez:

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

Aquí, *k* corresponde a la anchura del filtro; c_j corresponde a los coeficientes de convolución que dependen del orden de la derivada, del grado polinomial y de la anchura del filtro y pueden consultarse en tablas, y x_{i+j} corresponde a los valores de absorbancia del espectro de salida en la longitud de onda *i+j*.

Al principio y al final del espectro, se utilizan valores extrapolados para las longitudes de onda del filtro que se encuentran fuera del espectro (extrapolación horizontal).

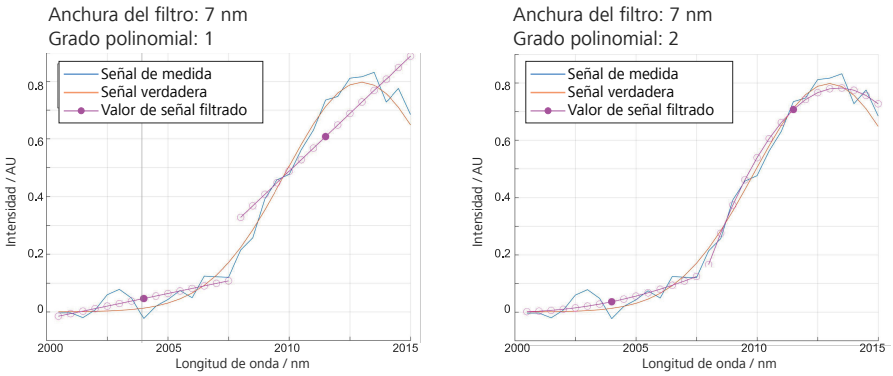


Figura 23 Filtrado de Savitzky-Golay con diferentes grados polinomiales

La *figura 23* ilustra el filtrado de Savitzky-Golay. La anchura del filtro es de 7 nm. Se muestran los polinomios para las longitudes de onda de 2004 nm y 2011,5 nm. Los nuevos valores son los puntos rellenados o, si se utilizan derivadas, sus derivadas. Todas las demás longitudes de onda se tratan de la misma manera.

La anchura del filtro define la gama de longitudes de onda en la que se ajusta cada polinomio. La convolución se pondera de modo que la influencia de los valores de absorbancia disminuya a ambos lados de la longitud de onda correspondiente.

Ajustes de parámetros

Se puede conseguir un mayor suavizado mediante:

- una derivada de orden inferior,
- una mayor anchura de filtro,
- un polinomio de grado inferior.

Un suavizado excesivo conduce a una pérdida de varianza relevante, lo que disminuye la capacidad predictiva del modelo.

4.3.1.4 SNV – Standard Normal Variate

SNV normaliza un único espectro a la varianza 1 y al valor medio 0. SNV normaliza los valores de absorbancia x_i para cada longitud de onda i dentro de una gama de longitudes de onda definida como sigue:

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

Aquí, m corresponde al valor medio y s a la desviación estándar de todos los valores de absorbancia dentro de la gama de longitudes de onda especificada.

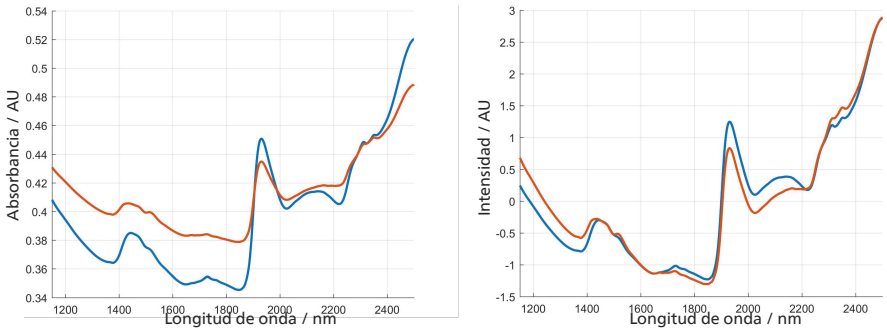


Figura 24 Espectros de absorción (a la izquierda) y espectros tratados con SNV (a la derecha).

La normalización elimina la varianza entre los espectros. Esto es útil si la varianza proviene de propiedades que no son de interés, como diferentes espesores de capa en muestras granulares o en polvo o en medios turbios.

Nota: Si se aplica una derivada después de la SNV, la varianza eliminada puede reaparecer parcialmente. Por lo tanto, las derivadas deberían aplicarse antes de la SNV. Si excepcionalmente se debe realizar la SNV antes de la derivada, se puede considerar la siguiente secuencia: detrend, SNV, derivada. Sin embargo, la secuencia habitual es: derivada, SNV, detrend *(véase "Secuencia con múltiples pasos de pretratamiento de datos", capítulo 4.3.1.7, página 49)*.

Parámetros

- **Gamas de longitudes de onda**
  Si los artefactos afectan a determinadas gamas de longitudes de onda (p. ej., debido a la saturación o a un fuerte ruido), esas gamas pueden excluirse.
  Se toman en cuenta exclusivamente las gamas de longitudes de onda definidas para el cálculo del valor medio y de la desviación estándar. La normalización subsiguiente se efectúa tanto en las gamas de longitudes de onda definidas como en las secciones intermedias. Se adopta el valor normalizado de la longitud de onda inicial adyacente para las longitudes de onda excluidas al principio del espectro. Se adopta el valor normalizado de la longitud de onda final adyacente para las longitudes de onda excluidas al final del espectro.
  En caso necesario, las longitudes de onda excluidas también se pueden excluir para el cálculo del modelo *(véase "Gama de longitudes de onda", capítulo 4.3.2, página 50)*.
  AVISO: A partir de la versión de OMNIS Software 4.6 se pueden definir varias gamas de longitudes de onda.

4.3.1.5 Detrend

Detrend ajusta un polinomio de segundo orden a todo el espectro usando el método de mínimos cuadrados. A continuación, detrend resta el polinomio del espectro.

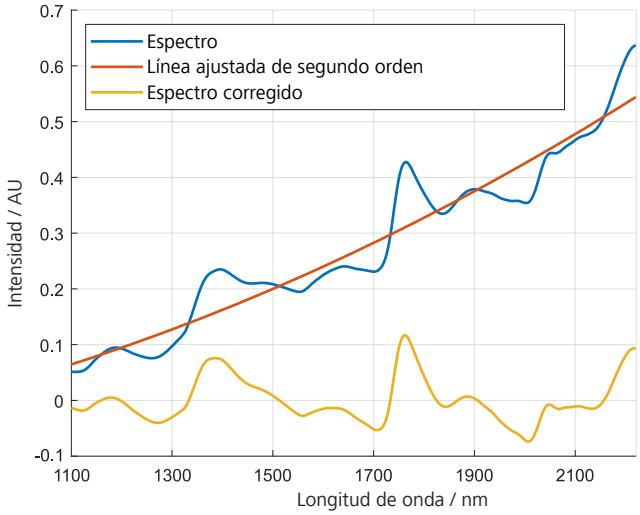


Figura 25 Detrend convierte el espectro azul en el espectro amarillo.

Detrend reduce los efectos de dispersión dependientes de la longitud de onda hasta pendientes cuadráticas de la línea base.

La figura anterior muestra un espectro (azul) en el que domina la tendencia. Si esta tendencia dominante es similar para todos los espectros, detrend puede funcionar bien. En otros casos, detrend tiende a eliminar

variaciones útiles. En tales casos, las derivadas probablemente sean la mejor opción.

Dado que se ajusta un polinomio diferente a cada espectro, puede surgir una varianza adicional indeseada. Normalmente, la SNV se aplica antes de detrend. Esto conduce a estimaciones más robustas de los coeficientes polinomiales.

Parámetros

- **Gamas de longitudes de onda**
  Si los artefactos afectan a determinadas gamas de longitudes de onda (p. ej., debido a la saturación o a un fuerte ruido), esas gamas pueden excluirse.
  Se ajusta un polinomio a todos los valores de intensidad de las gamas de longitudes de onda definidas. A continuación, el polinomio se resta del espectro en todas las gamas de longitudes de onda definidas. El valor de intensidad se ajusta a cero para todas las longitudes de onda excluidas.
  En caso necesario, las longitudes de onda excluidas también se pueden excluir para el cálculo del modelo *(véase "Gama de longitudes de onda", capítulo 4.3.2, página 50)*.
  AVISO: A partir de la versión de OMNIS Software 4.6 se pueden definir varias gamas de longitudes de onda.

4.3.1.6 Vista general del pretratamiento de datos

Pretratamiento	Objeto	Efectos positivos	Efectos negativos
Gap-Segment	Suavizado Se consigue un mayor suavizado con una derivada de menor orden, un mayor tamaño del segmento o una mayor distancia entre segmentos.	▪ El suavizado reduce el ruido de alta frecuencia.	▪ Un suavizado excesivo conduce a una pérdida de varianza relevante.
Derivadas con Gap-Segment	Corrección de línea base	▪ Derivada de primer orden: elimina los desplazamientos de la línea base. ▪ Derivada de segundo orden: elimina los desplazamientos de la línea base y las pendientes de la línea base.	▪ Amplifica el ruido. ▪ Modifica el aspecto del espectro.

Pretratamiento	Objeto	Efectos positivos	Efectos negativos
Savitzky-Golay	Suavizado Se obtiene un suavizado mayor con una derivada de menor orden, una mayor anchura de filtro o un polinomio de grado inferior.	▪ El suavizado reduce el ruido de alta frecuencia.	▪ Un suavizado excesivo conduce a una pérdida de varianza relevante.
Derivadas en Savitzky-Golay	Corrección de línea base	▪ Derivada de primer orden: elimina los desplazamientos de la línea base. ▪ Derivada de segundo orden: elimina los desplazamientos de la línea base y las pendientes de la línea base.	▪ Amplifica el ruido. ▪ Modifica el aspecto del espectro.
SNV – Standard Normal Variate	Corrección de la dispersión *	▪ Elimina los desplazamientos de la línea base.	▪ Se elimina la varianza relevante, dado el caso.
Detrend	Corrección de la dispersión *	▪ Elimina los desplazamientos de la línea base. ▪ Elimina las pendientes de la línea base y las pendientes cuadráticas de la línea base.	▪ Se elimina la varianza relevante, dado el caso. ▪ Puede producirse una varianza irrelevante.

* Nota: Deberían excluirse las gamas de longitudes de onda con artefactos (por ejemplo, saturación o ruido fuerte).

4.3.1.7 Secuencia con múltiples pasos de pretratamiento de datos

Al usar múltiples pasos de pretratamiento de datos, la secuencia puede ser crucial. La regla básica es la siguiente.

El filtro Gap-Segment o el filtro Savitzky-Golay deberían aplicarse preferentemente antes de la SNV, y esta antes de detrend.

Ejemplo con una derivada de primer orden y SNV: la derivada de primer orden convierte las pendientes de la línea base en desplazamientos de la línea base. Una SNV posterior elimina estos desplazamientos. Si se invierte la secuencia, la SNV no cambia las pendientes de la línea base. La derivada

de primer orden siguiente las convierte en desplazamientos de la línea base. Los desplazamientos permanecerían.

Ejemplo con una derivada de segundo orden y SNV: los desplazamientos de la línea base y las pendientes de la línea base se eliminan en todos los casos. Sin embargo, al aplicar la derivada de segundo orden y la SNV en la secuencia correcta se pueden eliminar las pendientes cuadráticas de la línea base. La derivada de segundo orden convierte las pendientes cuadráticas de la línea base en desplazamientos de la línea base. Una SNV posterior elimina estos desplazamientos. Si se invierte la secuencia, la SNV no cambia la pendiente cuadrática de la línea base. La derivada de segundo orden siguiente las convierte en desplazamientos de la línea base. Los desplazamientos permanecerían.

4.3.2 Gama de longitudes de onda

Después del pretratamiento de datos *(véase "Pretratamiento de datos", capítulo 4.3.1, página 41)*, se lleva a cabo el segundo paso de la parametrización: la selección de las gamas de longitudes de onda permite excluir las gamas que no son adecuadas para el propósito. En particular, las gamas de longitudes de onda con ruido o saturadas pueden afectar a los cálculos posteriores y deberían excluirse.

Ruido

El ruido se produce con altos valores de absorbancia cuando solo una pequeña cantidad de luz llega al detector. La figura siguiente muestra áreas con ruido.

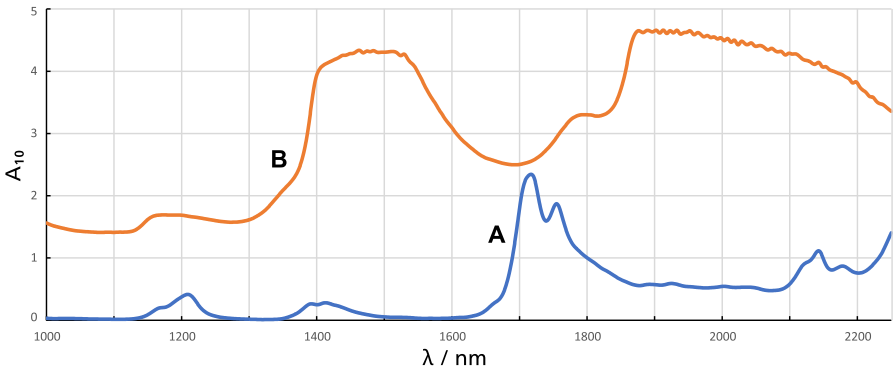


Figura 26 Ejemplo con gamas de longitudes de onda con ruido.

El espectro **A** tiene formas del pico normales. El espectro **B** tiene 2 áreas con ruido: una de 1400 a 1550 nm y otra por encima de 1870 nm. Estas zonas tienen mucho ruido y no se parecen a una curva de campana ni a una combinación de las mismas.

Saturación

La saturación del detector puede producirse cuando al detector llega una gran cantidad de luz, es decir, en caso de bajos valores de absorbancia.

- **OMNIS NIR Analyzer**
  El tiempo de integración siempre se ajusta automáticamente. Esto evita la saturación y minimiza el ruido *(véase "Tiempo de integración", página 9)*.
- **2060 The NIR**
  Si el tiempo de integración automático está activado, no se produce saturación.
  Si se activa el tiempo de integración manual, los tiempos de integración excesivamente largos pueden provocar la saturación del detector. Las zonas saturadas se producen en valores de absorbancia bajos, pero no siempre resulta fácil reconocerlas visualmente. Por ello, el tiempo de integración manual debería ajustarse con margen suficiente *(véase "Tiempo de integración", página 9)*.

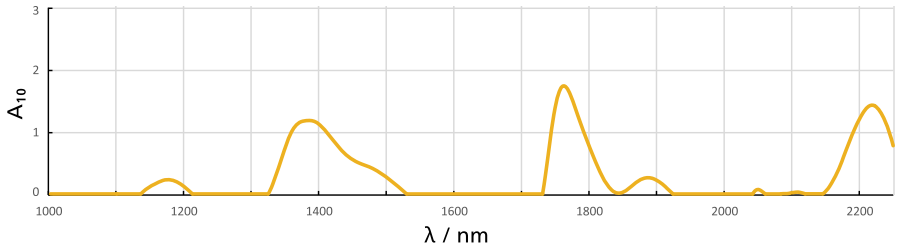


Figura 27 Ejemplo con gamas de longitudes de onda saturadas.

Otras razones

Hay otras razones para incluir o excluir gamas de longitudes de onda. La selección se puede basar en el conocimiento del parámetro de interés y sus respectivas bandas de absorción *(véase "La luz y su interacción con la materia", capítulo 2.1, página 3)*. Sin embargo, se debe tener en cuenta que la información relevante puede desplazarse a otras gamas de longitudes de onda según el tipo de pretratamiento.

Si durante el pretratamiento de datos se introdujeron anomalías al principio y al final de los espectros, se pueden excluir las longitudes de onda correspondientes.

