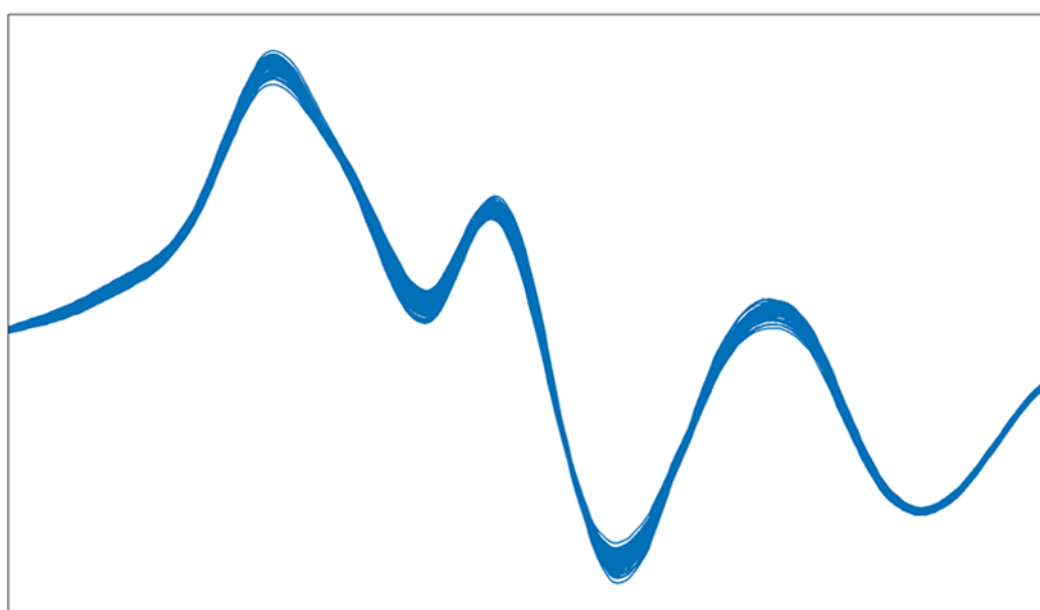


# Teoría de OMNIS NIR



Manual

8.0600.8101ES / v10 / 2026-04-08





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 / v10 /  
2026-04-08

Esta documentación está protegida con derechos de autor. Todos los derechos reservados.

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

|          |                                                                             |           |
|----------|-----------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Información general</b>                                                  | <b>1</b>  |
| 1.1      | Introducción .....                                                          | 1         |
| 1.2      | Marco conceptual .....                                                      | 1         |
| 1.3      | Acerca de la documentación .....                                            | 2         |
| 1.4      | Información adicional .....                                                 | 2         |
| <b>2</b> | <b>Luz de infrarrojo cercano y espectros</b>                                | <b>3</b>  |
| 2.1      | La luz y su interacción con la materia .....                                | 3         |
| 2.2      | Fundamentos matemáticos .....                                               | 7         |
| 2.2.1    | Ley de Beer-Lambert .....                                                   | 7         |
| 2.2.2    | Regresión lineal .....                                                      | 8         |
| 2.3      | Cómo se convierte la luz en un espectro .....                               | 9         |
| <b>3</b> | <b>Configuración del aparato</b>                                            | <b>13</b> |
| 3.1      | Calibración de las longitudes de onda .....                                 | 14        |
| 3.2      | Normalización de referencia .....                                           | 15        |
| 3.2.1    | OMNIS NIR Analyzer .....                                                    | 16        |
| 3.2.2    | 2060 NIR .....                                                              | 17        |
| 3.3      | Pruebas de rendimiento del aparato .....                                    | 26        |
| 3.3.1    | Pruebas de rendimiento del aparato ( OMNIS NIR Analyzer )<br>externas ..... | 30        |
| <b>4</b> | <b>Desarrollo del modelo</b>                                                | <b>32</b> |
| 4.1      | Muestras físicas .....                                                      | 34        |
| 4.2      | Análisis de componentes principales (PCA) .....                             | 37        |
| 4.3      | Preparación de datos .....                                                  | 41        |
| 4.3.1    | Pretratamiento de datos .....                                               | 41        |
| 4.3.2    | Gama de longitudes de onda .....                                            | 50        |
| 4.3.3    | Valores discrepantes espectrales .....                                      | 52        |
| 4.3.4    | Valor de referencia discrepante (para cuantificación) .....                 | 59        |
| 4.3.5    | División de conjunto de datos .....                                         | 60        |
| 4.4      | Cuantificación .....                                                        | 62        |
| 4.4.1    | Regresión PLS .....                                                         | 62        |
| 4.4.2    | Validación de los modelos de cuantificación .....                           | 65        |
| 4.4.3    | OMNIS Model Developer (OMD) .....                                           | 73        |
| 4.4.4    | Corrección de pendiente/del sector de eje de coordena-<br>das 'y' .....     | 75        |
| 4.5      | Identificación y verificación .....                                         | 77        |
| 4.5.1    | Máquina de soporte vectorial (SVM) .....                                    | 77        |

|            |                                                           |            |
|------------|-----------------------------------------------------------|------------|
| 4.5.2      | Predicción de la clase de producto de una muestra .....   | 80         |
| 4.5.3      | Validación de modelos de identificación .....             | 82         |
| <b>4.6</b> | <b>Cualificación .....</b>                                | <b>84</b>  |
| 4.6.1      | Cálculo de modelos de cualificación .....                 | 84         |
| 4.6.2      | Validación de modelos de cualificación .....              | 84         |
| <b>5</b>   | <b>Predicción .....</b>                                   | <b>86</b>  |
| <b>5.1</b> | <b>Cuantificación .....</b>                               | <b>86</b>  |
| 5.1.1      | Valor discrepante y monitorización del resultado .....    | 86         |
| <b>5.2</b> | <b>Identificación y verificación .....</b>                | <b>88</b>  |
| <b>5.3</b> | <b>Cualificación .....</b>                                | <b>89</b>  |
| <b>6</b>   | <b>Apéndice .....</b>                                     | <b>90</b>  |
| <b>6.1</b> | <b>Ejemplo de una regresión lineal .....</b>              | <b>90</b>  |
| <b>6.2</b> | <b>Algoritmo PCA .....</b>                                | <b>94</b>  |
| <b>6.3</b> | <b>Algoritmo PLS .....</b>                                | <b>96</b>  |
| <b>6.4</b> | <b>T<sup>2</sup> de Hotelling y residuos Q .....</b>      | <b>97</b>  |
| <b>6.5</b> | <b>Valores discrepantes espectrales – Algoritmo .....</b> | <b>98</b>  |
| <b>6.6</b> | <b>Valor de referencia discrepante – Algoritmo .....</b>  | <b>101</b> |

# 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  $\mathbf{x}^t$ .

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,  $\bar{Y}$ .

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

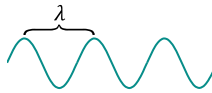
En las siguientes páginas se encuentra información adicional sobre el producto:

- 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  $\lambda$  (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  $\lambda$ :

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

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

La [figura 1](#) muestra diferentes gamas de radiación electromagnética.

