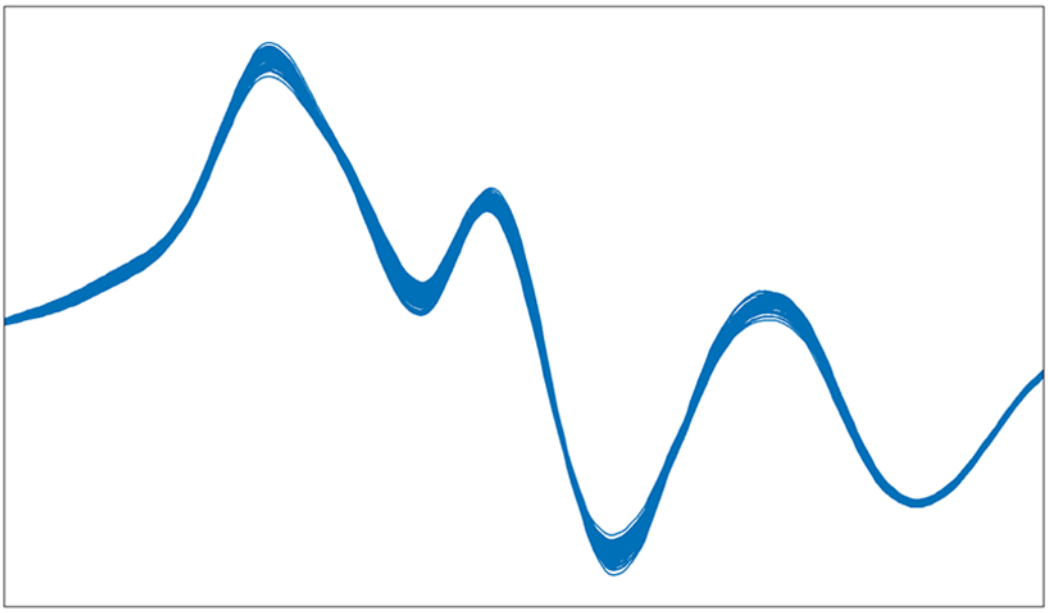
Teoria OMNIS NIR



Manual

8.0600.8101PT / v9 / 2025-10-10

Metrohm

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

Teoria OMNIS NIR

Manual

8.0600.8101PT / v9 /
2025-10-10

Technical Communication
Metrohm AG
CH-9100 Herisau

Esta documentação é um documento original.

Esta documentação foi cuidadosamente elaborada. No entanto, ainda pode conter erros. Nesse caso, solicita-se o envio de comunicação sobre eventuais erros ao endereço acima indicado.

Aviso de isenção de responsabilidade

Estão expressamente excluídas da garantia defeitos que não sejam da responsabilidade da Metrohm como armazenamento ou uso irregular, etc. As modificações não autorizadas do produto (por exemplo, conversões ou anexos) excluem qualquer responsabilidade por parte do fabricante pelos danos resultantes e suas consequências. As instruções e notas na documentação do produto da Metrohm devem ser rigorosamente seguidas. Caso contrário, a responsabilidade da Metrohm estará excluída.

Índice

1 Visão geral

1.1 Introdução

A espectroscopia de infravermelho próximo (ou espectroscopia NIR) é um método de análise rápido e sem reagente que não destrói a amostra e é adequado para um amplo espectro de amostras. Ela pode analisar ao mesmo tempo vários parâmetros e determinar tanto as propriedades físicas quanto químicas de um material. Alguns exemplos de propriedades são concentrações de analito, densidade, tamanho das partículas ou viscosidade intrínseca.

A espectroscopia NIR também permite a identificação de amostras desconhecidas (a partir do OMNIS Software 4.0) e a verificação de amostras (a partir do OMNIS Software 4.2).

A possibilidade de fazer a medição de amostras à distância sem destruí-las tem importância decisiva no controle de qualidade e no monitoramento de processos.

O manual descreve técnicas e algoritmos para registro, processamento e análise de espectros do infravermelho próximo, conforme implementados no OMNIS Software. O capítulo 2 esclarece brevemente como os sinais de medição são convertidos em espectros de absorção. O capítulo 3 aborda a calibração do equipamento. O capítulo 4 descreve o desenvolvimento de modelos capazes de prever os parâmetros de interesse (quantificação) ou se a amostra é parte de um produto (identificação). O capítulo 5 trata da previsão de amostras desconhecidas. O capítulo 6 traz um anexo com esclarecimentos sobre diversos algoritmos.

1.2 Âmbito conceitual

Os processos apresentados enquadram-se nos seguintes âmbitos:

1. **Calibração, padronização e testes de desempenho**
   Verificação de transferibilidade e confiabilidade dos espectros de absorção registrados com o equipamento.
2. **Desenvolvimento de modelo**
   Será desenvolvido um modelo para a previsão de um parâmetro quantitativo ou para a identificação de amostras.
   O desenvolvimento é baseado em amostras com parâmetros de interesse conhecidos ou no pertencimento de amostras a um produto.

3. **Análise de amostras**
   É registrado um espectro da amostra analisada. Com base no espectro, um modelo de quantificação fornece uma previsão quantitativa ou um modelo de identificação identifica ou verifica a amostra.
4. **Monitoramento**
   O monitoramento do modelo e do equipamento confirma se o sistema é adequado para a continuação da análise.

1.3 Informações sobre a documentação

Convenções de apresentação

Letras maiúsculas em negrito designam matrizes, letras minúsculas em negrito designam vetores. A transposição de uma matriz ou de um vetor é identificada por um "t" como expoente, por exemplo, $\mathbf{x}^{\mathbf{t}}$.

Grandezas escalares são representadas por letras minúsculas. O sinal de acento circunflexo (^) identifica valores estimados (calculados), por exemplo $\hat{y}$. Um traço identifica valores médios, por exemplo $\overline{y}$.

Uma grandeza escalar pode representar uma variável em função de um comprimento de onda, por exemplo, a absorvância A. A forma de escrever grandezas escalares pode ser trocada pela forma de escrever grandezas vetoriais.

1.4 Informações adicionais

Nas páginas a seguir, há disponíveis informações adicionais sobre o produto:

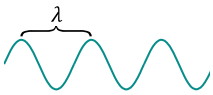
- Website da Metrohm *https://www.metrohm.com* – Visão geral da família de produtos, documentos PDF, dados do acessório e informações sobre aplicações.
- Ajuda do software OMNIS Software *https://guide.metrohm.com* – Informações individuais filtradas por temas sobre o software de controle.

2 Luz infravermelha e espectros

2.1 A luz e sua interação com a matéria

Um espectrômetro mede como uma amostra interage com a luz. A luz pode ser absorvida ou dispersada em diferentes intensidades. Essa interação depende das propriedades da luz, principalmente de seu comprimento de onda, e das propriedades do material, principalmente de sua estrutura molecular.

Comprimento de onda



A luz é uma radiação eletromagnética. Ela se movimenta como onda com campos elétricos e magnéticos oscilantes através do espaço. As ondas se espalham no espaço ao longo do tempo. Uma onda é caracterizada conforme seu comprimento de onda λ (p. ex. em nanômetros = 10^{-9} metros) e sua frequência (Hz).

O comprimento de onda é inversamente proporcional à frequência da onda. Ondas com altas frequências (mais oscilações por segundo) têm comprimentos de onda curtos e vice-versa. Devido a essa relação, uma onda pode ser descrita pelo seu comprimento de onda (nm) ou por sua frequência (Hz).

A luz pode trocar energia em unidades quânticas distintas, os fótons. A energia de um único fóton, *E*, depende de sua frequência *f* ou de seu comprimento de onda λ:

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

Nessa relação, *h* é a constante de Planck e *c* a velocidade da luz.

A figura 1 mostra diferentes faixas da radiação eletromagnética.

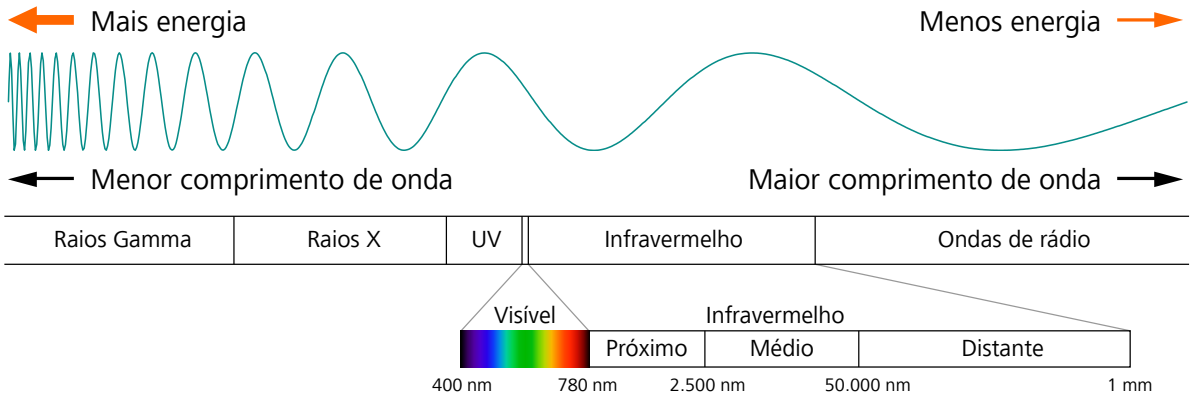


Figura 1 Faixas da radiação eletromagnética. A faixa de infravermelho próximo (NIR) está ao lado da faixa de luz visível. A faixa NIR compreende comprimentos de onda de 780 nm até 2.500 nm.

Fontes de radiação

As diferentes faixas de radiação eletromagnética têm diferentes fontes de radiação. Na faixa NIR, a fonte é a **radiação de calor**. Uma câmera de infravermelhos reconhece, por exemplo, as áreas do corpo humano com uma temperatura maior do que o ambiente.

Intensidade

A amplitude da onda eletromagnética determina a intensidade da luz. Quanto maior a amplitude, mais elevada é a intensidade. Na luz visível, a intensidade é percebida como claridade.

Interação entre luz e matéria

A seguir, é apresentado o processo de absorção da luz pelas moléculas.

A forma como a luz interage com a matéria depende da faixa de radiação eletromagnética. Se a energia for transferida da luz visível para as moléculas, por exemplo, os elétrons nas moléculas passam de um baixo nível de energia para um alto nível de energia (transição eletrônica). Na faixa infravermelha, ocorrem **transições vibracionais**. Ligações químicas, grupos funcionais e moléculas podem vibrar de diferentes formas, por exemplo, vibração de valência, vibração de deformação ou vibração de torção.

As moléculas só podem assumir modos de vibração distintos. À temperatura ambiente, a maioria das moléculas se encontra no modo fundamental de vibração (nível 0). As transições a partir do modo fundamental de vibração para estados excitados são denominadas conforme o esquema a seguir:

Transição vibracional *i* ➔ *j*	Nome
0 ➔ 1	Transição fundamental
0 ➔ 2	transição para primeiro sobretom
0 ➔ 3	transição para segundo sobretom

Um modo de vibração mais alto corresponde a um nível de energia mais alto. Para a transição do estado *i* para o estado excitado *j*, a molécula precisa absorver uma determinada energia de transição ΔE_{ij}.

A luz pode trocar a energia em partes de $E = hf$, sendo f a frequência da luz. A absorção da luz ocorre se a energia do fóton hf for igual à energia de transição ΔE_{ij}.

Os modos de oscilação permitidos dependem, entre outros fatores, da intensidade das ligações e da massa dos átomos envolvidos. Portanto, um determinado tipo de ligação pode ser correlacionado com energias de transição ou comprimentos de onda absorvidos característicos.

Para que a luz possa ser absorvida, outras condições precisam ser cumpridas. A transição vibracional deve deslocar a distribuição da carga de modo que o momento dipolar elétrico da molécula seja alterado. A probabilidade de uma absorção de energia depende da ordem de grandeza da alteração do momento dipolar ao longo da ligação química envolvida.

Uma transição vibracional pode causar uma alteração do momento dipolar tanto em moléculas polares quanto apolares ou em grupos funcionais. Moléculas homonucleares diatômicas, como N_2, não absorvem luz infravermelha.

A duração do estado de vibração excitado é limitada. Quando a molécula volta para um estado de vibração mais baixo, a energia é convertida em calor.

A faixa espectral NIR

Os comprimentos de onda correspondentes das transições fundamentais estão na faixa de infravermelhos médios. A faixa de infravermelhos próximos compreende transições de sobretom e bandas de combinação. *A figura 2* mostra as bandas de comprimento de onda absorvidas pelas diferentes moléculas e grupos funcionais.

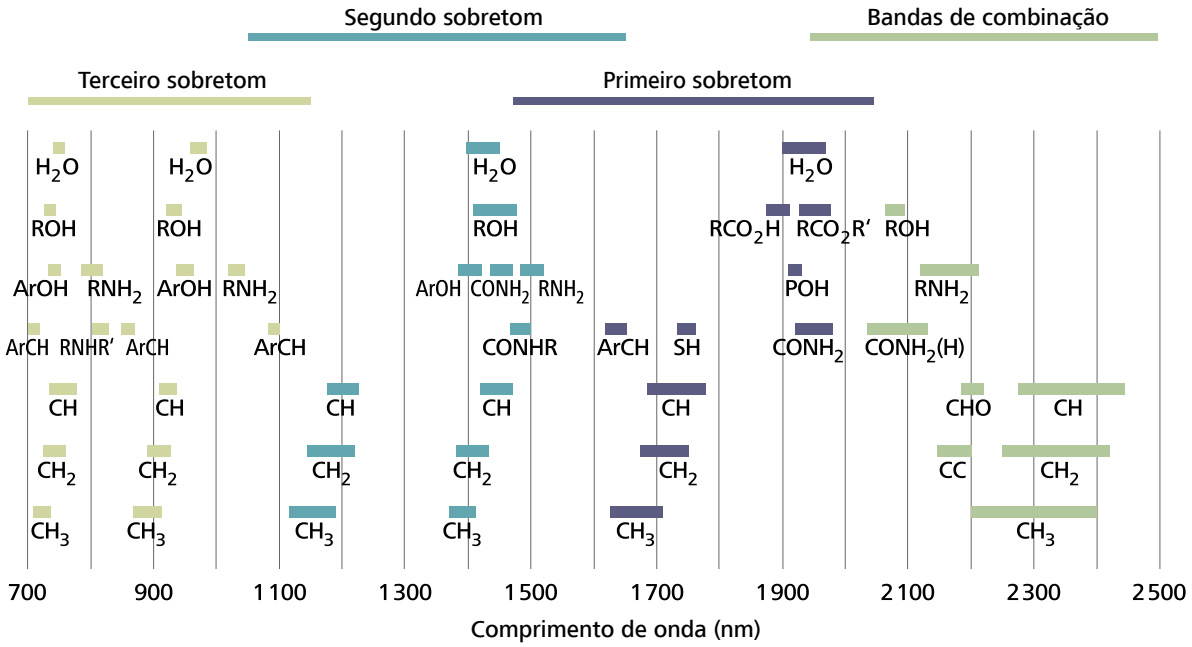


Figura 2 Bandas de absorção NIR

A transição fundamental é a transição com maior probabilidade de ocorrência, ela é a mais frequente. Transições de sobretom têm menor probabilidade de ocorrência. Portanto, a transição fundamental absorve mais luz do que transições de sobretom. De maneira geral, a absorção diminui a cada sobretom. Assim, os sobretons são adequados para moléculas com absorção mais intensa.

Duas ou mais vibrações fundamentais podem ser excitadas simultaneamente com uma única frequência de luz que corresponde à frequência combinada das vibrações fundamentais. As bandas de absorção correspondentes são designadas como **bandas de combinação**. Algumas bandas de combinação estão na faixa NIR, ou seja, entre 1.900 e 2.500 nm.

2.2 Princípios básicos matemáticos

2.2.1 Lei de Beer-Lambert

A lei de Beer-Lambert descreve como a absorção de luz por uma amostra homogênea depende das propriedades de uma substância que absorve luz na amostra.

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

Nessa relação, A corresponde à absorvância, ε ao coeficiente de extinção molar do absorvente (L/mol/cm), c à concentração do absorvente (mol/L) e l à espessura da camada de amostra (cm).

O coeficiente de extinção molar ε é uma constante que informa o quanto uma substância absorve. O coeficiente de extinção molar é específico para um determinado comprimento de onda *i* e uma determinada substância *j*. A absorvância total de uma mistura é a soma da absorvância de todas as substâncias contidas na mistura:

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

Neste cálculo, A_i corresponde à absorvância no comprimento de onda *i*, *N* à quantidade de substâncias na mistura, ε_{ij} ao coeficiente de extinção molar para o comprimento de onda *i* e a substância *j*, e c_j à concentração da substância *j*.

A lei de Beer-Lambert pressupõe uma relação linear entre a absorvância e a concentração, bem como uma relação linear entre a absorvância e o coeficiente de extinção molar. Esse comportamento linear é válido para várias situações.

Com base nos princípios da lei de Beer-Lambert, as medições espectroscópicas podem comprovar:

- Variações de concentração de um absorvente.
  Esta é a aplicação mais frequente.
- Variações nos fatores que influenciam o coeficiente de extinção molar.
  Temperatura, viscosidade, valor de pH ou a constante dielétrica do solvente podem influenciar o coeficiente de extinção molar. Em alguns casos, isso pode ser utilizado para medições espectroscópicas.

Efeitos de dispersão não estão relacionados à lei de Beer-Lambert. Algumas vezes, efeitos de dispersão podem ser usados para reconhecer variações de tamanho das partículas, por exemplo.

2.2.2 Regressão linear

Um comprimento de onda

No caso mais simples, uma mistura contém somente um absorvente. Segundo a lei de Beer-Lambert, a absorvância em um determinado comprimento de onda é linear em relação à concentração do absorvente.

Na figura 1, cada ponto representa uma amostra com uma concentração conhecida do absorvente (eixo x) e a absorvância medida (eixo y). Uma regressão linear resulta na reta de regressão **A**.

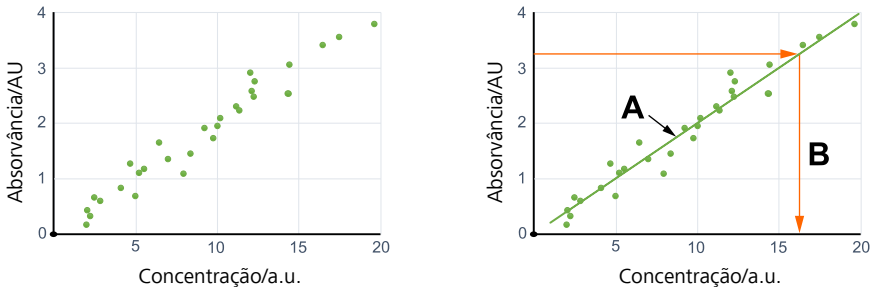


Figura 3 Relação entre valores de absorvância e concentrações

Para uma amostra com concentração desconhecida de absorvente, a concentração pode ser determinada da seguinte forma:

1. Medição da absorvância no comprimento de onda predefinido.
2. Utilização da reta de regressão para determinação da concentração (**B**).
   A reta de regressão é um modelo de quantificação simples para previsão do parâmetro de interesse, p. ex., da concentração de absorvente.

Porém, esse procedimento não funciona se a mistura contiver vários absorventes com diferentes concentrações.

Vários comprimentos de onda

Ao invés de uma medição da absorção em um único comprimento de onda, é possível medir a absorção em vários comprimentos de onda, o que resulta em um espectro. De maneira semelhante ao procedimento descrito acima, a relação entre os espectros e o parâmetro de interesse pode ser determinada por meio de uma regressão linear. Para vários comprimentos de onda, é necessária uma regressão linear múltipla (MLR).

É possível comprovar que a regressão linear múltipla também pode prever o parâmetro de interesse quando a mistura contém vários absorventes com diferentes concentrações *(ver capítulo 6.1, página 90)*.

Mas, para uma regressão linear múltipla, é necessário ter uma quantidade de amostras maior do que a quantidade de comprimentos de onda. Para

previsões espectroscópicas, podem ser utilizados outros métodos, como PCA ou PLS.

2.3 Como a luz é convertida em um espectro

Um espectrômetro (ou espectrofotômetro) consiste em uma fonte de luz e uma unidade detectora. A fonte de luz emite luz com um amplo espectro de comprimentos de onda, ou seja, luz policromática. A luz interage com a amostra. A seguir, o espectrômetro registra a luz restante como uma função do comprimento de onda.

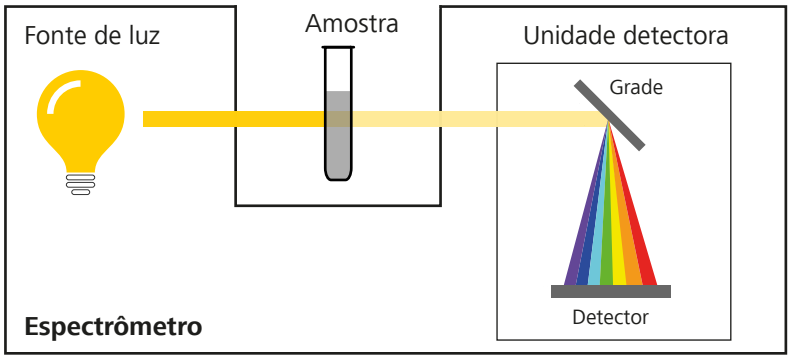


Figura 4 Um espectrômetro com fonte de luz e unidade detectora.

No espectrômetro, a luz é decomposta com o auxílio de uma grade em comprimentos de onda individuais. Um detector mede a luz, sendo que cada comprimento de onda atinge um sensor de arranjo linear em um determinado item ou pixel.

Uma **leitura** é uma medição por todos os pixels. Cada pixel gera um sinal fotoelétrico proporcional à intensidade de luz. Os sinais podem ser representados em um gráfico em relação aos pixels.

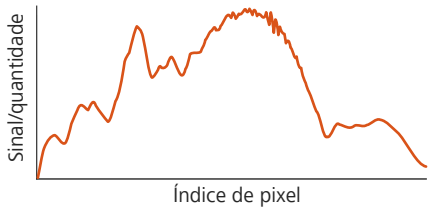


Figura 5 O espectro do sinal do detector em função dos pixels.

Tempo de integração

O tempo de integração é o intervalo de tempo em que o detector coleta a luz. Um tempo de integração maior aumenta o sinal.

Tempos de integração excessivamente longos levam a uma saturação do detector e, com isso, a uma perda de informações. Tempos de integração excessivamente curtos reduzem o sinal e, com isso, a relação sinal-ruído.

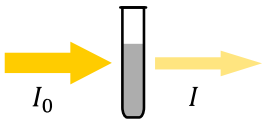
O **tempo de integração automático** assegura a iluminação ideal, ou seja, uma relação sinal-ruído ideal sem que ocorra uma saturação. Antes de cada leitura da amostra e cada leitura de referência, são executadas diversas medições. O tempo de integração é ajustado de modo que o sinal com a maior intensidade alcance aproximadamente 90% da área de reconhecimento.

Diferenças nos tempos de integração são consideradas nos próximos cálculos.

- **OMNIS NIR Analyzer**
  O tempo de integração sempre é ajustado automaticamente.
- **2060 The NIR**
  O tempo de integração pode ser ajustado manual ou automaticamente.
  O tempo de integração manual pode reduzir a duração da medição em medições semelhantes. Para evitar a saturação do detector em tempos de integração excessivamente longos, o tempo de integração deve ser ajustado com uma margem suficiente *(ver "Saturação", página 49)*.

Absorvância

Medições espectroscópicas determinam quanta luz uma amostra absorve ou dispersa. A amostra é submetida a um raio de luz. Um detector mede a luz dissipada de uma fonte de luz e a luz restante após a interação com a amostra.



A absorvância A é definida como o logaritmo decádico da relação entre a intensidade da luz antes da interação do raio de luz com a amostra (I_0) e a intensidade de luz depois da interação do raio de luz com a amostra (I):

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

Uma absorvância de 1 significa que 10% da luz passa através da amostra, enquanto uma absorvância de 2 significa que apenas 1% passa através.

A absorvância não é expressa em uma unidade.

Leitura de referência e leitura da amostra

Conforme a fórmula acima, são necessárias duas leituras para o cálculo de um espectro de absorção. As leituras medem um sinal fotoelétrico para cada pixel S:

- Uma **leitura de referência** mede o sinal S_0 antes da interação do raio de luz com a amostra.
- Uma **leitura da amostra** mede o sinal S depois da interação do raio de luz com a amostra.

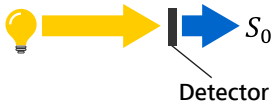
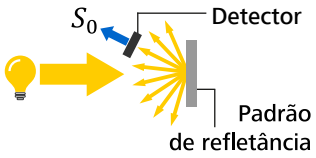
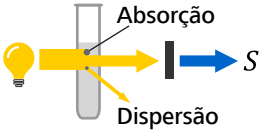
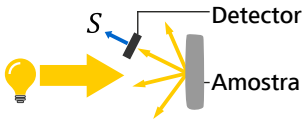
O sinal fotoelétrico medido em um pixel é proporcional ao valor médio da intensidade de luz pela área do pixel. Portanto, $S_0/S = I_0/I$. Assim, os sinais fotoelétricos podem ser utilizados para o cálculo da absorvância:

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

Transmissão e refletância

No **modo de transmissão**, é medida a luz que passa através da amostra. S_0 é medido na ausência da amostra. S é medido a partir da luz que passa através da amostra.

No **modo de refletância**, é medida a luz refletida pela amostra. Ao invés da amostra, um padrão de refletância é utilizado como referência. É ideal que o padrão de refletância reflita 100% da luz. Uma parte da luz refletida é conduzida pelo detector e fornece o sinal S_0. O sinal S é medido da mesma forma, porém com a amostra que reflete a luz.

	Transmissão	Refletância difusa
Sem amostra		
Com amostra		

A absorvância calculada A representa toda a luz que não alcançou o detector. Portanto, A contém não somente a luz absorvida pela amostra, como também:

- A luz que não alcançou o detector, porque foi dispersada para fora dele.
- A luz que foi dispersada incorretamente para o detector.

Espectro de absorção

O espectro de absorção é calculado com base na leitura de referência (sinal S_0) e na leitura da amostra (sinal S), segundo a fórmula acima.

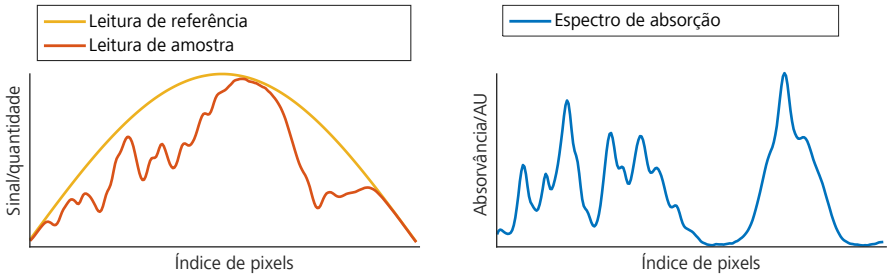


Figura 6 A leitura de referência e a leitura da amostra (esquerda), bem como o espectro de absorção (direita), em função do índice de pixels.

O cálculo acima pressupõe que o escaneamento de referência e o escaneamento de amostra usam os mesmos caminhos ópticos ou caminhos ópticos com propriedades ópticas semelhantes. Em ambientes de processo, é utilizada uma abordagem de referenciamento em vários níveis *(ver capítulo 3.2.2, página 17)*.

Dos pixels aos comprimentos de onda

A escala de pixels é convertida na escala de comprimentos de onda. O equipamento atribui a cada pixel um comprimento de onda exato, p. ex.:

Pixel 6 ➔ comprimento de onda 1.009,4 nm

O comprimento de onda exato de cada pixel é determinado pela calibração do comprimento de onda *(ver capítulo 3.1, página 14)*.