Las variaciones de los componentes químicos o las fluctuaciones de las condiciones ambientales pueden influir en determinadas gamas de longitudes de onda. Excluir estas gamas de longitudes de onda puede mejorar la robustez del modelo.

Es necesario tener cuidado al excluir gamas de longitudes de onda bien definidas. Las gamas de longitudes de onda que parecen no contener información pueden, en realidad, proporcionar información oculta e importante. Estas pueden ser útiles para detectar valores discrepantes o para procesar bandas de absorción interferentes. De hecho, las bandas de absorción interferentes son la razón principal para realizar medidas multivariantes *(véase "Ejemplo de una regresión lineal", capítulo 6.1, página 90)*.

4.3.3 Valores discrepantes espectrales

Se denomina valor discrepante espectral a un espectro que difiere de la mayoría de los demás espectros.

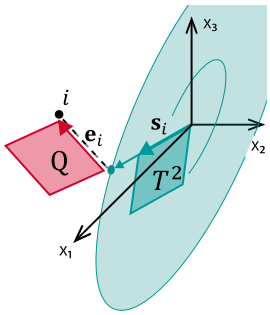
Los valores discrepantes se deberían comprobar cuidadosamente. Un valor discrepante puede distorsionar el modelo si, por ejemplo, la causa es una muestra contaminada o un error de medida. En este caso, el valor discrepante no debería tenerse en cuenta al calcular el modelo.

Por otro lado, un valor discrepante puede representar propiedades que no están bien contempladas por los demás espectros. En este caso, el valor discrepante realmente mejora el modelo. Si el valor discrepante parece ser una muestra válida, debería comprobarse si las muestras de calibración están distribuidas de modo uniforme en el rango de variación.

Las medidas adecuadas para reconocer los valores discrepantes son T^2 de Hotelling y los residuos Q.

T² de Hotelling y residuos Q

Al convertir los datos espectroscópicos en un espacio de componentes principales, los espectros pueden caracterizarse por sus distribuciones y residuos *(véase "Análisis de componentes principales (PCA)", capítulo 4.2, página 37)*. Lo mismo se aplica a la conversión en un espacio de variables latentes *(véase "Regresión PLS", capítulo 4.4.1, página 62)*.



Ejemplo: un espacio tridimensional de longitudes de onda (x_1, x_2, x_3) se convierte en un espacio bidimensional (verde). El punto *i* representa el espectro *i* y se proyecta desde el espacio tridimensional al espacio bidimensional. Esto da como resultado:

- un vector de distribución $\Sigma \mathbf{s}_i$ dentro del espacio bidimensional, o bien su vector de distribución de variables normalizadas $\mathbf{s}_i$, que representa la distancia de Mahalanobis.
- un residuo $\mathbf{e}_i$ dentro del espacio tridimensional.

A partir de $\mathbf{s}_i$ y $\mathbf{e}_i$ se pueden derivar las siguientes magnitudes *(véase "T² de Hotelling y residuos Q", capítulo 6.4, página 97)*:

- ***T²*** de Hotelling o simplemente ***T²*** es la distancia de Mahalanobis al cuadrado, es decir, la distancia normalizada al cuadrado desde el centro del modelo a la proyección ortogonal del espectro sobre el espacio de componentes principales o el espacio de variables latentes.
  En caso de que todas las distribuciones de un espectro correspondan al valor medio, T^2 es igual a 0 y el espectro se ubica en el centro del modelo. El modelo se puede ajustar mejor cerca del centro.
  Si se aleja mucho del centro, es posible que el modelo no se ajuste bien. Los valores de T^2 son altos. Un valor de T^2 alto indica un espectro extremo, por ejemplo, una muestra con una composición de componentes químicos extrema.

- El **residuo Q** es el residuo al cuadrado, es decir, la distancia ortogonal al cuadrado desde el espectro hasta el espacio de componentes principales o el espacio de variables latentes.
  Los residuos Q muestran las variaciones que el modelo no explica. Un residuo Q alto indica que es posible que el espectro no coincida con el modelo, por ejemplo, si la muestra medida contiene una sustancia diferente.

Detección de valores discrepantes espectrales

La detección de valores discrepantes espectrales identifica los espectros que se desvían de la totalidad de la población estadística.

1. La parametrización se tiene en cuenta del siguiente modo:
   a. A partir de la versión de OMNIS Software 4.2: El usuario decide si la parametrización (pretratamiento de datos y selección de longitud de onda) se tiene en cuenta o no.
   Los cambios posteriores en la parametrización no influyen en la división de conjunto de datos.
   b. A partir de la versión de OMNIS Software 3.3 hasta la versión de OMNIS Software 4.1: El usuario decide si el pretratamiento de datos se tiene en cuenta o no. La selección de longitud de onda y los cambios posteriores en el pretratamiento de datos no influyen en la división de conjunto de datos.
   c. Hasta la versión de OMNIS Software 3.2: El pretratamiento de datos se tiene en cuenta tal y como se definió en el momento de la detección de valores discrepantes. La selección de longitud de onda y los cambios posteriores en el pretratamiento de datos no influyen en la división de conjunto de datos.
2. La detección de valores discrepantes espectrales se basa en el modelo PCA de todos los espectros centrados en la media *(véase "Análisis de componentes principales (PCA)", capítulo 4.2, página 37)*. Se selecciona la cantidad de componentes principales, de modo que la varianza explicada sea al menos del 95%.
3. Para detectar los valores discrepantes espectrales, se usan los valores de T^2 de Hotelling y los residuos Q de los espectros. El algoritmo evalúa si el T^2 de Hotelling o el residuo Q del espectro analizado son el resultado de una variación aleatoria o sistemática.
   Encontrará una descripción del algoritmo en el apéndice *(véase "Valores discrepantes espectrales – Algoritmo", capítulo 6.5, página 98)*.

4.3.3.1 Gráfico de influencia

Así, el gráfico de influencia muestra las propiedades básicas de los espectros y ayuda a analizar los valores discrepantes espectrales.

Los fundamentos para el gráfico de influencia es un modelo PCA *(véase "Análisis de componentes principales (PCA)", capítulo 4.2, página 37)* o un modelo PLS:

- **Cuantificación**: El gráfico de influencia se basa de forma opcional en **PCA** o **PLS**.
  La regresión PLS, de forma similar a PCA, reduce los datos espectroscópicos a menos variables. PLS también tiene en cuenta los valores de referencia.
  Los componentes principales del PCA se denominan **variables latentes** en el PLS.
  (véase "Regresión PLS", capítulo 4.4.1, página 62)
- **Identificación**: Hay disponible un gráfico de influencia basado en **PCA** (a partir de la versión de OMNIS Software 4.3).

Tipos de valores discrepantes espectrales

El gráfico de influencia visualiza para cada espectro los valores de T^2 de Hotelling y el residuo Q *(véase "T² de Hotelling y residuos Q", página 52)*. T^2 de Hotelling y los residuos Q permiten reconocer distintos tipos de valores discrepantes espectrales:

- Los **valores discrepantes de T^2 de Hotelling**, también conocidos como valores discrepantes de palanca (en inglés: leverage outlier): un T^2 elevado significa que la proyección del espectro sobre el espacio de componentes principales (PCA) o el espacio de variables latentes (PLS) está lejos del centro del modelo.
- **Valor discrepante de residuos Q**: Un residuo Q elevado significa que el espectro está mal descrito por el modelo.

La *figura 28* muestra varios espectros en diferentes vistas:

- Gráfico de influencia a la izquierda: los residuos Q describen las variaciones que no son explicadas por el modelo, mientras que el T^2 de Hotelling tiene en cuenta las variaciones dentro del propio modelo.
  Las líneas discontinuas muestran los **valores críticos** o los **valores límite** para el nivel de significancia establecido *(véase "Valores discrepantes espectrales – Algoritmo", capítulo 6.5, página 98)*.
  Cuanto mayor sea el nivel de significancia, menores serán los valores límite, de modo que más puntos pueden quedar fuera de los valores límite.

- A la derecha: Espacio original ejemplar con 3 variables x_1, x_2, x_3, que se convierte en un espacio bidimensional con variables latentes a modo de ejemplo.
  Para los puntos 'A' a 'D', se muestra la distancia ortogonal al plano (líneas discontinuas), así como el punto modelado en el espacio de variables latentes (puntos verdes).

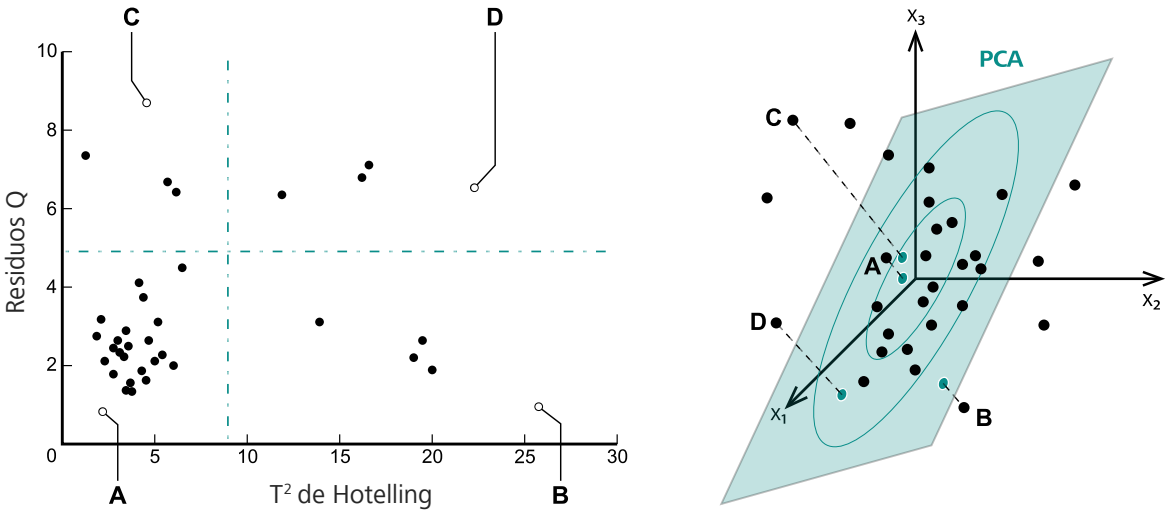


Figura 28 Gráfico de influencia (a la izquierda), espacio original y espacio de variables latentes (a la derecha). Cada espectro está representado por un punto en la figura a la izquierda y un punto en la figura a la derecha.

En ambas vistas se destacan 4 puntos con características diferentes:

- El espectro A tiene distribuciones bajas y residuos bajos. Está cerca del centro del modelo y es bien explicado por el modelo.
- El espectro B es un valor discrepante de T^2 de Hotelling. Se encuentra alejado del centro, pero es bien explicado por el modelo.
- El espectro C es un valor discrepante de residuos Q. Se encuentra lejos del centro y está mal explicado por el modelo.
- El espectro D es tanto un valor discrepante de T^2 de Hotelling como un valor discrepante de residuos Q. Está alejado del centro y solo es parcialmente explicado por el modelo.

El gráfico de influencia muestra cómo influyen los distintos espectros en el modelo. Dado que todas las variables latentes pasan por el centro, los espectros cercanos al punto central (por ejemplo, el espectro A) tienen pocas posibilidades de cambiar la dirección de las variables latentes. No tienen ningún valor de palanca. Cuanto mayor es la distancia con respecto al centro, mayor es el valor de palanca y el potencial de influir en el modelo. Algunos espectros logran realmente tirar del modelo en su dirección (espectro B), mientras que con otros esto solo ocurre hasta cierto punto (espectro D) o no ocurre en absoluto (espectro C).

En comparación con un modelo basado en todos los espectros, el cálculo de un modelo sin el espectro B probablemente cambiará el modelo más que uno sin el espectro D, y aún más que uno sin el espectro C. Es probable que el espectro B influya mucho en el modelo de cuantificación, para bien o para mal. La decisión de eliminar o no un posible valor discrepante en el cuadrante inferior derecho del gráfico de influencia requiere especial cuidado.

Lo ideal es que el modelo capte la varianza de una gran cantidad de espectros. No es deseable que el modelo esté marcado por solo unos pocos espectros. En la figura anterior, unos pocos espectros tienen grandes distancias con respecto al centro y también con respecto a la mayoría de los demás espectros. Esto es sospechoso. El modelo está siendo influenciado por unos pocos espectros. Se trata de potenciales valores discrepantes que deberían investigarse. Además, hay que asegurar que las muestras estén distribuidas uniformemente a lo largo del rango de variación.

Gráfico de influencia PCA y PLS

El gráfico de influencia PCA solo depende de los espectros. El gráfico de influencia PLS depende de los espectros y de los valores de referencia.

La siguiente tabla muestra cómo afectan los distintos ajustes a los gráficos de influencia PCA y PLS.

	Gráfico de influencia PCA	Gráfico de influencia PLS
Espectros	El modelo PCA subyacente se basa en todos los espectros del conjunto de calibración, el conjunto de validación y el conjunto de valores discrepantes.	El modelo PLS subyacente se basa en todos los espectros del conjunto de calibración. Sobre la base de este modelo PLS, se calculan los valores T^2 y del residuo Q de los espectros de los 3 conjuntos de datos y se representan en el gráfico.
Parametrización	Tiene en cuenta los pretratamientos de datos seleccionados y las gamas de longitudes de onda. Nota: La detección de valores discrepantes se basa en PCA y tiene en cuenta la parametrización según los ajustes de usuario y la versión de OMNIS Software *(véase "Detección de valores discrepantes espectrales", página 53)*.	Tiene en cuenta los pretratamientos de datos seleccionados y las gamas de longitudes de onda. Nota: La evaluación de los valores discrepantes durante la predicción se basa en PLS y tiene en cuenta los pretratamientos de datos y las gamas de longitudes de onda.

	Gráfico de influencia PCA	Gráfico de influencia PLS
Cantidad de variables	Usa la cantidad de componentes principales que logra una varianza explicada de al menos el **95%**.	Utiliza la cantidad de variables latentes seleccionada actualmente.
Nivel de significancia y valores críticos	Usa el nivel de significancia seleccionado actualmente para calcular y visualizar los valores críticos (líneas discontinuas). Identificación: Si la última división de conjunto de datos realizada se llevó a cabo sin efectuar la detección de valores discrepantes, el gráfico de influencia usa un nivel de significancia del 5%. Nota: Un aumento del nivel de significancia conduce a valores críticos más bajos y, por lo tanto, a más valores discrepantes en el desarrollo del modelo.	Usa el nivel de significancia seleccionado actualmente para calcular y visualizar los valores críticos (líneas discontinuas). Nota: Un aumento del nivel de significancia conduce a valores críticos más bajos y, por lo tanto, a más valores discrepantes en la predicción.
Valores de referencia (cuantificación)	Los valores de referencia no tienen influencia en el modelo PCA. Sin embargo, cada espectro puede tener asociado un valor de referencia discrepante y, por lo tanto, ser marcado como valor discrepante.	Los valores de referencia influyen en el modelo PLS y, por lo tanto, en el gráfico de influencia PLS. Además, cada espectro puede tener un valor de referencia discrepante asociado y, por lo tanto, ser etiquetado como valor discrepante.

Análisis de valores discrepantes

Al analizar los posibles valores discrepantes, deben tenerse en cuenta los siguientes factores:

- Un valor discrepante de T^2 de Hotelling indica una muestra con una composición extrema de componentes químicos en comparación con las demás muestras.
- Un valor discrepante de residuos Q puede indicar, por ejemplo, una muestra contaminada o un error en la adquisición del espectro.

Los posibles valores discrepantes deberían analizarse cuidadosamente. Los verdaderos valores discrepantes deben eliminarse de la lista de espectros. Las muestras válidas se deberían mantener. Si el conjunto de datos se vuelve a dividir, la detección de valores discrepantes puede encontrar posibles valores discrepantes que no se encontraron en la primera ejecución. Una razón posible es que el nuevo modelo PCA necesita menos componentes principales para alcanzar la varianza explicada del 95%. Si

los nuevos valores discrepantes resultan ser muestras válidas, deberían conservarse. En este caso, la división automática puede repetirse sin la detección de valores discrepantes.