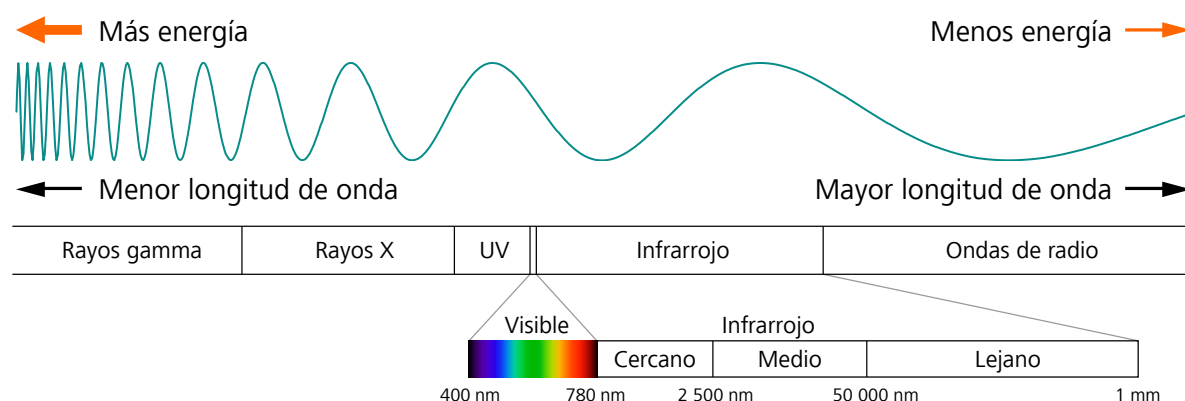


Figura 1 Gamas de radiación electromagnética. La región del infrarrojo cercano (NIR) se encuentra adyacente al rango de la luz visible. La NIR abarca longitudes de onda de 780 nm a 2500 nm.

### 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<br>$i \rightarrow j$ | Nombre                 |
|---------------------------------------------|------------------------|
| $0 \rightarrow 1$                           | Transición fundamental |
| $0 \rightarrow 2$                           | Primer sobretono       |
| $0 \rightarrow 3$                           | 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  $\Delta E_{ij}$ .

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  $\Delta E_{ij}$ .

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.

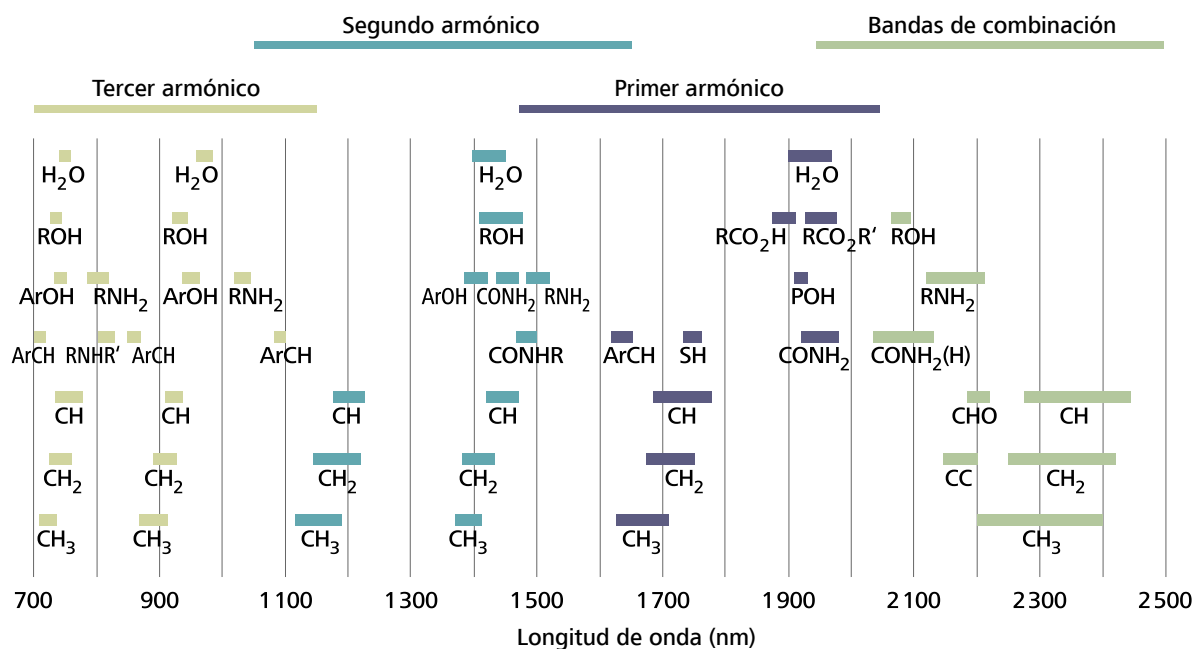


Figura 2 Bandas de absorción del NIR

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,  $\varepsilon$  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  $\varepsilon$  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;  $\varepsilon_{ij}$  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.



## 2.3 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.

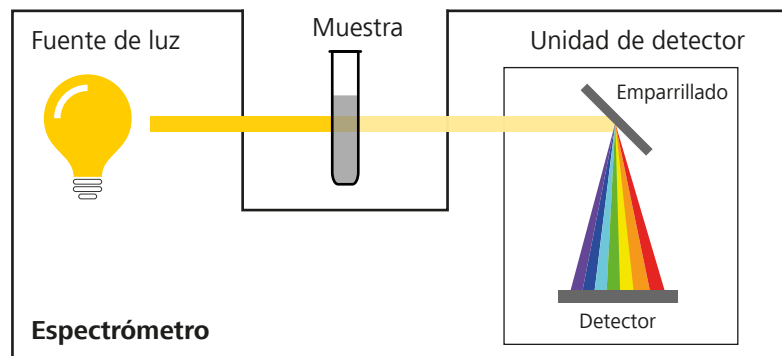


Figura 4 Un espectrómetro con fuente de luz y unidad de detección.

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.

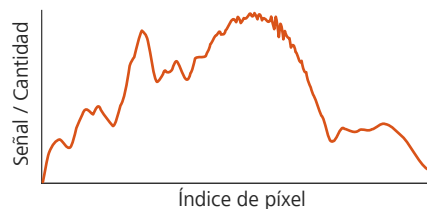


Figura 5 Espectro de la señal del detector como función de los píxeles.

### 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

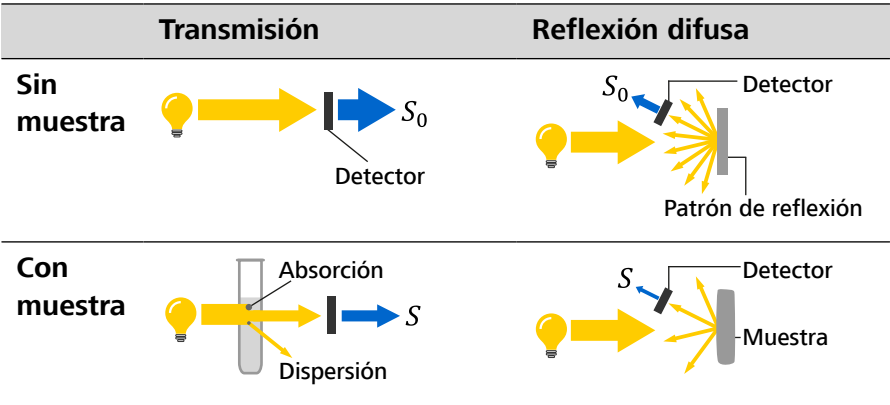


$$A = \log_{10} \frac{I_0}{I} = \log_{10} \frac{S_0}{S}$$

**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.



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.