Conversão da escala de comprimento de onda

O espectro é transposto por interpolação para a escala de comprimento de onda padrão:

1.000,0 nm, 1.000,5 nm, 1.001,0 nm, …

Figura 7 O espectro de absorção na escala de comprimento de onda.

3 Definições do equipamento

Os passos a seguir asseguram que sejam obtidos espectros idênticos para uma determinada amostra com uma determinada estrutura de medição da amostra, independentemente de elas terem sido capturadas em momentos diferentes, de ter sido utilizado o mesmo equipamento da família de produtos **OMNIS NIR Analyzer** ou não, ou de ter sido utilizado outro equipamento do tipo **2060 The NIR**.

Tanto o eixo x quanto o eixo y dos espectros devem ser considerados:

- **Eixo x**: calibração do comprimento de onda *(ver capítulo 3.1, página 14)*
- **Eixo y**: padronização de referência *(ver capítulo 3.2, página 15)*

Além disso, é necessário assegurar que o desempenho do equipamento corresponda aos requisitos, ou seja:

- Os **Testes de desempenho do equipamento** devem ter sido executados com sucesso antes de poder registrar espectros com o equipamento *(ver capítulo 3.3, página 26)*.

Padrões

Para a calibração do comprimento de onda e a Testes de desempenho do equipamento, o equipamento utiliza um **padrão de comprimento de onda** interno rastreável metrologicamente.

Além disso, ao utilizar o modo de refletância, é necessário um **padrão de refletância** orientado pelo tipo do equipamento para a padronização de referência e a Testes de desempenho do equipamento:

- **OMNIS NIR Analyzer**: padrão de refletância interno
- **2060 The NIR**: padrão de refletância externo

3.1 Calibração do comprimento de onda

A calibração do comprimento de onda normaliza os valores de comprimento de onda, ou seja, o eixo x dos espectros. Ela atribui cada pixel da matriz do fotodetector a um comprimento de onda.

A calibração do comprimento de onda utiliza um padrão de comprimento de onda interno rastreável metrologicamente. O padrão de comprimento de onda tem um espectro de absorção com picos definidos e posições de picos conhecidas.

O comando **CAL WL** executa os passos a seguir:

1. O espectro de absorção do padrão de comprimento de onda é registrado utilizando um caminho de referência interno na escala de pixels.
2. No espectro registrado, são identificadas as posições de pico com precisão de subpixels.
3. Utilizando as posições de pico medidas na escala de pixels e as posições nominais do padrão de comprimento de onda, é executada uma regressão polinomial.
4. O polinômio de regressão atribui cada pixel a seu respectivo comprimento de onda.

Os coeficientes de regressão são salvos no equipamento:

- **OMNIS NIR Analyzer**: para cada apresentação de amostras, é salvo um conjunto de coeficientes de regressão no equipamento. Isso significa que, para o OMNIS NIR Analyzer Liquid/Solid, deve ser executada uma calibração do comprimento de onda e uma validação em ambos os dispositivos.
- **2060 The NIR**: os coeficientes de regressão são específicos do equipamento. Para todos os canais, é utilizado o mesmo conjunto.

Validação da calibração do comprimento de onda

Após a execução de uma calibração do comprimento de onda, ela deve ser validada. O comando **VAL WL** executa os passos a seguir:

1. O espectro de absorção do padrão de comprimento de onda será registrado.
2. Validação dos comprimentos de onda:
    a. No espectro registrado, as posições de pico são identificadas na escala de comprimento de onda.
    b. São calculados os resíduos de comprimentos de onda entre as posições de pico medidas e as posições de pico conhecidas.
    c. Para cada pico, o resíduo de comprimento de onda deve estar dentro da tolerância para que ele seja aprovado no teste.

3. Validação das larguras de banda:
   a. No espectro registrado, as larguras de pico são determinadas.
   b. Os resíduos de larguras de banda são calculados entre as larguras de pico medidas e as larguras de pico conhecidas.
   c. Para cada pico, o resíduo de largura de banda deve estar dentro da tolerância para que ele seja aprovado no teste.
4. O status geral da validação é bem-sucedido quando todos os resíduos mencionados acima estão dentro da tolerância.

A validação deve ser executada com sucesso antes de poder registrar espectros com o equipamento.

3.2 Padronização de referência

A padronização de referência normaliza os valores de absorvância, ou seja, o eixo y dos espectros.

Determinação da absorvância

Para o cálculo da absorvância A de uma amostra, os sinais S_0 (leitura de referência) e S (leitura da amostra) são necessários *(ver capítulo 2.3, página 9)*:

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

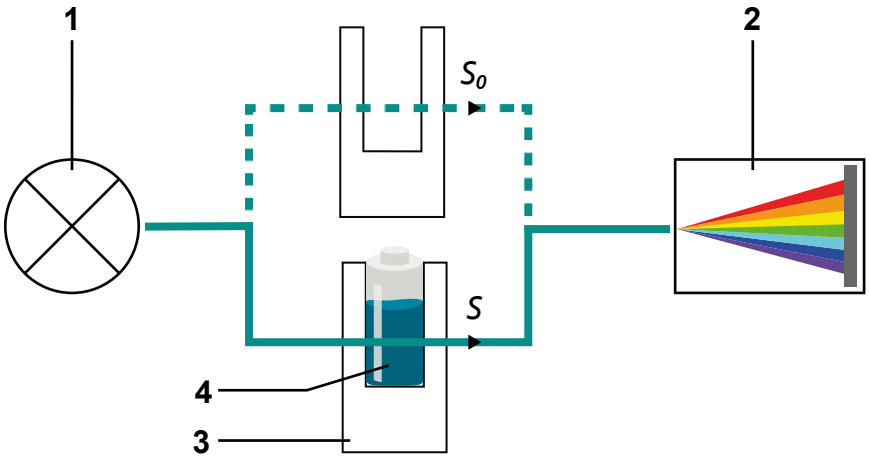


Figura 8 Caminho óptico no modo de transmissão (como exemplo com uma apresentação de amostras de líquidos).

Na *figura 8*, a luz emitida pela fonte de luz (**1**) passa através de um suporte de amostra (**3**) e chega ao detector (**2**).

O sinal de referência S_0 é medido sem a amostra, o sinal S com a amostra (**4**). Caso contrário, as propriedades ópticas de ambos os caminhos ópticos são idênticas e causam atenuação na mesma porcentagem para ambos os sinais. Isso não altera o resultado da fórmula acima.

A equação relaciona S com S_0, a referência. Ambos os sinais são igualmente importantes. Um desvio em ambos os sinais leva a outro valor de absorbância e, consequentemente, a outro espectro.

Tanto S como S_0 são influenciados por variações do equipamento e pelas condições ambientais. Para assegurar que essas influências se eliminem mutuamente, ambos os sinais devem ser medidos em um intervalo de tempo curto.

A implementação desse princípio depende do tipo do equipamento:

- **OMNIS NIR Analyzer**
  O espectro de absorção da amostra é calculado usando os sinais S e S_0 *(ver capítulo 3.2.1, página 16)*.
- **2060 The NIR**
  Em ambientes de processo, é impraticável utilizar o mesmo caminho óptico para as medições de S e S_0. Portanto, é necessário tomar outras medidas *(ver capítulo 3.2.2, página 17)*.

3.2.1 OMNIS NIR Analyzer

A padronização de referência é efetuada por meio da medição dos sinais S_0 e S, bem como pelo cálculo da absorvância A.

Registrar o espectro de uma amostra

Antes de poder registrar espectros com um dispositivo, os Testes de desempenho do equipamento para o dispositivo devem ser executados com sucesso *(ver capítulo 3.3, página 26)*.

1. A amostra deve estar pronta na apresentação de amostras.
2. Cada espectro de absorção é calculado com o espectro de referência registrado por último S_0. Para obter um valor atual para S_0, pode ser executado o comando **MEAS REF SPEC**.
   Ao utilizar apresentação de amostra de sólido, o equipamento insere automaticamente um padrão de refletância no caminho óptico. Esse padrão de refletância requer uma correção do sinal S_0.
3. Com o comando **MEAS SPEC**, a amostra é medida. Isso resulta no sinal S.
4. O software calcula A, a absorvância da amostra:
   $$A = \log_{10} \frac{S_0}{S}$$
   Neste caso S_0 corresponde ao sinal medido no caminho de referência e S sinal medido por meio da amostra.

3.2.2 2060 The NIR

Equipamentos do tipo **2060 The NIR** requerem uma padronização de referência externa.

Padronização de referência externa

A medição repetida do sinal S (com a amostra) e do sinal S_0 (sem a amostra) por caminhos ópticos com propriedades ópticas idênticas é demorada e susceptível a erros.

Portanto, são introduzidas mais duas trajetórias de raios *(ver figura 9, página 17)*:

- Uma **referência interna** no equipamento. O caminho de referência interno fornece o sinal S_{ref}, que pode ser medido de maneira simples.
- Outro caminho óptico externo em que as fibras ópticas estão ligadas a um **dispositivo de calibração**. Este caminho óptico fornece o sinal S_{fiber}.

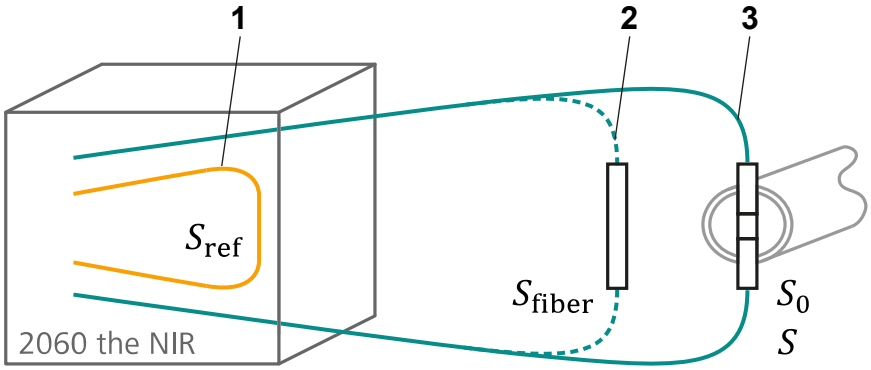


Figura 9 Caminhos ópticos como exemplo no modo de transmissão: caminho de referência interno (1), fibras ópticas externas ligadas a um dispositivo de calibração (2) e fibras ópticas externas ligadas à sonda com ou sem amostra (3). Os caminhos ópticos 2 e 3 representam a mesma fibra óptica simplesmente ligada de maneiras diferentes.

O dispositivo de calibração fixa as fibras ópticas constituindo, assim, o caminho de referência (**2**). No modo de transmissão, o ar serve como referência e transmite 100% da luz. No modo de refletância, o dispositivo de calibração também registra o padrão de refletância. Primeiramente, parte-se de um padrão de refletância ideal que reflita 100% da luz.

A absorvância A da amostra é calculada a partir dos sinais S_0 e S. Se os outros dois sinais adicionais S_{ref} e S_{fiber} forem adicionados ao numerador e ao denominador, o resultado permanece inalterado:

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

Essa equação pode ser convertida em:

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

Os 3 termos representam valores de absorvância e podem ser designados da seguinte forma:

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

A figura 10 ilustra como os sinais S_{ref}, S_{fiber}, S_0 e S são medidos.

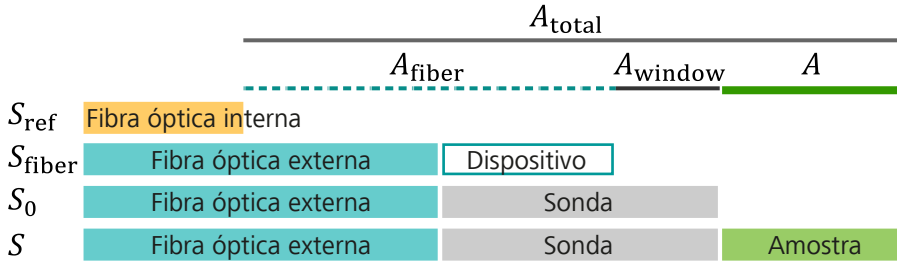


Figura 10 Padronização de referência externa

A_{total} é a absorvância da fibra óptica externa, da sonda e da amostra com relação à fibra óptica interna.

A_{fiber} é a absorvância da fibra óptica externa adicionada ao dispositivo de calibração com relação à fibra óptica interna.

A_{window} é a absorvância da sonda subtraída do dispositivo de calibração.

Eliminação de variações do ambiente

Para determinar a absorvância A da amostra, 3 valores de absorvância são medidos. A partir destes 3 valores, A é calculado conforme a equação acima:

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

Isso permite determinar 3 valores de absorvância para diferentes momentos. Dessa forma, é possível eliminar com facilidade variações no equipamento ou variações nas condições ambientais para cada uma das 3 determinações:

- A_{total} é determinado com cada medição de amostra. Ao fazer isso, S_{ref} e S devem ser medidos dentro de um intervalo de tempo curto para eliminar variações.
- A **correção da fibra óptica** A_{fiber} pode ser determinada mais raramente. Ao fazer isso, S_{ref} e S_{fiber} devem ser medidos dentro de um intervalo de tempo curto para eliminar variações.

- A **correção de janela** A_{window} pode ser determinada mais raramente, de maneira geral, somente uma vez após a instalação. Ao fazer isso, S_{fiber} e S_0 devem ser medidos dentro de um intervalo de tempo curto para eliminar variações.

Quando uma correção de janela é necessária?

Caso o dispositivo de calibração transmitir completamente as propriedades ópticas da sonda, A_{window} será igual a 0. Nesse caso, a correção de janela não é necessária.

Para o modo de transmissão, geralmente é necessária uma correção de janela. Para o modo de refletância, a correção de janela não é necessária. Porém, existem as seguintes exceções listadas na tabela:

Modo de medição	Sonda	Padronização de referência
Transmissão	Par de transmissão	Fibra óptica + janela
	Sonda de transmissão	
	Sonda de transfletância com fibras individuais	
Refletância	Sonda de refletância	Fibra óptica
	Sonda de transfletância com MicroBundle	Fibra óptica + janela

Para decidir se é necessária uma correção de janela, cada combinação de dispositivo de calibração e sonda deve ser examinada. Se o dispositivo de calibração *não* transmitir completamente as propriedades da sonda, um correção de janela é necessária.

Canais

Um equipamento do tipo **2060 The NIR** oferece mais canais. Cada canal pode ser ligado com outra configuração de fibra óptica e de sonda. Portanto, a padronização de referência deve ser executada separadamente para cada canal.

Todos os canais utilizam o mesmo caminho de referência interno, ou seja, o mesmo sinal S_{ref}. Um multiplexador comuta entre a referência interna e os diversos canais de medição e vice-versa.

Execução de uma correção da fibra óptica

Após a colocação em funcionamento ou quando a configuração da fibra óptica de um canal for alterada, deve ser executada uma correção da fibra óptica. Uma substituição da lâmpada ou uma alteração extrema nas condições ambientais também pode fazer com que seja recomendável executar uma nova padronização.

Nestes processos, é utilizado um material de referência:

- No modo de refletância, o material de referência é o padrão de refletância. Parte-se de um padrão de refletância não ideal (p. ex. 99%). O padrão de refletância tem um espectro de absorção nominal conhecido $A_{nominal}$.
- No modo de transmissão, o ar serve como referência. O espectro de absorção nominal é a uma linha zero ($A_{nominal} = 0$), por pressupõe-se que o ar não absorve luz.

A figura 11 ilustra procedimento a seguir:

1. As fibras ópticas externas devem ser conectadas ao dispositivo de calibração.
   No modo de refletância, o dispositivo de calibração é combinado com o padrão de refletância.
2. O comando **REF STD** com a interface **Fibra de vidro** executa as leituras a seguir:
   a. Uma leitura de referência interna fornece um valor para S_{ref}.
   b. Uma leitura externa mede as fibras ópticas externas, o dispositivo de calibração e o material de referência. Isso resulta no sinal S_{raw}.
3. O software calcula A_{raw} (**Espectro bruto medido**):
   $$A_{\mathrm{raw}} = \log_{10} \frac{S_{\mathrm{ref}}}{S_{\mathrm{raw}}}$$
   Neste cálculo, A_{raw} corresponde à absorvância das fibras ópticas externas, do dispositivo de calibração e do material de referência com relação ao caminho óptico interno.
4. O espectro de absorção nominal do material de referência, $A_{nominal}$, é exibido no software como **Espectro de referência**. O espectro de referência deve ser subtraído de A_{raw} para obter A_{fiber}:
   $$A_{\mathrm{fiber}} = A_{\mathrm{raw}} - A_{\mathrm{nominal}}$$
   Neste cálculo, A_{fiber} corresponde à absorvância das fibras ópticas e do dispositivo de calibração com relação ao caminho óptico interno.
   Aviso: no modo de transmissão, $A_{nominal} = 0$ e $A_{fiber} = A_{raw}$.
   A_{fiber} representa o **Espectro de correção** da fibra óptica.
5. A_{fiber} permanece inalterado até que outra correção da fibra óptica seja executada para o respectivo canal.

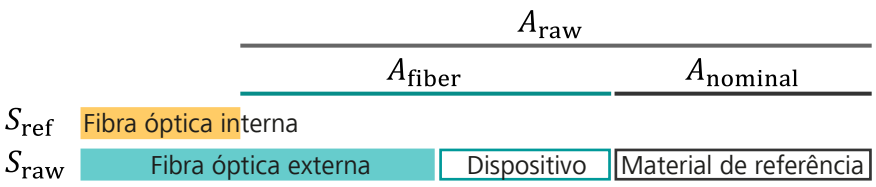


Figura 11 Correção da fibra óptica

Validação da correção da fibra óptica

A correção da fibra óptica deve ser validada com os mesmos parâmetros de medição e o mesmo dispositivo de calibração.

Nestes processos, é utilizado um material de referência:

- No modo de refletância, o material de referência é o padrão de refletância. Parte-se de um padrão de refletância não ideal (p. ex. 99%). O padrão de refletância tem um espectro de absorção nominal conhecido $A_{nominal}$.
- No modo de transmissão, o ar serve como referência. O espectro de absorção nominal é a uma linha zero ($A_{nominal}$ = 0), por pressupõe-se que o ar não absorve luz.

A figura 12 ilustra como os resíduos da validação são determinados:

1. As fibras ópticas externas devem ser conectadas ao dispositivo de calibração.
   No modo de refletância, o dispositivo de calibração é combinado com o padrão de refletância.
2. O comando **VAL REF STD** com a interface **Fibra de vidro** executa as leituras a seguir:
   a. Uma leitura de referência interna fornece um valor para S_{ref}.
   b. Uma leitura externa no respectivo canal mede as fibras ópticas externas, o dispositivo de calibração e o material de referência. Isso resulta no sinal S_{raw}.
3. O software calcula A_{raw} (**Espectro bruto medido**):
   $$A_{\mathrm{raw}} = \log_{10} \frac{S_{\mathrm{ref}}}{S_{\mathrm{raw}}}$$
4. A_{raw} é corrigido pelo espectro de correção de fibra óptica para eliminar a absorvância das fibras ópticas e do dispositivo de calibração:
   $$A_{\mathrm{corrected}} = A_{\mathrm{raw}} - A_{\mathrm{fiber}}$$
   $A_{corrected}$ é exibido no software como **Espectro corrigido medido**.
5. É ideal que $A_{corrected}$ corresponda ao **Espectro de referência** $A_{nominal}$. As diferenças entre os dois são calculadas como **Resíduos da validação**:
   $$A_{\mathrm{residual}} = A_{\mathrm{corrected}} - A_{\mathrm{nominal}}$$
   Aviso: no modo de transmissão, $A_{nominal}$ = 0 e $A_{residual}$ = $A_{corrected}$.

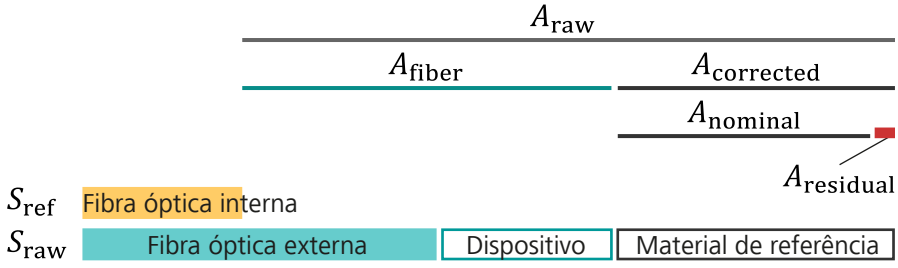


Figura 12 Resíduos para a validação da correção da fibra óptica

Para examinar os resíduos da validação, a faixa de comprimento de onda é dividida em vários segmentos. Para cada segmento, o valor médio dos

resíduos ao quadrado dividido pelos comprimentos de onda resulta no **resíduo RMS** (unidade: mAU):

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

Neste cálculo, f corresponde à quantidade de comprimentos de onda no segmento e $A_{\mathrm{residual}\ i}$ ao resíduo do comprimento de onda i.

Cada segmento deve manter uma tolerância predefinida para A_{RMS}. Se a tolerância for mantida por todos os segmentos, a validação geral é bem-sucedida.

A validação deve ser executada com sucesso antes de poder registrar espectros no respectivo canal com o equipamento.

Execução de uma correção de janela

Se for necessário executar uma correção de janela, ela deve ser feita após a colocação em funcionamento ou a cada alteração da configuração de sonda ou de fibra óptica de um canal. Alterações na sonda, como p. ex., sujeira, também podem fazer com que seja recomendável repetir a padronização.

Nestes processos, é utilizado um material de referência:

- No modo de refletância, o material de referência é o padrão de refletância. Parte-se de um padrão de refletância não ideal (p. ex. 99%). O padrão de refletância tem um espectro de absorção nominal conhecido A_{nominal}.
- No modo de transmissão, o ar serve como referência. O espectro de absorção nominal é a uma linha zero ($A_{\mathrm{nominal}} = 0$), por pressupõe-se que o ar não absorve luz.

A figura 13 ilustra procedimento a seguir:

1. Para obter um valor atual de A_{fiber}, uma correção da fibra óptica deve ser executada conforme descrito acima.
   Importante: a correção da fibra óptica deve ser efetuada antes da correção de janela.
2. As fibras ópticas externas deverão, então, ser conectadas sem uma amostra disponível na sonda. Caso necessário, um padrão de refletância pode ocupar o espaço da amostra.
3. O comando **REF STD** com a interface **Janela** executa as leituras a seguir:
   a. Uma leitura de referência interna fornece um valor para S_{ref}.
   b. Uma leitura externa no respectivo canal mede as fibras ópticas externas, a sonda e o material de referência. Isso resulta no sinal S_{probe}.

4. A absorvância A_{probe} com relação ao caminho óptico interno é:

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

5. O software calcula A_{raw} (**Espectro bruto medido**):

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

Neste cálculo, A_{raw} corresponde à absorvância da sonda e do material de referência com relação ao dispositivo de calibração.

6. O espectro de absorção nominal do material de referência, $A_{nominal}$, é exibido no software como **Espectro de referência**. O espectro de referência deve ser subtraído de A_{raw} para obter A_{window}:

$$A_{window} = A_{raw} - A_{nominal}$$

Neste cálculo, A_{window} corresponde à absorvância da sonda com relação ao dispositivo de calibração.
Aviso: no modo de transmissão, $A_{nominal}$ = 0 e A_{window} = A_{raw}.
A_{window} representa o **Espectro de correção** da janela.

7. A_{window} permanece inalterado até que outra correção de janela seja executada para o respectivo canal.

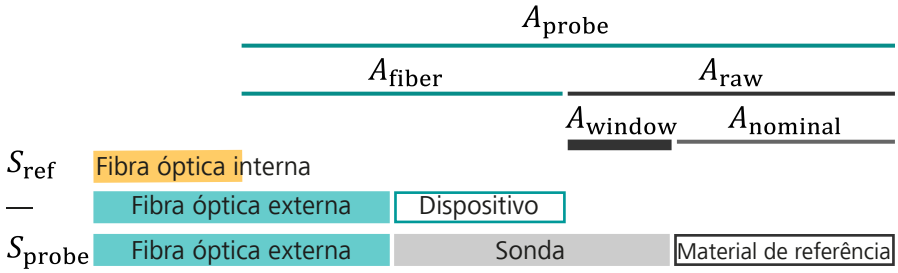


Figura 13 Correção de janela

Validação da correção de janela

A correção de janela deve ser validada com os mesmos parâmetros de medição e o mesmo dispositivo de calibração.

Nestes processos, é utilizado um material de referência:

- No modo de refletância, o material de referência é o padrão de refletância. Parte-se de um padrão de refletância não ideal (p. ex. 99%). O padrão de refletância tem um espectro de absorção nominal conhecido $A_{nominal}$.
- No modo de transmissão, o ar serve como referência. O espectro de absorção nominal é a uma linha zero ($A_{nominal}$ = 0), por pressupõe-se que o ar não absorve luz.

A figura 14 ilustra como os resíduos da validação são determinados:

1. As fibras ópticas externas deverão ser conectadas sem uma amostra disponível na sonda. Caso necessário, um padrão de refletância pode ocupar o espaço da amostra.

2. O comando **VAL REF STD** com a interface **Janela** executa as leituras a seguir:
   a. Uma leitura de referência interna fornece um valor para S_{ref}.
   b. Uma leitura externa no respectivo canal mede as fibras ópticas externas, a sonda e o material de referência. Isso resulta no sinal S_{probe}.
3. A absorvância A_{probe} com relação ao caminho óptico interno é:

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

4. O software calcula A_{raw} (**Espectro bruto medido**):

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

5. Por meio da subtração do espectro da correção de janela de A_{raw}, as diferenças de absorvância entre o dispositivo de calibração e a sonda são eliminadas:

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

   $A_{corrected}$ é exibido no software como **Espectro corrigido medido**.
6. É ideal que $A_{corrected}$ corresponda ao **Espectro de referência** $A_{nominal}$. As diferenças entre os dois são calculadas como **Resíduos da validação**:

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

   Aviso: no modo de transmissão, $A_{nominal} = 0$ e $A_{residual} = A_{corrected}$.

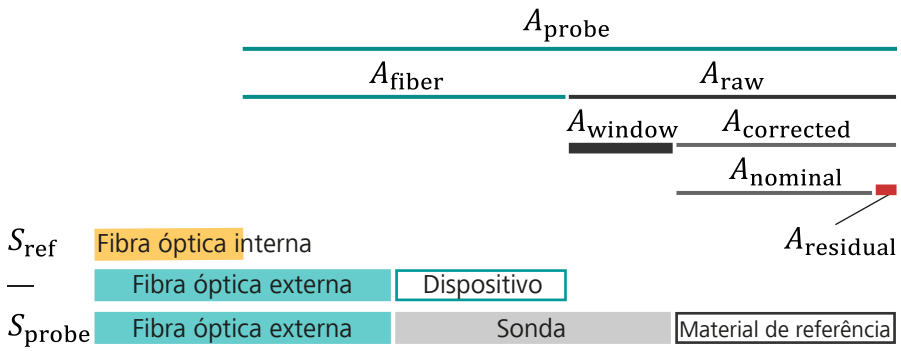


Figura 14 Resíduos para a validação da correção de janela

Para examinar os resíduos da validação, a faixa de comprimento de onda é dividida em vários segmentos. Para cada segmento, o valor médio dos resíduos ao quadrado dividido pelos comprimentos de onda resulta no **ruído RMS** (unidade: mAU):

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

Neste cálculo, *f* corresponde à quantidade de comprimentos de onda no segmento e $A_{residual\ i}$ ao resíduo do comprimento de onda *i*.

Cada segmento deve manter uma tolerância predefinida para A_{RMS}. Se a tolerância for mantida por todos os segmentos, a validação geral é bem-sucedida.