4.3.3.2 Gráfico de distribución

Los fundamentos para el gráfico de distribución es un modelo PCA o un modelo PLS:

- **Cuantificación**: El gráfico de distribución (a partir de la versión de OMNIS Software 3.0) se basa en **PLS** *(véase "Regresión PLS", capítulo 4.4.1, página 62)*.
- **Identificación**: El gráfico de distribución (a partir de la versión de OMNIS Software 4.3) se basa en **PCA** *(véase "Análisis de componentes principales (PCA)", capítulo 4.2, página 37)*.

Cada espectro tiene un valor de distribución para cada componente principal o variable latente. En el gráfico de distribución, cada espectro está representado por un punto. El eje x, por ejemplo, muestra la distribución de la primera variable latente, el eje 'y', por ejemplo, la distribución de la segunda variable latente. De la misma manera, se puede mostrar cualquier par de variables latentes.

Dado que los valores de absorbancia de cada variable de longitud de onda se han centrado en la media, las distribuciones de cada variable latente también están centradas en la media. Un punto cercano al centro del gráfico de distribución (0/0) representa un espectro del valor medio con respecto a las dos variables latentes visualizadas. Los puntos cercanos entre sí representan espectros similares, los puntos más alejados representan espectros disímiles en relación con las dos variables latentes mostradas.

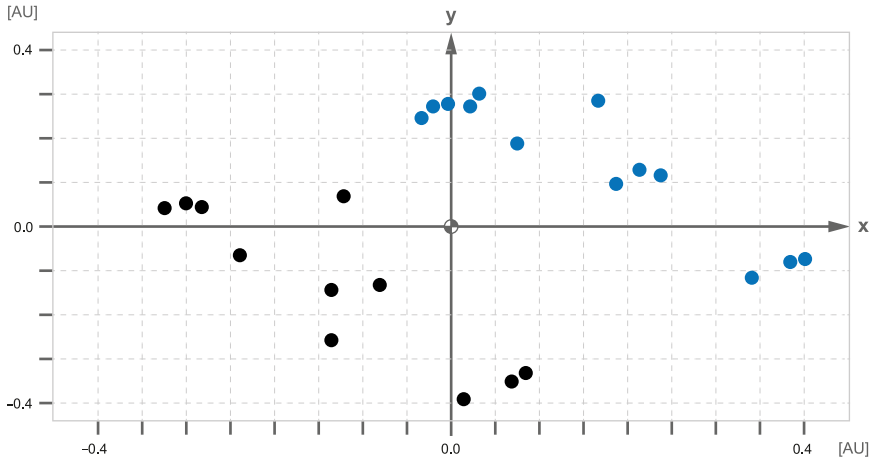


Figura 29 Gráfico de distribución para la variable latente 1 (eje x) y la variable latente 2 (eje 'y'). AU = unidad arbitraria.

La *figura 29* muestra un ejemplo de 2 conjuntos de datos medidos en condiciones diferentes. Las distribuciones están normalizadas, cada variable latente recibe la misma ponderación.

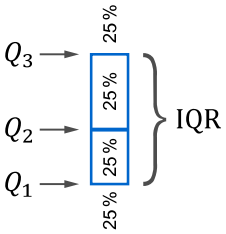
Las distribuciones de todos los componentes principales o todas las variables latentes de un espectro se pueden resumir en un valor individual (T^2 de Hotelling), que se muestra en el eje x del gráfico de influencia.

4.3.4 Valor de referencia discrepante (para cuantificación)

En los modelos de cuantificación, además de los valores discrepantes espectrales, también se determinan valores de referencia discrepantes. Los valores de referencia discrepantes indican anomalías en el valor de referencia.

Normalmente, los valores de referencia discrepantes son números transmitidos de forma incorrecta, por ejemplo, 143 en lugar de 14,3 o 15,9 en lugar de 51,9. La detección de valores discrepantes identifica estos errores de transmisión o transcripción basándose en un enfoque empírico. Solo los errores evidentes se marcan para un análisis más profundo.

Diagramas de caja



Los valores de referencia discrepantes se identifican mediante un método basado en diagramas de caja. Un **diagrama de caja** ordena los valores de referencia en orden ascendente. Los cuartiles dividen el conjunto de datos en 4 partes. Cada parte contiene el 25% de los valores de referencia.

El primer cuartil Q_1 separa el 25% de los valores más bajos del resto. Q_2 es la mediana y separa el 50% de los valores más bajos del resto. El tercer cuartil Q_3 separa el 75% de los valores más bajos del resto. Un rectángulo vertical representa el 50% intermedio de los datos, el rango intercuartílico (IQR, por sus siglas en inglés: interquartile range).

Los datos que se encuentran fuera de la caja IQR en una determinada cantidad se consideran posibles valores discrepantes y pueden representarse como pequeños círculos. Los valores límites inferior y superior para los valores discrepantes suelen definirse como 1,5 veces el IQR:

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

Aquí, Q_1 corresponde al primer cuartil, Q_3 al tercer cuartil e IQR al rango intercuartílico (Q_3 - Q_1).

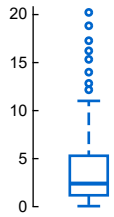
Para completar el diagrama de caja, los "bigotes" (whiskers) se extienden por encima y por debajo de la caja hasta los puntos más lejanos que no se marcan como posibles valores discrepantes.

Ajuste para distribuciones asimétricas

El diagrama de caja habitual supone una distribución aproximadamente simétrica de los datos. En el caso de distribuciones asimétricas, muchos valores de referencia regulares suelen marcarse como posibles valores

discrepantes. Por esta razón, OMNIS Software usa una versión adaptada del diagrama de caja que tiene en cuenta la asimetría de la distribución.

En el siguiente ejemplo, la distribución está sesgada hacia los valores de referencia más altos. Esto da lugar a 8 valores discrepantes en el rango alto. Después de ajustar los límites de los valores discrepantes, solo quedan 2 de ellos. Por otro lado, aparece un nuevo valor discrepante con un valor bajo.

Distribución asimétrica	Valores límite de valores discrepantes ajustados
	

Puede encontrar una descripción de la adaptación en el apéndice *(véase "Valor de referencia discrepante – Algoritmo", capítulo 6.6, página 101)*.

4.3.5 División de conjunto de datos

El conjunto de datos consiste en espectros y valores de referencia (cuantificación) o en espectros y nombres de productos asociados (identificación, verificación). El conjunto de datos sirve para desarrollar y validar el modelo. Por lo tanto, el conjunto de datos debería dividirse en un conjunto de calibración, un conjunto de validación y un conjunto de valores discrepantes. La división se puede efectuar de modo manual o automático.

Suponiendo que se dispone de suficientes espectros, se pueden usar para el conjunto de validación entre el 20 y el 30% de los datos totales, por ejemplo.

Algoritmo de distribución automática

La división de conjunto de datos automática garantiza que el conjunto de calibración y el conjunto de validación sean representativos de la totalidad e independientes entre sí. El objetivo es dividir los datos en dos conjuntos que cubran aproximadamente el mismo rango en el espacio PCA y presenten propiedades estadísticas similares. El algoritmo utilizado es una versión ligeramente modificada del algoritmo dúplex de R. D. Snee, *Validation of Regression Models: Methods and Examples*, Technometrics, Volumen 19, N.° 4 (Nov. 1977), pp. 415–428.

Réplicas

No debería haber réplicas en el conjunto de datos. El algoritmo no elimina réplicas ni pseudorréplicas, ni obliga a añadir réplicas de una muestra concreta al mismo conjunto de datos.

1. La parametrización se tiene en cuenta del siguiente modo:
    a. A partir de la versión de OMNIS Software 4.2: El usuario decide si la parametrización (pretratamiento de datos y selección de longitud de onda) se tiene en cuenta o no.
    Los cambios posteriores en la parametrización no influyen en la división de conjunto de datos.
    b. A partir de la versión de OMNIS Software 3.3 hasta la versión de OMNIS Software 4.1: El usuario decide si el pretratamiento de datos se tiene en cuenta o no. La selección de longitud de onda y los cambios posteriores en el pretratamiento de datos no influyen en la división de conjunto de datos.
    c. Hasta la versión de OMNIS Software 3.2: El pretratamiento de datos se tiene en cuenta tal y como se definió en el momento de la detección de valores discrepantes. La selección de longitud de onda y los cambios posteriores en el pretratamiento de datos no influyen en la división de conjunto de datos.
2. La división se basa en el modelo PCA de todos los espectros centrados en la media. Se selecciona la cantidad de componentes principales, de modo que la varianza explicada sea al menos del 95%.
3. A partir de las distribuciones de PCA, se calculan las distancias entre todos los pares de espectros posibles.
4. Los 2 espectros más alejados entre sí se asignan al conjunto de calibración.
5. De los espectros restantes, los 2 espectros más alejados entre sí se asignan al conjunto de validación.
6. De los espectros restantes, el espectro que esté más alejado de los ya contenidos en el conjunto de calibración se asigna a dicho conjunto.
7. De los espectros restantes, el espectro que esté más alejado de los ya contenidos en el conjunto de validación se asigna a dicho conjunto.
8. Se continúa con el cambio hasta que uno de los conjuntos de datos haya alcanzado el tamaño previsto. Los espectros restantes se asignan al otro conjunto de datos.

4.4 Cuantificación

4.4.1 Regresión PLS

OMNIS Software usa la regresión PLS para el cálculo de los modelos de cuantificación.

De forma similar a PCA, la **regresión de mínimos cuadrados parciales** (**regresión PLS**, del inglés *partial least squares regression*), reduce los datos espectroscópicos a un menor número de variables. Sin embargo:

- Los componentes principales de PCA se denominan **variables latentes** en el PLS. Las direcciones de las variables latentes y de los componentes principales suelen ser similares, pero no idénticas.
- Además de los espectros, la regresión PLS también tiene en cuenta los **valores de referencia**. Por lo tanto, se necesitan menos variables latentes para lograr una correlación suficiente con los valores de referencia, lo que se traduce en menos ruido.

El PLS extrae las **variables latentes** subyacentes de los datos espectroscópicos extensos y altamente redundantes. Las variables latentes también se conocen como variables ocultas, ya que no se miden directamente. Las variables latentes explican la mayor parte posible de la variación de los datos y, al mismo tiempo, modelan bien los valores de referencia.

Pasos preparatorios

Los pasos preparatorios para la regresión PLS son similares a los de PCA:

1. **Parametrización**: OMNIS Software aplica a los espectros el pretratamiento de datos especificado y la selección de longitudes de onda especificada.
2. **Centrado en la media**: para cada longitud de onda se calcula el valor de absorbancia medio y se resta del valor respectivo en cada espectro.
   Los valores de referencia también se centran en la media.

Conversión en variables latentes

Después de los pasos preparatorios, la regresión PLS convierte los datos espectrales en un espacio de variables latentes, teniendo en cuenta los valores de referencia. La *figura 30* muestra las 2 primeras variables latentes LV1 y LV2. Del mismo modo, se pueden visualizar las variables latentes LV3, LV4, etc.

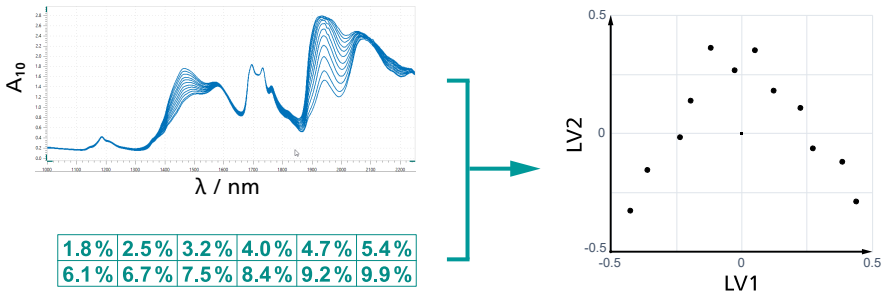
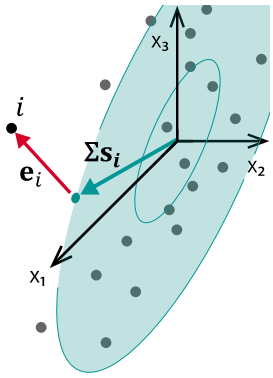


Figura 30 Conversión de espectros y valores de referencia en un espacio de variables latentes. Las distribuciones en el lado derecho se expresan en unidades arbitrarias.

La primera variable latente LV1 es la que mejor explica la varianza en los datos espectrales y, además, la que tiene la mayor correlación posible con los valores de referencia. Todas las variables latentes siguientes LV2, LV3, etc. explican mejor la varianza restante y tienen, además, la mayor correlación posible con los valores de referencia. Así, las primeras variables latentes explican la mayor parte de la varianza y maximizan la correlación, mientras que las demás contienen principalmente ruido y se pueden descartar.

Distribuciones y residuos



El PLS tiene magnitudes similares a las de PCA:

- **Distribuciones**: las distribuciones se miden en el espacio de las variables latentes. La proyección ortogonal de la muestra *i* en cada dirección de las variables latentes da como resultado el vector de distribución $\Sigma\mathbf{s}_i$, que representa la distancia euclídea desde el punto central.
  La **distancia de Mahalanobis** $\mathbf{s}_i$ es el vector de distribución de variables normalizadas que da a cada dirección la misma ponderación.
- **Residuos**: el vector residual $\mathbf{e}_i$ es el desplazamiento entre la muestra *i* y el espacio de la variable latente, medido en el espacio de la longitud de onda original.

Algoritmo PLS

El algoritmo PLS maximiza la covarianza entre los espectros y los valores de referencia *(véase "Algoritmo PLS", capítulo 6.3, página 96)*.

4.4.1.1 Cantidad de variables latentes

La selección de la cantidad de variables latentes en el modelo de cuantificación tiene una importancia fundamental para la capacidad predictiva del modelo. Si el número de variables latentes es demasiado bajo, no se registran las variaciones espectrales relevantes. Esto se conoce como **subajuste** y provoca predicciones menos precisas.

Si el número de variables latentes es demasiado alto, las muestras de calibración se modelan en exceso. El modelo capta variaciones espectra-

les irrelevantes (ruido). Esto se conoce como **sobreajuste** y da lugar a predicciones fluctuantes, inestables y menos precisas de las muestras desconocidas.

Para encontrar la cantidad de variables latentes óptima, debe lograrse un equilibrio entre los siguientes objetivos:

- Que el SEP se aproxime lo suficiente a su valor mínimo.
  Sin un conjunto de validación, se usa el SECV.
- El modelo de cuantificación debería usar la menor cantidad posible de variables latentes. En caso de duda, debe usarse el número más bajo.
- El diagrama de correlación del conjunto de validación se aproxima a su valor óptimo. De forma ideal, la pendiente se aproxima a 1, el sector de eje de coordenadas 'y' se aproxima a 0 y los puntos de datos tienen una dispersión mínima.
  Sin un conjunto de validación, se usan los valores de la validación cruzada *(véase "Validación cruzada", página 65)*.

Debe tenerse en cuenta que las figuras de mérito son solo valores estimados basados en las muestras de calibración y de validación disponibles. Por lo general, un modelo de cuantificación que está basado en menos variables latentes es más robusto.

4.4.1.2 Gráfico de contribución

El gráfico de contribución de OMNIS Software se basa en el modelo PLS *(véase "Algoritmo PLS", capítulo 6.3, página 96)*.

Las contribuciones PLS muestran cómo contribuyen las variables de longitud de onda originales (incluida la parametrización) a la formación de cada variable latente. Es irrelevante si las contribuciones son positivas o negativas.

Las contribuciones se calculan de forma que la primera variable latente capte la varianza que sea más significativa para el parámetro de referencia. Todas las variables latentes siguientes captan la varianza restante que es más significativa para el parámetro de referencia. Las contribuciones PLS que se desvían significativamente de 0 indican, por lo tanto, que las respectivas longitudes de onda son muy adecuadas para modelar el parámetro de referencia.