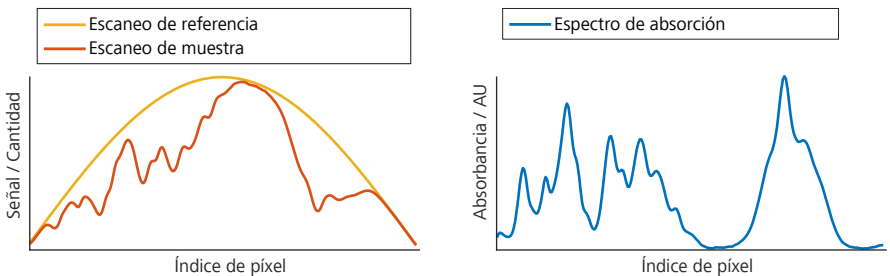


Figura 6 Escaneo de referencia y escaneo de muestra (a la izquierda), así como el espectro de absorción calculado (a la derecha) como función del índice de píxeles.



### 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 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 NIR:** patrón de reflexión externo



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):

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

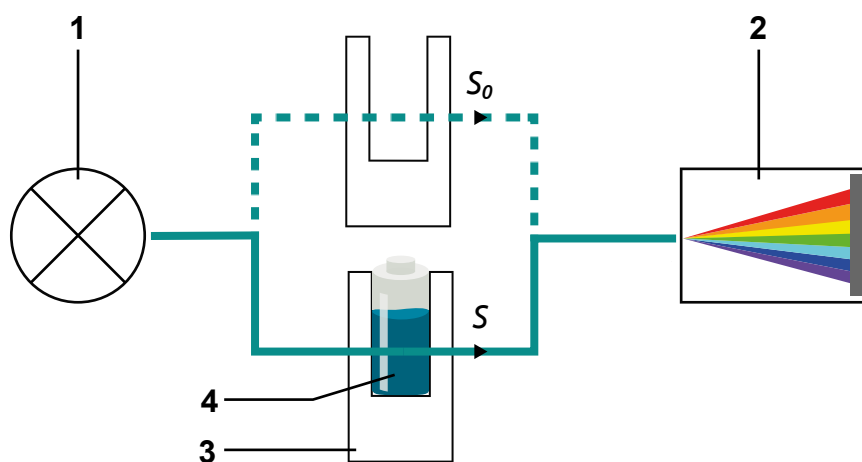


Figura 8 Camino óptico en el modo de transmisión (por ejemplo, con una presentación de muestras líquidas).

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



### 3.2.2 2060 NIR

Los aparatos del tipo **2060 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  $S_{\text{ref}}$ , 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  $S_{\text{fiber}}$ .

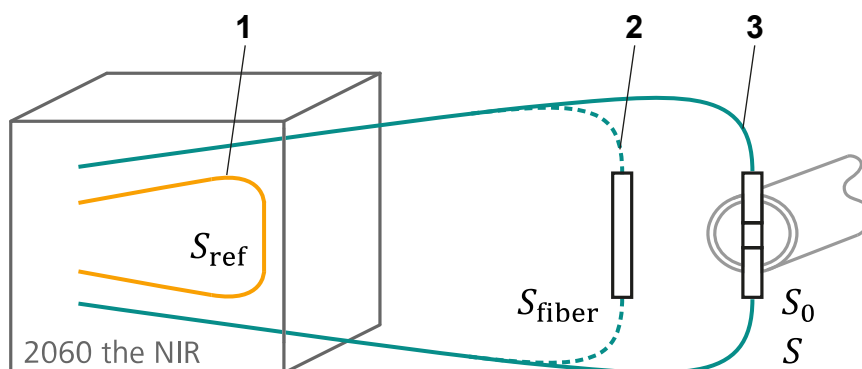


Figura 9 Caminos ópticos como ejemplo en el modo de transmisión: camino de referencia interna (1), fibra óptica externa conectada a un dispositivo de calibración (2) y fibra óptica externa conectada a la sonda con o sin muestra (3). Los caminos ópticos 2 y 3 representan la misma fibra óptica, que solo 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  $S_{\text{ref}}$  y  $S_{\text{fiber}}$  al numerador y al denominador, el resultado no se modifica:

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







### 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  $A_{\text{nominal}}$ .
- En el modo de transmisión, se usa el aire como referencia. El espectro de absorción nominal es una línea cero ( $A_{\text{nominal}} = 0$ ), 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  $S_{\text{ref}}$ .
  - 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  $S_{\text{raw}}$ .
3. El software calcula  $A_{\text{raw}}$  (**Espectro primario medido**):
 
$$A_{\text{raw}} = \log_{10} \frac{S_{\text{ref}}}{S_{\text{raw}}}$$
4.  $A_{\text{raw}}$  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:
 
$$A_{\text{corrected}} = A_{\text{raw}} - A_{\text{fiber}}$$
 $A_{\text{corrected}}$  se muestra en el software como **Espectro corregido medido**.
5. Idealmente,  $A_{\text{corrected}}$  debería coincidir con **Espectro de referencia**  $A_{\text{nominal}}$ . Las diferencias entre ambos se calculan como **Residuos de la validación**:

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

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



1. Para obtener un valor actual de  $A_{\text{fiber}}$ , 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  $S_{\text{ref}}$ .
  - 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  $S_{\text{probe}}$ .
4. La absorbancia  $A_{\text{probe}}$  con respecto al camino óptico interno es:
 
$$A_{\text{probe}} = \log_{10} \frac{S_{\text{ref}}}{S_{\text{probe}}}$$
5. El software calcula  $A_{\text{raw}}$  (**Espectro primario medido**):
 
$$A_{\text{raw}} = A_{\text{probe}} - A_{\text{fiber}}$$

Aquí,  $A_{\text{raw}}$  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,  $A_{\text{nominal}}$ , se muestra en el software como **Espectro de referencia**. El espectro de referencia se debe restar de  $A_{\text{raw}}$  para obtener  $A_{\text{window}}$ :
 
$$A_{\text{window}} = A_{\text{raw}} - A_{\text{nominal}}$$

Aquí,  $A_{\text{window}}$  corresponde a la absorbancia de la sonda, con respecto al dispositivo de calibración.

Nota: En el modo de transmisión, es  $A_{\text{nominal}} = 0$  y  $A_{\text{window}} = A_{\text{raw}}$ .  
 $A_{\text{window}}$  representa el **Espectro de corrección** de la ventana.
7.  $A_{\text{window}}$  permanece sin cambios hasta que se realiza una nueva corrección de la ventana para el canal correspondiente.