Registrar o espectro de uma amostra

Antes de poder registrar espectros no equipamento, os Testes de desempenho do equipamento devem ser executados com sucesso no respectivo canal *(ver capítulo 3.3, página 26)*.

O procedimento para captura do espectro de uma amostra é ilustrado na *figura 15*:

1. As fibras ópticas externas devem estar conectadas à sonda. Uma amostra deve estar disponível.
2. Cada espectro de absorção é calculado com o espectro de referência registrado por último S_{ref}. Para obter um valor atual para S_{ref}, pode ser executado o comando **MEAS REF SPEC**.
3. O comando **MEAS SPEC** mede a amostra, inclusive a sonda e as fibras ópticas. Isso resulta no sinal S.
4. O software calcula A_{total}, a absorvância da amostra, inclusive da sonda e das fibras ópticas com relação ao caminho óptico interno:
   $$A_{\mathrm{total}} = \log_{10} \frac{S_{\mathrm{ref}}}{S}$$
5. A seguir, a absorvância da amostra é calculada conforme descrito acima utilizando o espectro de correção da fibra óptica A_{fiber} e o espectro de correção de janela A_{window} do respectivo canal:
   $A = A_{\mathrm{total}} - A_{\mathrm{fiber}} - A_{\mathrm{window}}$
   A designa o espectro da amostra.

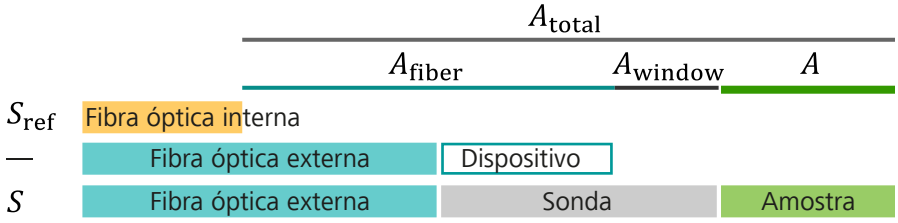


Figura 15 Registrar o espectro de uma amostra

3.3 Testes de desempenho do equipamento

Os Testes de desempenho do equipamento podem ser realizados por meio de caminhos ópticos internos e externos.

- **OMNIS NIR Analyzer**
  - Os **Testes de desempenho do equipamento** internos (obrigatórios): os testes internos usam o caminho de referência da respectiva apresentação de amostras. Os testes verificam os comprimentos de onda e o ruído do sinal.
    Antes dos testes, é necessário executar com sucesso e validar a calibração do comprimento de onda para a respectiva apresentação de amostras.
    Antes de poder registrar espectros com o equipamento nas respectivas apresentações de amostras, os testes internos devem ser executados com sucesso.
  - **Os Testes de desempenho do equipamento** externos (opcionais): os testes externos cumprem a validação de acordo com várias farmacopeias, como USP <856>, Ph.Eur 2.2.40 e JP 2.27. Comprimentos de onda, ruído do sinal e linearidade fotométrica são verificados *(ver capítulo 3.3.1, página 29)*.
- **2060 The NIR**
  Os testes podem usar o caminho óptico interno ou um dos caminhos ópticos externos. Os testes verificam os comprimentos de onda e o ruído do sinal.
  Antes dos testes, é necessário executar com sucesso e validar a calibração do comprimento de onda e a padronização de referência no respectivo canal.
  Os Testes de desempenho do equipamento devem ser executados com sucesso antes de poder registrar espectros no respectivo canal com o equipamento. Tolerâncias predefinidas devem ser respeitadas. As tolerâncias admissíveis dependem da configuração de fibra óptica do respectivo canal informadas nos dados específicos do equipamento (modo de medição, tipo de fibra e comprimento da fibra).

Teste de comprimento de onda

O teste de comprimento de onda examina a exatidão do comprimento de onda e a precisão do comprimento de onda. Para este propósito, é utilizado um padrão de comprimento de onda que possui um espectro de absorção com picos definidos e posições do pico conhecidas:

- **Interno**: o espectro de absorção do padrão de comprimento de onda interno, metrologicamente rastreável, é determinado através do caminho óptico interno:

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

  Nisso, A_{WL} corresponde à absorvância do padrão de comprimento de onda interno, S_{ref} ao sinal medido no caminho de referência interno e $S_{\mathrm{ref,WL}}$ ao sinal medido no caminho de referência interno com o padrão de comprimento de onda interno.
- **Externo** para dispositivos da família de produtos **OMNIS NIR Analyzer**: o teste de comprimento de onda externo é opcional *(ver "Teste de comprimento de onda externo", página 29)*.
- **Externo** para equipamentos do tipo **2060 The NIR**:
  As fibras ópticas externas devem ser conectadas por fibras individuais ao dispositivo de calibração e, com MicroBundle, ao padrão de refletância.
  O espectro de absorção do padrão de comprimento de onda interno, metrologicamente rastreável, é determinado através de um caminho óptico externo:

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

  Nisso, A_{WL} corresponde à absorvância do padrão de comprimento de onda interno, S_{ref} ao sinal medido no caminho de referência interno, $S_{\mathrm{fiber,WL}}$ ao sinal medido no caminho óptico do respectivo canal, sendo que as fibras estão ligadas ao dispositivo de calibração ou ao padrão de refletância, o padrão de comprimento de onda interno no caminho óptico é utilizado e A_{fiber} corresponde ao espectro de correção de fibra óptica da padronização de referência.

A_{WL} também contém a absorvância do padrão de reflexão, mas ela é irrelevante para os cálculos a seguir. É ideal que as posições de pico de A_{WL} sejam idênticas às posições nominais de pico do padrão de comprimento de onda.

A exatidão do comprimento de onda e a precisão do comprimento de onda são testadas da seguinte forma:

1. Uma série de espectros de absorção do padrão de comprimento de onda é registrada conforme descrito acima (A_{WL}).
2. Nos espectros registrados, as posições de pico são identificadas.
3. Para cada posição do pico, as seguintes estatísticas são calculadas sobre os espectros registrados:
   a. Valor médio (unidade: nm)
   b. Desvio padrão (unidade: nm)

4. **Exatidão**: em cada pico, a diferença entre a posição média do pico e a posição do pico nominal deve estar dentro da tolerância predefinida.
   ➨ **Aviso**: as posições nominais de pico podem variar ligeiramente de um teste para outro. Isso se deve às correções de temperatura das posições de pico. Uma correção de temperatura semelhante é executada na calibração do comprimento de onda. Ela contribui para que as medições feitas em uma determinada temperatura forneçam resultados comparáveis em todos os equipamentos.
5. **Precisão**: em cada pico, o desvio padrão deve estar dentro da tolerância predefinida.
6. Se as tolerâncias para todos os picos forem mantidas, o status geral do teste de comprimento de onda é bem-sucedido.

Teste de ruído

O ruído do sinal pode ser verificado internamente ou externamente:

- **Interno**: o ruído é determinado como a absorvância do caminho óptico interno em relação à absorbância de uma outra medição no mesmo caminho óptico:

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

  Nisso, $S_{\text{ref},1}$ e $S_{\text{ref},2}$ correspondem ao sinal medido no caminho de referência interno.
- **Externo** para dispositivos da família de produtos **OMNIS NIR Analyzer**: o teste de ruído externo é opcional *(ver "Testes de ruído externos", página 30)*.
- **Externo** para equipamentos do tipo **2060 The NIR**:
  As fibras ópticas externas devem ser conectadas por fibras individuais ao dispositivo de calibração e, com MicroBundle, ao padrão de refletância.
  O ruído é determinado como a diferença entre o espectro de absorção medido e o espectro de absorção nominal:

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

  Nisso, S_{ref} corresponde ao sinal medido no caminho de referência interno, S_{fiber} ao sinal medido no caminho óptico externo, sendo que as fibras estão ligadas ao dispositivo de calibração ou ao padrão de refletância, A_{fiber} corresponde ao espectro de correção de fibra óptica da padronização de referência e A_{nominal} ao espectro de absorção nominal do padrão de reflexão.
  Aviso: no modo de transmissão, $A_{\text{nominal}} = 0$.

É ideal que $A_{\text{noise}} = 0$.

O teste de ruído executa os seguintes passos:

1. Uma série de espectros de ruído é registrada conforme descrito acima (A_{noise}).
2. Os espectros de ruído são subdivididos em diversos segmentos de comprimento de onda.
3. Para cada espectro de ruído e cada segmento, são calculados 3 grandezas:
    a. ruído fotométrico (unidade: mAU)
    b. ruído pico a pico (unidade: mAU)
    c. viés de linha de base do ruído (unidade: mAU)
4. Para cada um das 3 grandezas em cada segmento, é calculado o valor médio pelos espectros de ruído registrados.
5. Se todos os valores médios estiverem dentro das tolerâncias predefinidas, o status geral do teste de ruído é bem-sucedido.

3.3.1 Testes de desempenho do equipamento externos (OMNIS NIR Analyzer)

Os equipamentos da família de produtos **OMNIS NIR Analyzer** podem ser validados de acordo com farmacopeias como a USP <856>, Ph.Eur 2.2.40 e JP 2.27 (a partir da versão do OMNIS Software 4.4). Esses testes exigem padrões de referência externos e metrologicamente rastreáveis. Os padrões de referência têm espectros de absorção nominais individuais medidos com um instrumento de referência à temperatura ambiente.

Teste de comprimento de onda externo

A exatidão do comprimento de onda e a precisão do comprimento de onda são testadas da seguinte forma:

1. O padrão de comprimento de onda externo (transmissão ou refletância) deve levado à posição.
2. Uma série de espectros de absorção do padrão de comprimento de onda externo é registrada:

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

    Nisso, A_{WL} corresponde à absorvância do padrão de comprimento de onda externo, S_{ref} ao sinal medido no caminho de referência interno e S_{WL} ao sinal medido por meio do padrão de comprimento de onda externo.
3. Nos espectros registrados, as posições de pico são identificadas.
4. Para cada posição do pico, as seguintes estatísticas são calculadas sobre os espectros registrados:
    a. Valor médio (unidade: nm)
    b. Desvio padrão (unidade: nm)
5. **Exatidão**: em cada pico, a diferença entre a posição média do pico e a posição do pico nominal deve estar dentro da tolerância predefinida.

6. **Precisão**: em cada pico, o desvio padrão deve estar dentro da tolerância predefinida.
7. Se as tolerâncias para todos os picos forem mantidas, o status geral do teste de comprimento de onda é bem-sucedido.

Testes de ruído externos

O ruído do sinal é testado uma vez em baixa corrente de fótons (baixo fluxo) e uma vez em alta corrente de fótons (alto fluxo):

1. O padrão de referência externo (transmissão ou refletância) para o teste Low-Flux ou o teste High-Flux deve ser levado à posição.
2. Uma série de espectros de ruído é registrada como a diferença entre os espectros de absorção medidos e o espectro de absorção nominal do padrão de referência:

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

   Nisso, S_{ref} corresponde ao sinal medido no caminho de referência interno, S_{ND} ao sinal medido por meio do padrão de referência externo e A_{nominal} ao espectro nominal do padrão de referência.
3. Os espectros de ruído são subdivididos em diversos segmentos de comprimento de onda.
4. Para cada espectro de ruído e cada segmento, são calculados 3 grandezas:
   a. ruído fotométrico (unidade: mAU)
   b. ruído pico a pico (unidade: mAU)
   c. viés de linha de base do ruído (unidade: mAU)
5. Para cada um das 3 grandezas em cada segmento, é calculado o valor médio pelos espectros de ruído registrados.
6. Se todos os valores médios estiverem dentro das tolerâncias predefinidas, o status geral do teste de ruído é bem-sucedido.

Linearidade fotométrica

O objetivo deste teste é demonstrar uma relação linear entre o grau de refletância (ou grau de transmitância) e a absorvância medida em toda a faixa de comprimento de onda:

1. São registrados espectros de absorção de 5 padrões de referência com um diferente grau de refletância (ou grau de transmitâncias).
2. A relação linear entre o grau de refletância (ou grau de transmitância) e a absorvância medida é garantida por uma regressão linear em vários comprimentos de onda.
3. Se os slopes e interceptações y de todas as retas de regressão estiverem dentro das tolerâncias predefinidas, o status geral do teste será bem-sucedido.

4 Desenvolvimento de modelo

É feita uma distinção entre os seguintes tipos de modelos:

- Um **Modelo de quantificação** descreve a dependência de um parâmetro de interesse (p. ex. teor de água) do espectro registrado das amostras.
- Um **Modelo de identificação** (a partir da versão do OMNIS Software 4.0) classifica as amostras com base nos espectros registrados em diferentes produtos (p. ex. diferentes tipos de grãos de café).
  Um produto designa uma determinada substância química com propriedades físicas específicas (p. ex. tamanho das partículas).

Para a análise de uma amostra desconhecida, é registrado um espectro da amostra. Conforme a aplicação, o espectro é utilizado da seguinte forma:

- Quantificação: com base no espectro, um modelo de quantificação cria uma previsão, p. ex., do teor de água da amostra.
- Identificação: com base no espectro, um modelo de identificação identifica a amostra, p. ex., como um café arábica.
- Verificação (a partir do OMNIS Software versão 4.2): com base no espectro, um modelo de identificação verifica, por exemplo, se a amostra é de café arábica ou não.

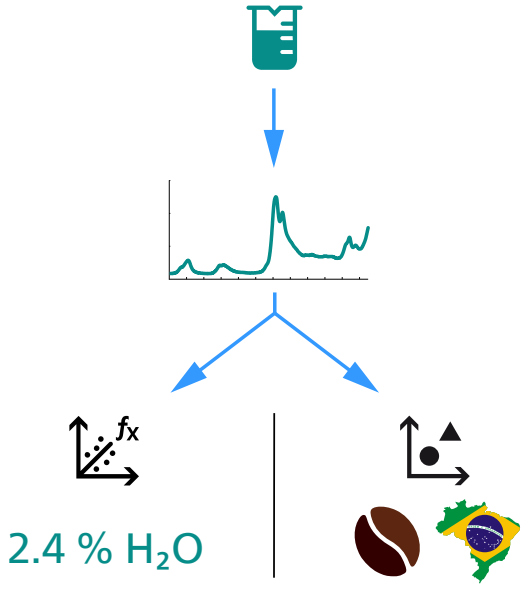


Figura 16 Quantificação (abaixo, à esquerda) e identificação (abaixo, à direita).

O desenvolvimento de um modelo para a análise de amostras compreende os seguintes passos:

1. **Coleta de amostras**
   São coletadas e processadas amostras físicas:
   a. Para cada amostra será registrado um espectro.
   b. Para a quantificação, um valor de referência para o parâmetro de interesse (p. ex., teor de água) será medido com um método de referência (p. ex., titulação). A medição de referência deve ser exata e precisa.
   c. Para a identificação, é necessário saber se a amostra pertence ao produto.
2. **Desenvolvimento de modelo**
   O desenvolvimento de modelo é efetuado em um processo iterativo que compreende os seguintes passos:
   a. Divisão do conjunto de dados em um conjunto de dados de calibração, um conjunto de dados de validação e um conjunto de dados de outliers.
   b. Aplicação do pré-tratamento de dados adequado e faixas de comprimento de onda aos espectros.
   c. Cálculo de um modelo com base no conjunto de dados de calibração.
   d. A validação do modelo assegura que o modelo cumpra os requisitos. A validação é baseada principalmente no conjunto de dados de validação que não foi utilizado para o desenvolvimento do modelo.
      Na quantificação, o modelo prevê o parâmetro de interesse para os espectros no conjunto de dados de validação. A seguir, os valores calculados são comparados com os valores de referência conhecidos.
      Para a identificação, o modelo atribuir os espectros a diversos produtos. A previsão de se as amostras pertencem aos produtos é comparada com os dados efetivos de pertencimento ao produto das amostras.
3. **Monitoramento**
   O monitoramento do modelo garante que a capacidade de previsão não diminua ao longo do tempo. Cada alteração do processo ou da amostra requer uma revalidação.

4.1 Amostras físicas

Para começar, é preciso coletar e analisar amostras físicas. A coleta adequada de amostras é um pré-requisito para o desenvolvimento de um modelo robusto. Para isso, é necessário observar vários aspectos.

Amplitude de variações

As amostras devem incluir variações típicas esperadas futuramente da amostra. As concentrações de todos os componentes químicos e os tamanhos das partículas devem incluir pelo menos a amplitude de variações esperada.

As amostras devem cobrir uma variação apropriada das condições e um período apropriado. Devem ser consideradas todas as variações, p. ex., variações de processos, variações sazonais ou variações das condições ambientais.

As amostras devem ser distribuídas uniformemente por toda a gama de variações. Na quantificação, essa faixa inclui os valores de referência. Se a faixa de valores de referência for de 1% até 10%, as amostras devem ser distribuídas uniformemente entre 1% e 10%.

Amostras de calibração e amostras de validação

Geralmente são utilizados 2 conjuntos de amostras:

- Conjunto de calibração: para as amostras usadas no desenvolvimento de modelos.
- Registro de dados de validação: para as amostras usadas na validação do modelo.

Ambos os conjuntos devem cobrir as variações esperadas. Para o registro de dados de validação, devem ser coletadas preferencialmente amostras independentes para verificar a robustez do modelo. Fatores como operação por diferentes pessoas, diferentes fornecedores ou diferentes equipamentos devem ser consideradas.

Se for impraticável coletar amostras para calibração e validação, uma alternativa possível é distribuir os espectros das amostras disponíveis em um conjunto de dados de calibração e um conjunto de dados de validação. Para garantir a independência dos conjuntos, deve ser utilizado o algoritmo de distribuição automático.

Assim que um modelo tiver sido desenvolvido, é possível considerar a utilização de amostras de outliers explícitas para verificar o reconhecimento de outliers.

Todas as amostras devem ser tratadas da mesma forma. Para o registro de espectros, deve ser utilizado o mesmo método com a mesma configuração de hardware e os mesmos parâmetros. Para a quantificação, deve

ser utilizado o mesmo método de referência com os mesmos parâmetros de medição.

Número de amostras

Quanto mais variações em condições, componentes químicos ou tamanhos das partículas forem incluídos, maior será a quantidade de amostras necessárias.

Quantificação: para que a análise estatística funcione corretamente, é necessário um número mínimo de pelo menos 50, sendo que o conjunto de calibração e o registro de dados de validação devem conter pelo menos de 20 a 25 amostras, cada um.

Identificação e verificação: para cada produto, as amostras devem incluir todas as variações esperadas. Os produtos podem ter um número diferente de amostras, a quantidade mínima é 3.

Com pelo menos 10 a 20 amostras (dependendo da quantidade de variações) é possível desenvolver um primeiro modelo sem registro de dados de validação. Se a validação cruzada (para modelos de quantificação) ou a validação interna (para modelos de identificação) indicarem que um modelo adequado pode ser criado, devem ser coletadas mais amostras de calibração e de validação para o desenvolvimento do modelo final.

Réplicas

Algumas vezes, há uma quantidade muito pequena de amostras em uma determinada área de determinação ou valores de referência. Para compensar isso, é possível tentar replicar essas amostras. Mas isso gera problemas. Se houver réplicas de uma amostra tanto no conjunto de calibração quanto no registro de dados de validação, as figuras de mérito podem levar a erros (excessivamente otimistas). Também devem ser evitadas réplicas de amostras no mesmo conjunto.

Método de referência (quantificação)

Na quantificação, é utilizado um método de referência para medição dos valores de referência. O **erro padrão do laboratório (SEL)** para o método de referência utilizado desempenha um papel importante no desenvolvimento de um modelo de quantificação. O SEL é o desvio padrão das diferenças entre as medições de amostras replicadas.

Frequentemente, o SEL é o maior erro que contribui para o erro padrão da predição (SEP) no método NIR *(ver "SEP – Erro padrão da predição", página 71)*. O SEL não deve superar o SEP necessário em pelo menos 0,7 vezes, preferencialmente em 0,5 vezes. A faixa dos valores de referência deve ser pelo menos 3 vezes, preferencialmente 5 vezes o SEL.

Uma possibilidade de redução do SEL é executar medições de referência repetidas vezes em cada amostra. O valor médio dos valores medidos deve ser definido como valor de referência para a amostra. Para cada

amostra, deve ser executada a mesma quantidade de medições de referência. As figuras de mérito são expressas relativamente a uma quantidade determinada de repetições de medições de referência. Uma quantidade diferente de repetições de medições de referência levaria a valores estimados incorretos das figuras de mérito e, portanto, deve ser evitada.

Temperatura de amostra

A temperatura de amostra influencia substancialmente os espectros de líquidos, água ou de outras pontes de hidrogênio. Os espectros de outros líquidos polares também podem ser influenciados, como por exemplo os espectros de sólidos que contiverem água, umidade ou solventes. Essas amostras devem ser medidas a uma temperatura definida.

Outliers

Algumas amostras podem ser definidas posteriormente como outliers. Um outlier é uma amostra que, por alguma razão, diferencia-se da maioria das amostras. Para evitar que outliers influenciem negativamente o modelo, as amostras identificadas como outliers não são incluídas no cálculo do modelo.

Há diversos tipos de outliers:

- **Outliers espectrais**
  Se o espectro registrado de uma amostra for diferente da maioria dos outros espectros, a amostra pode ser reconhecida como outlier espectral *(ver capítulo 4.3.3, página 51)*.
- **Outlier de valor de referência (quantificação)**
  Na quantificação, alguns valores de referência podem indicar anomalias e serem posteriormente reconhecidos como outliers de valor de referência *(ver capítulo 4.3.4, página 58)*.
  O OMNIS Model Developer (OMD) reconhece as respectivas amostras como outliers com base na detecção de outliers conforme ASTM D8321-22 *(ver capítulo 4.4.3, página 72)*.

4.2 Análise de componentes principais (PCA)

Os dados espectroscópicos das amostras de calibração contêm uma grande quantidade de variáveis (comprimentos de onda). As variáveis apresentam fortes correlações umas com as outras. Portanto, os dados são altamente redundantes. Para lidar com dados assim, são utilizados modelos de variáveis latentes, como PCA e PLS.

A **análise de componentes principais** (**PCA**, em inglês *principal component analysis*) concentra-se nos espectros, sem considerar os valores de referência.

O OMNIS Software utiliza a PCA durante o desenvolvimento de modelo para a divisão do conjunto de dados automática e para o reconhecimento de outliers espectrais.

Passos preparatórios

Os seguintes passos preparatórios são necessários:

1. **Pré-tratamento de dados**: o OMNIS Software aplica o pré-tratamento de dados definido aos espectros *(ver capítulo 4.3.1, página 40)*.
2. **Faixas de comprimento de onda**: o OMNIS Software aplica a seleção de comprimento de onda definida aos espectros *(ver capítulo 4.3.2, página 49)*.
3. **Centralização de valor médio**: para cada comprimento de onda, o valor de absorbância médio é calculado e subtraído de cada valor, em cada espectro.

Os primeiros componentes principais

Após os passos preparatórios, o PCA organiza novamente as informações nos dados de espectro e separa os dados relevantes do ruído. Para isso, o PCA transforma as variáveis de comprimento de onda em um novo espaço de variáveis, nos chamados **componentes principais** (**PC**, em inglês "Principal Components").

O PCA converte as informações relevantes a partir de diversas variáveis de comprimento de onda em somente alguns componentes principais. Para ilustrar esse conceito com base em um exemplo simples, vamos supor que existam somente 2 variáveis de comprimento de onda ao invés de milhares delas e essas 2 variáveis serão reduzidas a 1 componente principal.

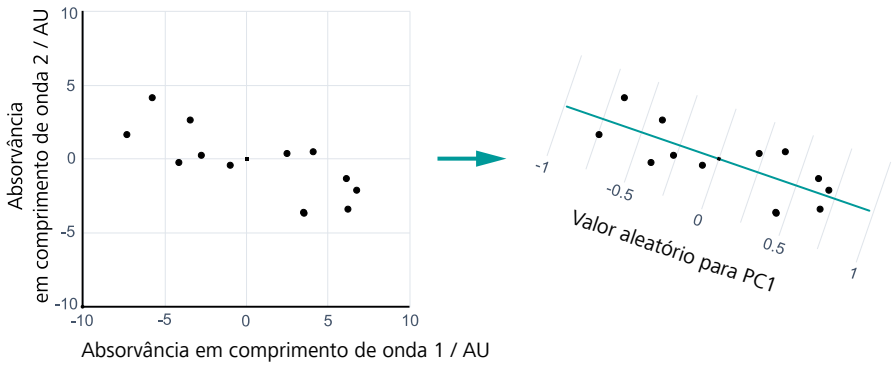


Figura 17 Pontos que representam os espectros em um espaço de comprimento de onda de 2 dimensões (esquerda). Os mesmos pontos em um espaço de componente principal com 1 dimensão (direita).

Na *figura 17* à esquerda, os eixos horizontais e verticais compõem o espaço de comprimento de onda original com 2 variáveis. Cada ponto

representa um espectro com somente 2 comprimentos de onda. O valor médio de todos os valores de comprimento de onda constitui o ponto zero.

À direita, a direção dos dados que explica a variância máxima constitui o componente principal PC1. Neste exemplo, PC1 é a única variável do espaço de componente principal. Consequentemente, as 2 variáveis originais são reduzidas a 1 variável.

Scores e resíduos

A figura 18 exibe as grandezas que caracterizam um espectro *i*:

- A distância s_i a partir do centro, medida no espaço de componente principal. No exemplo, s_i é medido com somente 1 componente principal na direção de PC1. A distância s_i é designada como o **score** do espectro *i*.
- O offset e_i do espaço de componente principal para o espectro. A distância e_i é designada como o **resíduo** do espectro *i*.

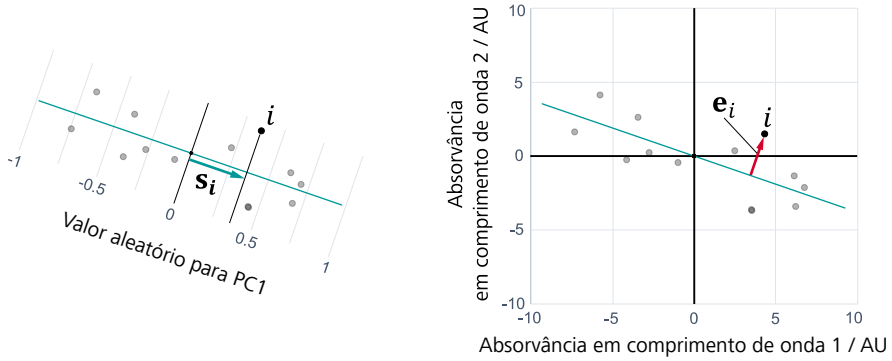


Figura 18 Espectro i com score (esquerda) e resíduo (direita).

O score s_i é medido no espaço de componente principal. O resíduo e_i é medido no espaço de comprimento de onda original.

Conversão em vários componentes principais

Geralmente é necessário mais de um componente principal para uma descrição adequada dos dados espectroscópicos.

Na *figura 19* há 3 variáveis originais x_1, x_2, x_3. Cada ponto representa um espectro com 3 comprimentos de onda.