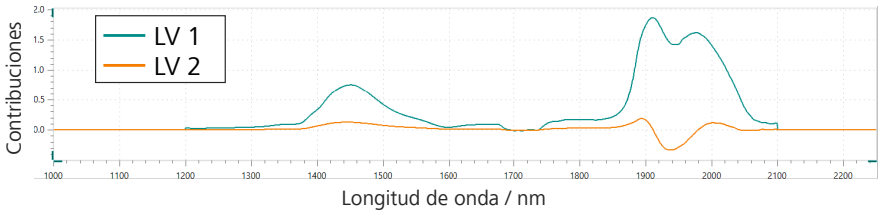


Figura 31 Gráfico de contribución para las variables latentes LV1 y LV2.

En la *figura 31*, la selección de la longitud de onda se limitó al rango de 1200 nm a 2100 nm. Por lo tanto, no se producen contribuciones fuera de este rango.

4.4.2 Validación de los modelos de cuantificación

En la validación se comprueba si el modelo cumple con todos los requisitos en cuanto a potencia y robustez. Para ello, el error esperado de la predicción debe estimarse de la forma más realista posible.

Por lo general, un modelo de cuantificación se valida del siguiente modo:

1. A partir de un número de muestras limitado y sin un conjunto de validación, se desarrolla un modelo y se prueba mediante validación cruzada (véase más adelante).
2. Con un mayor número de muestras, el conjunto de datos se divide en un conjunto de calibración para el desarrollo de un modelo y un conjunto de validación para validar el modelo.
3. Por último, las muestras para el conjunto de validación son recogidas y medidas otro día y, en lo posible, por una persona distinta y con un aparato diferente.

Validación cruzada

La validación cruzada se basa exclusivamente en el conjunto de calibración. Proporciona un valor estimado para cada muestra de calibración usando un modelo temporal que se creó sin la muestra en cuestión.

En la validación cruzada se emplea un procedimiento de múltiples iteraciones de una de las siguientes maneras:

- **Leave-One-Out (dejando una afuera)**
  En cada iteración de la validación cruzada Leave-One-Out (LOO-CV) se aparta 1 muestra, mientras que se crea un modelo a partir de las muestras restantes. Este modelo predice entonces el parámetro de interés para la muestra retirada. Esta predicción sirve como valor estimado para la muestra.
  El ciclo continúa hasta que cada muestra haya sido retirada una vez.

- **K-fold**
  En la validación cruzada K-fold, el conjunto de calibración se divide en *k* bloques de un tamaño lo más parecido posible.
  En cada iteración, se retira 1 bloque, mientras que se crea un modelo a partir de los bloques restantes. A continuación, este modelo predice el parámetro de interés para las muestras retiradas.
  El ciclo continúa hasta que cada bloque haya sido retirado una vez.
  Los bloques pueden seleccionarse de diferentes maneras:
  - **Fixed Blocks (DUPLEX)** (a partir de la versión de OMNIS Software 3.2): los bloques se seleccionan de forma reproducible basándose en un algoritmo dúplex. Cada predicción sirve directamente como valor estimado para la muestra respectiva.
  - **Random**: los bloques se seleccionan al azar. El procedimiento descrito arriba se repite varias veces. El conjunto de calibración se divide en *k* bloques cada vez. Finalmente, hay varios valores estimados para cada muestra. Su valor medio sirve como valor estimado para la muestra respectiva.

En general, se prefiere el método Leave-One-Out (dejando una afuera). Para grandes conjuntos de calibración, puede utilizarse el método de validación cruzada K-fold para ahorrar tiempo de cálculo. Un valor típico para *k* es 5.

4.4.2.1 Diagrama de correlación

El diagrama de correlación visualiza la correlación entre los valores de referencia y los valores calculados. Permite evaluar de un vistazo el modelo de cuantificación.

Los valores calculados se determinan del siguiente modo:

- Conjunto de validación y conjunto de valores discrepantes: predicción por el modelo de cuantificación
- Conjunto de calibración: valores estimados de la validación cruzada

En el diagrama de correlación, cada muestra está representada por un punto. El valor calculado de la muestra se encuentra en el eje x, el valor de referencia en el eje 'y'. Una recta de regresión muestra la relación sistemática entre las variables. Lo ideal sería una recta de regresión con una pendiente de 1, un sector de eje de coordenadas 'y' de 0 y todos los puntos situados sobre la línea. Esto significa que el valor calculado para cada muestra corresponde al valor de referencia.

Los errores sistemáticos y los errores aleatorios pueden distinguirse en función de las desviaciones con respecto al caso ideal. La posición de la recta de regresión muestra los errores sistemáticos. Las distancias de los puntos con respecto a las líneas de regresión indican los errores aleatorios.

El siguiente diagrama de correlación **A** muestra una buena correlación. Los demás diagramas muestran distintos tipos de errores, que se explican a continuación.

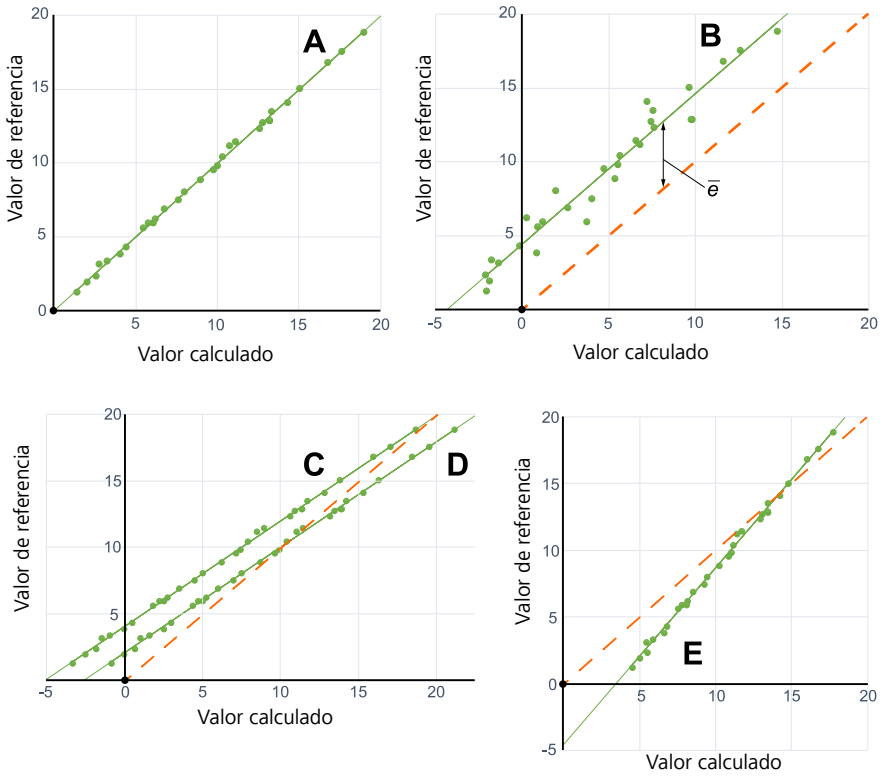


Figura 32 Diagramas de correlación. Cada punto representa una muestra e indica su valor de referencia, así como su valor calculado. La línea discontinua muestra la línea recta ideal de 45°.

Errores sistemáticos

Los errores sistemáticos son errores que se producen siempre y son repetibles para una aplicación específica. Los errores sistemáticos pueden corregirse. Se cuantifican mediante la desviación $\bar{e}$ y la pendiente *b* de las líneas de regresión:

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

Si la pendiente es igual a 1 y la desviación igual a 0, entonces no hay errores sistemáticos.

La **desviación** es el error medio entre los valores de referencia y los valores calculados:

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

Aquí, *n* corresponde al número de muestras; e_i al error de la *i*ésima muestra; y_i al valor de referencia de la *i*ésima muestra; $\hat{y}_i$ al valor calculado de la *i*ésima muestra; $\bar{y}$ al valor medio de los valores de referencia y $\bar{\hat{y}}$ al valor medio de los valores calculados.

Las líneas rectas **B** y **C** tienen una desviación positiva; la línea recta **E** tiene una desviación negativa.

Los errores con signos opuestos se contrarrestan entre sí. Por lo tanto, la desviación de la línea recta **D** es aproximadamente 0.

La **pendiente** de las líneas de regresión es:

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

Donde $s_{\hat{y}y}$ corresponde a la covarianza entre los valores de referencia y los valores calculados, y $s_{\hat{y}}^2$ corresponde a la varianza de los valores calculados.

La pendiente puede considerarse como un error que depende de las propiedades:

- $b > 1$ (línea recta **E**): cuanto mayor sea el valor calculado, mayor (más positivo) será el error que contribuye a la desviación.
- $b < 1$ (líneas rectas **C** y **D**): cuanto mayor sea el valor calculado, menor (más negativo) será el error que contribuye a la desviación.
- $b = 1$ (líneas rectas **A** y **B**): el error que contribuye a la desviación es constante.

El **sector de eje de coordenadas 'y'** de las líneas de regresión con el eje 'y' es $\bar{y} - b\bar{\hat{y}}$.

La pendiente y el sector de eje de coordenadas 'y' se calculan con los valores de referencia como variable dependiente (eje 'y') y los valores calculados como variable independiente (eje 'x').

Errores aleatorios

Si todos los puntos se encuentran directamente en las líneas de regresión, no hay errores aleatorios. Cuanto más dispersos estén los puntos, mayores serán los errores aleatorios.

En el diagrama de correlación **B**, los errores aleatorios son mayores que en los otros diagramas de correlación.

Visualización de los tipos de error

Las líneas rectas de las figuras anteriores muestran los siguientes tipos de error:

	Errores sistemáticos			Errores aleatorios
Línea recta	**Desviación**	**Pendiente**	**Sector de eje de coordenadas 'y'**	
A	~ 0	~ 1	~ 0	pequeño
B	> 0	~ 1	> 0	grande
C	> 0	< 1	> 0	pequeño
D	~ 0	< 1	> 0	pequeño
E	< 0	> 1	< 0	pequeño

4.4.2.2 Figuras de mérito

Las figuras de mérito expresan en cifras la concordancia entre los valores de referencia y los valores calculados. Los valores calculados se determinan mediante el modelo de cuantificación.

R^2 - Coeficiente de determinación

El **coeficiente de determinación R^2** (en inglés, coefficient of determination) mide la calidad del ajuste del modelo de cuantificación. Para un conjunto de datos específico, es la proporción de la variación del valor de referencia explicada por el modelo de cuantificación:

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

Donde SS_{reg} corresponde a la suma de cuadrados de la regresión (varianza de los valores calculados, es decir, varianza explicada); SS_{tot} a la suma total de cuadrados (varianza del valor de referencia); SS_{res} a la suma de cuadrados de los residuos (varianza residual, es decir, varianza no explicada); y_i al valor de referencia de la muestra *i*; $\hat{y}_i$ al valor calculado de la muestra *i* y $\bar{y}$ al valor medio de los valores de referencia.

El valor R^2 es una fracción de 1. Un R^2 de 1 significa que los valores calculados coinciden perfectamente con los valores de referencia. Un R^2 de 0,9 significa que el 90% de la varianza del valor de referencia se explica por los valores calculados y que el 10% no se explica.

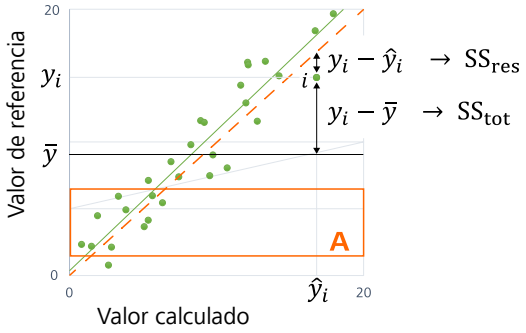


Figura 33 Los componentes para calcular R². La línea discontinua es la línea recta de 45°.

La figura anterior muestra un diagrama de correlación con una muestra *i* y su residuo de la regresión PLS. La varianza residual está contenida en SS_{res}, la varianza del valor de referencia en SS_{tot}.

Un valor alto de R² no garantiza un modelo de cuantificación útil ni predicciones precisas. El tamaño de R² depende directamente de la variación de los valores de referencia.

Una regresión con un rango menor de valores de referencia (rango **A**) tiene aproximadamente la misma variación residual, pero la variación del valor de referencia es menor. El valor R² resultante es menor.

Por lo tanto, la razón de un R² elevado podría ser un rango de valores de referencia exageradamente grande. Por otro lado, los datos de un proceso de fabricación, por ejemplo, pueden tener una gama de valores limitada, lo que conduce a un valor R² más bajo. Para evaluar la capacidad de predicción, se deberían considerar los errores estándar.

El valor R² absoluto debe considerarse con precaución. Es más significativo observar el grado de cambio con cada variable latente adicional *(véase "Cantidad de variables latentes", capítulo 4.4.1.1, página 63)*.

Dependiendo de los valores que se usen para el cálculo, se obtienen diferentes valores de R²:

- R²C (no se muestra en OMNIS Software): se calcula usando los valores calculados de los espectros en el conjunto de calibración.
- **R²CV**: se calcula usando los valores estimados de la validación cruzada de los espectros en el conjunto de calibración *(véase "Validación cruzada", página 65)*.
- **R²P**: se calcula usando los valores calculados de los espectros en el conjunto de validación.

Para el cálculo, OMNIS Software usa el cuadrado del coeficiente de correlación de Pearson de la muestra $r_{y,\hat{y}}$:

Coeficiente de determinación de la validación cruzada:

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

Donde y_i es el valor de referencia de la *i*-ésima muestra; $\bar{y}$ es el valor medio de los valores de referencia; $\hat{y}_{cv_i}$ es el valor estimado de la validación cruzada de la *i*-ésima muestra; $\bar{\hat{y}}_{cv}$ es el valor medio de los valores estimados de la validación cruzada y *n* es el número de muestras en el conjunto de calibración. Es importante notar que cada muestra en el conjunto de calibración tiene exactamente un valor estimado de la validación cruzada.

Coeficiente de determinación de la predicción:

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

Aquí, v_i corresponde al valor de referencia de la *i*-ésima muestra de validación; $\bar{v}$ al valor medio de los valores de referencia; $\hat{v}_i$ al valor calculado de la *i*-ésima muestra de validación; $\bar{\hat{v}}$ al valor medio de los valores calculados y v al número de muestras de validación.

SEC – Error estándar de la calibración

El **error estándar de la calibración (SEC)** se basa en el conjunto de calibración. El SEC se puede considerar como el valor estimado de mayor precisión de la predicción teórica. El SEC es la desviación estándar de los residuos de la regresión de mínimos cuadrados parciales (PLS):

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

Donde **e** corresponde al vector residual, que contiene todas las variaciones de los valores de referencia en el conjunto de calibración que no describe el modelo; *n* es el número de muestras de calibración y *k* es el número de variables latentes. El denominador *n*-*k*–1 es el número de grados de libertad del vector residual **e**.

En otras palabras, el SEC es la desviación estándar de las diferencias entre los valores de referencia y los valores calculados para las muestras del conjunto de calibración:

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

Donde y_i corresponde al valor de referencia de la *i*-ésima muestra de calibración; $\hat{y}_i$ al valor calculado de la *i*-ésima muestra de calibración; *n* es el número de muestras de calibración y *k* es el número de variables latentes.

En ocasiones, el SEC también se denomina RMSEC. El SEC incluye los errores aleatorios y los errores sistemáticos (pendiente y desviación).

SECV - Error estándar de la validación cruzada

El **error estándar de la validación cruzada (SECV)** se basa en el conjunto de calibración. El SECV estima la precisión de la predicción basándose en el conjunto de calibración y en un procedimiento de validación cruzada *(véase "Validación cruzada", página 65)*. El SECV puede usarse para realizar una primera evaluación del modelo o para determinar el número óptimo de variables latentes.