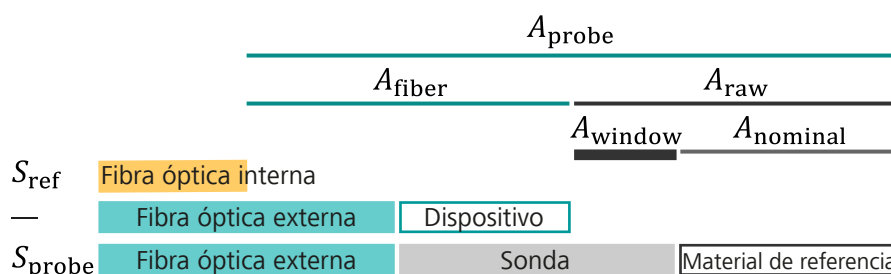


Figura 13 Corrección de la ventana

### 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.



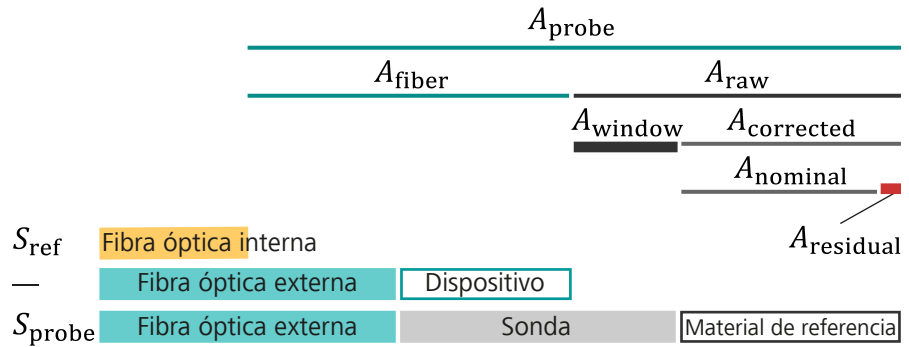


Figura 14 Residuos para la validación de la corrección de la ventana

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.

### Adquirir el espectro de una muestra

**i** 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:

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



- **2060 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 varias veces a través del camino óptico interno:

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

En este caso,  $A_{WL,x}$  corresponde al espectro de absorción del patrón de longitud de onda interno para la x-ésima medida,  $S_{\text{ref}}$  corresponde al espectro de referencia medido en el camino de referencia interno y  $S_{\text{ref},WL,x}$  corresponde al espectro de la x-ésima medida, medido en el camino de referencia interno con patrón de longitud de onda interno colocado.

- **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*).



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 varias veces mediante la absorbancia del camino óptico interno, en relación con la absorbancia de la medida de referencia en el mismo camino óptico:

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

En este caso,  $A_{\text{noise},x}$  corresponde al espectro de ruido de la x-ésima medida,  $S_{\text{ref}}$  corresponde al espectro de referencia medido en el camino de referencia interno y  $S_{\text{ref},x}$  corresponde al espectro de la x-ésima medida, medido en el mismo camino de referencia.

- **Externa** para aparatos de la familia de productos **OMNIS NIR Analyzer**: La prueba de ruido externa es opcional (*véase "Pruebas de ruido externas", página 31*).

- **Externa** para aparatos del tipo **2060 NIR**:

Las fibras ópticas externas se deben conectar al dispositivo de calibración en caso de transmisión y al patrón de reflexión en caso de reflexión.

El ruido se determina varias veces como la diferencia entre el espectro de absorción medido y el espectro de absorción nominal:

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

En este caso,  $A_{\text{noise},x}$  corresponde al espectro de ruido de la x-ésima medida,  $S_{\text{ref}}$  corresponde al espectro de referencia medido en el camino de referencia interno,  $S_{\text{fiber},x}$  corresponde al espectro de la x-ésima medida, medido en el camino óptico del respectivo canal, a cuyo efecto las fibras están conectadas al dispositivo de calibración o al patrón de reflexión,  $A_{\text{fiber}}$  corresponde al espectro de corrección de fibra de vidrio de la normalización de referencia y  $A_{\text{nominal}}$  corresponde al espectro de absorción nominal del patrón de reflexión.

Nota: En el modo de transmisión,  $A_{\text{nominal}}$  es = 0.

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

En la prueba de ruido se ejecutan los siguientes pasos:

1. El espectro  $S_{\text{ref},x}$  o  $S_{\text{fiber},x}$  se mide 10 veces, con lo cual se obtienen 10 espectros de ruido  $A_{\text{noise},x}$ .
2. Los espectros de ruido se dividen en diferentes segmentos de longitud de onda.



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 registran 10 espectros de ruido como la diferencia entre los espectros de absorción medidos y el espectro de absorción nominal del patrón de referencia:

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

En este caso,  $A_{\text{noise},x}$  corresponde al espectro de referencia de la  $x$ -ésima medida,  $S_{\text{ref}}$  corresponde al espectro de referencia medido en el camino de referencia interno,  $S_{\text{ND},x}$  corresponde al espectro de la  $x$ -ésima medida, medido mediante el patrón de referencia externo y  $A_{\text{nominal}}$  corresponde al espectro nominal del patrón de referencia.

3. Los 10 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. El estado general de la prueba es correcto si las pendientes y los sectores de eje de coordenadas 'y' de todas las rectas de regresión están dentro de las tolerancias predefinidas.



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.



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

### Número de muestras

Cuanto 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



## 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.

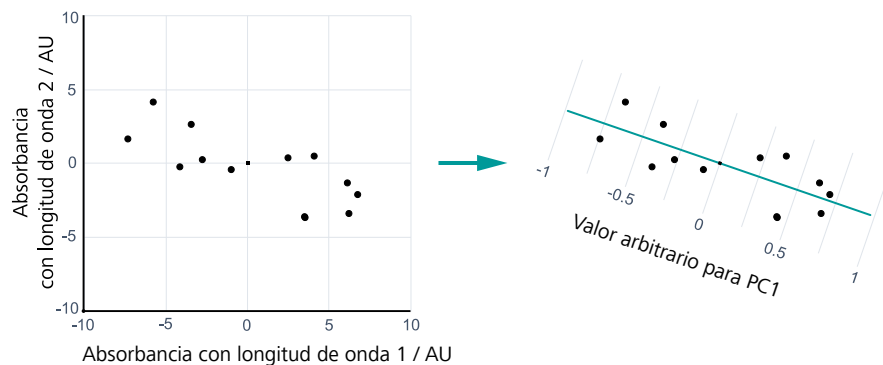


Figura 17 Puntos que representan espectros en un espacio bidimensional de longitudes de onda (a la izquierda). Los mismos puntos en un espacio unidimensional de componentes principales (a la derecha).

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  $i$ .