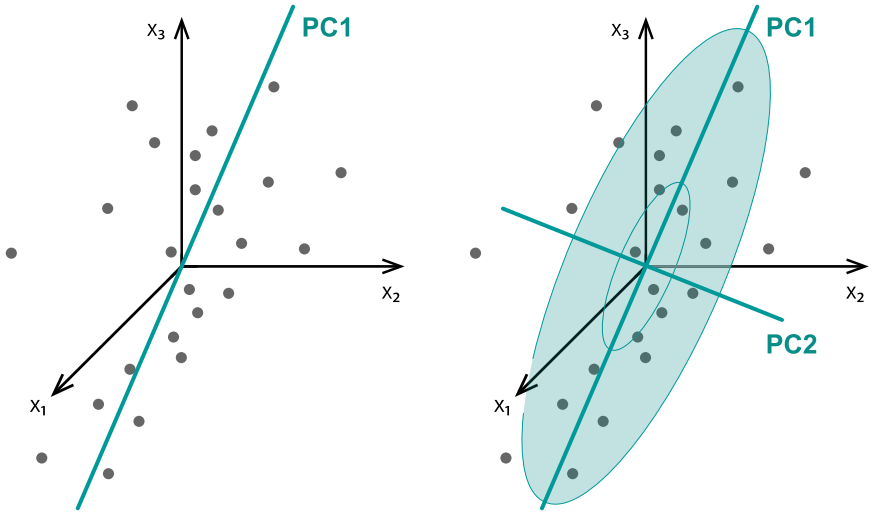


Figura 19 A 3 variáveis originais serão reduzidas a 1 componente principal (esquerda) ou 2 componentes principais (direita). PC1 e PC2 constituem um espaço de componente principal com 2 dimensões.

O primeiro componente principal PC1 está novamente na direção dos dados que explicam a variância máxima.

O segundo componente principal PC2 está na direção dos dados que explicam a variância máxima restante. Isso também é válido para todos os próximos componentes principais que descrevem a variância máxima restante. Portanto, os primeiros componentes principais são responsáveis pela maior parte da variância nos dados, enquanto os outros contêm principalmente ruídos e podem ser descartados. Desta forma, é possível reduzir a quantidade de variáveis.

Uma característica essencial do PCA é que todos os componentes principais são **ortogonais** (em um ângulo reto) em relação uns aos outros. Portanto, os scores não são correlacionados.

Distância de Mahalanobis

Conforme apresentado acima, o score de um espectro *i* é medido no espaço de componente principal, enquanto o resíduo é medido no espaço de comprimento de onda original.

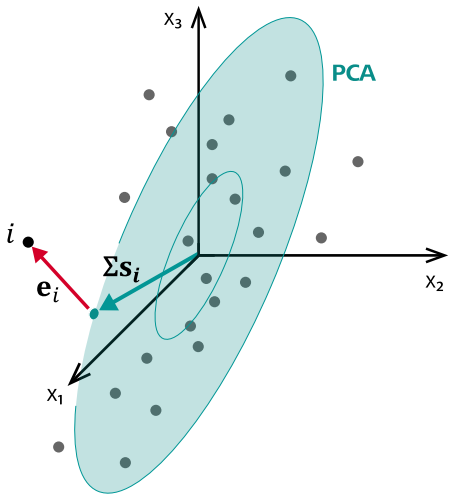
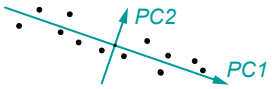


Figura 20 Score e resíduo do espectro i. O ponto verde é a projeção ortogonal do ponto i (que representa o espectro i) no espaço de componente principal.

Na *figura 20*, o vetor de score $\mathbf{\Sigma s}_i$ representa a distância absoluta (distância euclidiana) do ponto médio do modelo PCA em relação à projeção ortogonal do espectro no espaço de componente principal.



No exemplo, as distâncias euclidianas dos espectros na direção de PC1 são maiores do que na direção de PC2. A distribuição pode ser medida como **variância**. A variância em PC1 é maior do que em PC2.

O vetor de score normalizado $\mathbf{s}_i$ representa uma distância normalizada, a chamada **distância de Mahalanobis**. A distância de Mahalanobis considera a variância diferente nas diversas direções de componentes principais. Cada direção recebe a mesma ponderação. Portanto, uma pequena distância euclidiana em uma direção com pouca variância pode contar tanto quanto uma grande distância euclidiana em uma direção com maior variância.

Conversão de espectros com diversos comprimentos de onda

Os mesmos conceitos são válidos para a conversão de espectros com diversas variáveis de comprimento de onda em componentes principais. Na *figura 21*, cada espectro é representado por uma curva (esquerda) e um ponto (direita).

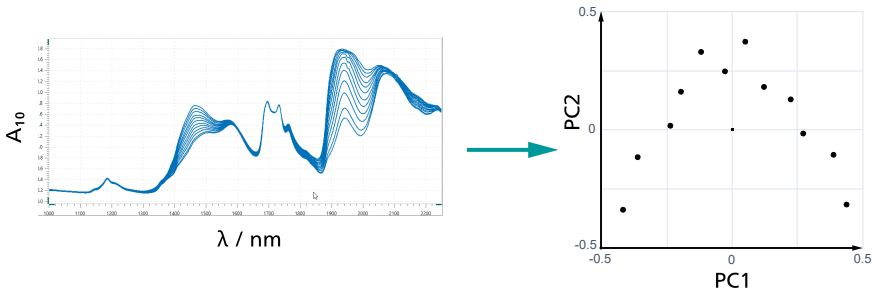


Figura 21 Conversão de dados de espectro em um espaço de componente principal. Os scores no lado direito são expressos em unidades aleatórias.

A figura à direita exibe os primeiros 2 componentes principais PC1 e PC2. Da mesma forma, os componentes principais seguintes PC3, PC4, etc. também podem ser visualizados.

Um modelo PCA utiliza uma quantidade fixa de componentes principais. Quanto mais componentes principais, mais variações espectrais relevantes o modelo explica. Ao mesmo tempo, o modelo também abrange mais variações espectrais irrelevantes (ruído). É necessário encontrar um equilíbrio.

Quando o OMNIS Software executa uma análise de componentes principais, a quantidade de componentes principais é selecionada de forma que a variância explicada seja de pelo menos 95%.

Algoritmo PCA

Há diversas possibilidades de converter os dados originais em um espaço de componente principal. O OMNIS Software executa uma decomposição em valores singulares *(ver capítulo 6.2, página 94)*.

4.3 Preparação de dados

4.3.1 Pré-tratamento de dados

Modelos espectroscópicos dependem da relação entre os valores de absorvância e os parâmetros de interesse (quantificação) ou o pertencimento a um produto (identificação, verificação). A **parametrização** dos espectros assegura que os espectros expressem da melhor forma possível essa relação. O objetivo é eliminar variâncias irrelevantes sem as perder informações importantes. Os artefatos e não linearidades serão corrigidos. Uma parametrização executada corretamente aumenta a exatidão e a robustez do modelo, assim como sua reprodutibilidade e a reprodutibilidade das previsões.

A parametrização será utilizada no conjunto de dados de calibração, no conjunto de dados de validação, no conjunto de dados de outliers e

em todas as futuras amostras desconhecidas que serão analisadas com o mesmo modelo.

O primeiro passo da parametrização é o **pré-tratamento de dados**. O pré-tratamento de dados é efetuado na sequência predefinida. No segundo passo da parametrização, as faixas de comprimento de onda relevantes podem ser definidas *(ver capítulo 4.3.2, página 49)*.

Minimização de ruídos

Os espectros podem conter diversos tipos de variações aleatórias ao redor do sinal. Alguns exemplos são ruídos de frequência elevada ocasionados pelo detector e os circuitos eletrônicos do equipamento ou ruídos de frequência baixa causados pelo desvio do equipamento durante as medições de leitura.

O espectrômetro fornece um espectro médio estimado a partir de uma série de medições individuais. Isso permite reduzir claramente ruídos de frequência elevada. Outra forma de minimizar ruídos é utilizar um filtro de alisamento. Esses filtros são baseados na ideia de que os ruídos têm frequência elevada e o sinal tem frequência baixa. Eles fazem uma aproximação do sinal utilizando os valores de absorvância próximos e reduzem o ruído por meio do cálculo de valor médio.

Correção de dispersão

A dispersão designa uma alteração na direção da luz causada pela interação com a amostra. A luz dispersada que não alcança o detector leva a variações de linha de base nos espectros.

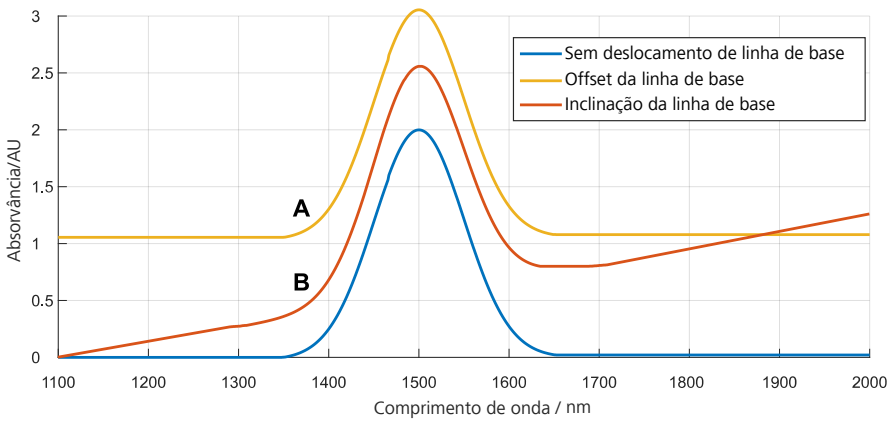


Figura 22 Os tipos mais importantes de deslocamento da linha de base são o offset da linha de base e a inclinação da linha base.

É possível diferenciar entre diversos tipos de deslocamentos de linha de base:

- Um fator aditivo constante que leva a um **offset da linha base** (espectro **A**).

- Um fator multiplicativo dependente do comprimento de onda que leva a uma **inclinação da linha base** (espectro **B**).
- Um fator multiplicativo de segunda ordem que leva a uma **inclinação quadrada da linha base**.
  Também podem ocorrer inclinações de linha de base de ordens maiores.
- Fatores multiplicativos dependentes da absorvância que levam a uma **intensificação**. Mas as intensificações são irrelevantes.

A dispersão é mais perceptível em amostras de sólidos. Os deslocamentos da linha de base resultantes podem ser utilizados para reconhecer alterações nos tamanhos das partículas ou outras alterações físicas.

Mas, se a intenção for reconhecer variações químicas, os deslocamentos de linha de base devem ser minimizados por meio de um pré-tratamento de dados adequado.

Pré-tratamentos de dados

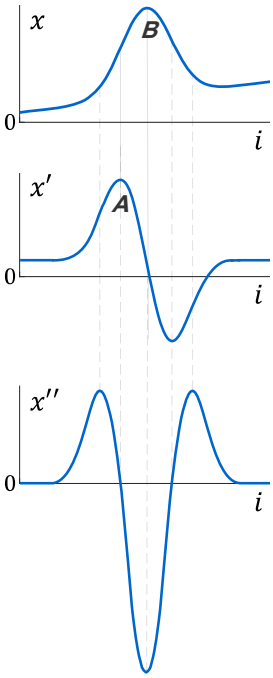
Cada um dos pré-tratamentos de dados a seguir é aplicado a um único espectro. Não são incluídos outros espectros no cálculo.

O pré-tratamento de dados altera os valores de sinais. Os números no eixo y não têm mais significado.

Os pré-tratamentos de dados aplicados são transformações lineares. Portanto, continua a ser válida a lei de Beer-Lambert.

4.3.1.1 Derivadas

Podem ser executadas derivadas com o filtro Gap-Segment ou com o filtro Savitzky-Golay.



A derivada de um espectro descreve o slope ou o quanto a curva aumenta em cada ponto. O slope é a taxa de alteração do espectro de saída.

No espectro, x_i é a absorvância no comprimento de onda *i*. A derivada de primeira ordem x_i' apresenta o slope do espectro no comprimento de onda *i*. No ponto mais alto do espectro de saída, a derivada de primeira ordem tem o máximo (**A**). No ponto do espectro de saída com um pico (**B**), a derivada de primeira ordem é igual a 0.

A derivada de primeira ordem elimina offsets da linha de base e converte inclinações da linha de base em offsets da linha base.

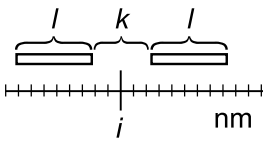
A derivada de segunda ordem x_i'' corresponde ao slope da derivada de primeira ordem no comprimento de onda *i*. Picos positivos no espectro original (**B**) se tornam picos negativos e vice-versa.

A derivada de segunda ordem elimina offsets da linha de base e inclinações da linha base do espectro original.

Recomenda-se ter cuidado quando o espectro tiver uma quantidade muito grande de ruídos. Cada derivada piora significativamente a relação sinal-ruído. Por esta razão, as derivadas são combinadas com uma função de alisamento no filtro Gap-Segment ou no filtro Savitzky-Golay.

4.3.1.2 Gap-Segment

O filtro Gap-Segment alisa o espectro. Opcionalmente, o filtro Gap-Segment executa uma derivada de primeira ordem ou de segunda ordem. O cálculo depende de se as derivadas serão ou não utilizadas:



- **Ordem de derivação 0**: para cada comprimento de onda *i*, o filtro Gap-Segment calcula o valor médio de 2 segmentos com o tamanho de segmento *l*, p. ex. 10 nm. Os 2 segmentos são separados por uma distância *k*, p. ex. 5 nm.
- **Ordem de derivada 1**: para a derivada de primeira ordem, os valores médios dos 2 segmentos são calculados separadamente . A seguir, é calculada a diferença entre os dois segmentos.
- **Ordem de derivada 2**: a derivada de segunda ordem pode ser calculada da mesma forma que a derivada de primeira ordem.

No início e no final do espectro, são calculados *l* + *k*/2 comprimentos de onda utilizando os valores de zero para comprimentos de onda de segmentos fora do espectro.

No início e no fim do espectro, os valores de zero são utilizados para os comprimentos de onda de segmentos fora do espectro.

O alisamento pode ser acompanhado por um ligeiro deslocamento do pico e um viés.

Configuração de parâmetros

Um alisamento mais intenso é obtido por:

- uma derivada de menor ordem,
- um tamanho de segmento maior,
- uma distância entre segmentos maior.

Um alisamento excessivo leva a uma perda de variância relevante que diminui a capacidade de previsão do modelo.

4.3.1.3 Savitzky-Golay

Assim como o filtro Gap-Segment, o filtro Savitzky-Golay também alisa o espectro e executa opcionalmente uma derivada de primeira ou de segunda ordem. Mas o filtro Savitzky-Golay utiliza outro método de alisamento.

Para cada comprimento de onda *i*, o filtro Savitzky-Golay adapta um polinômio de menor ordem na faixa do respectivo comprimento de onda. O valor do polinômio no comprimento de onda *i* é o valor alisado. Se for efetuada uma derivada, o valor da derivada é utilizado.

Uma soma ponderada dos valores adjacentes calcula tudo novamente:

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

Neste cálculo, *k* corresponde à largura de filtro, c_j ao coeficiente de dobra dependente da ordem de derivação, do grau de polinômio e da largura do filtro e pode ser consultado em tabelas, e x_{i+j} corresponde ao valor de absorvância do espectro de saída no comprimento de onda *i+j*.

No início e no fim do espectro, os valores de extrapolados (extrapolação horizontal) são utilizados para os comprimentos de onda de filtro fora do espectro.

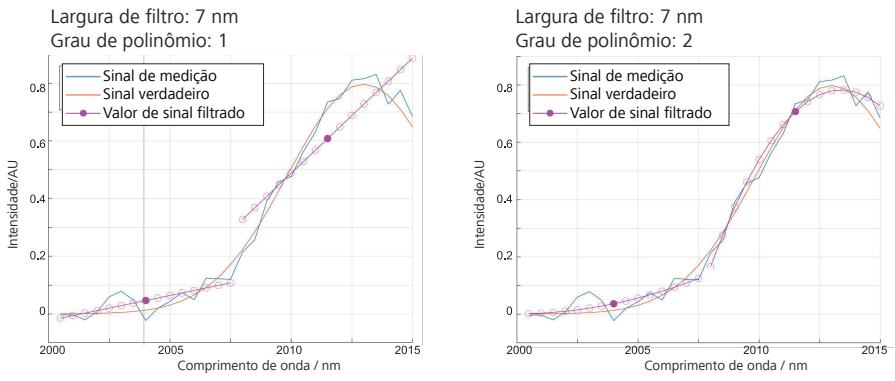


Figura 23 Filtro Savitzky-Golay com diferentes graus de polinômios

A figura 23 ilustra o filtro Savitzky-Golay. A largura de filtro é 7 nm. Os polinômios para os comprimentos de onda 2.004 nm e 2.011,5 nm são apresentados. Os novos valores são os pontos preenchidos ou, se forem utilizadas derivadas, a derivada dos pontos. Todos os outros comprimentos de onda serão tratados da mesma forma.

A largura de filtro define a faixa de comprimento de onda em que cada polinômio é adequado. A dobra é ponderada de forma que a influência dos valores de absorvância seja diminuída para ambos os lados do respectivo comprimento de onda.

Configuração de parâmetros

Um alisamento mais intenso é obtido por:

- uma derivada de menor ordem,
- uma largura de filtro maior,
- um grau de polinômio menor.

Um alisamento excessivo leva a uma perda de variância relevante que diminui a capacidade de previsão do modelo.

4.3.1.4 SNV – Standard Normal Variate

O SNV normaliza um espectro individual para a variância 1 e o valor médio 0. O SNV normaliza os valores de absorvância x_i para cada comprimento de onda i dentro de uma faixa de comprimento de onda definida da seguinte forma:

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

Neste cálculo, m corresponde ao valor médio e s ao desvio padrão de todos os valores de absorvância dentro da faixa de comprimento de onda definida.

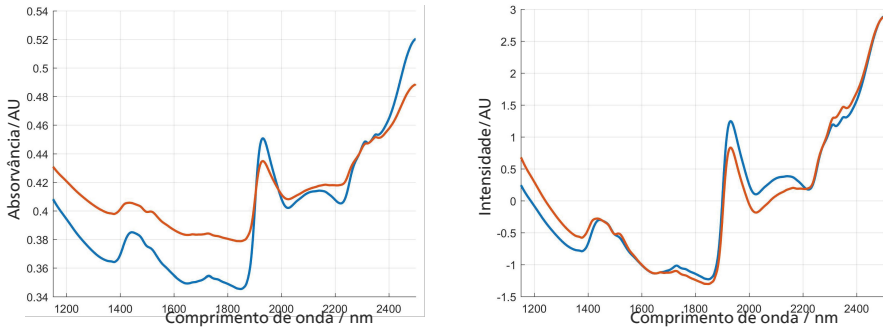


Figura 24 Espectros de absorção (esquerda) e espectros tratados com SNV (direita).

Por meio da normalização, a variância entre os espectros é eliminada. Isso é importante caso a variância seja resultante de propriedades que não sejam de interesse, p. ex., diferentes espessuras de camadas em amostras granulares, em pó ou em meios turvos.

Aviso: se após o SNV for utilizada uma derivada, a variância eliminada pode voltar parcialmente. Portanto, as derivadas devem ser utilizadas antes do SNV. Se o SNV precisar ser executado antes da derivada em casos excepcionais, é recomendável considerar a seguinte sequência:

detrend, SNV, derivada. Mas a sequência usual é: derivada, SNV, detrend *(ver capítulo 4.3.1.7, página 48)*.

Parâmetros

- **Faixas de comprimento de onda**
  Se artefatos prejudicarem determinadas faixas de comprimento de onda (p. ex., por saturação ou ruídos intensos), essas áreas podem ser excluídas.
  Apenas as faixas de comprimento de onda definidas são levadas em consideração no cálculo do valor médio e do desvio padrão. A normalização subsequente é realizada tanto nas faixas de comprimento de onda definidas quanto nas áreas intermediárias. Para comprimentos de onda excluídos no início do espectro, adotar o valor normalizado do comprimento inicial de onda adjacente. Para comprimentos de onda excluídos no final do espectro, adotar o valor normalizado do comprimento final de onda adjacente.
  Se necessário, comprimentos de onda excluídos também podem ser excluídos do cálculo do modelo *(ver capítulo 4.3.2, página 49)*.
  Aviso: a partir da versão 4.6 do OMNIS Software, é possível definir diversas faixas de comprimento de onda.

4.3.1.5 Detrend

O detrend utiliza o método do menor quadrado para adaptar um polinômio de segunda ordem a todo o espectro. A seguir, o detrend subtrai o polinômio do espectro.

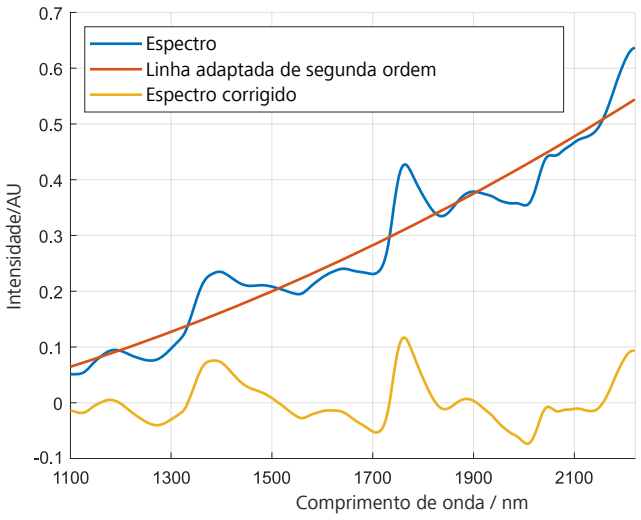


Figura 25 O detrend converte o espectro azul no espectro amarelo.

O detrend reduz efeitos de dispersão dependentes de comprimentos de onda até inclinações quadradas da linha base.

A figura acima exibe um espectro (azul) dominado pela tendência (Trend). Se essa tendência dominante for semelhante em todos os espectros, o

detrend pode funcionar bem. Em outros casos, o detrend pode remover variações úteis. Nestes casos, provavelmente as derivadas serão uma opção melhor.

Como é feita a adaptação de um polinômio próprio para cada espectro, pode ocorrer uma variância adicional de interferência. Normalmente, o SNV é utilizado antes do detrend. Isso proporciona valores estimados mais robustos dos coeficientes de polinômio.

Parâmetros

- **Faixas de comprimento de onda**
  Se artefatos prejudicarem determinadas faixas de comprimento de onda (p. ex., por saturação ou ruídos intensos), essas áreas podem ser excluídas.
  Um polinômio é ajustado a todos os valores de intensidade das faixas de comprimento de onda definidas. Em seguida, o polinômio é subtraído do espectro em todas as faixas de comprimento de onda definidas. Definir o valor de intensidade como zero para todos os comprimentos de onda excluídos.
  Se necessário, comprimentos de onda excluídos também podem ser excluídos do cálculo do modelo *(ver capítulo 4.3.2, página 49)*.
  Aviso: a partir da versão 4.6 do OMNIS Software, é possível definir diversas faixas de comprimento de onda.

4.3.1.6 Visão geral dos pré-tratamentos de dados

Pré-tratamento	Finalidade	Efeitos positivos	Efeitos negativos
Gap-Segment	Alisamento Um alisamento mais intenso é alcançado com uma ordem de derivação menor, um tamanho de segmento maior ou uma distância entre segmentos maior.	▪ Reduz ruídos de frequência elevada.	▪ Um alisamento excessivo leva a uma perda de variâncias relevantes.
Derivadas com Gap-Segment	Correção de linha de base	▪ Derivada de primeira ordem: elimina offsets da linha base. ▪ Derivada de segunda ordem: elimina offsets da linha base e inclinações da linha de base.	▪ Amplifica o ruído. ▪ Altera a aparência do espectro.

Pré-tratamento	Finalidade	Efeitos positivos	Efeitos negativos
Savitzky--Golay	Alisamento Um alisamento mais intenso é alcançado com uma ordem de derivação menor, uma largura de filtro maior ou um grau de polinômio menor.	▪ Reduz ruídos de frequência elevada.	▪ Um alisamento excessivo leva a uma perda de variâncias relevantes.
Derivadas em Savitzky--Golay	Correção de linha de base	▪ Derivada de primeira ordem: elimina offsets da linha base. ▪ Derivada de segunda ordem: elimina offsets da linha base e inclinações da linha de base.	▪ Amplifica o ruído. ▪ Altera a aparência do espectro.
SNV – Standard Normal Variate	Correção de dispersão *	▪ Elimina offsets da linha base.	▪ Variâncias relevantes podem ser removidas.
Detrend	Correção de dispersão *	▪ Elimina offsets da linha base. ▪ Elimina inclinações da linha base e inclinações quadradas da linha base.	▪ Variâncias relevantes podem ser removidas. ▪ Podem ocorrer variâncias irrelevantes.

* Aviso: faixas de comprimento de onda com artefatos (p. ex. saturação ou ruído intenso) devem ser excluídas.

4.3.1.7 Sequências de vários passos de pré-tratamento de dados

Ao utilizar diversos passos de pré-tratamento de dados, a ordem pode ser decisiva. A regra fundamental é a seguinte.

Preferencialmente, Gap-Segment ou Savitzky-Golay devem ser utilizados antes de SNV e SNV antes de detrend.

Exemplo com uma derivada de terceira ordem e SNV: a derivada de primeira ordem converte inclinações da linha de base em offsets da linha base. Uma SNV subsequente elimina esses offsets. Se a sequência for invertida, a SNV não altera as inclinações da linha base. A derivada de primeira ordem subsequente converte as inclinações em offsets da linha base. Os offsets permaneceriam.

Exemplo com uma derivada de segunda ordem e SNV: offsets da linha de base e inclinações da linha base são eliminadas em cada caso. A aplicação da derivada de segunda ordem e de SNV na sequência correta possibilita a eliminação de inclinações die quadradas da linha base. A derivada de segunda ordem converte inclinações quadradas da linha base em offsets da linha de base. Uma SNV subsequente elimina esses offsets. Se a sequência for invertida, a SNV não altera a inclinação quadrada da linha base. A derivada de segunda ordem subsequente converte as inclinações em offsets da linha base. Os offsets permaneceriam.

4.3.2 Faixas de comprimento de onda

Após o pré-tratamento de dados *(ver capítulo 4.3.1, página 40)*, é efetuado o segundo passo da parametrização: a seleção das faixas de comprimento de onda possibilita a exclusão de áreas que não são adequadas para a finalidade. Principalmente faixas de comprimento de onda com ruídos ou saturações podem prejudicar os cálculos subsequentes e devem ser excluídas.

Ruído

Ruídos ocorrem em elevados valores de absorvância, quando somente uma pequena quantidade de luz alcança o detector. A seguinte figura mostra faixas com ruídos.

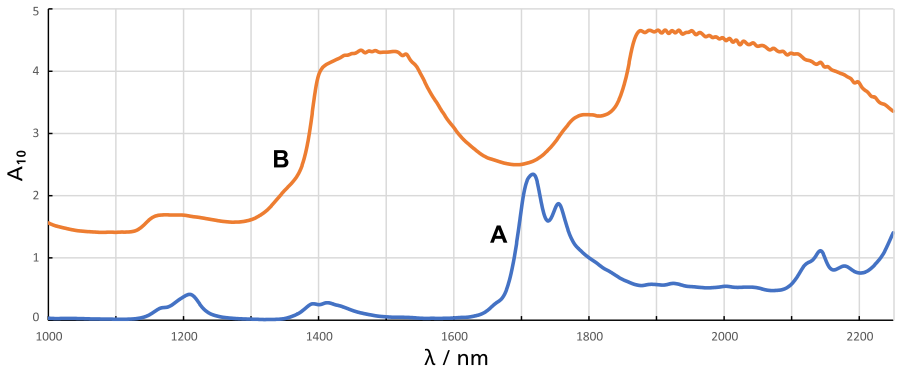


Figura 26 Exemplo com faixas de comprimento de onda com ruídos.

O espectro **A** indica formas do pico normais. O espectro **B** tem 2 faixas com ruídos: uma de 1.400 até 1.550 nm e uma acima de 1.870 nm. As faixas têm ruídos intensos e não são semelhantes a uma curva de sino ou a uma combinação de várias destas curvas.