El SECV es la desviación estándar de las diferencias entre los valores de referencia y los valores estimados de la validación cruzada para las muestras del conjunto de calibración.

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

Donde y_i es el valor de referencia de la *i*-ésima muestra; $\hat{y}_{\mathrm{cv}_i}$ es el valor estimado de la validación cruzada de la *i*-ésima muestra y *n* es el número de muestras en el conjunto de calibración.

El SECV incluye todos los errores: los errores aleatorios y los errores sistemáticos (pendiente y desviación).

Otros enfoques usan valores separados para un SECV corregido por la desviación y para el SECV sin corregir, que se denomina RMSECV. Con una desviación baja, estos valores son similares.

SEP - Error estándar de la predicción

El **error estándar de la predicción (SEP)** se basa en el conjunto de validación. Por lo tanto, el SEP proporciona el valor estimado más realista de precisión de la predicción.

El SEP es la desviación estándar de las diferencias entre los valores de referencia y los valores calculados para las muestras del conjunto de validación:

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

Donde v_i corresponde al valor de referencia de la *i*-ésima muestra de validación; $\hat{v}_i$ al valor calculado de la *i*-ésima muestra de validación y v al número de muestras de validación.

El SEP incluye todos los errores: los errores aleatorios y los errores sistemáticos (pendiente y desviación).

Otros enfoques usan valores separados para un SEP corregido por la desviación y para el SEP sin corregir, que se denomina RMSEP. Con una desviación baja, estos valores son similares.

Interpretación de las figuras de mérito

Las figuras de mérito son valores estimados. Ejemplo: el valor SEP es un *valor estimado* de una desviación estándar, sobre la base de las muestras existentes, y también tiene su propia desviación estándar. Cuanto mayor sea el número de muestras de validación, más fiable será el valor estimado.

El SEP debería ser comparable con el SECV y el SEC. Si hay diferencias excesivas, esto puede indicar un sobreajuste *(véase "Cantidad de variables latentes", capítulo 4.4.1.1, página 63)*. Como regla general, las diferencias no deberían superar el 20%.

Los errores estándar también deben considerarse en relación con el error estándar del laboratorio (SEL) para el método de referencia. La exactitud del método NIR es aceptable si el SEP es de 1,4 a 2,0 veces mayor que el SEL. Un SEP mayor puede ser aceptado siempre que cumpla con los requisitos.

Un SEC o SECV inferior al SEL para el método de referencia indica un sobreajuste.

Las relaciones mencionadas anteriormente con el SEL, suponen que el SEL se basa en el número correcto de medidas de referencia repetidas para cada muestra *(véase "Método de referencia (para cuantificación)", página 35)*.

4.4.3 OMNIS Model Developer (OMD)

El desarrollo de un modelo de cuantificación es exigente, lleva mucho tiempo y requiere cierta pericia. El OMNIS Model Developer (OMD, desde la versión de OMNIS Software 4.0) automatiza el desarrollo y proporciona modelos de cuantificación bien optimizados.

Funcionamiento

Un muestreo adecuado es un requisito *(véase "Muestras físicas", capítulo 4.1, página 34)*. Como entrada para el OMD, se usa un conjunto de datos que consta de espectros y los valores de referencia correspondientes.

El OMD identifica los valores discrepantes espectrales con un nivel de significancia del 5% y usando el algoritmo que figura en el apéndice, sin tener en cuenta el pretratamiento de datos *(véase "Detección de valores discrepantes espectrales en el desarrollo del modelo", página 99)*.

La división del conjunto de datos depende del número de espectros que queden después de la detección de valores discrepantes:

Número de espectros	Método de validación cruzada	Conjunto de validación
> 99	K-fold (5 bloques, DUPLEX)	25% de los espectros
30–99	K-fold (5 bloques, DUPLEX)	—
< 30	Leave-One-Out (dejando una afuera)	—

Tras la división, se determinan los valores discrepantes adicionales en el conjunto de calibración según ASTM D8321-22.

La evaluación y la clasificación de los modelos se realizan usando diversas métricas. El OMD optimiza el pretratamiento de datos, la selección de longitudes de onda y el número de variables latentes, con lo que se busca un equilibrio entre el riesgo de sobreajuste y el riesgo de subajuste.

Resultado

El resultado del OMD es una lista de modelos clasificados según su capacidad predictiva. La **capacidad predictiva** se calcula en función de la complejidad del modelo, las figuras de mérito y el tamaño del conjunto de datos.

La lista está codificada por colores para facilitar la selección del modelo preferido:

- Verde: buena capacidad predictiva.
  Si el número de muestras es suficientemente grande, el modelo funcionará bien con todas las muestras desconocidas del mismo tipo. Las figuras de mérito proporcionan un valor estimado fiable de los errores futuros.
- Amarillo: capacidad predictiva media.
  Si el número de muestras es suficientemente grande, es probable que el modelo funcione bien. Sin embargo, las figuras de mérito podrían arrojar datos demasiado optimistas para muestras futuras. Se recomienda realizar una validación por separado.
- Rojo: capacidad predictiva insuficiente.
  El modelo presenta deficiencias importantes. No debería utilizarse.

Los modelos del mismo color se ordenan según un criterio de información que favorece una compensación equilibrada entre un error de predicción estimado bajo y un número reducido de variables latentes.

Optimizar la parametrización

En lugar de crear todo el modelo automáticamente, también la parametrización solo se puede optimizar. Los ajustes actuales (por ejemplo, la

división de conjunto de datos, el método de validación cruzada) no se modifican, pero no influyen en la optimización.

4.4.4 Corrección de pendiente/del sector de eje de coordenadas 'y'

La corrección de pendiente/del sector de eje de coordenadas 'y' permite corregir los errores sistemáticos (desviación, pendiente) al aplicar un modelo de cuantificación.

Las posibles causas de errores sistemáticos en el conjunto de calibración son:

- Errores sistemáticos en el modelo de cuantificación. Por ejemplo, valores discrepantes no reconocidos o un número de muestras insuficiente.
- Errores sistemáticos en el método de medición espectroscópica.
- Errores sistemáticos en el método de medición de referencia.

Si se producen errores sistemáticos en el conjunto de validación, otras causas posibles son:

- Cambios en el procedimiento de medición espectroscópica, por ejemplo, en el aparato.
- Cambios en el método de medición de referencia, por ejemplo, nuevo laboratorio, nuevo equipo.
- Cambios en las muestras, p. ej., durante el manejo, el almacenamiento o el transporte.

La corrección de desviación y, aún más, la corrección de pendiente/del sector de eje de coordenadas 'y' deberían usarse con precaución.

Si los errores sistemáticos no son significativos, no debería aplicarse ninguna corrección. Si los errores son significativos, se deberían examinar minuciosamente. Si es posible, se debería corregir la causa de los errores. Si los errores no pueden corregirse por una razón válida, puede aplicarse una corrección de desviación o una corrección de pendiente/del sector de eje de coordenadas 'y'.

Se necesitan al menos 20 muestras para un valor estimado fiable de la desviación. Se necesitan al menos 30 muestras para un valor estimado fiable de la pendiente.

Corrección de desviación

El siguiente diagrama de correlación muestra una corrección de desviación. La pendiente de las líneas de regresión originales **F** no se modifica. Tras la corrección (recta de regresión **G**), los errores positivos y los negativos se contrarrestan entre sí.

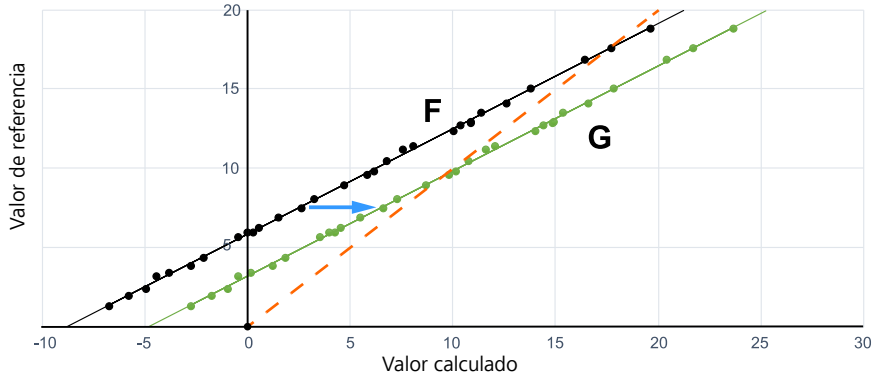


Figura 34 Corrección de desviación

Corrección de pendiente/del sector de eje de coordenadas 'y'

El siguiente diagrama de correlación muestra una corrección de pendiente/del sector de eje de coordenadas 'y'. La recta de regresión original **H** se corrige mediante la pendiente y el sector de eje de coordenadas 'y'. Como resultado, se corrigen tanto la pendiente como la desviación (recta de regresión **K**).

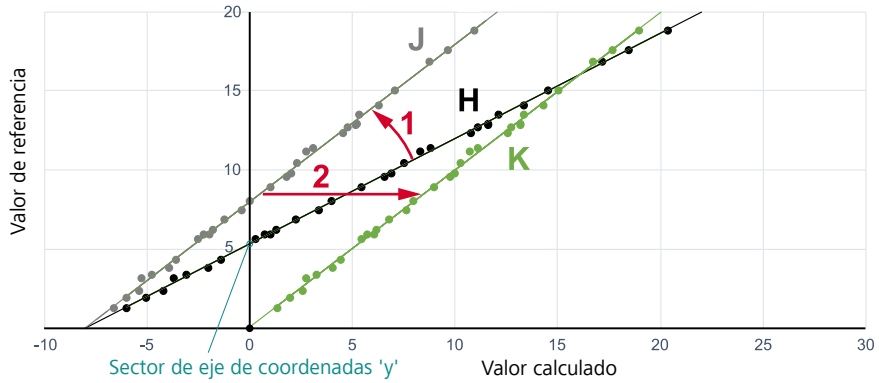


Figura 35 Corrección de pendiente/del sector de eje de coordenadas 'y'

SEP

Las muestras incluidas en el conjunto de valores de corrección constituyen la base para la corrección de pendiente/del sector de eje de coordenadas 'y'. Basándose en estas muestras, OMNIS Software calcula los siguientes **errores estándar de la predicción (SEP)**. Los denominadores tienen en cuenta los grados de libertad correspondientes:

Tipo de corrección	SEP
No corregido	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n}}$
Corrección de desviación	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-1}}$
Corrección de pendiente/del sector de eje de coordenadas 'y'	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-2}}$

En este caso, v_i corresponde al valor de referencia de la muestra *i*-ten en el conjunto de valores de corrección, $\hat{v}_i$ corresponde al valor predicho de la muestra *i*-ten en el conjunto de valores de corrección y *n* corresponde a la cantidad de muestras en el conjunto de valores de corrección.

El SEP incluye todos los errores: los errores aleatorios y los errores sistemáticos (pendiente y desviación).

4.5 Identificación y verificación

4.5.1 Máquina de soporte vectorial (SVM)

Los modelos de identificación (a partir de la versión de OMNIS Software 4.0) usan máquinas de soporte vectorial (SVM) para la clasificación entre diferentes productos. Una máquina de soporte vectorial es un algoritmo de aprendizaje automático supervisado. Basándose en las muestras de calibración, aprende a asignar nuevas muestras a un producto.

Para que resulte más sencillo, a continuación se describe la clasificación entre 2 productos. El concepto es escalable; el modelo final permite clasificar entre cualquier número de productos.

Clasificación lineal

La *figura 36 (a la izquierda)* muestra datos de entrada con 2 variables.

Para simplificar, los espectros parametrizados se representan en un espacio de variables bidimensional. Cada punto representa un espectro, el color indica la pertenencia al producto.

Los productos son linealmente separables. El algoritmo SVM crea un hiperplano entre los productos (figura de la derecha).

Los hiperplanos son una generalización de los planos del espacio tridimensional a espacios de cualquier dimensión. La dimensión de un hiperplano es una menos que la dimensión del espacio que lo rodea. Un hiperplano en un espacio bidimensional es una línea.

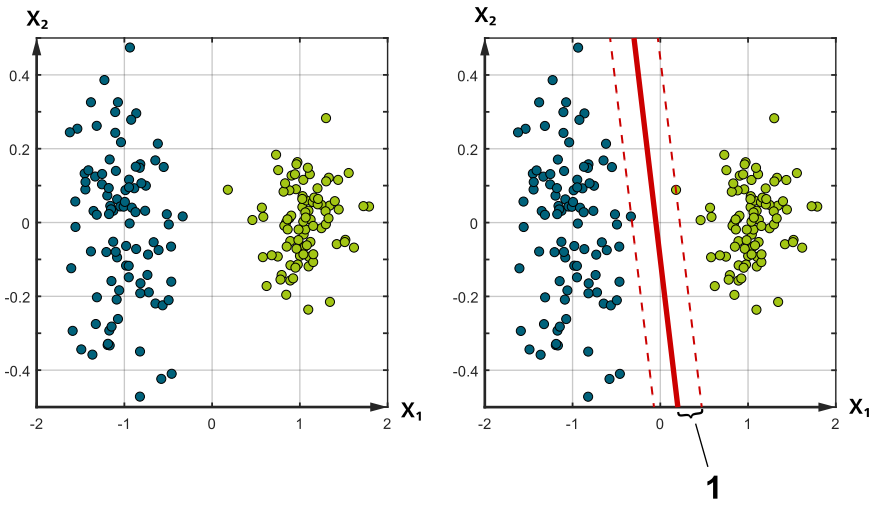


Figura 36 Los datos de entrada (a la izquierda) y el hiperplano creado por SVM (línea roja sólida a la derecha). Los valores se proporcionan en unidades arbitrarias.

El algoritmo SVM maximiza el margen (**1**) entre el hiperplano y el punto más cercano en cada lado. Los nuevos espectros pueden mapearse en el mismo espacio y asignarse a un producto dependiendo de qué lado del hiperplano caigan.

Para definir el hiperplano, el algoritmo SVM solo considera los puntos más cercanos a los puntos del producto opuesto. Estos puntos o vectores apoyan la formación del hiperplano y se llaman vectores de soporte.

Si los puntos no son linealmente separables, por ejemplo, debido a un valor discrepante, aún se puede determinar un hiperplano de clasificación lineal. En este caso, un algoritmo de optimización encuentra un equilibrio entre aumentar el margen desde el hiperplano hasta los vectores de soporte en cada lado y garantizar que todos los puntos estén en el lado correcto del hiperplano. Un parámetro de regularización controla este equilibrio y, por lo tanto, la posición final del hiperplano.

Clasificación no lineal

En la *figura 37 (a la izquierda),* los productos no son linealmente separables. Para separar los productos se necesita un clasificador no lineal.

Una función kernel lineal o no lineal transforma los datos en un espacio de características de mayor dimensión. La transformación se realiza de tal manera que los datos en el espacio de características puedan separarse linealmente por un hiperplano.

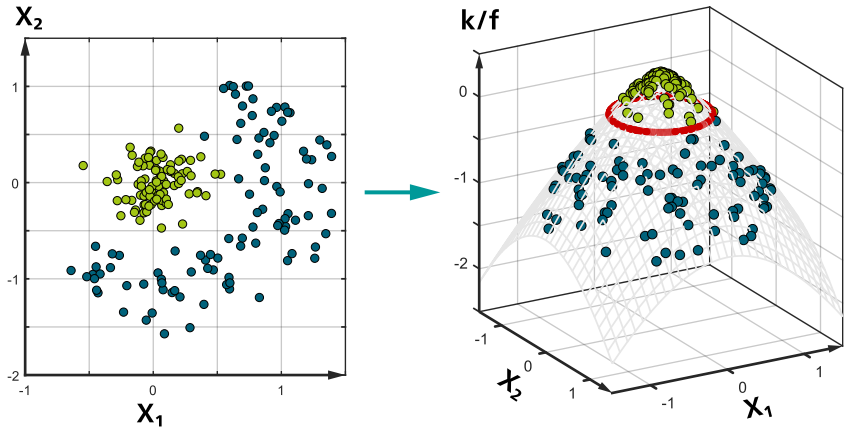


Figura 37 Productos no linealmente separables (a la izquierda). Una dimensión adicional k/f (característica kernel) facilita la separación (a la derecha). Los valores se proporcionan en unidades arbitrarias.

En la figura, los datos se transforman del espacio bidimensional al espacio tridimensional. Los puntos de un producto se elevan por encima del plano original, mientras que los puntos del otro producto se desplazan hacia abajo. El hiperplano lineal entre los productos es un plano bidimensional muy similar al plano original. Considerado en 2 dimensiones, el límite de decisión es una línea no lineal, en este caso la línea circular roja.

También en este caso, el hiperplano sirve como clasificador. Una nueva muestra se puede asignar a un producto en función del lado del hiperplano en el que caiga su espectro.

OMNIS Software usa un kernel de función de base radial para transformar los datos en el espacio de características. El kernel utiliza un parámetro de escalado que controla el grado de falta de linealidad.

Selección de parámetros

Deben seleccionarse valores adecuados para el parámetro de regularización, que controla la posición del hiperplano, y para el parámetro de escalado, que controla el grado de falta de linealidad.

Ambos parámetros afectan a la capacidad de generalización de la máquina de soporte vectorial, es decir, hasta qué punto puede generalizarse a partir de los espectros de calibración a nuevos espectros desconocidos. Por ejemplo, si el parámetro de escalado permite un alto grado de falta de linealidad, el hiperplano puede estar demasiado ajustado a los espectros de calibración (sobreajuste). Si el parámetro de escalado solo permite un bajo grado de falta de linealidad, el hiperplano puede no estar suficientemente ajustado a los espectros de calibración (subajuste).

Para una buena generalización se utiliza la **búsqueda en cuadrícula**. El algoritmo encuentra la mejor combinación de parámetros con el menor número de intentos posible:

1. Se utiliza un conjunto de combinaciones de parámetros predefinidas.
2. Para algunas combinaciones de parámetros seleccionadas, la SVM aprende una regla de clasificación que distingue entre los espectros de calibración de los distintos productos con la máxima exactitud.
3. Para cada regla de clasificación, se usa una validación cruzada para estimar el grado de acierto en la asignación de productos.
4. En función de los resultados de la validación cruzada, se seleccionan otras combinaciones de parámetros.
   La SVM vuelve a aprender las reglas de clasificación correspondientes y un método de validación cruzada estima la exactitud de la clasificación.
   Tras unas cuantas repeticiones, se determina la combinación final de parámetros.
5. Con la combinación final de parámetros y todos los espectros de calibración, la SVM evalúa la regla de clasificación definitiva.

Probabilidades

Basándose en la regla de clasificación, el modelo de identificación calcula la probabilidad de que una muestra determinada pertenezca o no a un determinado producto. La probabilidad depende de las distancias respecto a las poblaciones de los productos y de los parámetros del modelo.

Las probabilidades así calculadas permiten un control de la asignación de las muestras a los productos *(véase "Asignación de una muestra (a partir de la versión 4.4. de OMNIS Software)", página 80)*.

Para reducir al mínimo las identificaciones falsamente positivas, a partir de la versión 4.4 de OMNIS Software se pone a punto también un modelo de cualificación para cada producto *(véase "Cualificación", capítulo 4.6, página 84)*.

4.5.2 Predicción de la clase de producto de una muestra

Para la identificación de una determinada muestra, el modelo de identificación proporciona una probabilidad por producto *(véase "Probabilidades", página 80)*.

Cada probabilidad es un valor individual entre 0% y 100%. Los valores no suman 100% entre los productos.
Los valores deben considerarse relativos entre sí, lo que permite establecer una comparación entre los diferentes productos.

Asignación de una muestra (a partir de la versión 4.4. de OMNIS Software)

La evaluación se efectúa utilizando un **umbral de probabilidad** ajustable y cualificaciones para productos específicos:

1. Para cada producto cuya probabilidad supere el umbral de probabilidad, se ejecuta una cualificación de la muestra con el modelo de cualificación correspondiente. Si la cualificación falla, la probabilidad para el producto correspondiente se establece en cero.
2. Evaluación con las probabilidades modificadas del paso 1:
   a. Si ninguna probabilidad supera el umbral de probabilidad, la identificación falla (estado de identificación **No identificado**).
   b. Si una sola probabilidad supera el umbral de probabilidad, la muestra se identifica correctamente y se asigna al producto correspondiente (estado de identificación **Identificado**).
   c. Si varias probabilidades superan el umbral de probabilidad, la predicción es ambigua y la identificación falla (estado de identificación **Ambiguo**).

La tabla 1 muestra un ejemplo de evaluación con diferentes umbrales de probabilidad. En este ejemplo, ambas cualificaciones para los productos A y C son correctas.

Tabla 1 Ejemplo de evaluación con diferentes umbrales de probabilidad (a partir de la versión 4.4 de OMNIS Software)

Probabilidad	Umbral de probabilidad	Cualificación	Resultado de la identificación
Producto A: 87% Producto B: 71% Producto C: 68% Producto D: 30%	90%	–	No identificado
	80%	Producto A: correcta → 87%	Producto A
	70%	Producto A: correcta → 87% Producto B: fallida → 0%	Producto A
	60%	Producto A: correcta → 87% Producto B: fallida → 0% Producto C: correcta → 68%	Ambiguo

Asignación de una muestra (versión 4.0 a 4.3 de OMNIS Software)

Las probabilidades de la muestra se evalúan mediante un **umbral de probabilidad** ajustable:

- Si ninguna probabilidad supera el umbral de probabilidad, la identificación falla (estado de identificación **No identificado**).
- Si una sola probabilidad supera el umbral de probabilidad, la muestra se identifica correctamente y se asigna al producto correspondiente (estado de identificación **Identificado**).
- Si varias probabilidades superan el umbral de probabilidad, la predicción es ambigua y la identificación falla (estado de identificación **Ambiguo**).

La tabla 2 muestra un ejemplo de evaluación con diferentes umbrales de probabilidad.

Tabla 2 Ejemplo de evaluación con diferentes umbrales de probabilidad (versión 4.0 a 4.3 de OMNIS Software)

Probabilidad	Umbral de probabilidad	Resultado de la identificación
Producto A: 87%	90%	No identificado
Producto B: 72%	80%	Producto A
Producto C: 68%	70%	Ambiguo

4.5.3 Validación de modelos de identificación

Por lo general, un modelo de identificación se valida del siguiente modo:

1. A partir de un número de muestras limitado y sin un conjunto de validación, se desarrolla un modelo y se prueba con las muestras en un conjunto de calibración.
2. Si el número de muestras es suficiente, el conjunto de datos se puede dividir en un conjunto de calibración y un conjunto de validación. Las muestras del conjunto de validación no se usan para el desarrollo del modelo.
3. Por último, las muestras para un conjunto de validación externo son recogidas y medidas otro día y, en lo posible, por una persona distinta y con un aparato diferente.

Para cada muestra, se compara la clase de producto predicha con la clase de producto real correspondiente. Si coinciden, la predicción es correcta (= acertada), de lo contrario es incorrecta (= fallida).

Validación

OMNIS Software muestra las siguientes métricas para los modelos de identificación, que indican hasta qué punto funcionan los modelos. En el escenario ideal, todos los valores estarían en 100%.

El *% de éxito (total)* mide la *exactitud*, la corrección global de un modelo. Este porcentaje responde a la pregunta: ¿cuántas de las muestras puede identificar correctamente el modelo?

$$\%\ \text{de éxito (total)} = \frac{\text{clasificaciones correctas}}{\text{todas las clasificaciones}}$$

% de éxito (total) asigna a cada muestra el mismo peso. Por lo tanto, los productos con más muestras tienen un mayor impacto sobre el porcentaje que los productos con menos muestras.

Para cada producto existen números similares para el *% de éxito*.

Mejorar la identificación

Para mejorar el modelo se pueden realizar las siguientes acciones:

- **Ajustar el umbral de probabilidad**
  - Si muchas predicciones son ambiguas o aparecen muchas probabilidades del 0,0%, se puede elevar el umbral de probabilidad.
  - Si no se identifican muchas muestras, porque no se alcanzó el umbral de probabilidad, este umbral puede rebajarse.
- **Ajustar la parametrización**
  Determinar un pretratamiento de datos y una selección de longitud de onda más adecuados.
- **Usar jerarquías de modelos**
  Una jerarquía de modelos permite estructurar de forma jerárquica los modelos de identificación.
  Por ejemplo: un modelo de identificación con 4 productos diferentes puede tener problemas para distinguir entre productos similares como fructosa y glucosa. Si la fructosa y la glucosa se asignan a un grupo de productos llamado "Azúcar", el modelo puede diferenciar entre azúcar y los otros dos productos. Si una muestra se identifica como "azúcar", otro modelo se hace cargo de diferenciar entre fructosa y glucosa. Como este modelo es más especializado, puede distinguir más fácilmente entre los productos similares.

Muestras no identificadas

Si, por error, las muestras no se identifican:

- Verificar si ha ocurrido alguna anomalía en las muestras o en el procesamiento de muestras.
- Verificar el umbral de probabilidad y, en caso necesario, rebajarlo.
- Verificar el pretratamiento de datos y la selección de longitud de onda.
- Revisar los espectros de las muestras no identificadas en el gráfico de distribución:
  - Si los espectros no están ya incluidos en el modelo: agregar los espectros al conjunto de validación del respectivo producto.
  - En el gráfico de distribución, comparar las distribuciones de los espectros que se van a verificar con las distribuciones de los espectros en el conjunto de calibración.

  Si las muestras no son valores discrepantes, pero sus variaciones en el conjunto de calibración están subrrepresentadas, se debería ampliar el conjunto de calibración según corresponda.

4.6 Cualificación

Los modelos de cualificación establecen distinciones entre un grupo de muestras y otras muestras (a partir de la versión 4.4 de OMNIS Software). Estos modelos son adecuados para diferenciar, por ejemplo, entre muestras utilizables (muestras positivas) y muestras no utilizables (muestras negativas).

4.6.1 Cálculo de modelos de cualificación

El cálculo de un modelo de cualificación se lleva a cabo de forma similar al de un modelo de identificación *(véase "Máquina de soporte vectorial (SVM)", capítulo 4.5.1, página 77)*. No obstante, el conjunto de calibración para la cualificación contiene únicamente una sola clase de muestras (muestras positivas).

Una máquina de soporte vectorial (SVM) transforma los datos de entrada en un espacio de mayor dimensión. Un parámetro de regularización establece la posición del hiperplano. mientras que un parámetro de escalado determina el grado de falta de linealidad. Mediante una búsqueda en cuadrícula se determina una proyección adecuada. El límite de decisión para esta proyección constituye la base para el modelo de cualificación.

4.6.2 Validación de modelos de cualificación

Procedimiento

Por lo general. un modelo de cualificación se desarrolla y se valida gradualmente. Así, el número de muestras se va aumentando poco a poco:

1. Conjunto de validación positivo:
   a. Si hay un número limitado de muestras positivas, al principio no se crea ningún conjunto de validación positivo.
   b. Tan pronto como se disponga de un número suficiente de muestras positivas se crea un conjunto de validación positivo mediante la división de conjunto de datos realizada automáticamente.

   Conjunto de validación negativo:
   a. Las muestras negativas recogidas se asignan al conjunto de validación negativo.
   b. La determinación automática de espectros negativos puede emplearse para detectar valores discrepantes espectrales y asignarlos al conjunto de validación negativo *(véase "Valores discrepantes espectrales – Algoritmo", capítulo 6.5, página 98)*.
2. Por último, las muestras para el conjunto de validación positivo y el conjunto de validación negativo son recogidas y medidas otro día y, en lo posible, por una persona distinta y con un aparato diferente.

Las muestras del conjunto de validación positivo y del conjunto de validación negativo no se usan para el cálculo del modelo.

Validación

El modelo de cualificación determina un resultado (positivo o negativo) para cada muestra. Se espera un resultado positivo para las muestras del conjunto de calibración y del conjunto de validación positivo. Se espera un resultado negativo para las muestras del conjunto de validación negativo. Si el resultado coincide con lo esperado, la predicción es correcta (= acertada), de lo contrario es incorrecta (= fallida).

Para los modelos de cualificación, OMNIS Software muestra los siguientes valores:

El *% de éxito (total)* mide la corrección global de un modelo.

$$\%\ de\ éxito\ (total) = \frac{predicciones\ correctas}{todas\ las\ predicciones}$$

Existen números similares para el *% de éxito* para cada conjunto de datos. En el escenario ideal, todos los valores estarían en 100%.

Mejorar la cualificación

Mediante una parametrización más adecuada (pretratamiento de datos y selección de longitud de onda) se puede mejorar el modelo de cualificación.

Muestras no cualificadas

Si, por error, las muestras no se cualifican:

- Verificar si ha ocurrido alguna anomalía en las muestras o en el procesamiento de muestras.
- Verificar el pretratamiento de datos y la selección de longitud de onda.
- Verificar los espectros de las muestras no cualificadas en el gráfico de distribución:
  - Si los espectros no están ya incluidos en el modelo: agregar los espectros al conjunto de validación positivo.
  - En el gráfico de distribución, comparar las distribuciones de los espectros que se van a verificar con las distribuciones de los espectros en el conjunto de calibración.

  Si las muestras no son valores discrepantes, pero sus variaciones en el conjunto de calibración están subrrepresentadas, se debería ampliar el conjunto de calibración según corresponda.

5 Predicción

5.1 Cuantificación

En la predicción de una muestra con un parámetro de interés desconocido, se procede del siguiente modo:

1. Se registra el espectro de la muestra.
2. El modelo de cuantificación aplica el mismo pretratamiento de datos y las mismas gamas de longitudes de onda que aplica a los espectros en el conjunto de calibración.
3. Teniendo en cuenta el espectro resultante, el modelo predice el parámetro de interés.

Al calcular el modelo de cuantificación, los espectros de calibración y los valores de referencia se centraron en la media. Por supuesto, esto se tiene en cuenta en el procedimiento descrito anteriormente.

5.1.1 Valor discrepante y monitorización del resultado

En la predicción de un parámetro de interés cuantitativo, se pueden reconocer distintos tipos de valores discrepantes y monitorizar los resultados:

Valor discrepante/Monitorización		Causa (Ejemplo)
Valores discrepantes espectrales	T^2 de Hotelling	La concentración de un componente químico está fuera de la gama de concentración de las muestras de calibración.
	Residuos Q	La muestra contiene componentes que no están presentes en las muestras de calibración.
Valor discrepante Nearest Neighbor (opcional, a partir de la versión de OMNIS Software 4.2)		La muestra presenta variaciones que no están representadas en esta combinación en las muestras de calibración.
Monitorización del resultado (opcional)		El valor del resultado del parámetro de interés está fuera de los criterios de aceptación seleccionados por el cliente.

Valores discrepantes espectrales

Los valores discrepantes espectrales se determinan de la siguiente manera:

1. El software calcula los valores de T^2 de Hotelling y de residuos Q para el espectro teniendo en cuenta el modelo de cuantificación (modelo PLS).
2. Si el valor de T^2 o de residuos Q del espectro es mayor que el valor crítico correspondiente calculado por el modelo, la muestra se considera como valor discrepante con respecto al modelo aplicado *(véase "Evaluación de valores discrepantes durante la predicción (cuantificación)", página 100)*.

Los valores residuales T^2 y Q están disponibles como variables en OMNIS Software. Usted puede comparar con los valores en el gráfico de influencia PLS del modelo de cuantificación. Las líneas discontinuas indican los valores críticos.

Valor discrepante Nearest Neighbor

(a partir de la versión de OMNIS Software 4.2)

Lo ideal es que las muestras de calibración cubran todas las combinaciones posibles de variaciones de la muestra. En realidad, algunas combinaciones se dan con más frecuencia, otras no. En consecuencia, las muestras de calibración están distribuidas de forma desigual en el espacio de variables latentes. En algunas partes hay muchas muestras de calibración, en otras hay lagunas.

En caso de que el espectro de una muestra desconocida caiga en una laguna entre las muestras de calibración, el resultado de la predicción puede ser inválido o inexacto. Para reconocer estos casos, se calcula la distancia *D* de la muestra desconocida *i* a cada muestra de calibración *u*:

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

Aquí $\mathbf{s_i}$ corresponde a las distribuciones de la muestra desconocida *i* y $\mathbf{s_u}$ a las distribuciones de la muestra de calibración *u*. Las distribuciones están normalizadas y son ortogonales.

La distancia más pequeña es la distancia a la muestra de calibración más cercana y se denomina **Nearest Neighbor Distance** (NND):

Si el valor NND supera un determinado valor límite NND, la muestra desconocida se denomina Valor discrepante Nearest Neighbor.

El valor límite NND se determina del siguiente modo:

1. Para cada muestra de calibración se determina un valor NND. Este valor corresponde a la distancia a la más cercana de las muestras de calibración restantes.
2. El valor NND máximo de todas las muestras de calibración es el valor límite NND.

El valor NND de la muestra desconocida y el valor límite NND están disponibles como variables en OMNIS Software.

Monitorización del resultado

En OMNIS Software, la monitorización del resultado puede usarse para definir límites de indicación y límites de intervención para la sección de resultados de la predicción. De manera opcional, se pueden definir las acciones que se pondrán en marcha cuando se infrinjan los límites definidos.

5.2 Identificación y verificación

Al identificar o verificar una muestra se procede del siguiente modo:

1. Se registra el espectro de la muestra.
2. El modelo de identificación aplica el mismo pretratamiento de datos y la misma selección de longitud de onda que se utilizan para los espectros en el conjunto de calibración.
3. El modelo de identificación evalúa el espectro *(véase "Predicción de la clase de producto de una muestra", capítulo 4.5.2, página 80)*.
4. Si el modelo de identificación forma parte de una jerarquía de modelos y otro modelo de identificación está vinculado al producto determinado, se ejecuta este modelo. Si uno o varios modelos de cuantificación están vinculados al producto determinado, se ejecutan estos modelos.
5. Se muestra el resultado de la identificación o de la verificación; en el caso de una jerarquía de modelos, también pueden mostrarse los resultados de la cuantificación.

Estado de la identificación

- Identificado
  La identificación se ha efectuado correctamente.
- Ambiguo
  Varios productos superan el umbral de probabilidad. La identificación ha fallado.
- No identificado
  Ningún producto supera el umbral de probabilidad. La identificación ha fallado.

Estado de la verificación

- Satisfactorio
  La muestra fue identificada correctamente y el resultado corresponde al producto esperado.
- Fallo
  La verificación ha fallado.

5.3 Cualificación

En la cualificación de una muestra se procede del siguiente modo:

1. Se registra el espectro de la muestra.
2. El modelo de cualificación aplica el mismo pretratamiento de datos y la misma selección de longitud de onda que se utilizan para los espectros en el conjunto de calibración.
3. Teniendo en cuenta el espectro resultante, el modelo cualifica la muestra.
4. Se muestra el resultado de cualificación.

Estado de cualificación

- Satisfactorio
- Fallo

6 Apéndice

6.1 Ejemplo de una regresión lineal

Regresión lineal univariada

En el caso más simple, una mezcla tiene solo un absorbedor y el espectro solo tiene un pico. Las muestras con diferentes concentraciones del absorbedor tienen picos con diferentes valores de absorbancia *(véase "Ley de Beer-Lambert", capítulo 2.2.1, página 7)*.

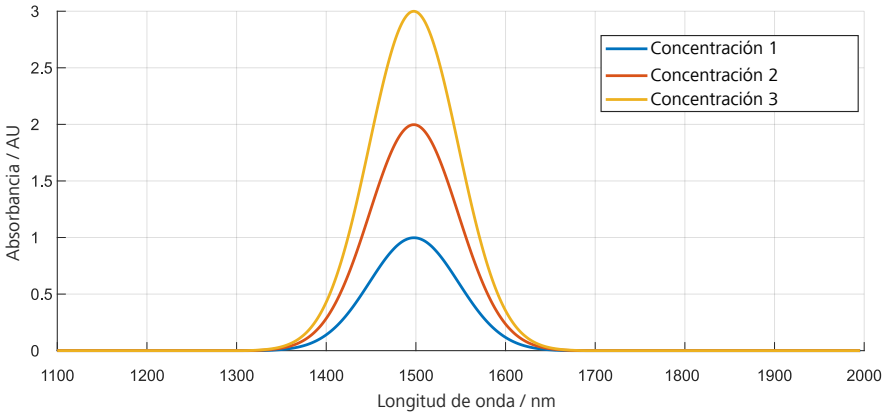


Figura 38 Datos modelados para 3 muestras con un pico a 1500 nm.

Los 3 valores de absorbancia medidos a 1500 nm pueden representarse gráficamente frente a la concentración del absorbedor.

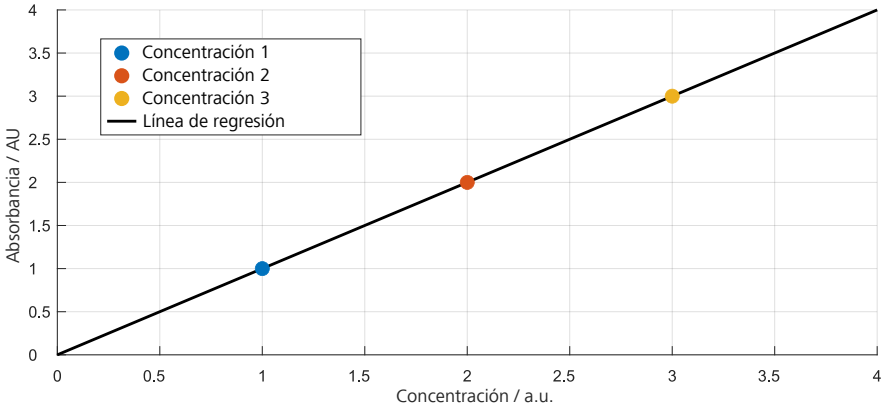


Figura 39 Relación entre valores de absorbancia y concentraciones.

Según la ley de Beer-Lambert, la relación entre los valores de absorbancia y las concentraciones es lineal. Con el conjunto de 3 muestras se puede realizar una regresión lineal, que da como resultado una recta de regresión como la representada.

Dado que solo hay 1 variable (1 longitud de onda), se trata de una regresión lineal univariada. Esta regresión puede emplearse como modelo de cuantificación. Para una muestra con una concentración desconocida, se mide la absorbancia A a 1500 nm. A partir de las líneas de regresión se obtiene entonces la concentración correspondiente c del absorbedor:

$$c = bA$$

El coeficiente *b* es constante e idéntico a la pendiente de las líneas de regresión.

Debe tenerse en cuenta que todas las muestras deben contener el mismo absorbedor con el mismo coeficiente de extinción molar. Asimismo, todas las medidas de absorción deben realizarse con idénticos espesores de capa.

Regresión lineal multivariante

Las mezclas reales contienen más de un absorbedor. El espectro registrado es la suma de todos los espectros de los absorbedores *(véase "Ley de Beer-Lambert", capítulo 2.2.1, página 7)*.

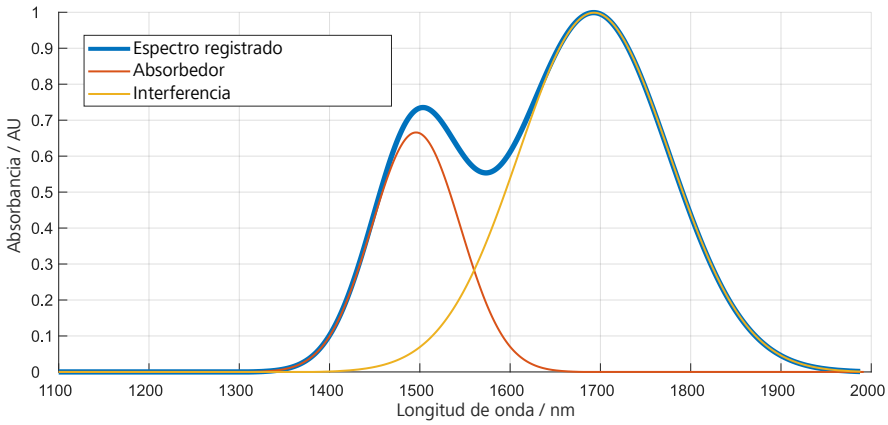


Figura 40 Datos modelados con 2 componentes. El absorbedor (línea roja) debe cuantificarse.

El espectro registrado (línea azul) es la suma del espectro del absorbedor puro y un espectro interferente superpuesto. A 1500 nm, el valor de absorbancia medido se compone no solo de la absorbancia del absorbedor, sino también de la absorbancia del interferente. El parámetro de interés no puede ser cuantificado con una única medida a 1500 nm. También es imposible saber si había un interferente presente y si la medida es fiable.

¿Qué sucede si se miden 2 longitudes de onda, por ejemplo, a 1500 nm y 1700 nm? La absorbancia medida en la longitud de onda 1, A_1, es la suma de la señal del absorbedor pura A_1^a (índice a = absorbedor) y la señal interferente pura A_1^f (índice f = interferente). Lo mismo se aplica a la absorbancia medida en la longitud de onda 2, A_2:

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

Donde $\varepsilon_1{}^{\mathrm{a}}$ y $\varepsilon_1{}^{\mathrm{f}}$ corresponden a los coeficientes de extinción molar en la longitud de onda 1 para el absorbedor o bien el interferente, y c_{a} y c_{f} corresponden a las concentraciones para el absorbedor y el interferente, respectivamente.

En las ecuaciones anteriores, el espesor de capa l se excluye de la ley de Beer-Lambert. Esto simplifica el cálculo algebraico más adelante. Es importante destacar que el espesor de capa debe ser constante en todas las muestras. Por lo tanto, las absorbancias en las ecuaciones son absorbancias por cm.

Las ecuaciones se pueden escribir en forma de matriz de la siguiente manera:

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

Por lo tanto:

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

La solución para la concentración del absorbedor da:

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

Esto significa que la concentración del absorbedor puede calcularse, incluso en presencia de un interferente, midiendo la absorbancia en dos longitudes de onda y multiplicando cada absorbancia por una constante.

Las constantes se refieren a los coeficientes de extinción molar y pueden consultarse en tablas. Sin embargo, esto nunca ocurre en la realidad. En su lugar, se determinan mediante un proceso de calibración y la solución de un sistema de ecuaciones lineales usando técnicas de regresión lineal multivariante como la regresión PLS. Por lo tanto, las constantes se denominan coeficientes de regresión b_1 y b_2:

$$c = b_1 A_1 + b_2 A_2$$

Más de dos absorbedores

Como se ha mostrado anteriormente, con 1 absorbedor es suficiente para determinar la absorbancia en 1 longitud de onda. Con 2 absorbedores, es suficiente para determinar la absorbancia en 2 longitudes de onda.

Esto se puede generalizar. Más absorbedores necesitan más valores de absorbancia A_i en diferentes longitudes de onda i. Se trata aún de una relación lineal:

$$c = b_1A_1 + b_2A_2 + \cdots + b_nA_n$$

Desarrollo de un modelo de cuantificación

Para que la ecuación anterior sea capaz de predecir la concentración en muestras desconocidas, primero se deben determinar b_1, b_2, etc. Para ello se necesita un paso de calibración. Se miden varias muestras con diferentes concentraciones del parámetro de interés.

De acuerdo con la terminología usada más adelante para el PCA y el PLS, *c* se puede sustituir por *y*; asimismo *A* se puede sustituir por *x*. La ecuación anterior se escribe entonces de la siguiente manera para cada muestra de calibración:

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

Donde *n* corresponde al número de muestras; *f* a la cantidad de longitudes de onda; y_1 al valor de referencia para la muestra 1, medido con el método de referencia (por ejemplo, titulación); $x_{1,1}$ a la absorbancia medida de la muestra 1 en la longitud de onda 1, y $\{x_{1,1} \ldots x_{1,f}\}$ al espectro de la muestra 1, medido en *f* longitudes de onda. $b_1 \ldots b_n$ corresponden a los coeficientes de regresión y $e_1 \ldots e_n$ a los términos de error, que muestran hasta qué punto los coeficientes de regresión modelan los datos medidos.

En una forma matricial más compacta, esto da como resultado:

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

X se define como la matriz $f \times n$. $\mathbf{X^t}$ es la matriz **X** transpuesta, es decir, se intercambian las filas y las columnas para obtener la matriz $n \times f$ anterior. El vector de predicción **p** corresponde a los coeficientes de regresión **b** anteriores.

El vector de predicción **p** permite predecir el parámetro de interés de una nueva muestra a partir de su espectro **x**. El valor calculado $\hat{y}$ es:

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

La finalidad de la regresión lineal multivariante es determinar coeficientes de regresión que den como resultado términos de error mínimos. Sin embargo, la regresión lineal múltiple (MLR) necesitaría un mayor número de muestras de calibración que la cantidad de longitudes de onda. Otro obstáculo es la elevada correlación entre las variables.

Para predicciones espectroscópicas, se pueden usar otros métodos. Con el PCA, se puede reducir considerablemente la cantidad de datos y se puede eliminar la correlación por completo. Con una regresión PLS se tienen en cuenta también los valores de referencia de las muestras.

6.2 Algoritmo PCA

El **análisis de componentes principales** (**PCA**, del inglés *principal component analysis*) se usa para la división de conjunto de datos automática y la detección de valores discrepantes espectrales *(véase "Análisis de componentes principales (PCA)", capítulo 4.2, página 37)*.

Para el análisis de componentes principales de los espectros parametrizados y centrados en la media, OMNIS Software realiza una **descomposición en valores singulares** (**SVD**, del inglés *singular value decomposition*). Descompone la matriz de datos original **X** (los espectros del conjunto de calibración) en 3 matrices y calcula n componentes principales:

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

Donde **X** corresponde a los datos espectrales originales (una matriz $f \times n$ con f longitudes de onda y n muestras); **L** a las contribuciones (una matriz $f \times n$); Σ a los valores singulares (una matriz $n \times n$) y **S** a las distribuciones (una matriz $n \times n$). La ecuación se puede representar en forma de gráfico *(véase figura 41, página 95)*.

Las contribuciones **L** proyectan los ejes del espacio de longitudes de onda sobre los ejes del espacio de componentes principales. Los vectores columna de **L** son los ejes principales o las direcciones principales.

Las distribuciones **S** son las proyecciones de los espectros originales sobre los ejes principales. Cada vector fila de **S** contiene un espectro específico.

Σ es una matriz diagonal de los valores singulares σ_i. Los valores singulares describen la varianza explicada por cada componente principal. Por convención, se ordenan de forma que $\sigma_1 > \sigma_2 > \ldots > \sigma_n$.

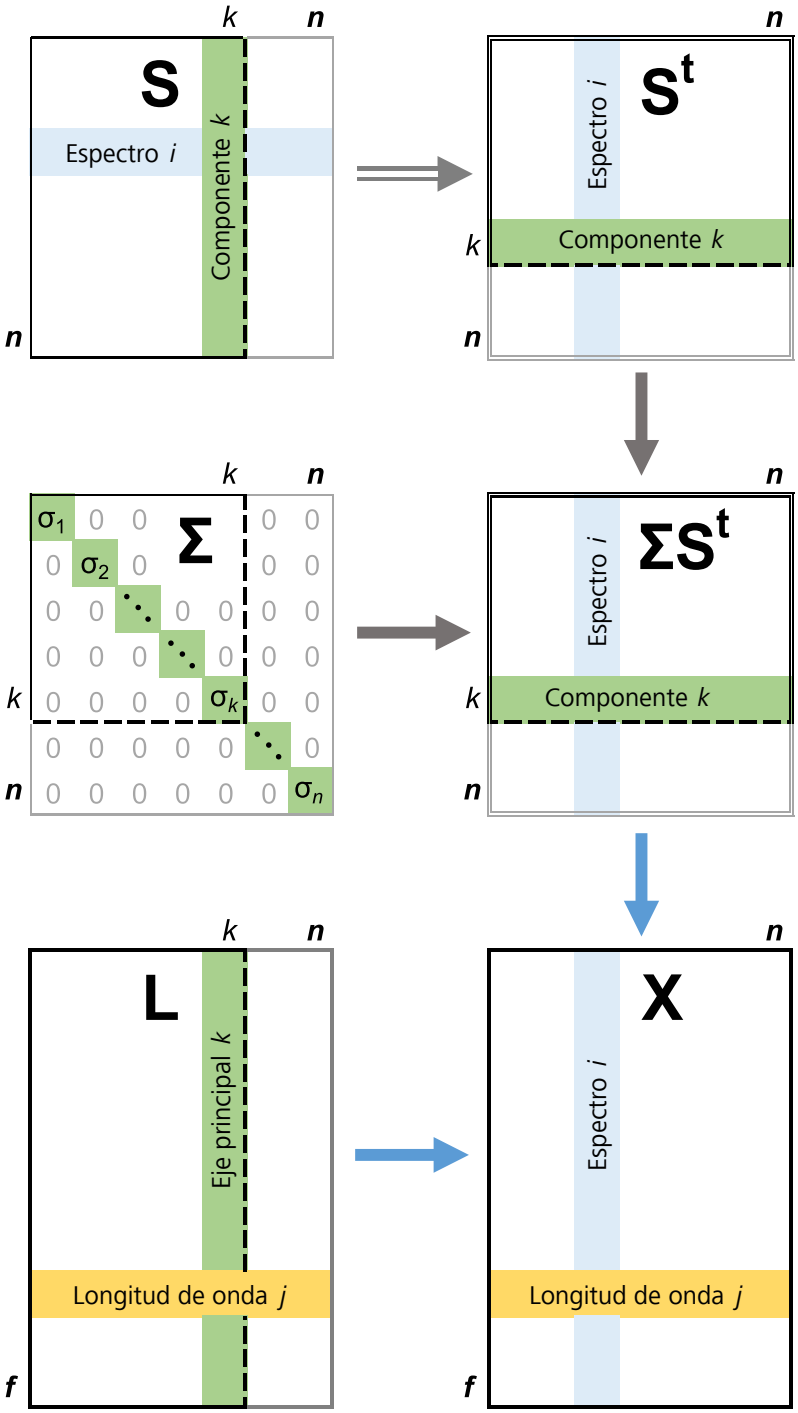


Figura 41 *Ecuación de la descomposición del valor singular en forma gráfica. Las líneas discontinuas muestran las matrices truncadas para un modelo con k componentes principales. La información de los componentes principales omitidos n-k pasa a la matriz residual (una matriz f × n, no representada).*

Matriz residual

Un modelo PCA solo usa algunos de los *n* componentes principales calculados. En la *figura 41*, se trata de los primeros *k* de *n* componentes principales. Los datos originales **X** se pueden dividir en datos descritos por el modelo y datos no descritos por el modelo:

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

Donde **X** corresponde a los datos espectrales originales (una matriz $f \times n$); $\mathbf{L_a}$ (una matriz $f \times k$) a las primeras columnas *k* de **L**; $\mathbf{\Sigma}_a$ (una matriz diagonal $k \times k$) a los primeros *k* valores singulares; $\mathbf{S_a}$ (una matriz $n \times k$ con *k* componentes principales) a las primeras *k* columnas de **S**, y **E** a la matriz residual (una matriz $f \times n$), que contiene todas las variaciones espectrales en **X**, que no pueden ser descritas por el modelo.

Por lo general, es $k \ll n \ll f$. Por ejemplo: $k = 3$ componentes principales, $n = 100$ muestras y $f = 2500$ longitudes de onda.

Cada columna $\mathbf{e}_i$ de la matriz residual **E** muestra la distancia ortogonal del espectro *i* al espacio PCA, llamado **residuo**. Cuantos más componentes principales use el modelo, menor será el residuo.

6.3 Algoritmo PLS

Para calcular el modelo de cuantificación se usa la **regresión de mínimos cuadrados parciales** (**regresión PLS**, del inglés, *partial least squares regression*) *(véase "Regresión PLS", capítulo 4.4.1, página 62)*.

PLS incluye dos bloques de datos:

- La matriz **X** parametrizada y centrada en la media (los espectros).
- El vector **y** centrado en la media (los valores de referencia).

PLS descompone la matriz **X** en 2 matrices:

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

Donde **X** (una matriz $f \times n$ con *f* longitudes de onda y *n* muestras) corresponde a los espectros pretratados y centrados en la media; **L** a las contribuciones (una matriz $f \times k$ con *k* variables latentes); **S** a las distribuciones (una matriz $n \times k$), y **Z** a la matriz residual (una matriz $f \times n$), que contiene todas las variaciones espectrales en **X** que no pueden ser descritas por el modelo.

Mientras que en el PCA la matriz de distribución **S** explica la varianza de **X**, en el PLS la matriz de distribución **S** explica la covarianza entre **X** e **y**. PLS maximiza la covarianza explicada por las distribuciones. Esto significa que las distribuciones no solo explican mejor la varianza de **X**, sino que también tienen la mayor correlación posible con los valores de referencia.

Para maximizar la covarianza entre **X** e **y**, el algoritmo PLS intercambia los datos entre **X** e **y**. Por lo tanto, **X** e **y** se fusionan en un único sistema integrado. Las distribuciones **S** se someten a una regresión frente a los valores de referencia **y** para obtener los coeficientes de regresión **b**:

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

donde **e** es el vector residual que contiene todas las variaciones del valor de referencia en **y** que no pueden ser descritas por el modelo.

Predicción

A partir de los coeficientes de regresión **b** se puede determinar el vector de predicción **p**. Para predecir el parámetro de interés $\hat{y}$ de una nueva muestra se usa el vector de predicción **p** y el espectro pretratado y centrado en la media **x**:

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

OMNIS Software implementa el PLS con el algoritmo SIMPLS y un único conjunto de valores de referencia (PLS-1).

6.4 T² de Hotelling y residuos Q

Los *T*² de Hotelling y los residuos Q caracterizan los espectros en un modelo PCA o PLS. Son especialmente útiles para identificar posibles valores discrepantes *(véase "T² de Hotelling y residuos Q", página 52)*.

T² de Hotelling

La distancia de Mahalanobis es una medida de cuánto se desvía un espectro con respecto al centro del modelo. La distancia se normaliza. Cada componente principal o variable latente recibe el mismo peso.

Si suponemos que los espectros o las distribuciones se distribuyen normalmente, las distancias de Mahalanobis al cuadrado, MD², siguen una distribución *T*² de Hotelling:

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

La distancia de Mahalanobis al cuadrado para el espectro *i* es la suma de cuadrados de las distribuciones normalizadas para los primeros *k* componentes principales o variables latentes:

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

Donde $\mathbf{s_i}$ corresponde a la fila *i*-ésima de la matriz de distribución truncada **S**; $s_{i,a}$ corresponde a la distribución normalizada para el espectro *i* y el componente principal (o variable latente) *a*, y *k* corresponde al número de variables latentes o componentes principales empleados.

MD² se puede denominar simplemente T^2. Un espectro con $T^2 = 0$ se proyecta sobre el centro del modelo centrado en la media, es decir, todas las distribuciones se sitúan en la media. Cuanto más se aleje la proyección de un espectro sobre el hiperplano del punto central, mayor será el valor de T^2. El modelo se puede ajustar mejor cerca del centro. Si se aleja mucho del centro, es posible que el modelo no se ajuste bien.

Se puede definir un **nivel de significancia** de, por ejemplo, el 5% para una prueba de hipótesis que reconozca los valores discrepantes *(véase "Valores discrepantes espectrales – Algoritmo", capítulo 6.5, página 98)*.

Residuo Q

El residuo Q del espectro *i* es la distancia al cuadrado del modelo PCA o PLS con respecto al espectro:

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

Donde $\mathbf{e_i}$ corresponde a la *i*-ésima columna de la matriz residual; **E**, $e_{i,j}$ al residuo del espectro *i* y *f* al número de longitudes de onda.

Los residuos Q muestran las variaciones que el modelo no explica. Un residuo Q alto indica que, posiblemente, el espectro no coincida con el modelo, por ejemplo, si la muestra medida contiene una sustancia diferente.

Se puede definir un **nivel de significancia** de, por ejemplo, el 5% para una prueba de hipótesis que reconozca los valores discrepantes *(véase "Valores discrepantes espectrales – Algoritmo", capítulo 6.5, página 98)*.

6.5 Valores discrepantes espectrales – Algoritmo

La detección de valores discrepantes espectrales identifica los espectros que se desvían de la totalidad de la población estadística *(véase "Detección de valores discrepantes espectrales", página 53)*. El algoritmo evalúa si los valores para T^2 de Hotelling o para el residuo Q del espectro analizado son el resultado de una variación aleatoria o sistemática.

Detección de valores discrepantes espectrales en el desarrollo del modelo

1. La parametrización se tiene en cuenta del siguiente modo:
    a. A partir de la versión de OMNIS Software 4.2: El usuario decide si la parametrización (pretratamiento de datos y selección de longitud de onda) se tiene en cuenta o no.
    Los cambios posteriores en la parametrización no influyen en la división de conjunto de datos.
    b. A partir de la versión de OMNIS Software 3.3 hasta la versión de OMNIS Software 4.1: El usuario decide si el pretratamiento de datos se tiene en cuenta o no. La selección de longitud de onda y los cambios posteriores en el pretratamiento de datos no influyen en la división de conjunto de datos.
    c. Hasta la versión de OMNIS Software 3.2: El pretratamiento de datos se tiene en cuenta tal y como se definió en el momento de la detección de valores discrepantes. La selección de longitud de onda y los cambios posteriores en el pretratamiento de datos no influyen en la división de conjunto de datos.
2. La detección de valores discrepantes espectrales se basa en el modelo PCA de todos los espectros centrados en la media del listado de espectros *(véase "Análisis de componentes principales (PCA)", capítulo 4.2, página 37)*. El espectro que se va a analizar también se registra en el modelo PCA. Se selecciona la cantidad de componentes principales, de modo que la varianza explicada sea al menos del 95%.

3. **Valor discrepante de T² de Hotelling**:
   A partir de los valores de T^2 de Hotelling *(véase "T² de Hotelling", página 97)*, se realiza una prueba de hipótesis según H. Hotelling, *The Generalization of Student's Ratio*, The Annals of Mathematical Statistics vol. 2, N.º 3 (ago. 1931), pp. 360–378.
   a. La hipótesis nula consiste en que el valor de T^2 del espectro que se va a analizar coincide con los valores T^2 del modelo PCA. Si la hipótesis nula es cierta, T^2 sigue una distribución T^2 de Hotelling. La distribución puede expresarse como una distribución F escalada:
   $$T^2 \sim \frac{k(n-1)}{n-k} F_{k,n-k}$$
   Donde k corresponde al número de componentes principales; n al número de espectros y $F_{k,n-k}$ a la distribución F con los parámetros k y $n-k$.
   b. El nivel de significancia ajustable controla la probabilidad de rechazar la hipótesis nula si esta es verdadera (lo que se conoce como error de tipo I). El valor por defecto es 5%.
   c. A partir de esta distribución y del nivel de significancia, se calcula un valor crítico para T^2.
   d. Si el valor de T^2 para el espectro es mayor que el valor crítico, se rechaza la hipótesis nula y el espectro se etiqueta como un posible valor discrepante.
4. **Valor discrepante de residuos Q**
   A partir de los residuos Q *(véase "Residuo Q", página 98)*, se realiza una prueba de hipótesis según J. E. Jackson y G. S. Mudholkar, *Control Procedures for Residuals Associated With Principal Component Analysis*, Technometrics vol. 21, N.° 3 (ago. 1979), pp. 341–349.
   a. La hipótesis nula dice que el residuo Q del espectro que se va a analizar coincide con los valores de los residuos Q del modelo PCA. Si la hipótesis nula es verdadera, se puede generar una estadística de prueba de residuo Q de distribución aproximadamente normal.
   b. El nivel de significancia ajustable controla la probabilidad de rechazar la hipótesis nula si esta es verdadera (lo que se conoce como error de tipo I). El valor por defecto es 5%.
   c. A partir de esta estadística de prueba y del nivel de significancia, se calcula un valor crítico para Q.
   d. Si el valor del residuo Q para el espectro es mayor que el valor crítico, se rechaza la hipótesis nula y el espectro se etiqueta como un posible valor discrepante.

Evaluación de valores discrepantes durante la predicción (cuantificación)

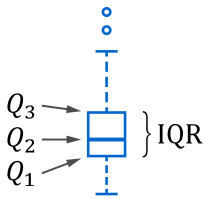
En la cuantificación, hay disponible una evaluación de valores discrepantes durante la predicción:

1. Pasos preparatorios:
   a. El modelo de cuantificación usa el mismo algoritmo mencionado anteriormente. Sin embargo, la base es el modelo de cuantificación (modelo PLS) con todos los espectros centrados en la media del conjunto de calibración. El pretratamiento de datos, las gamas de longitudes de onda y el número de variables latentes se tienen en cuenta tal como se especifica en el modelo de cuantificación.
   b. A partir del nivel de significancia especificado, el modelo de cuantificación calcula los valores críticos para T² y Q (línea discontinua en el gráfico de influencia PLS). Los valores críticos se guardan en el modelo de cuantificación.
2. Durante la predicción, los valores de T^2 y Q del espectro se calculan sobre la base del modelo de cuantificación.
3. Si el valor T^2 o el valor Q del espectro es mayor al valor crítico respectivo, la muestra se considera un potencial valor discrepante en relación con el modelo de cuantificación usado.

6.6 Valor de referencia discrepante – Algoritmo

Los diagramas de caja permiten reconocer valores discrepantes en los valores de referencia *(véase "Valor de referencia discrepante (para cuantificación)", capítulo 4.3.4, página 59)*.

Para tener en cuenta la asimetría de la distribución, los límites de los valores discrepantes se ajustan con los siguientes cálculos.



El **Medcouple** (MC) mide la asimetría de los valores de referencia. El cálculo comienza con la mediana del diagrama de caja, Q_2. Con todos los posibles pares (en inglés, *couples*) de la mitad superior e inferior de los valores de referencia, se calcula una función. La mediana de los resultados es el Medcouple:

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

Donde Q_2 corresponde al segundo cuartil, que define la línea central en el diagrama de caja e y_i, y_j a un par de valores de referencia.

El Medcouple siempre se encuentra entre –1 y 1. Para una distribución simétrica, MC = 0. Una distribución asimétrica con MC > 0 está sesgada hacia los valores de referencia más altos, mientras que con MC < 0 está sesgada hacia los valores de referencia más bajos.

El cálculo de los **límites ajustados para los valores discrepantes** depende de hacia qué lado esté desplazada la distribución:

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

En una distribución simétrica (MC = 0), las distancias entre los valores límite y la caja son de 1,5 IQR.

La función exponencial permite una detección precisa y robusta de valores discrepantes en diversas distribuciones con diferentes grados de asimetría, como lo demostraron empíricamente M. Hubert y E. Vandervieren en *An adjusted boxplot for skewed distributions*, Computational Statistics & Data Analysis, vol. 52, N.° 12 (agosto de 2008), pp. 5186–5201.

El porcentaje esperado de valores discrepantes marcados es de aproximadamente el 1% y es bastante similar al porcentaje del diagrama de caja estándar para la distribución normal. Nota: Este porcentaje es independiente del nivel de significancia.