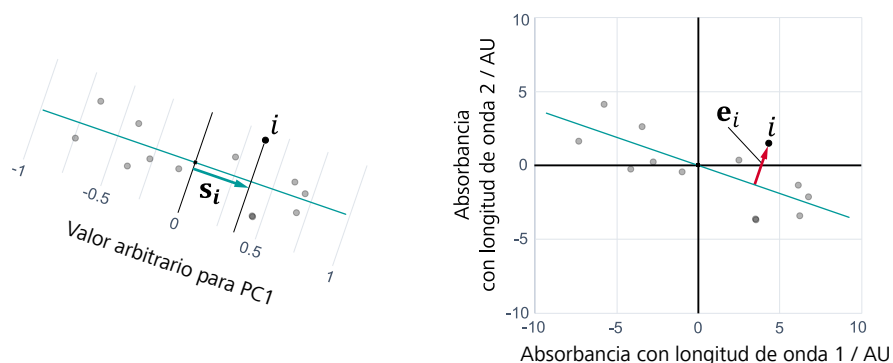


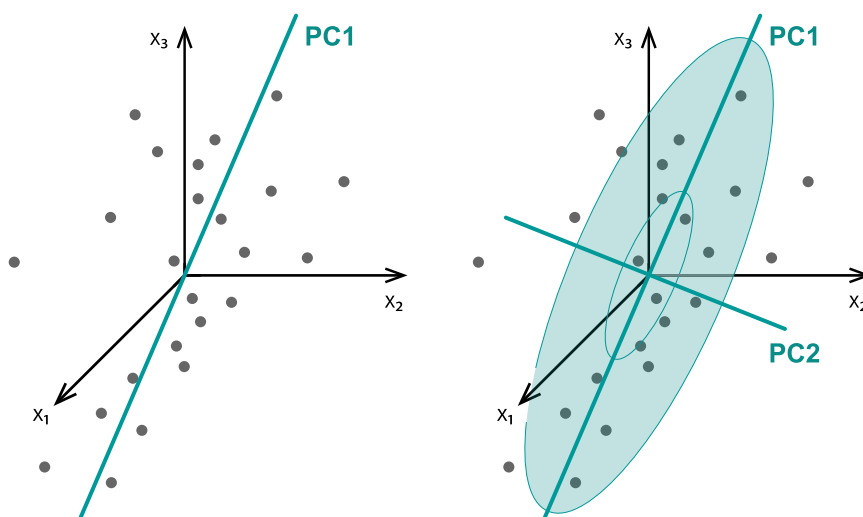
Figura 18 Espectro  $i$  con distribución (a la izquierda) y residuo (a la derecha).

**i** 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.



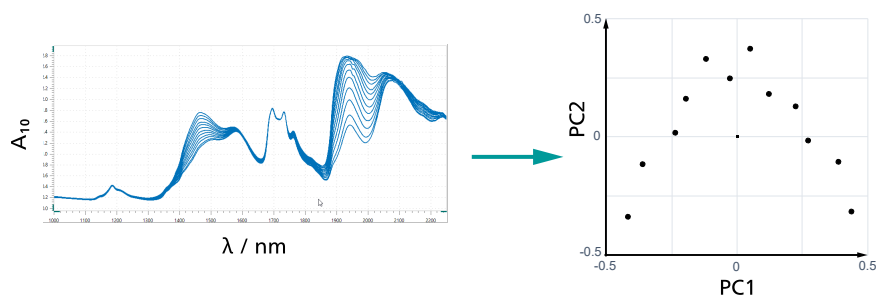


Figura 21 Conversión de datos espectrales en un espacio de componentes principales. Las distribuciones en el lado derecho se expresan en unidades arbitrarias.

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.

**i** 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.



- 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 ampliaciones 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

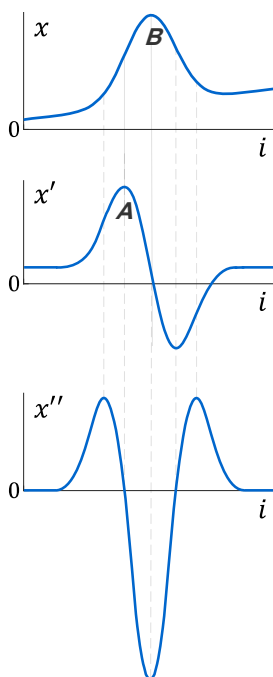
**i** Todos los siguientes pretratamientos de datos se aplican a un único espectro. No se incluye ningún otro espectro en los cálculos.

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

**i** 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.



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).

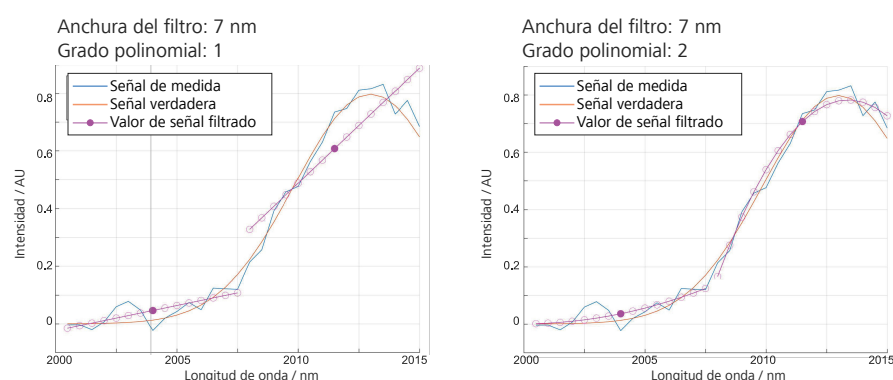


Figura 23 Filtrado de Savitzky-Golay con diferentes grados polinomiales

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.

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



## 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.

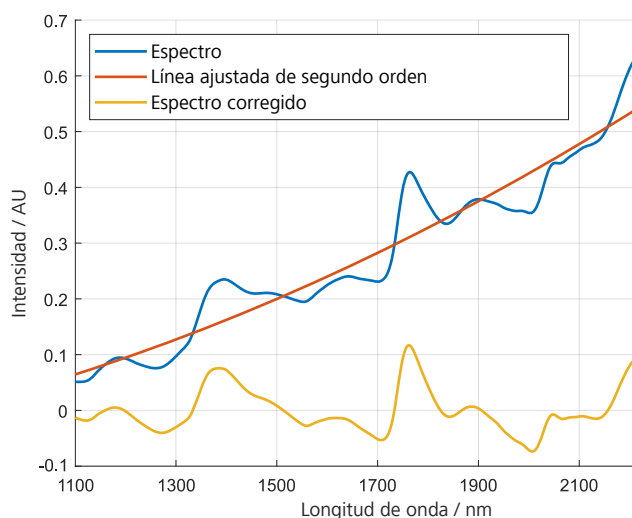


Figura 25 Detrend convierte el espectro azul en el espectro amarillo.