Saturação

Pode ocorrer uma saturação quando uma grande quantidade de luz atingir o detector, ou seja, com valores de absorvância baixos.

- **OMNIS NIR Analyzer**
  O tempo de integração sempre é ajustado automaticamente. Isso evita saturação e minimiza o ruído *(ver "Tempo de integração", página 9)*.

- **2060 The NIR**
  Se o tempo de integração automático estiver ativado, não ocorre saturação.
  Se o tempo de integração manual estiver ativado, podem ocorrer saturações do detector em longos tempos de integração. Áreas saturadas ocorrem com valores de absorvância baixos, mas nem sempre são fáceis de reconhecer visualmente. Portanto, o tempo de integração manual deve ser ajustado com uma margem suficiente *(ver "Tempo de integração", página 9)*.

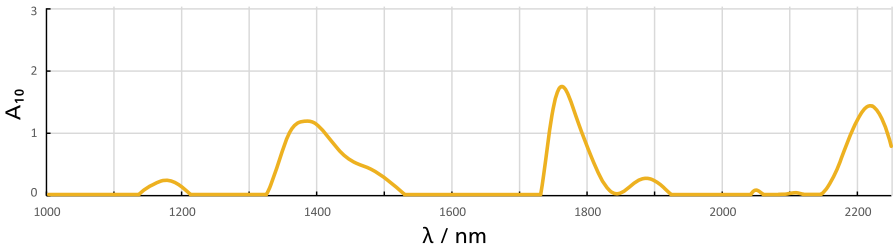


Figura 27 Exemplo com faixas de comprimento de onda saturadas.

Outras razões

Há também outras razões para incluir ou excluir faixas de comprimento de onda. A seleção pode ser relacionada a conhecimentos disponíveis sobre os parâmetros de interesse e suas respectivas faixas de absorção *(ver capítulo 2.1, página 3)*. Ao selecionar, é necessário considerar que as informações relevantes podem ser deslocadas para outras faixas de comprimento de onda conforme o tipo do pré-tratamento.

Se anomalias tiverem sido introduzidas no início e no fim dos espectros durante o pré-tratamento de dados, os respectivos comprimentos de onda podem ser excluídos.

Variações em componentes químicos ou nas condições ambientais podem influenciar determinadas faixas de comprimento de onda. A exclusão dessas faixas de comprimento de onda pode aprimorar a robustez do modelo.

Ao excluir faixas de comprimento de onda com formatos bem definidos, é necessário ter cuidado. Áreas que aparentemente não contêm informações podem se revelar decisivas por fornecerem importantes informações ocultas. Elas podem ser úteis para reconhecer outliers e no processamento de bandas de absorção causadoras de interferência. E as bandas de absorção causadoras de interferência são a razão principal para a execução de medições com variâncias múltiplas *(ver capítulo 6.1, página 90)*.

4.3.3 Outliers espectrais

Um espectro diferente da maioria dos outros espectros é denominado "outlier espectral".

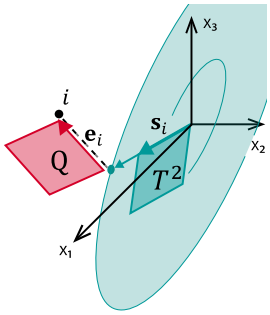
Os outliers devem ser verificados cuidadosamente. O modelo pode ser deformado por um outlier causado por uma amostra contaminada ou um erro de medição, por exemplo. Nesses casos, o outlier não deve ser considerado no cálculo do modelo.

Por outro lado, um outlier pode representar propriedades que não são cobertas de maneira adequada por outros espectros. Nesses casos, o outlier até aprimora o modelo. Se o outlier parecer ser uma amostra válida, é necessário verificar se as amostras de calibração podem ser distribuídas uniformemente pela amplitude de variação.

Medidas apropriadas para o reconhecimento de outliers são Hotelling T^2 e resíduos Q.

Hotelling T² e resíduos Q

Ao converter dados espectroscópicos em um espaço de componentes principais, os espectros podem ser caracterizados por seus scores e resíduos*(ver capítulo 4.2, página 35)*. Isso também é válido para a conversão em um espaço de variáveis latentes *(ver capítulo 4.4.1, página 61)*.



Exemplo: um espaço de comprimento de onda com 3 dimensões (x_1, x_2, x_3) é convertido em um espaço com 2 dimensões (verde). O ponto *i* representa o espectro *i* e é projetado a partir do espaço de 3 dimensões no espaço de 2 dimensões. A partir disso, obtém-se:

- um vetor de score $\Sigma\mathbf{s}_i$ dentro do espaço de 2 dimensões ou seu vetor de score normalizado $\mathbf{s}_i$, que representa a distância de Mahalanobis.
- um resíduo $\mathbf{e}_i$ dentro do espaço de 3 dimensões.

A partir de $\mathbf{s}_i$ e $\mathbf{e}_i$, podem ser derivadas as seguintes grandezas *(ver capítulo 6.4, página 97)*:

- **Hotelling T^2** ou T^2 é a distância de Mahalanobis ao quadrado, ou seja, o quadrado da distância normalizada do ponto central do modelo em relação à projeção ortogonal do espectro no espaço de componentes principais ou no espaço das variáveis latentes.
  Se todos os scores de um espectro corresponderem ao valor médio, T^2 é igual a 0 e o espectro está no ponto central do modelo. Quanto mais próximo do ponto central, melhor é o modelo.
  Em distâncias grandes em relação ao ponto central, o modelo provavelmente não é bom. Os valores T^2 são altos. Um valor T^2 alto indica um espectro extremo que, p. ex., representa uma amostra com uma composição extrema dos componentes químicos.

- O **resíduo Q** é o resíduo quadrado, p. ex., a distância ortogonal quadrada do espectro em relação ao espaço de componentes principais ou ao espaço de variáveis latentes.
  Os resíduos Q mostram variações que não podem ser esclarecidas pelo modelo. Um alto resíduo Q mostra que o espectro possivelmente não é adequado para o modelo, p. ex., quando uma amostra medida contiver outra substância.

Reconhecimento de outliers espectrais

O reconhecimento de outliers espectrais identifica espectros que desviam em relação à população.

1. A parametrização é levada em consideração da seguinte forma:
   a. A partir do OMNIS Software versão 4.2: o usuário decide se a parametrização (pré-tratamento de dados e seleção de comprimento de onda) é aplicada ou não.
   Alterações posteriores na parametrização não têm influência na divisão do conjunto de dados.
   b. A partir do OMNIS Software versão 3.3 até a versão do OMNIS Software 4.1: o usuário decide se o pré-tratamento de dados será considerado ou não. A seleção de comprimento de onda e as alterações posteriores do pré-tratamento de dados não têm influência na divisão do conjunto de dados.
   c. Com a versão do OMNIS Software 3.2: o pré-tratamento de dados é considerado do modo como foi definido no momento do reconhecimento de outliers. A seleção de comprimento de onda e as alterações posteriores do pré-tratamento de dados não têm influência na divisão do conjunto de dados.
2. O reconhecimento de outliers espectrais tem como base o modelo PCA de todos os espectros centralizados no valor médio *(ver capítulo 4.2, página 35)*. A quantidade de componentes principais é escolhida de forma que a variância declarada seja de pelo menos 95%.
3. Para o reconhecimento de outliers espectrais, são utilizados os valores de Hotelling T^2 e os resíduos Q dos espectros. O algoritmo avalia se Hotelling T^2 ou o resíduo Q do espectro verificado é o resultado de uma variação aleatória ou sistemática.
   Uma descrição do algoritmo pode ser encontrada no anexo *(ver capítulo 6.5, página 99)*.

4.3.3.1 Gráfico de influência

Assim, o gráfico de influência mostra as propriedades fundamentais dos espectros e contribui para a análise de outliers espectrais.

A base para o gráfico de influência é um modelo PCA *(ver capítulo 4.2, página 35)* ou um modelo PLS:

- **Quantificação**: o gráfico de influência é baseado em **PCA** ou em **PLS**.
  Assim como a PCA, a regressão PLS reduz os dados espectroscópicos a poucas variáveis. Mas a PLS também considera os valores de referência. Os componentes principais da PCA são denominados **variáveis latentes** na PLS.
  (ver capítulo 4.4.1, página 61)
- **Identificação**: um gráfico de influência baseado em **PCA** está disponível (a partir da versão 4.3 do OMNIS Software).

Tipos de outliers espectrais

Para cada espectro, o gráfico de influência permite visualizar os valores para Hotelling T^2 e resíduo Q *(ver "Hotelling T² e resíduos Q", página 51)*. Hotelling T^2 e resíduos Q permitem detectar diversos tipos de outliers espectrais:

- Outliers **Hotelling T^2**, também denominados de outliers de alavanca (em inglês, "leverage outlier"): um elevado T^2 significa que a projeção do espectro no espaço de componentes principais (PCA) ou no espaço de variáveis latentes (PLS) está muito longe do ponto central do modelo.
- **Outliers resíduos Q**: um resíduo Q elevado significa que o espectro é descrito inadequadamente pelo modelo.

A figura 28 mostra vários espectros em diversas exibições:

- Links de gráficos de influência: os resíduos Q descrevem as variações não explicadas pelo modelo, enquanto Hotelling T^2 consideram as variações dentro do próprio modelo.
  As linhas tracejadas mostram os **valores críticos** ou **valores-limites** para o nível de significância definido *(ver capítulo 6.5, página 99)*. Quanto maior o nível de significância, menores são os valores limite e, portanto, mais pontos podem ficar fora dos valores limite.
- À direita: espaço original exemplar com 3 variáveis, x_1, x_2, x_3, que será convertido em um espaço de 2 dimensões com variáveis latentes, como exemplo.
  Para os pontos A até D, é representada a distância ortogonal ao nível (linhas tracejadas) e o ponto modelado no espaço de variáveis latentes (pontos verdes).

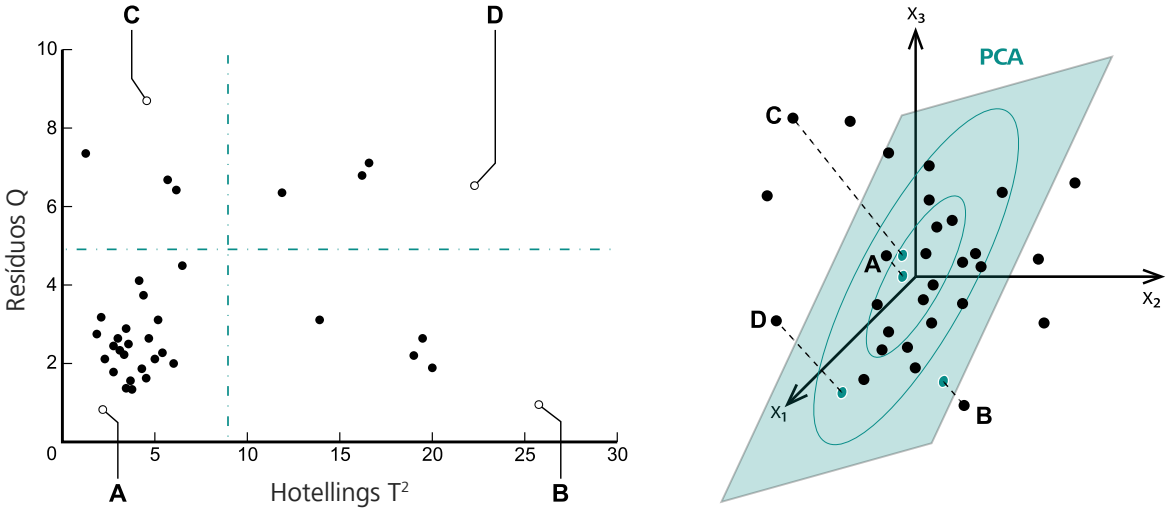


Figura 28 Gráfico de influência (à esquerda), espaço original e espaço de variáveis latentes (à direita). Cada espectro é representado por um ponto na figura à esquerda e um ponto na figura à direita.

Em ambas as visualizações, 4 pontos com diferentes características são destacados:

- O espectro A tem scores menores e resíduos menores. Ele está próximo ao ponto médio do modelo e é bem explicado pelo modelo.
- O espectro B é um outlier Hotelling T^2. Ele está longe do ponto central, mas é bem explicado pelo modelo.
- O espectro C é um outlier resíduos Q. Ele está longe do ponto central e é explicado inadequadamente pelo modelo.
- O espectro D é tanto um outlier Hotelling T^2 quanto um outlier resíduos Q. Ele está longe do ponto central e é explicado parcialmente pelo modelo.

O gráfico de influência mostra como diferentes espectros influenciam o modelo. Como todas as variáveis latentes passam pelo ponto central, os espectros próximos ao ponto médio (como o espectro A, p. ex.) praticamente não têm chance de alterar a direção das variáveis latentes. Eles não têm nenhum valor de alavanca. Quanto maior for a distância em relação ao ponto central, maior é o valor de alavanca e o potencial de influenciar o modelo. Alguns espectros realmente conseguem "puxar" o modelo em sua direção (espectro B), enquanto outros só conseguem fazer isso parcialmente (espectro D) ou não conseguem (espectro C).

Comparado com um modelo que tem como base todos os espectros, o cálculo de um modelo sem o espectro B provavelmente altera o modelo mais do que um cálculo sem o espectro D e ainda mais do que sem o espectro C. O espectro B provavelmente influencia fortemente o modelo de quantificação – positivamente ou negativamente. É necessário ter

extremo cuidado ao decidir se um potencial outlier no quadrante inferior direito do gráfico de influência deve ser excluído ou não.

É ideal que o modelo compreenda a variância de uma quantidade maior de espectros. Não é desejável que o modelo seja marcado somente por poucos espectros. Na figura acima, poucos aspectos têm grandes distâncias em relação ao ponto médio e também à maioria dos outros espectros. Isso é suspeito. O modelo é influenciado por poucos espectros. Há um potencial outlier que deve ser examinado. Além disso, é necessário assegurar que as amostras sejam distribuídas uniformemente por toda a gama de variações.

Gráfico de influência PCA e gráfico de influência PLS

O gráfico de influência PCA depende somente dos espectros. O gráfico de influência PLS depende dos espectros e dos valores de referência.

A tabela a seguir mostra como diferentes configurações se refletem no gráfico de influência PCA e no gráfico de influência PLS.

	Gráfico de influência PCA	Gráfico de influência PLS
Espectros	O modelo PCA utilizado como fundamento tem como base todos os espectros no conjunto de dados de calibração, no conjunto de dados de validação e no conjunto de dados de outliers.	O modelo PLS utilizado como fundamento tem como base todos os espectros no conjunto de dados de calibração. Com base neste modelo PLS, são calculados e representados no gráfico os valores T^2 e resíduos Q dos espectros em todos os 3 conjuntos de dados.
Parametrização	Considera os pré-tratamentos de dados e faixas de comprimento de onda selecionados. Aviso: a detecção de outliers é baseada em PCA e leva em consideração a parametrização de acordo com a configuração do usuário e a versão do OMNIS Software *(ver "Reconhecimento de outliers espectrais", página 52)*.	Considera os pré-tratamentos de dados e faixas de comprimento de onda selecionados. Aviso: a avaliação de outliers na previsão é baseada em PLS e considera os pré-tratamentos de dados e as faixas de comprimento de onda.
Quantidade de variáveis	Utiliza a quantidade de componentes principais que alcançou uma variância declarada de pelo menos **95%**.	Utiliza a quantidade atualmente selecionada de variáveis latentes.

	Gráfico de influência PCA	Gráfico de influência PLS
Nível de significância e valores críticos	Utiliza o nível de significância atualmente selecionado para o cálculo e a visualização dos valores críticos (linhas tracejadas). Identificação: se a última divisão do conjunto de dados foi realizada sem detecção de outliers, o gráfico de influência usa um nível de significância de 5%. Aviso: um aumento do nível de significância leva a menores valores críticos e, com isso, mais outliers no desenvolvimento de modelos.	Utiliza o nível de significância atualmente selecionado para o cálculo e a visualização dos valores críticos (linhas tracejadas). Aviso: um aumento do nível de significância leva a menores valores críticos e, com isso, a mais outliers na previsão.
Valores de referência (quantificação)	Os valores de referência não têm qualquer influência sobre o modelo PCA. Porém, cada espectro pode ter um respectivo outlier de valor de referência e, por isso, ser identificado como outlier.	Os valores de referência influenciam o modelo PLS e, com isso, o gráfico de influência PLS. Além disso, cada espectro pode ter um respectivo outlier de valor de referência e, por isso, ser identificado como outlier.

Análise de outliers

Na análise de potenciais outliers, devem ser considerados os fatores a seguir:

- Um outlier Hotelling T^2 indica uma amostra com composição extrema dos componentes químicos em relação a outras amostras.
- Um outlier resíduos Q pode, por exemplo, indicar uma amostra contaminada ou um erro no registro do espectro.

Os potenciais outliers devem ser examinados cuidadosamente. Outliers verdadeiros devem ser removidos da lista de espectros. Amostras válidas devem ser mantidas. Se o conjunto de dados for novamente distribuído, o reconhecimento de outliers pode encontrar potenciais outliers que não foram encontrados no primeiro ciclo. Uma possível razão para isso é que o novo modelo PCA precisa de menos componentes principais para alcançar a variância declarada de 95%. Se os novos outliers encontrados forem amostras válidas, eles devem ser mantidos. Neste caso, a distribuição automática pode ser repetida sem o reconhecimento de outliers.

4.3.3.2 Gráfico de scores

A base para o gráfico de scores é um modelo PCA ou um modelo PLS:

- **Quantificação**: o gráfico de scores (a partir da versão 3.0 do OMNIS Software) é baseado no **PLS** *(ver capítulo 4.4.1, página 61)*.
- **Identificação**: o gráfico de scores (a partir da versão 4.3 do OMNIS Software) é baseado no **PCA** *(ver capítulo 4.2, página 35)*.

Cada espectro tem um valor de score para cada componente principal ou variável latente. No gráfico de scores, cada espectro é representado por um ponto. O eixo x mostra, p. ex., o score da primeira variável latente e o eixo y, p. ex., p score da segunda variável latente. Cada par de variáveis latentes pode ser exibido de mesma forma.

Como os valores de absorbância de cada variável de comprimento de onda foram centralizados em relação ao valor médio, os scores de cada variável latente também são centralizados em relação ao valor médio. Um ponto próximo ao valor médio do gráfico de scores (0/0) representa um espectro médio em relação a ambas as variáveis latentes. Pontos próximos uns dos outros representam espectros semelhantes, pontos distantes uns dos outros representam espectros diferentes no que diz respeito às duas variáveis latentes exibidas.

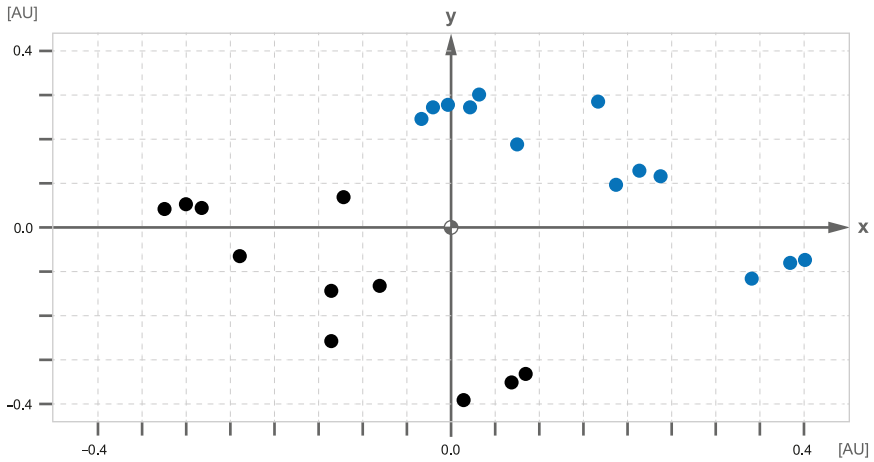


Figura 29 Gráfico de scores para a variável latente 1 (eixo x) e a variável latente 2 (eixo y). AU = unidade aleatória.

A figura 29 mostra um exemplo de 2 conjuntos de dados que foram medidos em condições diferentes. Os scores são normalizados, cada variável latente recebe a mesma ponderação.

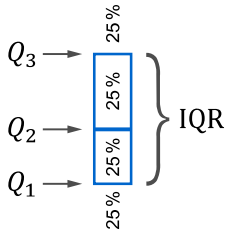
Os scores de todos os componentes principais ou variáveis latentes de um espectro podem ser resumidos em um valor individual (Hotelling T^2), que é exibido no eixo x do gráfico de influência.

4.3.4 Outlier de valor de referência (quantificação)

Em modelos de quantificação, além dos outliers espectrais também são estimados outliers de valor de referência. Os outliers de valor de referência mostram anomalias no valor de referência.

Tipicamente, os outliers de valor de referência são números estimados incorretamente, p. ex., 143 ao invés de 14,3 ou 15,9 ao invés de 51,9. O reconhecimento de outliers identifica esses erros de transmissão ou transcrição com base em uma análise empírica. Somente erros claros serão identificados para análise posterior.

Box plots



Outliers de valor de referência são determinados com o auxílio de um método que tem como base box plots. Um **box plot** organiza os valores de referência em ordem crescente. Quartis dividem o conjunto de dados em 4 partes. Cada parte contém 25% dos valores de referência.

O primeiro quartil Q_1 separa os valores 25% menores do restante. Q_2 é a mediana e separa os valores 50% menores do restante. O terceiro quartil, Q_3, separa os valores 75% menores do restante. Um quadrado vertical representa os 50% médios dos dados, a distância entre quartis (em inglês, "interquartile range", IQR).

Dados fora da caixa IQR em um determinado valor são considerados potenciais outliers e podem ser apresentados como pequenos círculos. Os valores-limites superior e inferior para outliers frequentemente são definidos como 1,5 vezes o IQR definido:

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

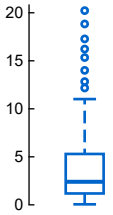
Neste cálculo, Q_1 corresponde ao primeiro quartil, Q_3 ao terceiro quartil e IQR à distância entre quartis ($Q_3 - Q_1$).

Para completar o box plot, as linhas acima e abaixo da caixa são traçadas até os pontos mais distantes que não estão identificados como potenciais outliers.

Adaptação para distribuições assimétricas

O box plot mais utilizado parte de uma distribuição quase simétrica dos dados. Em distribuições assimétricas, geralmente muitos valores de referência regulares são marcados como potenciais outliers. Por esta razão, o OMNIS Software usa uma versão adaptada de box plot que considera a distribuição assimétrica.

No exemplo a seguir, a distribuição é deslocada em direção aos valores de referência mais altos. Isso resulta em 8 outliers com valores altos. Após a adaptação dos valores-limites de outliers, apenas 2 deles permanecem. Por outro lado, aparece um novo outlier com valor mais baixo.

Distribuição assimétrica	Valores-limites de outliers adaptados
	

Uma descrição da adaptação pode ser encontrada no anexo *(ver capítulo 6.6, página 101)*.

4.3.5 Divisão do conjunto de dados

O conjunto de dados é composto por espectros e valores de referência (quantificação) ou por espectros e seus respectivos nomes de produto (identificação, verificação). O conjunto de dados é utilizado para o desenvolvimento e a validação do modelo. Portanto, o conjunto de dados deve ser dividido em um conjunto de dados de calibração, um conjunto de dados de validação e um conjunto de dados de outliers. A distribuição pode ser efetuada manual ou automaticamente.

Assumindo que haja espectros suficientes disponíveis, podem ser utilizados, por exemplo, de 20 a 30% dos dados gerais para o conjunto de dados de validação.

Algoritmo de distribuição automático

Na divisão do conjunto de dados automática, ocorre uma verificação para que o conjunto de dados de calibração e o conjunto de dados de validação sejam representativos em relação à população e independentes um do outro. O objetivo é distribuir os dados em dois conjuntos que cubram quase a mesma área no espaço PCA e tenham propriedades estatísticas semelhantes. O último algoritmo utilizado é um algoritmo duplex ligeiramente modificado de R. D. Snee, *Validation of Regression Models: Methods and Examples*, Technometrics, volume 19, nº. 4 (novembro de 1977), pág. 415–428.

Réplicas

Não deve haver réplicas no conjunto de dados. O algoritmo não remove réplicas ou pseudo-réplicas nem obriga que as réplicas de uma determinada amostra sejam adicionadas ao mesmo conjunto de dados.

1. A parametrização é levada em consideração da seguinte forma:
    a. A partir do OMNIS Software versão 4.2: o usuário decide se a parametrização (pré-tratamento de dados e seleção de comprimento de onda) é aplicada ou não.
    Alterações posteriores na parametrização não têm influência na divisão do conjunto de dados.
    b. A partir do OMNIS Software versão 3.3 até a versão do OMNIS Software 4.1: o usuário decide se o pré-tratamento de dados será considerado ou não. A seleção de comprimento de onda e as alterações posteriores do pré-tratamento de dados não têm influência na divisão do conjunto de dados.
    c. Com a versão do OMNIS Software 3.2: o pré-tratamento de dados é considerado do modo como foi definido no momento do reconhecimento de outliers. A seleção de comprimento de onda e as alterações posteriores do pré-tratamento de dados não têm influência na divisão do conjunto de dados.
2. A distribuição tem como base o modelo PCA de todos os espectros centralizados no valor médio. A quantidade de componentes principais é escolhida de forma que a variância declarada seja de pelo menos 95%.
3. A partir dos scores PCA, são calculadas as distâncias entre todos os possíveis pares de espectros.
4. Os 2 espectros mais distantes um do outro serão atribuídos ao conjunto de dados de calibração.
5. Dentre os espectros restantes, os 2 espectros mais distantes um do outro serão atribuídos ao conjunto de dados de validação.
6. Dentre os espectros restantes, o espectro mais distante dos espectros já contidos no conjunto de dados de calibração será atribuído ao conjunto de dados de calibração.
7. Dentre os espectros restantes, o espectro mais distante dos espectros já contidos no conjunto de dados de validação será atribuído ao conjunto de dados de validação.
8. A alternância continua até que um dos conjuntos de dados alcance seu tamanho previsto. Os espectros restantes serão atribuídos a outro conjunto de dados.

4.4 Quantificação

4.4.1 Regressão PLS

O OMNIS Software utiliza a regressão PLS para o cálculo de modelos de quantificação.

Assim como a PCA, a **redução parcial de quadrados mínimos** (**regressão PLS**, em inglês *partial least squares regression*) reduz os dados espectroscópicos a poucas variáveis. Porém:

- Os componentes principais na PCA são denominados **variáveis latentes** na PLS. As direções das variáveis latentes e dos componentes principais geralmente são parecidas, mas não idênticas.
- Além dos espectros, a regressão PLS também considera os **valores de referência**. Portanto, menos variáveis latentes são necessárias para alcançar uma correlação suficiente com os valores de referência, o que leva a menos ruídos.

A partir de dados espectroscópicos amplos e altamente redundantes, a PLS extraí as **variáveis latentes** subjacentes. As variáveis latentes também são denominadas "variáveis ocultas", pois elas não são medidas diretamente. As variáveis latentes explicam uma parte tão grande quanto possível da variação dos dados e, simultaneamente, modelam bem os valores de referência.

Passos preparatórios

Os passos preparatórios para a regressão PLS são semelhantes aos da PCA:

1. **Parametrização**: o OMNIS Software aplica o pré-tratamento de dados e a seleção de comprimentos de onda definidos aos espectros.
2. **Centralização de valor médio**: para cada comprimento de onda, o valor de absorbância médio é calculado e subtraído de cada valor, em cada espectro.
   Os valores de referência também são centralizados em relação ao valor médio.