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



| Pretratamiento                       | Objeto                                                                                                                                    | Efectos positivos                                                                                                                                                                                                                             | Efectos negativos                                                                                                                                     |
|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Savitzky-Golay</b>                | Suavizado<br>Se obtiene un suavizado mayor con una derivada de menor orden, una mayor anchura de filtro o un polinomio de grado inferior. | <ul style="list-style-type: none"> <li>El suavizado reduce el ruido de alta frecuencia.</li> </ul>                                                                                                                                            | <ul style="list-style-type: none"> <li>Un suavizado excesivo conduce a una pérdida de varianza relevante.</li> </ul>                                  |
| Derivadas en Savitzky-Golay          | Corrección de línea base                                                                                                                  | <ul style="list-style-type: none"> <li>Derivada de primer orden: elimina los desplazamientos de la línea base.</li> <li>Derivada de segundo orden: elimina los desplazamientos de la línea base y las pendientes de la línea base.</li> </ul> | <ul style="list-style-type: none"> <li>Amplifica el ruido.</li> <li>Modifica el aspecto del espectro.</li> </ul>                                      |
| <b>SNV</b> – Standard Normal Variate | Corrección de la dispersión *                                                                                                             | <ul style="list-style-type: none"> <li>Elimina los desplazamientos de la línea base.</li> </ul>                                                                                                                                               | <ul style="list-style-type: none"> <li>Se elimina la varianza relevante, dado el caso.</li> </ul>                                                     |
| <b>Detrend</b>                       | Corrección de la dispersión *                                                                                                             | <ul style="list-style-type: none"> <li>Elimina los desplazamientos de la línea base.</li> <li>Elimina las pendientes de la línea base y las pendientes cuadráticas de la línea base.</li> </ul>                                               | <ul style="list-style-type: none"> <li>Se elimina la varianza relevante, dado el caso.</li> <li>Puede producirse una varianza irrelevante.</li> </ul> |

\* 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



### ■ 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 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*).

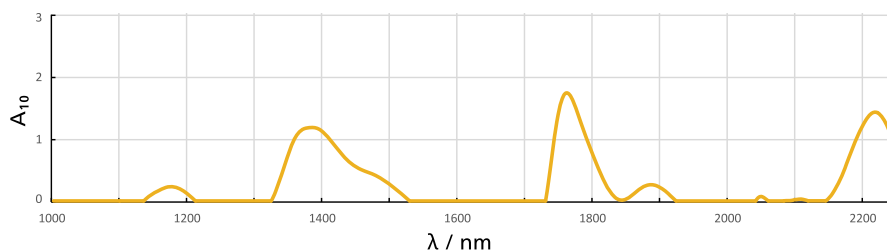


Figura 27 Ejemplo con gamas de longitudes de onda saturadas.

### 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.

**i** 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*).



- 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*).



- 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).

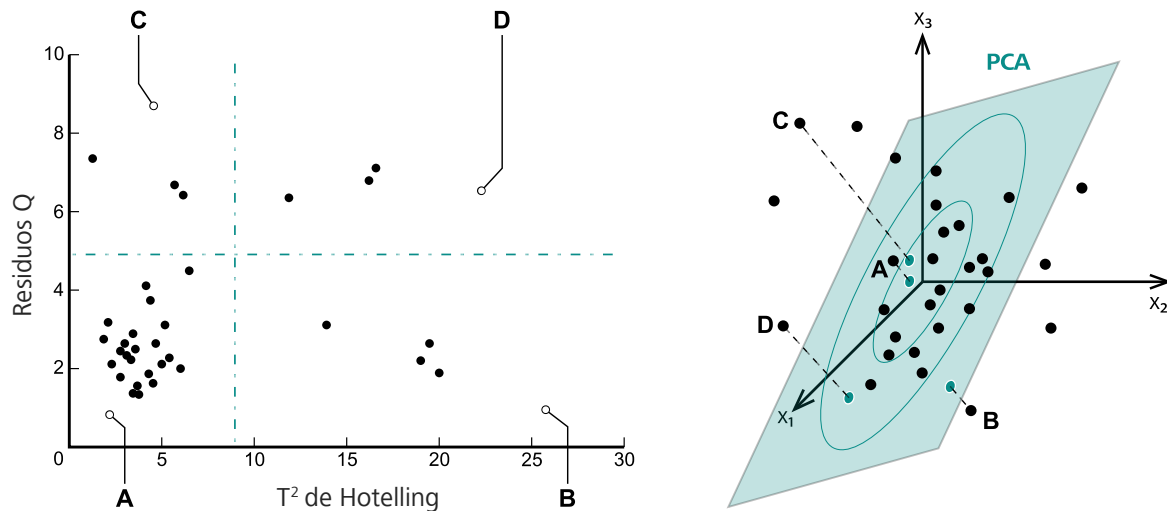


Figura 28 Gráfico de influencia (a la izquierda), espacio original y espacio de variables latentes (a la derecha). Cada espectro está representado por un punto en la figura a la izquierda y un punto en la figura a la derecha.

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).







**i** 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.

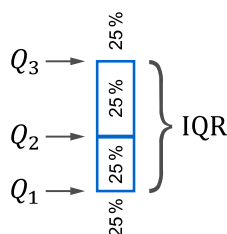
#### 4.3.4 Valor de referencia discrepante (para cuantificación)

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:

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

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









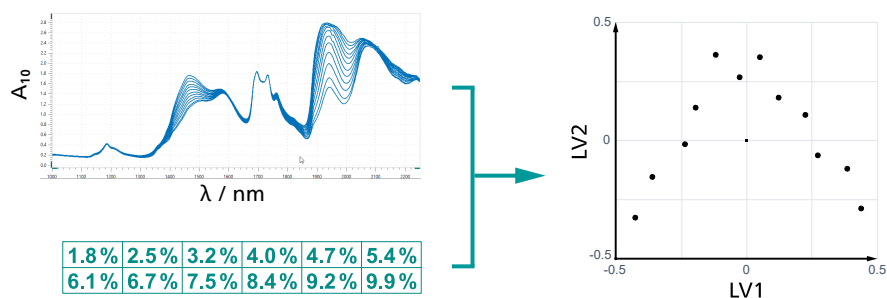


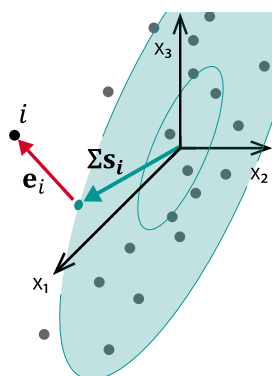
Figura 30 Conversión de espectros y valores de referencia en un espacio de variables latentes. Las distribuciones en el lado derecho se expresan en unidades arbitrarias.

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  $\Sigma s_i$ , que representa la distancia euclídea desde el punto central. La **distancia de Mahalanobis**  $s_i$  es el vector de distribución de variables normalizadas que da a cada dirección la misma ponderación.
- **Residuos:** el vector residual  $e_i$  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-



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.



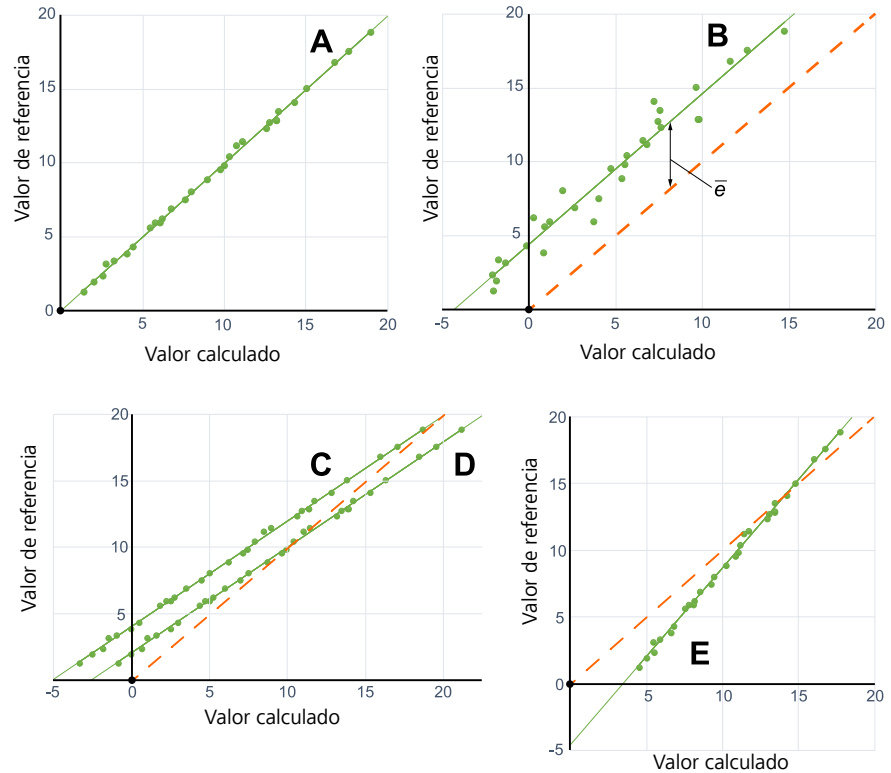


Figura 32 Diagramas de correlación. Cada punto representa una muestra e indica su valor de referencia, así como su valor calculado. La línea discontinua muestra la línea recta ideal de 45°.

### 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.



| Línea<br>recta | Errores sistemáticos |           |                                          | Errores<br>aleatorios |
|----------------|----------------------|-----------|------------------------------------------|-----------------------|
|                | Desvia-<br>ción      | Pendiente | Sector de<br>eje de coor-<br>denadas '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<sup>2</sup> - Coeficiente de determinación**

El **coeficiente de determinación R<sup>2</sup>** (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:

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

Donde  $SS_{\text{reg}}$  corresponde a la suma de cuadrados de la regresión (varianza de los valores calculados, es decir, varianza explicada);  $SS_{\text{tot}}$  a la suma total de cuadrados (varianza del valor de referencia);  $SS_{\text{res}}$  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.



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

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:

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

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):

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

Donde  $\mathbf{e}$  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  $\mathbf{e}$ .

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.



**i** 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*).



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.

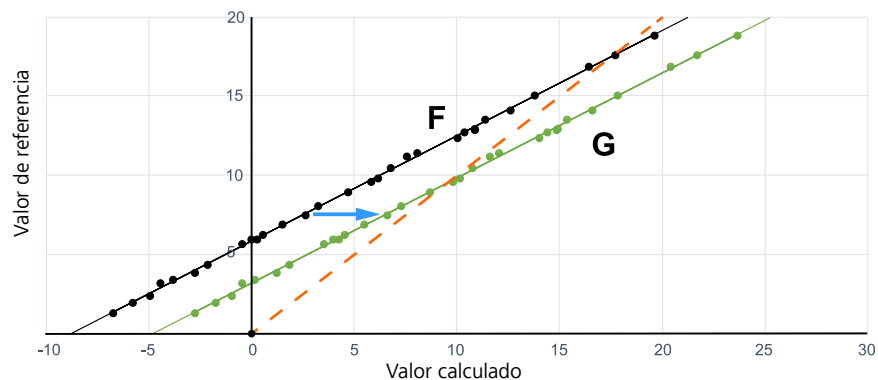


Figura 34 Corrección de desviación

### 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**).

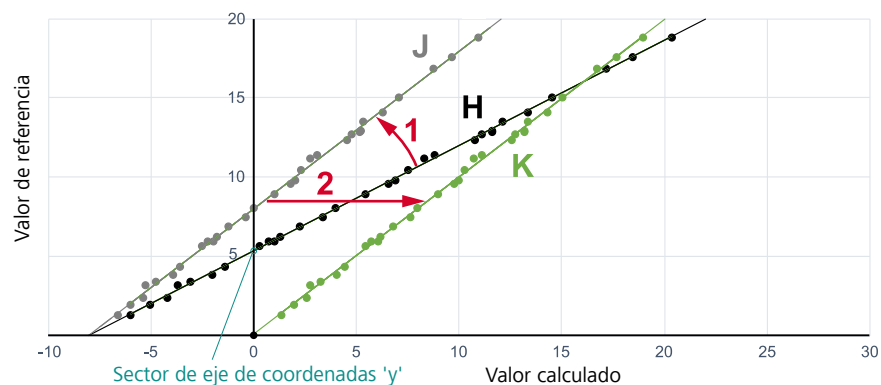


Figura 35 Corrección de pendiente/del sector de eje de coordenadas 'y'

### 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                                                 | $SEP = \sqrt{\frac{\sum_{i=1}^n (v_i - \hat{v}_i)^2}{n}}$     |
| Corrección de desviación                                     | $SEP = \sqrt{\frac{\sum_{i=1}^n (v_i - \hat{v}_i)^2}{n - 1}}$ |
| Corrección de pendiente/del sector de eje de coordenadas 'y' | $SEP = \sqrt{\frac{\sum_{i=1}^n (v_i - \hat{v}_i)^2}{n - 2}}$ |

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.

**i** 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.

**i** 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.

**i** 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).



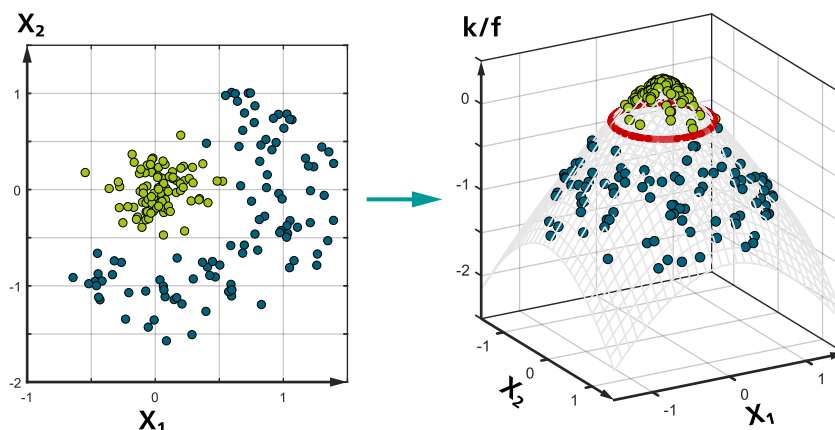


Figura 37 Productos no linealmente separables (a la izquierda). Una dimensión adicional  $k/f$  (característica kernel) facilita la separación (a la derecha). Los valores se proporcionan en unidades arbitrarias.

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. 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% | 90%                    | –                          | No identificado                |
| Producto B: 71% | 80%                    | Producto A: correcta → 87% | Producto A                     |
| Producto C: 68% | 70%                    | Producto A: correcta → 87% | Producto A                     |
| Producto D: 30% | 60%                    | Producto B: fallida → 0%   |                                |
|                 |                        | Producto A: correcta → 87% | Ambiguo                        |
|                 |                        | Producto B: fallida → 0%   |                                |
|                 |                        | Producto C: correcta → 68% |                                |

### 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**).



## 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 subrepresentadas, se debería ampliar el conjunto de calibración según corresponda.



**i** 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.

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

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 subrepresentadas, se debería ampliar el conjunto de calibración según corresponda.



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).

**i** 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  $u$ :

$$D = \sqrt{(\mathbf{s}_i - \mathbf{s}_u)^t (\mathbf{s}_i - \mathbf{s}_u)}$$

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.