Conversão em variáveis latentes

Após os passos preparatórios, a regressão PLS converte os dados espectrais em um espaço de variáveis latentes considerando os valores de referência. *A figura 30* mostra as 2 primeiras variáveis latentes LV1 e LV2. Da mesma forma, podem ser visualizadas as variáveis latentes LV3, LV4, etc.

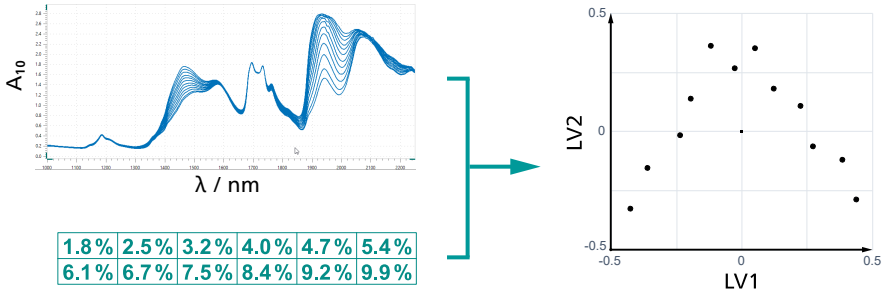
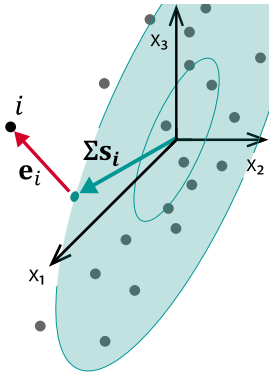


Figura 30 Conversão de espectros e valores de referência em um espaço de variáveis latentes. Os scores no lado direito são expressos em unidades aleatórias.

A primeira variável latente, LV1, explica melhor a variância nos dados espectrais e também indica a maior correlação possível com os valores de referência. Todas as variáveis latentes a seguir, LV2, LV3, etc., explicam melhor a variância restante e também indicam a maior correlação possível com os valores de referência. Portanto, as primeiras variáveis latentes explicam a maior parte da variância e maximizam a correlação, enquanto as outras contêm principalmente ruídos e podem ser descartadas.

Scores e resíduos



A PLS tem grandezas semelhantes à PCA:

- **Scores**: os scores são medidos no espaço de variáveis latentes. A projeção ortogonal da amostra *i* em cada direção das variáveis latentes resulta no vetor de score $\mathbf{\Sigma s}_i$, que representa a distância euclidiana em relação ao ponto médio.
  A **distância de Mahalanobis $\mathbf{s}_i$** é o vetor de score que confere a mesma ponderação para cada direção.
- **Resíduos**: o vetor de resíduo $\mathbf{e}_i$ é o offset entre a amostra *i* e o espaço de variáveis latentes medido no espaço de comprimento de onda original.

Algoritmo PLS

O algoritmo PLS maximiza a covariância entre os espectros e os valores de referência *(ver capítulo 6.3, página 96)*.

4.4.1.1 Número de variáveis latentes

A escolha da quantidade de variáveis latentes em um modelo de quantificação é fundamental para a capacidade de previsão do modelo. Mas, se a quantidade de variáveis latentes for pequena demais, algumas variações espectrais relevantes não serão registradas. Isso é denominado **adaptação insuficiente** e leva a previsões menos exatas.

Se a quantidade de variáveis latentes for muito grande, as amostras são modeladas excessivamente bem. O modelo abrange variações espectrais irrelevantes (ruídos). Isso é denominado **adaptação excessiva** e leva a

variações, instabilidades e previsões menos exatas de amostras desconhecidas.

Para encontrar a quantidade ideal de variáveis latentes, é necessário alcançar um equilíbrio entre os seguintes objetivos:

- A SEP deve ser o mais próxima possível de seu valor mínimo.
  Sem um conjunto de dados de validação, é utilizada a SECV.
- O modelo de quantificação deve utilizar a menor quantidade possível de variáveis latentes. Em caso de dúvidas, a menor quantidade deve ser utilizada.
- O diagrama de correlação para o conjunto de dados de validação deve ser próximo ao ideal. É ideal que o slope seja próximo a 1, a interceptação y próxima a 0 e os pontos de dados indiquem uma dispersão mínima.
  Sem um conjunto de dados de validação, são utilizados os valores da validação cruzada *(ver "Validação cruzada", página 64)*.

É necessário observar que as figuras de mérito são somente valores estimados com base nas amostras de calibração e amostras de validação disponíveis. De maneira geral, um modelo de quantificação é baseado em menos variáveis latentes mais robustas.

4.4.1.2 Gráfico de loading

O gráfico de loading no OMNIS Software é baseado no modelo PLS *(ver capítulo 6.3, página 96)*.

Loadings PLS mostram como as variáveis de comprimento de onda originais (inclusive a parametrização) contribuem para compor cada variável latente. É irrelevante se os loadings são positivos ou negativos.

Os loadings são calculados de modo que a primeira variável latente registre a variância mais expressiva para o parâmetro de referência. Todas as variáveis latentes a seguir registram a variância restante mais expressiva para o parâmetro de referência. Loadings PLS com grandes desvios em relação a 0 indicam, portanto, que os respectivos comprimentos de onda são adequados para modelar o parâmetro de referência.

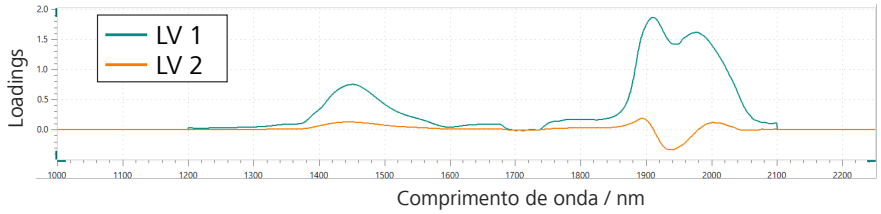


Figura 31 Gráfico de loading para as variáveis latentes LV1 e LV2.

Na *figura 31*, a seleção de comprimento de onda foi restringida à área de 1.200 nm até 2.100 nm. Por isso, não ocorrem loadings fora desta área.

4.4.2 Validação de modelos de quantificação

A validação verifica se o modelo cumpre os requisitos de desempenho e robustez. Para isso, os erros esperados da previsão devem ser estimados de maneira tão realista quanto possível.

Um modelo de quantificação geralmente é validado da seguinte forma:

1. Partindo de uma quantidade limitada de amostras e sem um conjunto de dados de validação, um modelo é desenvolvido e testado por meio de validação cruzada (ver abaixo).
2. Com uma quantidade maior de amostras, o conjunto de dados é dividido em um conjunto de dados de calibração para o desenvolvimento de um modelo e um conjunto de dados de validação para a validação do modelo.
3. Por fim, amostras para o conjunto de dados de validação são coletadas e medidas em um dia diferente, se possível por uma pessoa diferente e usando um equipamento diferente.

Validação cruzada

A validação cruzada como base exclusivamente o conjunto de dados de calibração. Ela gera um valor estimado para cada amostra de calibração utilizando um modelo temporário elaborado sem as respectivas amostras de calibração.

Na validação cruzada, é utilizado um procedimento de várias rodadas em uma das seguintes formas:

- **Leave-One-Out**
  Na validação cruzada Leave-One-Out (LOO-CV), a cada rodada 1 amostra é colocada de volta enquanto as amostras restantes servem de base para elaborar um modelo. Esse modelo prevê o parâmetro de interesse para a amostra colocada de volta. Essa previsão serve como valor estimado para a amostra.
  O ciclo continua até que cada amostra tenha sido colocada de volta uma vez.

- **K-fold**
  Na validação cruzada K-fold, o conjunto de dados de calibração é dividido em *k* blocos de tamanho o mais parecido possível.
  A cada rodada, 1 bloco é colocado de volta, enquanto os outros blocos servem de base para elaborar um modelo. Esse modelo prevê o parâmetro de interesse para as amostras colocadas de volta.
  O ciclo continua até cada bloco ter sido colocado de volta uma vez.
  A seleção dos blocos pode ser efetuada de diversas formas:
  - **Fixed Blocks (DUPLEX)** (a partir da versão do OMNIS Software 3.2): os blocos são reprodutíveis com base em um algoritmo duplex. Cada previsão serve diretamente como valor estimado para a respectiva amostra.
  - **Random**: os blocos são selecionados aleatoriamente. O procedimento descrito acima é repetido várias vezes. O conjunto de dados de calibração é dividido a cada vez de uma forma diferente em *k* blocos. Ao fim, há vários valores estimados para cada amostra. O valor médio de todos esses valores serve como valor estimado para a respectiva amostra.

Geralmente, é preferível usar Leave-One-Out. Com grandes conjuntos de dados de calibração, o método de validação cruzada K-fold pode ser utilizado para reduzir o tempo de cálculo. Um valor típico para *k* é 5.

4.4.2.1 Diagrama de correlação

O diagrama de correlação permite visualizar a correlação entre os valores de referência e os valores calculados. Ele possibilita fazer uma avaliação rápida do modelo de quantificação.

Os valores calculados são estimados do seguinte modo:

- Conjunto de dados de validação e conjunto de dados de outliers: previsão pelo modelo de quantificação
- Conjunto de dados de calibração: valores estimados da validação cruzada

No diagrama de correlação, cada amostra é representada por um ponto. O valor calculado da amostra pode ser encontrado no eixo x, o valor de referência no eixo y. Uma reta de regressão exibe a relação sistêmica entre as variáveis. É ideal que a reta de regressão tenha um slope de 1, uma interceptação y de 0 e todos os pontos estejam na reta. Isso significa, o valor calculado para cada amostra corresponde ao valor de referência.

Com base nos desvios da situação ideal, é possível diferenciar entre erros sistêmicos e erros de aleatoriedade. A posição da reta de regressão exibe os erros sistêmicos. As distâncias entre os pontos e a reta de regressão indicam erros de aleatoriedade.

O seguinte diagrama de correlação **A** mostra uma boa correlação. Os outros diagramas mostram diversos tipos de erros que serão esclarecidos a seguir.

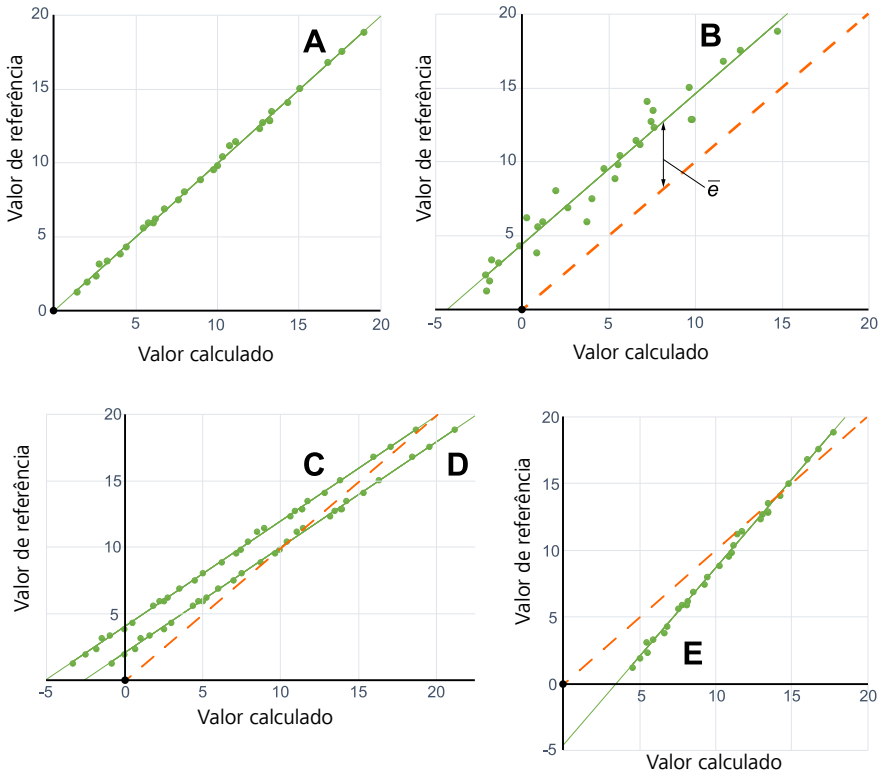


Figura 32 Diagramas de correlação. Cada ponto representa uma amostra e mostra seu valor de referência e seu valor calculado. A linha tracejada mostra a reta ideal de 45°.

Erros sistêmicos

Erros sistêmicos são erros que ocorrem sempre e podem ser reproduzidos em uma determinada aplicação. Os erros sistêmicos podem ser corrigidos. Eles são quantificados pelo viés $\bar{e}$ e pelo slope *b* da reta de regressão:

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

Se o slope for igual a 1 e o viés igual a 0, não há erros sistêmicos.

O **viés** é o erro médio entre os valores de referência e os valores calculados:

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

Neste cálculo, *n* corresponde ao número de amostras, e_i ao erro da amostra *i*, y_i ao valor de referência da amostra *i*, $\hat{y}_i$ ao valor calculado da amostra *i*, $\bar{y}$ ao valor médio dos valores de referência e $\bar{\hat{y}}$ ao valor médio dos valores calculados.

As retas **B** e **C** têm um viés positivo, a reta **E** tem um viés negativo.

Erros com sinal contrário se anulam. Portanto, o viés das retas **D** é próximo a 0.

O **slope** das retas de regressão é:

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

Neste cálculo, $s_{\hat{y}y}$ corresponde à covariância entre valores de referência e valores calculados e $s_{\hat{y}}^2$ à variância dos valores calculados.

O slope pode ser considerado como um erro dependente de características:

- $b > 1$ (reta **E**): quanto maior o valor calculado, maior (mais positivo) é o erro que contribui para o viés.
- $b < 1$ (retas **C** e **D**): quanto maior o valor calculado, menor (mais negativo) é o erro que contribui para o viés.
- $b = 1$ (retas **A** e **B**): o erro que contribui para o viés é constante.

A **interceptação y** das retas de regressão com o eixo y é $\bar{y} - b\bar{\hat{y}}$.

O slope e a interceptação y são calculados com os valores de referência como variáveis dependentes (eixo y) e o valores calculados são calculados como variáveis independentes (eixo x).

Erros de aleatoriedade

Se todos os pontos estiverem diretamente na reta de regressão, não há erros de aleatoriedade. Quanto mais espalhados os pontos estiverem, maiores são os erros de aleatoriedade.

No diagrama de correlação **B**, os erros de aleatoriedade são maiores do que nos outros diagramas de correlação.

Visualização dos tipos de erros

As retas nas figuras acima mostram os seguintes tipos de erros:

	Erros sistêmicos			**Erros de aleatoriedade**
Reta	**Viés**	**Slope**	**Interceptação Y**	
A	~ 0	~ 1	~ 0	pequeno
B	> 0	~ 1	> 0	grande
C	> 0	< 1	> 0	pequeno
D	~ 0	< 1	> 0	pequeno
E	< 0	> 1	< 0	pequeno

4.4.2.2 Figuras de mérito

Figuras de mérito expressam numericamente a correspondência entre os valores de referência e os valores calculados. Os valores calculados são estimados pelo modelo de quantificação.

R^2 – Coeficiente de determinação

O **coeficiente de determinação R^2** (em inglês, "coefficient of determination") mede a capacidade de adaptação do modelo de quantificação. Para um determinado conjunto de dados, ele é a parcela de variação de referência explicada pelo modelo de quantificação:

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

Neste cálculo, SS_{reg} corresponde à soma da regressão dos quadrados (variância dos valores calculados, ou seja, a variância explicada), SS_{tot} à soma dos quadrados total (variância do valor de referência), SS_{res} à soma residual dos quadrados (variância residual, ou seja, a variância não explicada), y_i ao valor de referência da amostra *i*, $\hat{y}_i$ ao valor calculado da amostra *i* e $\bar{y}$ ao valor médio dos valores de referência.

O valor R^2 é uma fração de 1. Um R^2 de 1 significa que os valores calculados correspondem perfeitamente aos valores de referência. Um R^2 de 0,9 significa que 90% da variância dos valores de referência é explicada pelos valores calculados e 10% não é explicada.

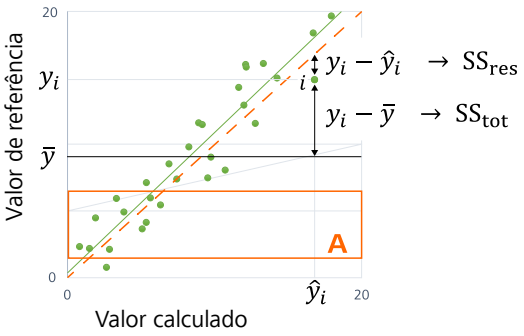


Figura 33 Os componentes para o cálculo de R^2. A linha tracejada é a reta de 45°.

A figura acima mostra um diagrama de correlação com uma amostra *i* e seu resíduo de uma regressão PLS. A variação residual está contida em SS_{res} e a variância dos valores de referência em SS_{tot}.

Um valor R² elevado não garante que um modelo de quantificação possa ser utilizado ou que as previsões sejam exatas. O tamanho de R² depende diretamente da variação dos valores de referência.

Uma regressão com uma faixa de valores de referência menor (faixa **A**) tem aproximadamente a mesma variação residual, mas a variação dos valores de referência é menor. O valor de R² resultante é menor.

A razão para um R² alto poderá, então, ser uma faixa de valores de referência maior do que o real. Por outro lado, dados de um processo de fabricação podem indicar, por exemplo, uma faixa de valores limitada que leve a um valor R² mais baixo. Para avaliar a capacidade de previsão, os erros padrão devem ser incluídos.

O valor R² absoluto deve ser considerado com cuidado. O grau de alteração é mais expressivo a cada variável latente adicional *(ver capítulo 4.4.1.1, página 62)*.

Conforme os valores incluídos no cálculo, é possível obter diferentes valores R²:

- R²C (não exibido no OMNIS Software): calculado com os valores calculados dos espectros no conjunto de dados de calibração.
- **R²CV**: calculado com os valores estimados da validação cruzada dos espectros no conjunto de dados de calibração *(ver "Validação cruzada", página 64)*.
- **R²P**: calculado com os valores calculados dos espectros no conjunto de dados de validação.

Para o cálculo, o OMNIS Software utiliza o quadrado do coeficiente de correlação de amostras de Pearson $r_{y,\hat{y}}$:

Coeficiente de determinação da validação cruzada:

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

Neste cálculo, y_i corresponde ao valor de referência da amostra *i*, $\bar{y}$ ao valor médio dos valores de referência, $\hat{y}_{\mathrm{cv}_i}$ ao valor estimado da validação cruzada da amostra *i*, $\bar{\hat{y}}_{\mathrm{cv}}$ ao valor médio dos valores estimados da validação cruzada e *n* à quantidade de amostras no conjunto de dados de calibração. É necessário observar que cada amostra no conjunto de dados de calibração tem exatamente um valor estimado da validação cruzada.

Coeficiente de determinação da previsão:

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

Neste cálculo, v_i corresponde ao valor de referência da amostra de validação *i*, $\overline{v}$ ao valor médio dos valores de referência, $\hat{v}_i$ ao valor calculado da amostra de validação *i*, $\overline{\hat{v}}$ ao valor médio dos valores calculados e v à quantidade de amostras de validação.

SEC – Erro padrão da calibração

O **erro padrão da calibração (SEC)** é baseado no conjunto de dados de calibração. O SEC pode ser visto como valor estimado para a exatidão de previsão teoricamente melhor. O SEC é o desvio padrão dos resíduos da regressão por mínimos quadrados parciais (PLS):

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

Neste cálculo, **e** corresponde ao vetor de resíduos que contém todas as variações de referência do conjunto de dados de calibração que não são descritas pelo modelo, *n* corresponde à quantidade de amostras de calibração e *k* às variáveis latentes. O denominador *n*-*k*–1 é o número dos graus de liberdade do vetor de resíduos **e**.

Em outras palavras: o SEC é o desvio padrão das diferenças entre os valores de referência e os valores calculados para as amostras no conjunto de dados de calibração:

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

Neste cálculo, y_i corresponde ao valor de referência da amostra de calibração *i*, $\hat{y}_i$ ao valor calculado da amostra de calibração *i*, *n* à quantidade de amostras de calibração e *k* à quantidade de variáveis latentes.

Algumas vezes, o SEC também é denominado RMSEC. O SEC contém os erros de aleatoriedade e os erros sistêmicos (slope e viés).

SECV – Erro padrão da validação cruzada

O **erro padrão da validação cruzada (SECV)** é baseado no conjunto de dados de calibração. O SECV estima a exatidão de precisão com base no conjunto de dados de calibração e em um procedimento de validação cruzada *(ver "Validação cruzada", página 64)*. O SECV pode ser utilizado para uma primeira avaliação do modelo ou para uma estimativa da quantidade ideal de variáveis latentes.

O SECV é o desvio padrão das diferenças entre os valores de referência e os valores estimados da validação cruzada para as amostras do conjunto de dados de calibração.

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

Neste cálculo, y_i corresponde ao valor de referência da amostra *i*, $\hat{y}_{\mathrm{cv}i}$ ao valor estimado da validação cruzada da amostra *i* e *n* à quantidade de amostras no conjunto de dados de calibração.

O SECV contém todos os erros: os erros de aleatoriedade e os erros sistêmicos (slope e viés).

Outras abordagens utilizam valores separados para um SECV com correção de viés e um SECV sem correção, posteriormente denominado RMSECV. Com um viés baixo, esses valores são semelhantes.

SEP – Erro padrão da predição

O **erro padrão da predição (SEP)** é baseado no conjunto de dados de validação. Portanto, o SEP fornece o valor estimado mais realista em relação à exatidão de previsão.

O SEP é o desvio padrão da diferença entre os valores de referência e os valores calculados para as amostras no conjunto de dados de validação:

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

Neste cálculo, v_i corresponde ao valor de referência da amostra de validação *i*, $\hat{v}_i$ ao valor calculado da amostra de validação *i* e v à quantidade de amostras de validação.

O SEP contém todos os erros: os erros de aleatoriedade e os erros sistêmicos (slope e viés).

Outras abordagens utilizam valores separados para um SEP com correção de viés e um SEP sem correção, posteriormente denominado RMSEP. Com um viés baixo, esses valores são semelhantes.

Interpretação das figuras de mérito

As figuras de mérito são valores estimados. Exemplo: o valor SEP é um *valor estimado* de um desvio padrão com base nas amostras disponíveis e também tem seu próprio desvio padrão. Quanto maior o número de amostras de validação, mais confiável é o valor estimado.

O SEP não deve ser comparável ao SECV e ao SEC. Diferenças muito grandes podem indicar uma adaptação excessiva *(ver capítulo 4.4.1.1, página 62)*. Como regra fundamental, as diferenças não devem ser mais de 20%.

Os erros padrão também devem ser considerados em relação ao erro padrão do laboratório (SEL) para o método de referência. A precisão do

método NIR é aceitável quando o SEP é de 1,4 até 2,0 vezes maior do que o SEL. Um SEP maior pode ser aceito, contanto que ele cumpra os requisitos.

Um SEC ou SECV abaixo do SEL para o método de referência indica uma adaptação excessiva.

As relações ao SEL mencionadas acima pressupõem que o SEL seja baseado na quantidade correta de medições de referência para cada amostra *(ver "Método de referência (quantificação)", página 34)*.

4.4.3 OMNIS Model Developer (OMD)

O desenvolvimento de um modelo de quantificação impõe elevados requisitos, é demorado e requer conhecimentos técnicos. O OMNIS Model Developer (OMD, a partir da versão do OMNIS Software 4.0) automatiza o desenvolvimento e fornece bons modelos de quantificação otimizados.

Modo de funcionamento

A coleta adequada de amostras é um pré-requisito *(ver capítulo 4.1, página 33)*. Um conjunto de dados composto por espectros e os respectivos valores de referência serve como entrada para o OMD.

O OMD estima outliers espectrais com um nível de significância de 5% e o algoritmo explicado no anexo, sem considerar o pré-tratamento de dados *(ver "Reconhecimento de outliers espectrais no desenvolvimento de modelos", página 99)*.

A divisão do conjunto de dados depende fortemente da quantidade de espectros que permanecem após o reconhecimento de outliers:

Número de espectros	Método de validação cruzada	Conjunto de dados de validação
> 99	K-fold (5 blocos, DUPLEX)	25% dos espectros
30–99	K-fold (5 blocos, DUPLEX)	—
< 30	Leave-One-Out	—

Após a divisão, outliers adicionais são estimados no conjunto de dados de calibração conforme ASTM D8321-22.

A avaliação e a ordenação dos modelos são efetuadas com base em diversos indicadores. O OMD otimiza o pré-tratamento de dados, a seleção de comprimento de onda e a quantidade de variáveis latentes procurando um equilíbrio entre o risco de uma adaptação excessiva e o risco de uma adaptação insuficiente.

Resultado

O resultado do OMD é uma lista de modelos organizados conforme suas capacidades de previsão. A **capacidade de previsão** é calculada com

base na complexidade do modelo, nas figuras de mérito e no tamanho do conjunto de dados.

A lista possui uma codificação por cores para facilitar a seleção do modelo preferível:

- Verde: boa capacidade de previsão.
  Se a quantidade de amostras for suficientemente grande, o modelo funcionará bem com todas as amostras desconhecidas do mesmo tipo. As figuras de mérito proporcionam valores estimados confiáveis para erros futuros.
- Amarelo: média capacidade de previsão.
  Se a quantidade de amostras for suficientemente grande, o modelo provavelmente funcionará bem. As figuras de mérito podem ser muito otimistas para amostras no futuro. É recomendado executar uma validação específica.
- Vermelho: capacidade de previsão insuficiente.
  O modelo tem sérias desvantagens. Ele não deve ser utilizado.

Os modelos da mesma cor são organizados conforme um critério de informação que prioriza um equilíbrio ao combinar baixos erros de previsão e uma quantidade pequena de variáveis latentes.

Otimizar a parametrização

Em vez de criar todo o modelo automaticamente, apenas a parametrização pode ser otimizada. As configurações atuais (p. ex., divisão do conjunto de dados, método de validação cruzada) permanecem inalteradas, mas não têm influência na otimização.

4.4.4 Correção da interceptação do eixo y / slope

A correção da interceptação do eixo y / slope permite a correção de erros sistêmicos (viés, slope) na aplicação de um modelo de quantificação.

Possíveis causas de erros sistêmicos no conjunto de dados de calibração são:

- Erros sistêmicos no modelo de quantificação. Por exemplo, outliers não detectados ou uma quantidade insuficiente de amostras.
- Erros sistêmicos no procedimento espectroscópico de medição.
- Erros sistêmicos no procedimento de medição de referência.

Se ocorrerem erros sistêmicos no conjunto de dados de validação, outras causas podem ser consideradas:

- Alterações no processo espectroscópico de medição, p. ex., do equipamento.
- Alterações no procedimento de medição de referência, p. ex., novo laboratório, nova infraestrutura.
- Alterações nas amostras, p. ex., por manuseio, armazenamento ou transporte.

A correção de viés e, principalmente, a correção da interceptação do eixo y / slope, devem ser aplicadas com cautela.

Se os erros sistêmicos não forem significantes, não se deve aplicar qualquer correção. Se os erros forem significantes, eles devem ser examinados profundamente. Se possível, a causa dos erros deve ser eliminada. Somente se os erros não puderem ser solucionados por uma razão justificável, pode ser aplicada uma correção de viés ou uma correção da interceptação do eixo y / slope.

Para um valor estimado confiável do viés, são necessárias pelo menos 20 amostras. Para um valor estimado confiável do slope, são necessárias pelo menos 30 amostras.

Correção de viés

O diagrama de correlação a seguir mostra uma correção de viés. O slope da reta de regressão original **F** permanece inalterado. Após a correção (reta de regressão **G**), os erros positivos e os erros negativos se anulam mutuamente.

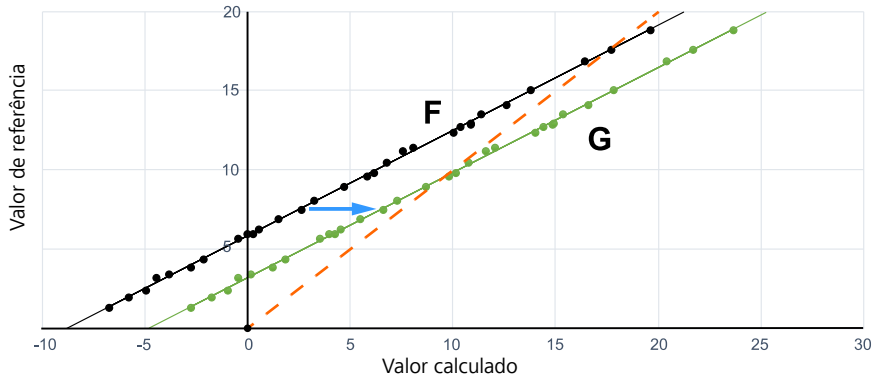


Figura 34 Correção de viés

Correção da interceptação do eixo y / slope

O diagrama de correlação a seguir mostra uma correção da interceptação do eixo y / slope. A reta de regressão original **H** é corrigida no slope e na interceptação y. Consequentemente, tanto o slope quanto o viés são corrigidos (reta de regressão **K**).

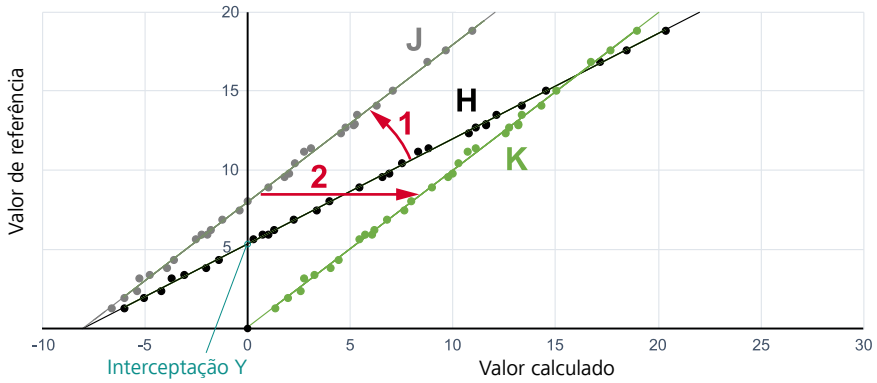


Figura 35 Correção da interceptação do eixo y / slope

SEP

As amostras no conjunto de dados de correção formam a base para a correção da interceptação do eixo y / slope. Com base nessas amostras, o OMNIS Software calcula os seguintes **erros padrão da predição (SEP)**. Os denominadores levar em consideração os graus de liberdade correspondentes:

Tipo da correção	SEP
Não corrigido	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n}}$
Correção de viés	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-1}}$
Correção da interceptação do eixo y / slope	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-2}}$

De fato, aqui v_i corresponde ao valor de referência da *i*-ésima amostra no conjunto de dados de correção, $\hat{v}_i$ ao valor previsto da *i*-ésima amostra no conjunto de dados de correção, e *n* ao número de amostras no conjunto de dados de correção.

O SEP contém todos os erros: os erros de aleatoriedade e os erros sistêmicos (slope e viés).

4.5 Identificação e verificação

4.5.1 Support Vector Machine (SVM)

Os modelos de identificação (a partir da versão do OMNIS Software 4.0) utilizam Support Vector Machines (SVM) para a classificação entre diferentes produtos. Uma Support Vector Machine é um algoritmo de aprendizagem de máquinas monitorado. Com base nas amostras de calibração, ela aprende a atribuir novas amostras a um produto.

Por questões de simplicidade, a classificação a seguir descreve 2 produtos. O conceito pode ser ampliado para que o modelo final classifique uma quantidade desejada de produtos.

Classificação linear

A figura 36 (à esquerda) mostra dados de entrada com 2 variáveis.

Por questões de simplicidade, os espectros parametrizados são apresentados em um espaço de variáveis com 2 dimensões. Cada ponto representa um espectro, as cores informam o pertencimento aos produtos.

Os produtos podem ser separados de maneira linear. O algoritmo SVM cria um hiperplano entre os produtos (figura à direita).

Hiperplanos são uma generalização de planos do espaço com 3 dimensões em espaços de um número qualquer de dimensões. A dimensão de um hiperplano é menor em um do que a dimensão do espaço ao redor dela. Um hiperplano em um espaço com 2 dimensões é uma linha.

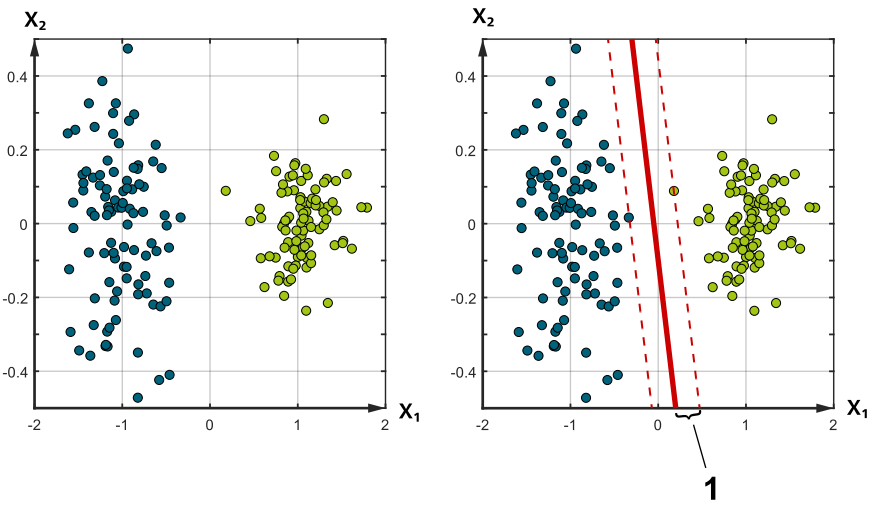


Figura 36 Os dados de entrada (à esquerda) e o hiperplano criado pela SVM (linha à direita traçada em vermelho). Os valores são informados em unidades aleatórias.

O algoritmo da SVM maximiza o intervalo (**1**) entre o hiperplano e o ponto mais próximo em cada lado. Novos espectros podem ser ilustrados no mesmo espaço e um produto pode ser atribuído, dependendo do lado do hiperplano em que o ponto estiver.

Para a definição do hiperplano, o algoritmo da SVM considera somente os pontos que estão mais próximos dos pontos do produto em frente. Esses pontos ou vetores fundamentam a composição do hiperplano e são denominados vetores de suporte.

Mesmo se os pontos não puderem ser separados de maneira linear – devido a um outlier, por exemplo –, ainda é possível determinar um hiperplano para classificação linear. Neste caso, um algoritmo de otimização faz concessões para combinar o aumento do intervalo do hiperplano e os vetores de suporte em cada lado, garantindo que todos os pontos estejam no lado correto do hiperplano. Um parâmetro de regularização controla essas concessões e, assim, a posição final do hiperplano.

Classificação não linear

Na *figura 37 (à esquerda)*, os produtos não podem ser separados de maneira linear. Para a separação dos produtos, é necessário um classificador não linear.

Uma função Kernel linear ou não linear transforma os dados em um espaço de características de altas dimensões. A transformação é executada de modo que os dados no espaço de características possam ser separados de maneira linear por um hiperplano.

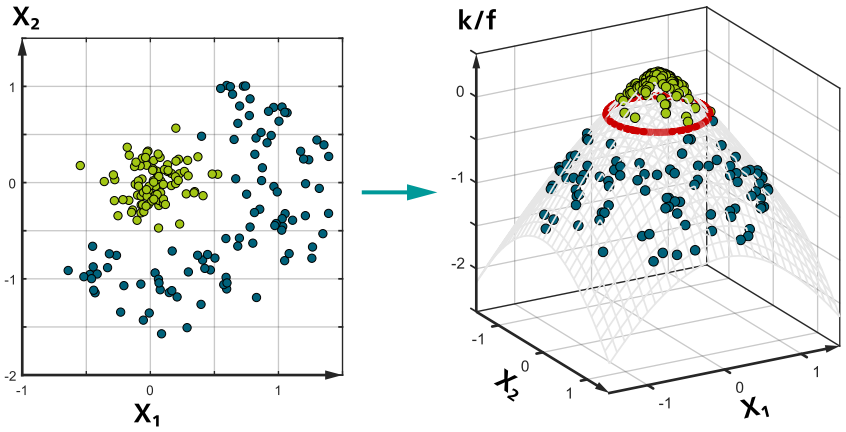


Figura 37 Produtos separáveis de maneira não linear (esquerda). Uma dimensão adicional k/f (característica Kernel) facilita a separação (à direita). Os valores são informados em unidades aleatórias.

Na figura, os dados do espaço com 2 dimensões são convertidos no espaço com 3 dimensões. Os pontos de um produto são elevados acima do plano original, enquanto os pontos do outro produto são deslocados para baixo. O hiperplano linear entre os produtos é um plano com 2 dimensões muito parecidas com o plano original. Considerado em 2 dimensões, o limite de decisão é uma linha não linear, neste caso a linha circular vermelha.

Também neste caso o hiperplano serve como classificador. Uma nova amostra pode ser atribuída a um produto conforme o lado do hiperplano em que seu espectro está.

O OMNIS Software utiliza um Kernel com função de base radial para transformar os dados no espaço de características. O Kernel utiliza parâmetros de escala que controla o grau da não linearidade.

Seleção de parâmetros

Devem ser escolhidos valores adequados para o parâmetro de regularização que controla a posição do hiperplano e para o parâmetro de escala que controla o grau da não linearidade.

Ambos os parâmetros se refletem na capacidade de generalização da Support Vector Machine, ou seja, no quão bem ela pode generalizar a partir dos espectros de calibração para novos espectros desconhecidos. Se o parâmetro de escala possibilitar, por exemplo, um alto grau de não linearidade, o hiperplano provavelmente está muito fortemente adaptado aos espectros de calibração (adaptação excessiva). Se o parâmetro de escala possibilitar somente um baixo grau de não linearidade, o hiperplano provavelmente não está adaptado suficientemente aos espectros de calibração (adaptação insuficiente).

Para uma boa generalização, a **busca em grade** é utilizada. O algoritmo encontra com o mínimo possível de tentativas a melhor combinação de parâmetros:

1. É utilizado um conjunto de combinações de parâmetros predefinidos.
2. Para uma combinação de parâmetros selecionada, a SVM aprende uma regra de classificação que diferencia os espectros de calibração dos diferentes produtos com exatidão máxima.
3. Para cada regra de classificação, uma validação cruzada é utilizada para estimar o quanto a atribuição ao produto é bem-sucedida.
4. Com base nos resultados da validação cruzada, são selecionadas outras combinações de parâmetros.
   A SVM aprende novamente as respectivas regras de classificação e um método de validação cruzada estima a exatidão da classificação. Após algumas repetições, a combinação de parâmetros final é estimada.
5. Com a combinação de parâmetros final e todos os espectros de calibração, a SVM avalia a regra de classificação final.

Probabilidades

Com base na regra de classificação, o modelo de identificação calcula a probabilidade de uma determinada amostra pertencer ou não a um determinado produto. A probabilidade depende das distâncias em relação à população dos produtos e dos parâmetros do modelo.

As probabilidades calculadas desta forma permitem um controle da atribuição das amostras aos produtos *(ver "Atribuição de uma amostra (a partir da versão do OMNIS Software 4.4)", página 79)*.

Para minimizar as identificações de falsos positivos, a partir da versão do OMNIS Software 4.4, também é treinado um modelo de qualificação para cada produto *(ver capítulo 4.6, página 83)*.

4.5.2 Previsão do pertencimento ao produto de uma amostra

Para a identificação de uma amostra específica, o modelo de identificação fornece uma probabilidade por produto *(ver "Probabilidades", página 79)*.

Cada probabilidade é um valor individual entre 0 % e 100 %. Os valores somados são relativos aos produtos, não a 100%.
Os valores devem ser vistos em relação uns aos outros, o que permite uma comparação dos diferentes produtos.

Atribuição de uma amostra (a partir da versão do OMNIS Software 4.4)

A avaliação é realizada com a ajuda de um **limiar de probabilidade** e com qualificações para produtos individuais:

1. Para cada produto cuja probabilidade está acima do limiar de probabilidade, é realizada uma qualificação da amostra usando o modelo de qualificação correspondente. Se a qualificação falhar, a probabilidade do produto correspondente será definida como zero.
2. Avaliação com as probabilidades modificadas do passo 1:
    a. Se nenhuma probabilidade estiver acima do limiar de probabilidade, a identificação falhará (status de identificação **Não identificado**).
    b. Se uma única probabilidade estiver acima do limiar de probabilidade, a amostra será identificada com sucesso e atribuída ao produto correspondente (status de identificação **Identificado**).
    c. Se várias probabilidades estiverem acima do limiar de probabilidade, a previsão será ambígua e a identificação falhou (status de identificação **Ambíguo**).

Tabela 1 mostra um exemplo de avaliação com diferentes limiares de probabilidade. Neste exemplo, as qualificações dos produtos A e C são executadas com sucesso.

Tabela 1 Exemplo de avaliação com diferentes limiares de probabilidade (a partir da versão do OMNIS Software 4.4)

Probabilidade	Limiar de probabilidade	Qualificação	Resultado de identificação
Produto A: 87% Produto B: 71% Produto C: 68% Produto D: 30%	90%	–	Não identificado
	80%	Produto A: bem-sucedido → 87%	Produto A
	70%	Produto A: bem-sucedido → 87% Produto B: falhou → 0 %	Produto A
	60%	Produto A: bem-sucedido → 87% Produto B: falhou → 0% Produto C: bem-sucedido → 68%	Ambíguo

Atribuição de uma amostra (a partir da versão do OMNIS Software 4.0 a 4.3)

As probabilidades da amostra são avaliadas com ajuda de um **limiar de probabilidade**:

- Se nenhuma probabilidade estiver acima do limiar de probabilidade, a identificação falhará (status de identificação **Não identificado**).
- Se uma única probabilidade estiver acima do limiar de probabilidade, a amostra será identificada com sucesso e atribuída ao produto correspondente (status de identificação **Identificado**).

- Se várias probabilidades estiverem acima do limiar de probabilidade, a previsão será ambígua e a identificação falhou (status de identificação **Ambíguo**).

Tabela 2 mostra um exemplo de avaliação com diferentes limiares de probabilidade.

Tabela 2 Exemplo de avaliação com diferentes limiares de probabilidade (a partir da versão do OMNIS Software 4.0 a 4.3)

Probabilidade	Limiar de probabilidade	Resultado de identificação
Produto A: 87%	90 %	Não identificado
Produto B: 72%	80 %	Produto A
Produto C: 68%	70 %	Ambíguo

4.5.3 Validação de modelos de identificação

Um modelo de identificação geralmente é validado da seguinte forma:

1. Partindo de uma quantidade limitada de amostras e sem um conjunto de dados de validação, um modelo é desenvolvido e testado com as amostras no conjunto de dados de calibração.
2. Com uma quantidade suficiente de amostras, o conjunto de dados é dividido em um conjunto de dados de calibração e um conjunto de dados de validação. As amostras no conjunto de dados de validação não serão utilizadas para o desenvolvimento do modelo.
3. Por fim, amostras para um conjunto de dados de validação externo são coletadas e medidas em um dia diferente, se possível por uma pessoa diferente e usando um equipamento diferente.

Para cada amostra, a previsão de pertencimento ao produto é comparada ao respectivo pertencimento efetivo ao produto. Se eles corresponderem, a previsão é correta (= bem-sucedida), caso contrário, ela é incorreta (= falhou).

Validação

Para o modelo de identificação, o OMNIS Software mostra as seguintes grandezas que informam o quão bem os modelos funcionam. Em casos ideais, todos os indicadores são 100%.

Bem-sucedido % (total) mede a *exatidão*, o quanto o modelo é correto. O percentual responde à pergunta: quantas das amostras podem identificar o modelo corretamente?

$$\textit{Bem-sucedido \% (total)} = \frac{\textit{classificação correta}}{\textit{todas as classificações}}$$

Bem-sucedido % (total) atribui a cada amostra o mesmo peso. Portanto, os produtos com várias amostras têm maior influência sobre o percentual do que os produtos com menos amostras.

Números semelhantes para *Bem-sucedido %* estão disponíveis para cada produto.

Aprimorar a identificação

As seguintes ações podem contribuir para o aprimoramento do modelo:

- **Adaptação do limiar de probabilidade**
  - Se muitas previsões forem ambíguas ou ocorrerem muitas probabilidades de 0,0%, o limiar de probabilidade pode ser aumentado.
  - Se muitas amostras não forem identificadas porque o limiar de probabilidade não foi atingido, o limiar de probabilidade pode ser reduzido.
- **Adaptação da parametrização**
  Estimar seleção de comprimento de onda e pré-tratamentos de dados mais adequados.
- **Utilização de hierarquias de modelos**
  Uma hierarquia de modelos possibilita uma estruturação hierárquica de modelos de identificação.
  Exemplo: um modelo de identificação com 4 produtos diferentes não pode diferenciar entre produtos semelhantes, como frutose e glucose. Se frutose e glucose forem reunidas em um grupo de produto, "Açúcar", o modelo poderá diferenciar entre açúcar e os dois outros produtos. Se uma amostra for identificada como açúcar, outro modelo assume a diferenciação entre frutose e glucose. Como este segundo modelo é mais especializado, ele pode diferenciar mais facilmente entre os produtos semelhantes.

Amostras não identificadas

Se as amostras não forem identificadas por engano:

- Verificar as amostras e tratamento de amostra para detectar anomalias.
- Verificar o limiar de probabilidade e diminuir se necessário.
- Verificar o pré-tratamentos de dados e a seleção de comprimento de onda.

- Verificar os espectros das amostras não identificadas no gráfico de scores:
  - Se os espectros ainda não estiverem incluídos no modelo: adicionar os espectros ao conjunto de dados de validação do respectivo produto.
  - No gráfico de score, comparar os scores dos espectros a serem testados com os scores dos espectros no conjunto de dados de calibração.

  Se as amostras não forem discrepantes, mas suas variações estiverem sub-representadas no conjunto de dados de calibração, o conjunto de dados de calibração deverá ser estendido adequadamente.

4.6 Qualificação

Modelos de qualificação (a partir da versão do OMNIS Software 4.4) distinguem um grupo de amostras de outras amostras. Esses modelos são adequados, por exemplo, para distinguir amostras utilizáveis (amostras positivas) de amostras inutilizáveis (amostras negativas).

4.6.1 Cálculo de modelos de qualificação

O cálculo de um modelo de qualificação é semelhante ao modelo de identificação *(ver capítulo 4.5.1, página 76)*. Entretanto, o conjunto de dados de calibração para qualificação contém apenas um tipo de amostra (amostras positivas).

Uma Support Vector Machine (SVM) transforma os dados de entrada em um espaço de dimensão superior. Um parâmetro de regularização determina a posição do hiperplano, enquanto um parâmetro de escala determina o grau de não linearidade. Uma pesquisa em grade determina uma projeção adequada. O limite de decisão para esta projeção forma a base para o modelo de qualificação.

4.6.2 Validação de modelos de qualificação

Procedimento

Um modelo de qualificação geralmente é desenvolvido e validado passo a passo. O número de amostras é aumentado gradualmente:

1. Conjunto de dados de validação positivo:
   a. Se um número limitado de amostras positivas for encontrado, nenhum conjunto de dados de validação positivo será criado inicialmente.
   b. Quando um número suficiente de amostras positivas estiver disponível, um conjunto de dados de validação positivo será criado usando a divisão automática do conjunto de dados.

   Conjunto de dados de validação negativo:
   a. As amostras negativas coletadas são atribuídas ao conjunto de dados de validação negativo.
   b. A detecção automática de espectros negativos pode ser usada para detectar outliers espectrais e atribuí-los ao conjunto de dados de validação negativa *(ver capítulo 6.5, página 99)*.
2. Por fim, amostras para o conjunto de dados de validação positivo e negativo são coletadas e medidas em um dia diferente, se possível por uma pessoa diferente e usando um equipamento diferente.

As amostras no conjunto de dados de validação positivo e negativo não serão utilizadas para o cálculo do modelo.

Validação

Para cada amostra, o modelo de qualificação determina um resultado (positivo ou negativo). Um resultado positivo é esperado para as amostras no conjunto de dados de calibração e no conjunto de dados de validação positivo. Um resultado negativo é esperado para as amostras no conjunto de dados de validação negativo. Se o resultado corresponder à expectativa, a previsão está correta (= bem-sucedida), caso contrário, está incorreta (= falhou).

O OMNIS Software exibe os seguintes valores para modelos de qualificação:

Bem-sucedido % (total) mede a exatidão total de um modelo.

$$\textit{Bem-sucedido \% (total)} = \frac{\textit{previsões corretas}}{\textit{todas previsões}}$$

Números semelhantes para *Bem-sucedido %* estão disponíveis para cada conjunto de dados. Em casos ideais, todos os indicadores são 100%.

Melhorar a qualificação

Uma parametrização mais apropriada (pré-tratamento de dados e seleção de comprimento de onda) pode melhorar o modelo de qualificação.

Amostras não qualificadas

Se as amostras não forem qualificadas incorretamente:

- Verificar as amostras e tratamento de amostra para detectar anomalias.

- Verificar o pré-tratamentos de dados e a seleção de comprimento de onda.
- Verificar os espectros das amostras não qualificadas no gráfico de scores:
  - Se os espectros ainda não estiverem incluídos no modelo: adicionar os espectros ao conjunto de dados de validação positivo.
  - No gráfico de score, comparar os scores dos espectros a serem testados com os scores dos espectros no conjunto de dados de calibração.

  Se as amostras não forem discrepantes, mas suas variações estiverem sub-representadas no conjunto de dados de calibração, o conjunto de dados de calibração deverá ser estendido adequadamente.

5 Previsão

5.1 Quantificação

Para a previsão de uma amostra com parâmetros de interesse desconhecidos, o procedimento ocorre conforme descrito a seguir:

1. O espectro da amostra é registrado.
2. O modelo de quantificação utiliza o mesmo pré-tratamento de dados e a mesma faixa de comprimento de onda que os espectros no conjunto de dados de calibração.
3. Com base no espectro resultante, o modelo prevê os parâmetros de interesse.

No cálculo do modelo de quantificação, os espectros de calibração e os valores de referência foram centralizados em relação ao valor médio. Naturalmente, isso é considerado no procedimento descrito acima.

5.1.1 Outlier e monitoramento de resultado

Na previsão de um parâmetro de interesse quantitativo, diferentes tipos de outliers podem ser reconhecidos e os resultados monitorados:

Outlier/monitoramento		Causa (exemplo)
Outliers espectrais	Hotelling T^2	A concentração de um componente químico está fora da faixa de concentração das amostras de calibração.
	Resíduos Q	A amostra contém componentes que não estão presentes nas amostras de calibração.
Outlier Nearest Neighbor (opcional, a partir da versão 4.2 do OMNIS Software)		A amostra possui variações de amostra que não estão representadas nesta combinação nas amostras de calibração.
Monitoramento de resultado (opcional)		O valor de resultado do parâmetro de interesse está fora dos critérios de aceitação escolhidos pelo cliente.

Outliers espectrais

Os outliers espectrais são determinados do seguinte modo:

1. O software calcula os valores de Hotelling T^2 e resíduos Q para o espectro com base no modelo de quantificação (modelo PLS).
2. Se o valor de T^2 ou valor de resíduos Q do espectro for maior do que o respectivo valor crítico calculado do modelo, a amostra é marcada como outlier em relação ao modelo utilizado *(ver "Avaliação de outliers na previsão (quantificação)", página 100)*.

Os valores de T^2 e valores de resíduos estão disponíveis como variáveis no OMNIS Software. Eles podem ser comparados com os valores no gráfico de influência PLS do modelo de quantificação. As linhas tracejadas informam os valores críticos.

Outlier Nearest Neighbor

(a partir da versão do OMNIS Software 4.2)

Idealmente, as amostras de calibração cobrem todas as combinações possíveis de variações de amostra. Na realidade, algumas combinações ocorrem com maior frequência, outras nem sequer ocorrem. Consequentemente, as amostras de calibração estão distribuídas de forma desigual no espaço das variáveis latentes. Em algumas áreas há muitas amostras de calibração, mas existem lacunas entre elas.

Se o espectro de uma amostra desconhecida cair em uma lacuna entre as amostras de calibração, o resultado de previsão poderá ser inválido ou impreciso. Para detectar tais casos, a distância *D* da amostra desconhecida *i* para cada amostra de calibração *u* é calculada:

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

Aqui $\mathbf{s_i}$ corresponde aos scores da amostra desconhecida *i* e $\mathbf{s_u}$ corresponde aos scores da amostra de calibração *u*. Os scores estão normalizados e ortogonais.

A menor distância é a distância até a amostra de calibração mais próxima e é chamada de **Nearest Neighbor Distance** (NND):

Se o valor NND exceder um determinado valor limite NND, a amostra desconhecida é chamada de outlier Nearest Neighbor.

O valor limite NND é determinado da seguinte forma:

1. Um valor NND é determinado para cada amostra de calibração. Este valor corresponde à distância até a amostra de calibração restante mais próxima.
2. O valor NND máximo de todas as amostras de calibração é o valor limite NND.

O valor NND da amostra desconhecida e o valor limite NND estão disponíveis como variáveis no OMNIS Software.

Monitoramento de resultado

No OMNIS Software, o monitoramento de resultados pode ser usado para definir limites de advertência e limites de intervenção para a faixa de resultados de previsão. Opcionalmente, podem ser definidas ações que serão iniciadas se os valores estiverem fora dos limites definidos.

5.2 Identificação e verificação

A identificação ou verificação de uma amostra, o procedimento ocorre conforme descrito a seguir:

1. O espectro da amostra é registrado.
2. O modelo de identificação utiliza o mesmo pré-tratamento de dados e a mesma seleção de comprimento de onda que os espectros no conjunto de dados de calibração.
3. O modelo de identificação avalia o espectro de *(ver capítulo 4.5.2, página 79)*.
4. Caso o modelo de identificação faça parte de uma hierarquia de modelos e outro modelo de identificação esteja vinculado ao produto identificado, este modelo é executado. Caso um ou mais modelos de quantificação estejam vinculados ao produto determinado, esses modelos são executados.
5. O resultado de identificação ou o resultado de verificação é exibido e, no caso de uma hierarquia de modelo, possivelmente também os resultados da quantificação.

Status de identificação

- Identificado
  A identificação foi concluída com sucesso.
- Ambíguo
  Vários produtos ultrapassam o limiar de probabilidade. A identificação falhou.
- Não identificado
  Nenhum produto ultrapassa o limiar de probabilidade. A identificação falhou.

Status da verificação

- Bem-sucedido
  A amostra foi identificada com sucesso e o resultado corresponde ao produto esperado.
- Falhou
  A verificação falhou.

5.3 Qualificação

Na qualificação de uma amostra, o procedimento ocorre conforme descrito a seguir:

1. O espectro da amostra é registrado.
2. O modelo de qualificação utiliza o mesmo pré-tratamento de dados e a mesma seleção de comprimento de onda que os espectros no conjunto de dados de calibração.
3. Com base no espectro resultante, o modelo qualifica a amostra.
4. O resultado de qualificação é exibido.

Status de qualificação

- Bem-sucedido
- Falhou

6 Anexo

6.1 Exemplo de uma regressão linear

Regressão linear simples

No caso mais simples, uma mistura contém somente um absorvente e um espectro contém somente um pico. Amostras com diferentes concentrações de absorvente possuem picos com diferentes valores de absorbância *(ver capítulo 2.2.1, página 7)*.

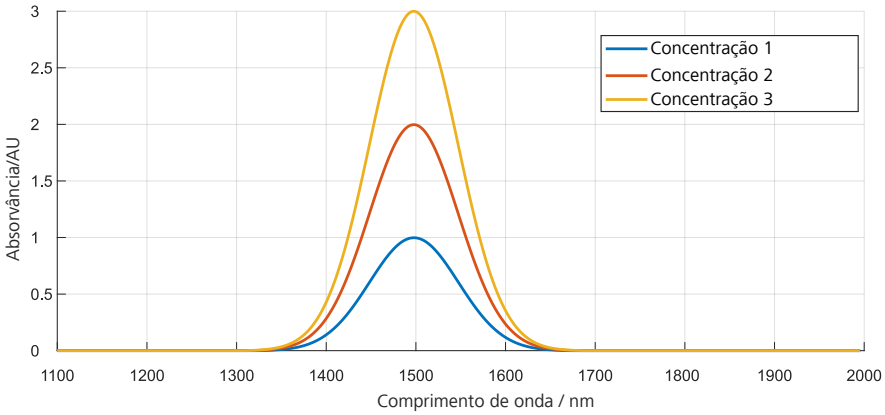


Figura 38 Dados modelados para 3 amostras com um pico em 1.500 nm.

Os 3 valores de absorbância medidos em 1.500 nm podem ser representados em relação à concentração de absorvente.

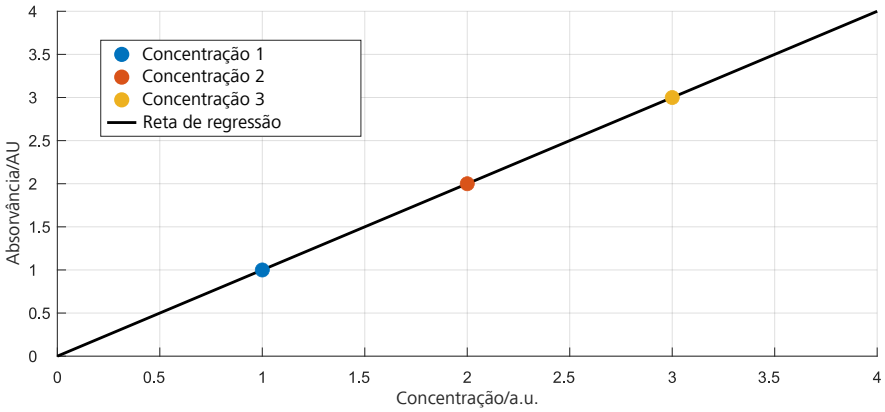


Figura 39 Relação entre valores de absorvância e concentrações.

Conforme a lei de Beer-Lambert, a relação entre os valores de absorbância e as concentrações é linear. Com um conjunto de 3 amostras, pode ser executada uma regressão linear que gera a reta de regressão ilustrada na figura.

Como há somente 1 variável (1 comprimento de onda), esta é uma regressão linear simples. Esta regressão pode ser utilizada como modelo de quantificação. Em uma amostra com concentração desconhecida, a absorvância A é medida em 1.500 nm. A partir da reta de regressão, é possível obter a respectiva concentração c de absorvente:

$$c = bA$$

O coeficiente *b* é constante e idêntico ao slope da regra de regressão.

É necessário observar que todas as amostras precisam conter o mesmo absorvente com o mesmo coeficiente de extinção molar. Além disso, todas as medições de absorção devem ser executadas com espessuras de camada idênticas.

Regressão linear múltipla

Misturas reais possuem mais de um absorvente. O espectro registrado é a soma de todos os espectros de absorventes *(ver capítulo 2.2.1, página 7)*.

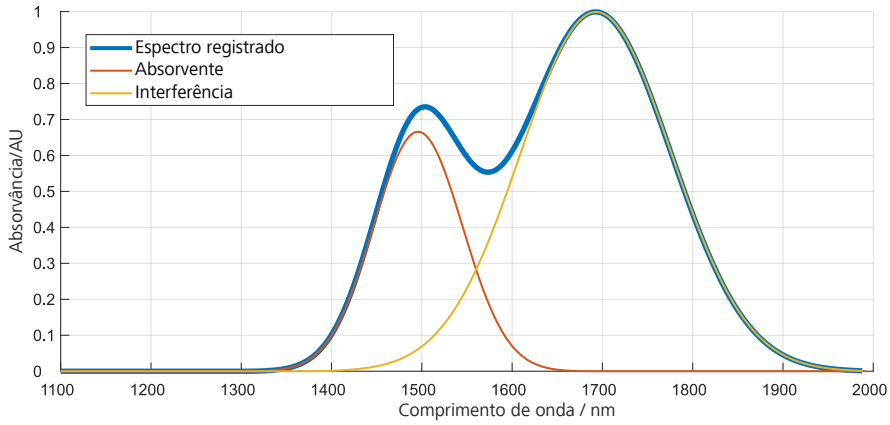


Figura 40 Dados modelados com 2 componentes. O absorvente (linha vermelha) deve ser quantificado.

O espectro registrado (linha azul) é a soma do espectro de absorvente puro e um espectro de interferência sobreposto. Em 1.500 nm, o valor de absorbância medido é composto não somente da absorvância do absorvente, mas também da absorvância da interferência. O parâmetro de interesse não pode ser quantificado em 1.500 nm com uma única medição. Também é impossível saber se houve uma interferência e se a medição é confiável.

O que acontece quando 2 comprimentos de onda, p. ex., em 1.500 nm e 1.700 nm são medidos? A absorvância medida no comprimento de onda 1, A_1, é a soma do sinal de absorvância puro $A_1{}^a$ (índice a = absorvente) e do sinal de interferência puro $A_1{}^f$ (índice f = interferência). Isso também é válido para a absorvância medida no comprimento de onda 2, A_2:

$$A_1 = A_1^a + A_1^f = \varepsilon_1^a c_a + \varepsilon_1^f c_f$$
$$A_2 = A_2^a + A_2^f = \varepsilon_2^a c_a + \varepsilon_2^f c_f$$

Neste cálculo, $\varepsilon_1{}^{a}$ e $\varepsilon_1{}^{f}$ correspondem aos coeficientes de extinção molar no comprimento de onda 1 para o absorvente ou para a interferência, c_a e c_f correspondem às concentrações do absorvente e do causador da interferência.

Nas equações a seguir, a espessura da camada l é excluída da lei de Beer-Lambert. Isso facilita os cálculos algébricos posteriormente. Naturalmente, a espessura da camada deve ser igual em todos os produtos. Portanto, as absorvâncias nas equações são absorvâncias por cm.

As equações podem ser escritas em formato de matriz do seguinte modo:

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

Portanto:

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

A solução para concentração do absorvente resulta em:

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

Assim, a concentração do absorvente também pode ser calculada quando houver disponível um causador de interferência ao medir a absorvância em dois comprimentos de onda e multiplicar cada absorvância por uma constante.

A constante se refere aos coeficientes de extinção molar e podem ser consultadas em tabelas. Mas essa situação nunca ocorre na realidade. Ao invés disso, elas são determinadas por um passo de calibração e a solução do sistema de equações lineares por uma regressão linear múltipla como a regressão PLS. Portanto, as constantes são designadas como coeficientes de regressão b_1 e b_2:

$$c = b_1 A_1 + b_2 A_2$$

Mais do que 2 absorventes

Como mostrado acima, basta 1 absorvente para determinar a absorvância em 1 comprimento de onda. E 2 absorventes são suficientes para determinar a absorvância com 2 comprimentos de onda.

Isso pode ser generalizado. Mais absorventes requerem mais valores de absorvância A_i com diferentes comprimentos de onda i. Este relacionamento ainda permanece um relacionamento linear:

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

Desenvolvimento de um modelo de quantificação

Antes da equação acima poder prever a concentração em amostras desconhecidas, os coeficientes b_1, b_2, etc. devem ser calculados. Para isso, é necessário um passo de calibração. Serão medidas várias amostras com diferentes concentrações do parâmetro de interesse.

Em conformidade com a terminologia posteriormente usada para PCA e PLS, *c* pode ser substituído por *y* e *A* por *x*. Depois disso, a equação acima pode ser escrita da seguinte forma para cada amostra de calibração:

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

Neste cálculo, *n* corresponde à quantidade de amostras, *f* à quantidade de comprimentos de onda, y_1 ao valor de referência para a amostra 1 medido com o método de referência (p. ex. titulação), $x_{1,1}$ corresponde à absorvância medida da amostra 1 com o comprimento de onda 1 e $\{x_{1,1} \ldots x_{1,f}\}$ ao espectro da amostra 1 medido com *f* comprimentos de onda. Além disso, $b_1 \ldots b_n$ correspondem aos coeficientes de regressão e $e_1 \ldots e_n$ correspondem aos termos de erro que mostram o quão bem os coeficientes de regressão modelam os dados medidos.

Em forma de matriz compacta, isso resulta em:

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

X está definido como matriz $f \times n$. $\mathbf{X^t}$ é a matriz transposta **X**, ou seja, as linhas e colunas são trocadas para obter a matriz acima $n \times f$. O vetor de previsão **p** corresponde aos coeficientes de regressão acima **b**.

O vetor de previsão **p** possibilita a previsão do parâmetro de interesse para uma nova amostra com base em seu espectro **x**. O valor calculado $\hat{y}$ é:

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

A tarefa da regressão linear múltipla é determinar os coeficientes de regressão que resultam em termos de erro mínimos. Mas uma regressão linear múltipla (MLR) iria requerer uma quantidade maior de amostras de calibração do que de comprimentos de onda. Outro obstáculo é a alta correlação entre as variáveis.

Para previsões espectroscópicas, podem ser utilizados outros métodos. Com o PCA, a quantidade de dados pode ser significativamente reduzida e a correlação completamente eliminada. Com uma regressão PLS, são considerados adicionalmente os valores de referência das amostras.

6.2 Algoritmo PCA

A **análise de componentes principais** (**PCA**, em inglês, *principal component analysis*) é utilizada para a divisão do conjunto de dados automática e para a detecção de outliers espectrais *(ver capítulo 4.2, página 35)*.

Para a análise de componentes principais dos espectros parametrizados e centralizados em relação ao valor médio, o OMNIS Software executa uma **decomposição em valores singulares** (**SVD**, em inglês, *singular value decomposition*). Ele decompõe a matriz de dados original **X** (os espectros do conjunto de dados de calibração) em 3 matrizes e calcula n componentes principais:

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

Neste cálculo, **X** corresponde aos dados do espectro originais (uma matriz $f \times n$ com f comprimentos de onda e n amostras), **L** corresponde aos loadings (uma matriz $f \times n$), Σ aos valores singulares (uma matriz $n \times n$), e **S** aos scores (uma matriz $n \times n$). Esta equação pode ser representada em forma de gráfico *(ver figura 41, página 95)*.

Os loadings **L** projetam os eixos do espaço de comprimento de onda nos eixos do espaço de componentes principais. Os vetores das colunas de **L** são os eixos principais ou as direções principais.

Os scores **S** são s projeções dos espectros originais nos eixos principais. Cada vetor de linha de **S** contém um determinado espectro.

Σ é uma matriz diagonal dos valores singulares σ_i. Os valores singulares descrevem a variância explicada por cada componente principal. Por convenção, eles são organizados de modo que $\sigma_1 > \sigma_2 > \ldots > \sigma_n$.

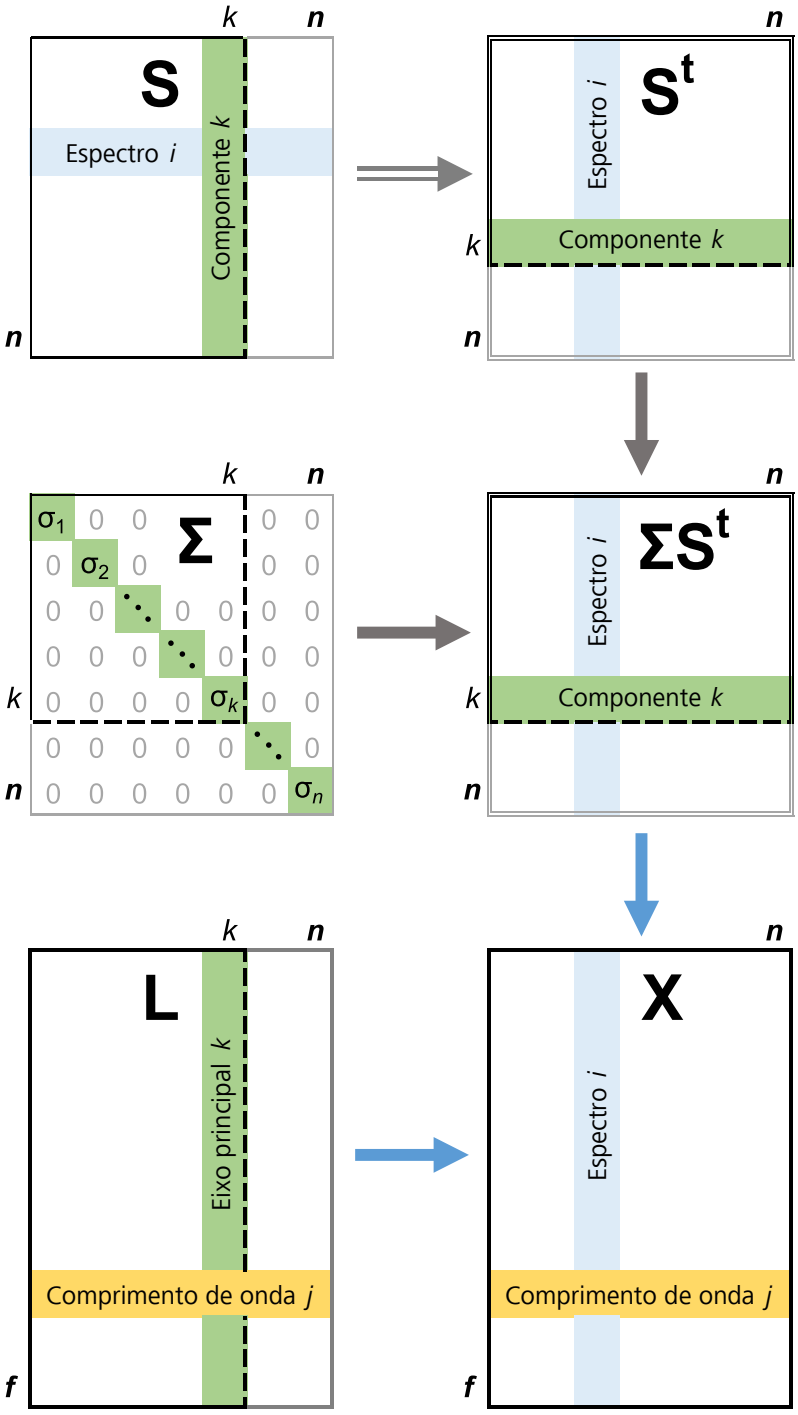


Figura 41 *A equação da decomposição de valores singulares em forma de gráfico. As linhas tracejadas mostram as matrizes reduzidas para um modelo com k componentes principais. As informações nos componentes principais n-k excluídos são integradas à matriz residual (uma matriz f × n, não incluída na ilustração).*

Matriz residual

Um modelo PCA utiliza somente o primeiro par dos *n* componentes principais calculados. Na *figura 41*, são os primeiros *k* de *n* componentes principais. Os dados originais **X** podem ser divididos em dados descritos pelo modelo e dados não descritos pelo modelo:

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

Neste cálculo, **X** corresponde aos dados do espectro originais (uma matriz $f \times n$), $\mathbf{L_a}$ (uma matriz $f \times k$) às primeiras *k* colunas de **L**, $\mathbf{\Sigma}_a$ (uma matriz diagonal $k \times k$) aos primeiros valores singulares *k*, $\mathbf{S_a}$ (uma matriz $n \times k$ com *k* componentes principais) corresponde às primeiras colunas *k* de **S** e **E** à matriz residual (uma matriz $f \times n$) que contém todas as variações espectrais em **X** que não podem ser descritas pelo modelo.

Geralmente, $k \ll n \ll f$. Por exemplo: $k = 3$ componentes principais, $n = 100$ amostras e $f = 2.500$ comprimentos de onda.

Cada coluna $\mathbf{e_i}$ da matriz residual **E** mostra a distância ortogonal do espectro *i* ao espaço PCA, chamado de **resíduo**. Quanto mais componentes principais o modelo utilizar, menor será o resíduo.

6.3 Algoritmo PLS

A **regressão por mínimos quadrados parciais** (**regressão PLS**, em inglês *partial least squares regression*) é utilizada para o cálculo do modelo de quantificação *(ver capítulo 4.4.1, página 61)*.

A PLS inclui dois blocos de dados:

- A matriz parametrizada e centralizada em relação aos valores médios **X** (os espectros).
- O vetor **y** centralizado em relação aos valores médios (os valores de referência).

O PLS decompõe a matriz **X** em 2 matrizes:

$$\mathbf{X} = \mathbf{LS^t} + \mathbf{Z}$$

Neste cálculo, **X** (uma matriz $f \times n$ com *f* comprimentos de onda e *n* amostras) corresponde aos espectros pré-tratados e centralizados em relação ao valor médio, **L** corresponde aos loadings (uma matriz $f \times k$ com *k* variáveis latentes), **S** aos scores (uma matriz $n \times k$) e **Z** à matriz residual (uma matriz $f \times n$) que contém todas as variações espectrais em **X** que não podem ser descritas pelo modelo.

Enquanto no PCA a matriz de scores **S** explica a variância de **X**, no PLS a matriz de scores **S** explica a covariância entre **X** e **y**. O PLS maximiza a covariância explicada pelos scores. Isso significa que os scores não

somente explicam melhor a variância de **X**, como também têm a maior correlação possível com os valores de referência.

Para maximizar a covariância entre **X** e **y**, o algoritmo PLS realiza a troca de dados entre **X** e **y**. Portanto, **X** e **y** se misturam em um único sistema integrado. Neste processo, é feita a regressão dos scores **S** em relação aos valores de referência **y** para obter o coeficiente de regressão **b**:

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

Sendo **e** o valor residual que contém todas as variações de referência em **y** que não podem ser descritas pelo modelo.

Previsão

A partir dos coeficientes de regressão **b**, é possível determinar o vetor de previsão **p**. Para a previsão do parâmetro de interesse $\hat{y}$ de uma nova amostra, são utilizados o vetor de previsão **p** e o espectro pré-tratado e centralizado em relação valor médio **x**:

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

O OMNIS Software implementa o PLS com o algoritmo SIMPLS e um único conjunto de valores de referência (PLS-1).

6.4 Hotelling T² e resíduos Q

Hotelling *T*² e resíduos Q caracterizam os espectros em um modelo PCA ou PLS. Eles contribuem principalmente para identificar possíveis outliers *(ver "Hotelling T² e resíduos Q", página 51)*.

Hotelling T²

A distância de Mahalanobis é uma medida do tamanho do desvio de um espectro em relação ao ponto central do modelo. A distância é normalizada. Cada componente principal ou variável latente recebe o mesmo peso.

Partindo do pressuposto que os espectros ou os scores estão em distribuição normal, os quadrados das distâncias de Mahalanobis, MD², de uma distribuição Hotelling *T*² são os seguintes:

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

Os quadrados da distância de Mahalanobis para o espectro *i* é a soma dos quadrados dos scores normalizados para os primeiros *k* componentes principais ou variáveis latentes:

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

Neste cálculo, $\mathbf{s_i}$ corresponde à linha *i* da matriz de scores reduzida **S**, $s_{i,a}$ corresponde ao score normalizado do espectro *i* e do componente principal (ou da variável latente) *a* e *k* corresponde à quantidade utilizada de componentes principais ou variáveis latentes.

MD² pode ser designada simplesmente como T^2. Um espectro com $T^2 = 0$ se projeta no ponto médio do modelo centralizado em relação ao valor de referência, ou seja, todos os scores estão no centro. Quanto mais distante a projeção de um espectro no hiperplano estiver em relação ao ponto central, maior é o valor T^2. Quanto mais próximo do ponto central, melhor é o modelo. Em distâncias grandes em relação ao ponto central, o modelo provavelmente não é bom.

Pode ser definido um **nível de significância** de p. ex. 5% para um teste de hipótese que reconhece outliers *(ver capítulo 6.5, página 99)*.

Resíduo Q

O resíduo Q do espectro *i* é o quadrado da distância do modelo PCA ou PLS em relação ao espectro:

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

Neste cálculo, $\mathbf{e_i}$ corresponde à coluna *i* da matriz residual **E**, $e_{i,j}$ corresponde ao resíduo para o espectro *i* e *f* à quantidade de comprimentos de onda.

Os resíduos Q mostram variações que não podem ser esclarecidas pelo modelo. Um alto resíduo Q mostra que o espectro possivelmente não é apropriado ao modelo, p. ex., quando uma amostra medida contiver outra substância.

Pode ser definido um **nível de significância** de p. ex. 5% para um teste de hipótese que reconhece outliers *(ver capítulo 6.5, página 99)*.

6.5 Outliers espectrais – Algoritmo

O reconhecimento de outliers espectrais identifica espectros que desviam em relação à população *(ver "Reconhecimento de outliers espectrais", página 52)*. O algoritmo avalia se os valores para Hotelling T^2 ou para resíduos Q do espectro verificado são o resultado de uma variação aleatória ou sistemática.

Reconhecimento de outliers espectrais no desenvolvimento de modelos

1. A parametrização é levada em consideração da seguinte forma:
   a. A partir do OMNIS Software versão 4.2: o usuário decide se a parametrização (pré-tratamento de dados e seleção de comprimento de onda) é aplicada ou não.
   Alterações posteriores na parametrização não têm influência na divisão do conjunto de dados.
   b. A partir do OMNIS Software versão 3.3 até a versão do OMNIS Software 4.1: o usuário decide se o pré-tratamento de dados será considerado ou não. A seleção de comprimento de onda e as alterações posteriores do pré-tratamento de dados não têm influência na divisão do conjunto de dados.
   c. Com a versão do OMNIS Software 3.2: o pré-tratamento de dados é considerado do modo como foi definido no momento do reconhecimento de outliers. A seleção de comprimento de onda e as alterações posteriores do pré-tratamento de dados não têm influência na divisão do conjunto de dados.
2. O reconhecimento de outliers espectrais é baseado no modelo PCA de todos os espectros centralizados em relação ao valor médio na lista de espectros *(ver capítulo 4.2, página 35)*. O espectro a ser testado também é registrado no modelo PCA. A quantidade de componentes principais é escolhida de forma que a variância declarada seja de pelo menos 95%.

3. Outlier **Hotelling T^2**:
   Com base nos valores T^2 *(ver "Hotelling T²", página 97)*, é executado um teste de hipóteses conforme H. Hotelling, *The Generalization of Student's Ratio*, The Annals of Mathematical Statistics volume 2, nº 3 (agosto de 1931), pág. 360–378.
   a. A hipótese nula é que o valor T^2 do espectro a examinar corresponde aos valores T^2 do modelo PCA. Se a hipótese nula for verdadeira, T^2 segue uma distribuição Hotelling T^2. A distribuição pode ser expressa como distribuição F em escala:
   $$T^2 \sim \frac{k(n-1)}{n-k} F_{k,n-k}$$
   Sendo k a quantidade de componentes principais, n a quantidade de espectros e $F_{k,n-k}$ a distribuição F com os parâmetros k e n-k.
   b. O nível de significância ajustável controla a probabilidade de rejeitar a hipótese nula caso ela seja verdadeira (procedimento conhecido como "erro tipo I" ou "falso positivo"). O valor padrão é 5%.
   c. Com base nesta distribuição e no nível de significância, é calculado um valor crítico para T^2.
   d. Se o valor T^2 para o espectro for maior do que o valor crítico, a hipótese nula será rejeitada e o espectro será identificado como potencial outlier.
4. **Outlier resíduos Q**
   Com base nos resíduos Q *(ver "Resíduo Q", página 98)*, é executado um teste de hipótese conforme J. E. Jackson e G. S. Mudholkar, *Control Procedures for Residuals Associated With Principal Component Analysis*, Technometrics volume 21, nº 3 (agosto de 1979), pág. 341–349.
   a. A hipótese nula é que o resíduo Q do espectro a examinar corresponde aos resíduos Q do modelo PCA. Se a hipótese nula for verdadeira, pode ser elaborada uma estatística de teste de resíduo Q com distribuição quase normal.
   b. O nível de significância ajustável controla a probabilidade de rejeitar a hipótese nula caso ela seja verdadeira (procedimento conhecido como "erro tipo I" ou "falso positivo"). O valor padrão é 5%.
   c. Com base nessa estatística de teste e no nível de significância, é calculado um valor crítico para Q.
   d. Se o valor do resíduo Q para o espectro for maior do que o valor crítico, a hipótese nula será rejeitada e o espectro será identificado como potencial outlier.

Avaliação de outliers na previsão (quantificação)

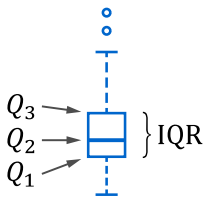
Na quantificação, há disponível uma avaliação de outliers durante a previsão:

1. Passos preparatórios:
   a. O modelo de quantificação utiliza o mesmo algoritmo descrito acima. Porém, a base é o modelo de quantificação (modelo PLS) com todos os espectros do conjunto de dados de calibração centralizados em relação ao valor médio. O pré-tratamento de dados, faixas de comprimento de onda e a quantidade de variáveis latentes são considerados no modelo de quantificação, conforme informado.
   b. Com base no nível de significância definido, o modelo de quantificação calcula os valores críticos para T^2 e Q (linhas tracejadas no gráfico de influência PLS). Os valores críticos são armazenados no modelo de quantificação.
2. Na previsão, os valores T^2 e Q do espectro são calculados com base no modelo de quantificação.
3. Se o valor T^2 ou o valor Q do espectro for maior do que o respectivo valor crítico, a amostra é considerada como potencial outlier em relação ao modelo de quantificação utilizado.

6.6 Outlier de valor de referência – Algoritmo

Os box plots possibilitam reconhecer outliers nos valores de referência *(ver capítulo 4.3.4, página 58)*.

Para considerar a assimetria da distribuição, os valores-limites de outlier são adaptados com os seguintes cálculos.



O **medcouple** (MC) mede a assimetria dos valores de referência. O cálculo começa com a mediana do box plot, Q_2. Com todos os possíveis pares (em inglês, *couples*) da metade superior e da metade inferior dos valores de referência, é calculada uma função. A mediana dos resultados é o medcouple:

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

Neste cálculo, Q_2 corresponde ao segundo quartil que define a linha central no box plot e y_i, y_j correspondem a um par de valores de referência.

O medcouple é sempre um valor entre –1 e 1. Em uma distribuição simétrica, MC = 0. Uma distribuição assimétrica com MC > 0 é deformada em direção aos maiores valores de referência, com MC < 0 ela é deformada em direção aos menores valores de referência.

O cálculo dos **valores-limites de outlier adaptados** depende do lado em direção ao qual a distribuição está deslocada:

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

Em uma distribuição simétrica (MC = 0), as distâncias entre os valores de referência e a caixa são 1,5 IQR.

A função exponencial possibilita um reconhecimento de outlier exato e robusto em diferentes distribuições com diferentes assimetrias, conforme comprovado empiricamente por M. Hubert e E. Vandervieren, *An adjusted boxplot for skewed distributions*, Computational Statistics & Data Analysis volume 52, nº 12 (agosto de 2008), pág. 5186–5201.

O percentual esperado de outliers marcados é de aproximadamente 1% e o percentual do box plot padrão para a distribuição normal é muito semelhante. Aviso: este percentual depende do nível de significância.