## 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



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*).

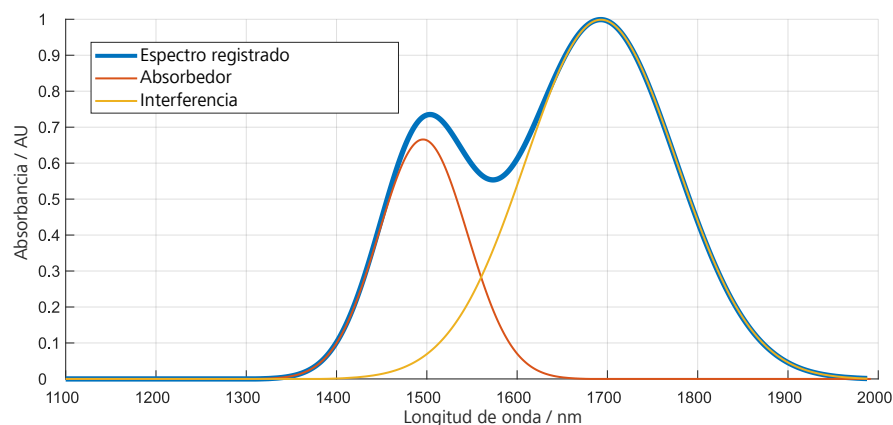


Figura 40 Datos modelados con 2 componentes. El absorbedor (línea roja) debe cuantificarse.

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,  $A_2$ :



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

### 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  $\{x_{1,1} \dots x_{1,f}\}$  al espectro de la muestra 1, medido en  $f$  longitudes de onda.  $b_1 \dots b_n$  corresponden a los coeficientes de regresión y  $e_1 \dots e_n$  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:

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

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

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

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

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.



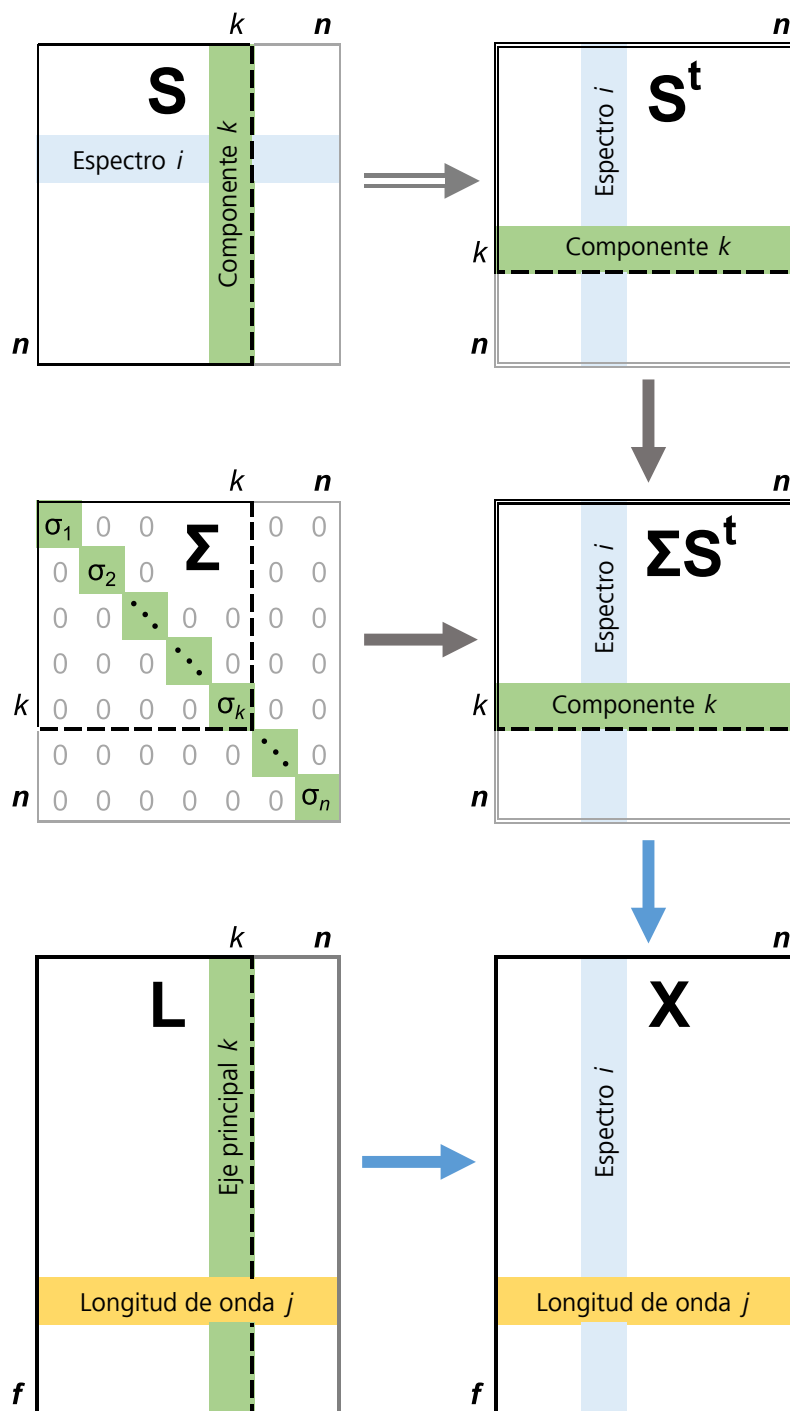


Figura 41 Ecuación de la descomposición del valor singular en forma gráfica. Las líneas discontinuas muestran las matrices truncadas para un modelo con  $k$  componentes principales. La información de los componentes principales omitidos  $n-k$  pasa a la matriz residual (una matriz  $f \times n$ , no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



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

$$\mathbf{y} = \mathbf{S}\mathbf{b} + \mathbf{e}$$

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

### Predicción

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

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

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

## 6.4 T<sup>2</sup> 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 "T<sup>2</sup> de Hotelling y residuos Q", página 52*).

### T<sup>2</sup> 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,  $MD^2$ , siguen una distribución  $T^2$  de Hotelling:

$$MD^2 \sim T^2$$

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:

$$MD_i^2 = \mathbf{s}_i \mathbf{s}_i^t = \sum_{a=1}^k s_{i,a}^2$$

Donde  $\mathbf{s}_i$  corresponde a la fila  $i$ -ésima de la matriz de distribución truncada  $\mathbf{S}$ ;  $s_{i,a}$  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.



### **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%.

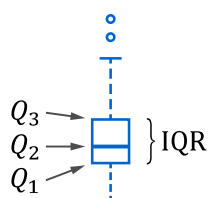


1. Pasos preparatorios:
  - a. El modelo de cuantificación usa el mismo algoritmo mencionado anteriormente. Sin embargo, la base es el modelo de cuantificación (modelo PLS) con todos los espectros centrados en la media del conjunto de calibración. El pretratamiento de datos, las gamas de longitudes de 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  $T^2$  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:

$$MC = \text{med}_{y_i \leq Q_2 \leq y_j} \frac{(y_j - Q_2) - (Q_2 - y_i)}{y_j - y_i}$$

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:

