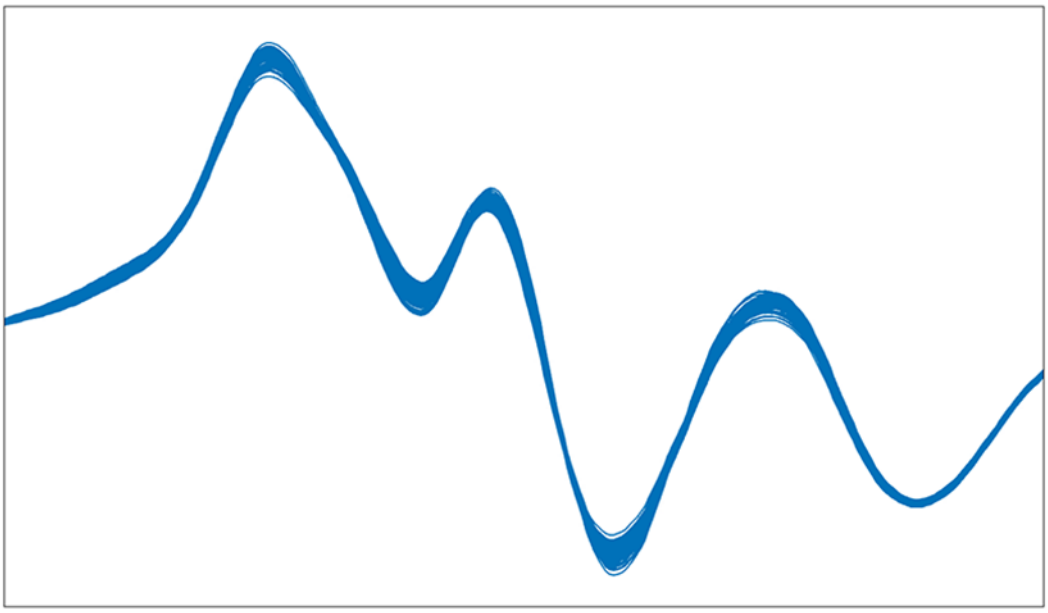
OMNIS NIR Theory



Manual

8.0600.8101EN / v9 / 2025-10-10

Metrohm

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

OMNIS NIR Theory

Manual

8.0600.8101EN / v9 /
2025-10-10

Technical Communication
Metrohm AG
CH-9100 Herisau

This documentation is an original document.

This documentation has been prepared with great care. However, errors can never be entirely ruled out. Please send comments regarding possible errors to the address above.

Disclaimer

Deficiencies arising from circumstances that are not the responsibility of Metrohm, such as improper storage or improper use, etc., are expressly excluded from the warranty. Unauthorized modifications to the product (e.g., conversions or attachments) exclude any liability on the part of the manufacturer for resulting damage and its consequences. Instructions and notes in the Metrohm product documentation must be strictly followed. Otherwise, Metrohm's liability is excluded.

Table of contents

1 Overview

1.1 Introduction

Near-infrared (NIR) spectroscopy is a non-destructive, fast, and reagent-free analysis method, that is suitable for a wide spectrum of samples. It can analyze multiple parameters at once, and determine both physical and chemical properties of a material. These include, for example, analyte concentrations, density, particle size, and intrinsic viscosity, among others.

The NIR spectroscopy also enables the identification of unknown samples (starting with OMNIS Software 4.0) and the verification of samples (starting with OMNIS Software 4.2).

The ability to measure samples remotely and without destruction is key to quality control and process monitoring.

The manual describes techniques and algorithms for acquiring, processing, and analyzing near-infrared spectra as implemented in the OMNIS Software. Chapter 2 looks briefly at how the measured signals are converted into absorption spectra. Chapter 3 covers the calibration of the instrument. Chapter 4 describes the development of models that can predict the parameter of interest (quantification) or the product membership (identification). Chapter 5 covers the prediction of unknowns. Chapter 6 forms the appendix with explanation of different algorithms.

1.2 Conceptual framework

The procedures presented fit into the following framework:

1. **Calibration, standardization, and performance tests**
   Ensuring the transferability and reliability of the absorption spectra acquired by the instrument.
2. **Model development**
   A model is developed for the prediction of a quantitative parameter or for the identification of samples.
   The development is based on samples with known parameters of interest or known product membership.
3. **Sample analysis**
   A spectrum is acquired from the sample being analyzed. Based on the spectrum, a quantification model provides a quantitative prediction, or an identification model identifies or verifies the sample.
4. **Monitoring**
   The monitoring of the model and the instrument confirms that the system is suitable for subsequent analysis.

1.3 About the documentation

Symbols and conventions

Boldface capital letters represent matrices, boldface lowercase letters represent vectors. The transpose of a matrix or vector is indicated by a superscript t, for example $\mathbf{x}^{\mathbf{t}}$.

Scalars are represented by lowercase letters. The caret (^) indicates estimated (calculated) values, for example $\hat{y}$. A dash indicates mean values, for example $\overline{y}$.

A scalar may represent a wavelength-dependent variable, for example the absorbance A. The scalar notation may be used interchangeably with the vector notation.

1.4 Further information

Additional information on the product is available on the following pages:

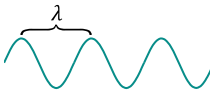
- Metrohm website *https://www.metrohm.com* – Overview of product family, documents as PDF, details of accessories, and information on applications.
- Help for the OMNIS Software *https://guide.metrohm.com* – Thematically filtered information on OMNIS Software.

2 Near-infrared light and spectra

2.1 Light and its interaction with matter

A spectrometer measures how a sample interacts with light. Light can be absorbed or scattered to varying extent. The interaction depends on the properties of the light, namely its wavelength, and on the characteristics of the material, especially its molecular structure.

Wavelength



Light is electromagnetic radiation. The light travels through space as a wave with oscillating electric and magnetic fields. The waves spread out in space and time. Accordingly, a wave is characterized by its wavelength λ (e.g., in nanometers = 10^{-9} meter) and its frequency (Hz).

The wavelength is inversely proportional to the frequency of the wave. Waves with higher frequencies (more oscillations per second) have shorter wavelengths and vice versa. Due to this relationship, the wave can be described with either the wavelength (nm) or the frequency (Hz).

Light can exchange energy in discrete quantum entities called photons. The energy of a single photon, E, depends on its frequency f or its wavelength λ:

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

Here h corresponds to Planck's constant and c to the speed of light.

Figure 1 shows different ranges of electromagnetic radiation.

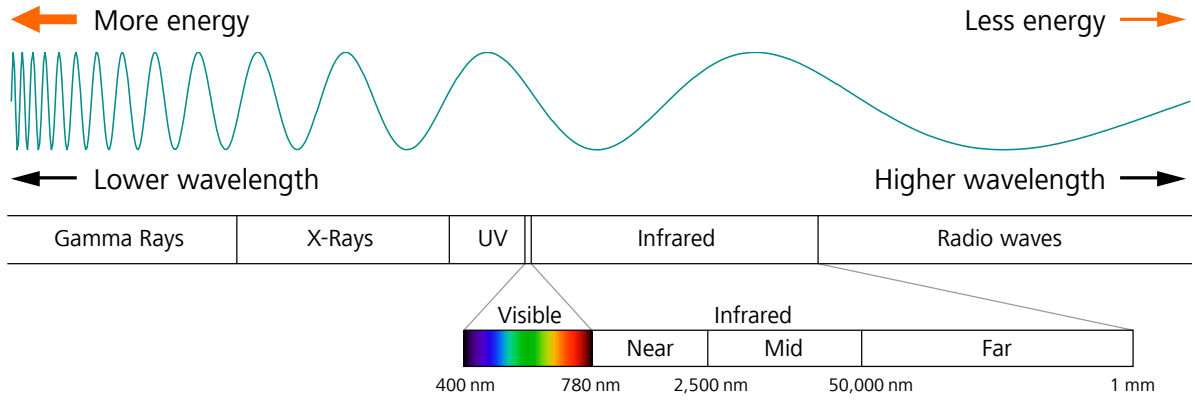


Figure 1 Regions of electromagnetic radiation. The near-infrared (NIR) region is next to the region of visible light. NIR includes wavelengths from 780 nm to 2,500 nm.

Sources of radiation

Different regions of electromagnetic radiation have different radiation sources. In the NIR region, the source is **thermal radiation**. An infrared camera, for instance, detects the human body because the body has a higher temperature than the surrounding area.

Intensity

The amplitude of the electromagnetic wave determines the intensity of the light. The higher the amplitude, the higher the intensity. For visible light, the intensity is perceived as brightness.

Interaction of light and matter

The process of the absorption of light by molecules is presented in the following.

The way light interacts with matter varies, depending on the range of the electromagnetic radiation. For instance, if energy is transferred from visible light to molecules, then electrons in the molecules switch from a lower energy level to a higher energy level (electronic transition). In the infrared region, **vibrational transitions** take place. Chemical bonds, functional groups and molecules can vibrate in various ways, i.e., as stretching vibration, deformation vibration or torsional vibration.

Molecules are able to take on only discrete vibrational states. At room temperature, most molecules are in their fundamental vibrational state (level 0). Transitions from the fundamental vibrational state to excited states are named according to the following scheme:

Vibrational transition ***i*** **➔** ***j***	**Name**
0 ➔ 1	Fundamental transition
0 ➔ 2	First overtone transition
0 ➔ 3	Second overtone transition

A higher vibrational state corresponds to a higher energy level. For the transition from state *i* to the excited state *j*, the molecule must absorb a certain transition energy ΔE_{ij}.

Light can swap energy in portions of $E = hf$, for which f is the frequency of the light. The absorption of light takes place if the photon energy hf is equal to the transition energy ΔE_{ij} ist.

The permitted vibrational states depend, among other things, on the bond strengths and the mass of the atoms involved. A particular type of bond can therefore be associated with characteristic transition energies or absorbed wavelengths.

Further conditions must be met for light to be absorbed. The vibrational transition must shift the charge distribution in such a way that the electric dipole moment of the molecule changes. The probability of an energy absorption depends on the magnitude of the change in the dipole moment along the chemical bond involved.

A vibrational transition can cause a dipole moment change in both polar and nonpolar molecules and functional groups. Homodinuclear molecules such as N_2 do not absorb infrared light.

The duration of the excited vibrational state is limited. When the molecule falls back to a lower vibrational state, the energy is transformed into heat.

The NIR spectral range

The wavelengths corresponding to the fundamental transitions are located in the mid-infrared region. The near-infrared region covers overtone transitions and combination bands. *Figure 2* shows the wavelength bands that are absorbed by various molecules and functional groups.

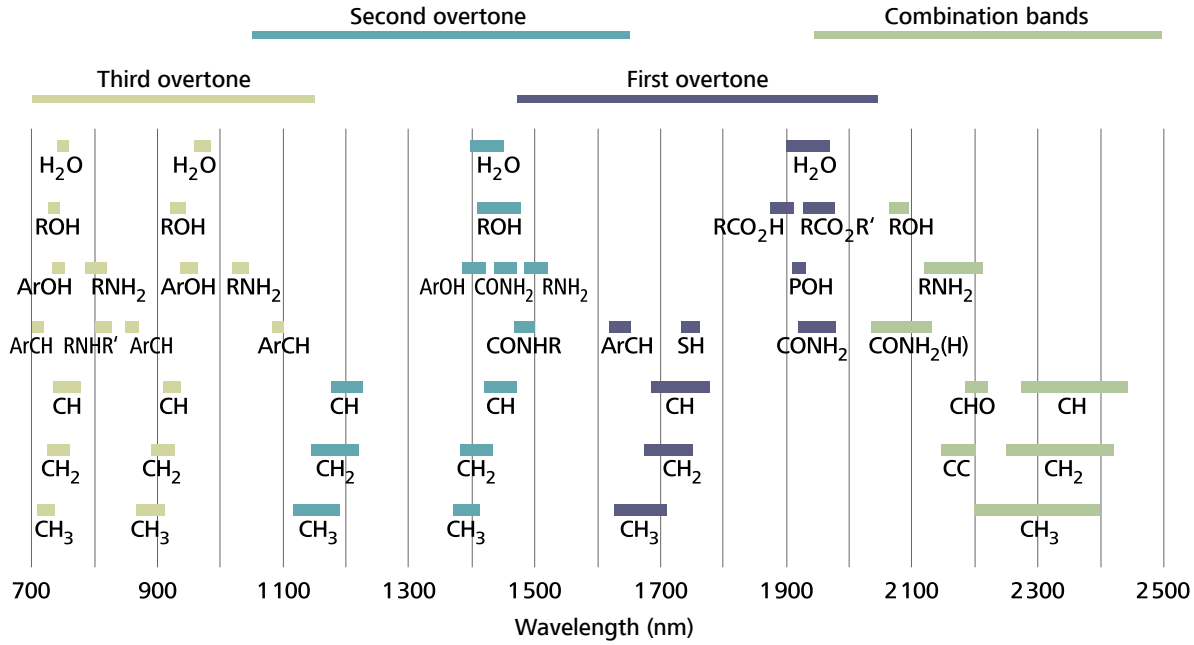


Figure 2 NIR absorption bands

The fundamental transition is the most likely transition; it happens the most often. Overtone transitions are less likely. The fundamental transition therefore absorbs more light than overtone transitions do. Generally speaking, absorption decreases with each overtone. This makes the overtones suitable for strongly absorbing molecules.

Two or more fundamental vibrations can be excited simultaneously from a single light frequency that equals the combined frequencies of the fundamental vibrations. The corresponding absorption bands are called

combination bands. Some combination bands are located in the NIR range, namely between 1,900 and 2,500 nm.

2.2 Mathematics fundamentals

2.2.1 Beer-Lambert law

The Beer-Lambert law describes how the absorption of light through a homogeneous sample depends on the properties of an absorbing substance in the sample.

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

Here A corresponds to the absorbance, ε to the molar absorptivity of the absorber (L/mol/cm), c to the concentration of the absorber (mol/L), and l to the pathlength of the sample (cm).

Molar absorptivity ε is a constant that indicates how much a substance absorbs. The molar absorptivity is specific for a given wavelength *i* and a given substance *j*. The total absorbance of a mixture is the sum of the absorbance of all of the substances contained in the mixture:

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

Here A_i corresponds to the absorbance at wavelength *i*, *N* to the number of substances in the mixture, ε_{ij} to the molar absorptivity for wavelength *i*, and substance *j*, and c_j to the concentration of substance *j*.

The Beer-Lambert law assumes a linear relationship between the absorbance and the concentration, and also a linear relationship between the absorbance and the molar absorptivity. This linear behavior is valid for many situations.

Spectroscopic absorption measurements are able to detect the following, based on the Beer-Lambert law:

- Fluctuations in the concentration of an absorber.
  This is the most common application.
- Fluctuations in the factors that affect molar absorptivity.
  Temperature, viscosity, pH value or the dielectric constant of the solvent may affect molar absorptivity. In some cases, this can be used for spectroscopic measurements.

Scattering effects are not elucidated by the Beer-Lambert law. Sometimes scattering effects can be used, for instance, to detect variations in particle size.

2.2.2 Linear regression

Single wavelength

In the simplest case, a mixture has only one absorber. According to the Beer-Lambert law, the absorbance at a particular wavelength is linear to the concentration of the absorber.

In Figure 1, each point represents a sample with a known concentration of the absorber (x-axis) and the measured absorbance (y-axis). A linear regression yields the regression line **A**.

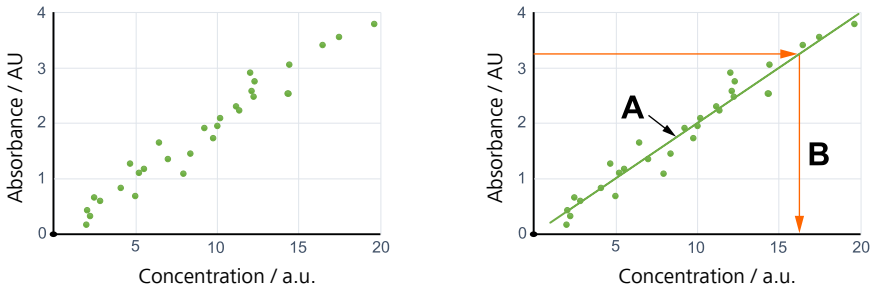


Figure 3 Relationship between absorbance values and concentrations

For a sample with unknown concentration of the absorber, the concentration can be determined as follows:

1. Measure the absorbance at the given wavelength.
2. Use the regression line to determine the concentration (**B**).
   The regression line is a simple quantification model to predict the parameter of interest, e.g., the concentration of the absorber.

However, this procedure will fail if the mixture contains multiple absorbers with varying concentrations.

Multiple wavelengths

Instead of measuring the absorption at a single wavelength, the absorbance across multiple wavelengths can be measured, thus resulting in a spectrum. Similar as above, a linear regression can extract the relationship between the spectra and the parameter of interest. For multiple wavelengths, multiple linear regression (MLR) is required.

It can be shown that multiple linear regression can predict the parameter of interest even when the mixture contains multiple absorbers with varying concentrations *(see "Linear regression example", chapter 6.1, page 84)*.

However, multiple linear regression would require a number of samples greater than the number of wavelengths. For spectroscopic predictions, other methods such as PCA or PLS can be used.

2.3 How light is converted into a spectrum

A spectrometer (or spectrophotometer) consists of a light source and a detector unit. The light source emits light with a broad spectrum of wavelengths, or in other words polychromatic light. The light interacts with the sample. The spectrometer then detects the remaining light as a function of the wavelength.

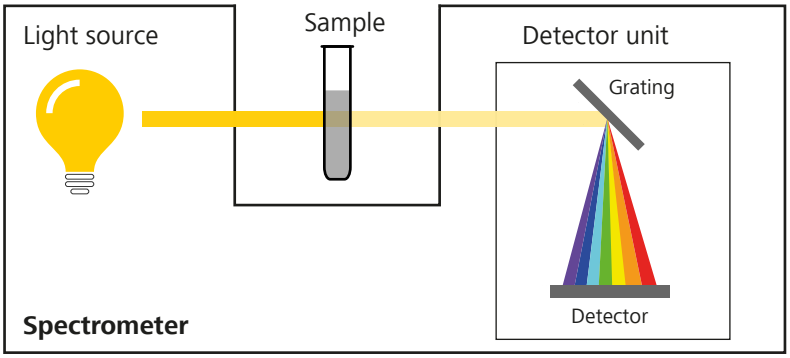


Figure 4 A spectrometer with light source and detector unit.

In the spectrometer, a grating disperses the light into discrete wavelengths. A detector measures the light, with each wavelength hitting a different element – or pixel – in a line sensor.

A **scan** is a measurement across all pixels. Each pixel generates a photoelectric signal that is proportional to the light intensity. The signals can be plotted against the pixels.

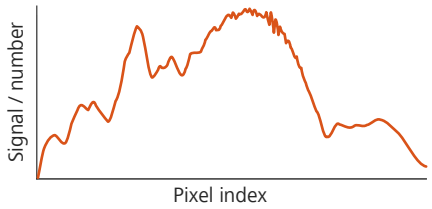


Figure 5 Spectrum of the detector signal as a function of pixels.

Integration time

Integration time is the time span during which the detector collects light. A higher integration time increases the signal.

Integration times that are too long lead to saturation of the detector and loss of information. Too short integration times reduce the signal and thus the signal-to-noise ratio.

An **automatic integration time** ensures optimal illumination, i.e., an optimal signal-to-noise ratio, without any occurrence of saturation. Multiple measurements are performed before each sample scan and each

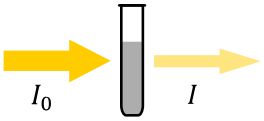
reference scan. The integration time is set so that the signal with the highest intensity reaches about 90% of the detectable range.

Differences in integration times are accounted for in further calculations.

- **OMNIS NIR Analyzer**
  The integration time is always set automatically.
- **2060 The NIR**
  The integration time can be set manually or automatically.
  The manual integration time can save measurement time for repeated, similar measurements. To avoid saturation of the detector with excessively long integration times, the integration time should be set with sufficient leeway *(see "Saturation", page 47)*.

Absorbance

Spectroscopic measurements determine how much light a sample absorbs or scatters. The sample is exposed to a light beam. A detector measures the light that is radiated by a light source and the remaining light after the interaction with the sample.



The absorbance A is defined as the decimal logarithm of the ratio between the light intensity before the beam of light interacts with the sample (I_0) and the light intensity after the beam of light interacts with the sample (I):

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

An absorbance of 1 means that 10% of the light passes through the sample, while an absorbance of 2 means that 1% passes through.

There is no measurement unit for absorbance.

Reference scan and sample scan

According to the above formula, two scans are needed to calculate an absorption spectrum. The scans measure a photoelectric signal for each pixel S:

- A **reference scan** measures the signal S_0 before the beam of light interacts with the sample.
- A **sample scan** measures the signal S after the beam of light interacts with the sample.

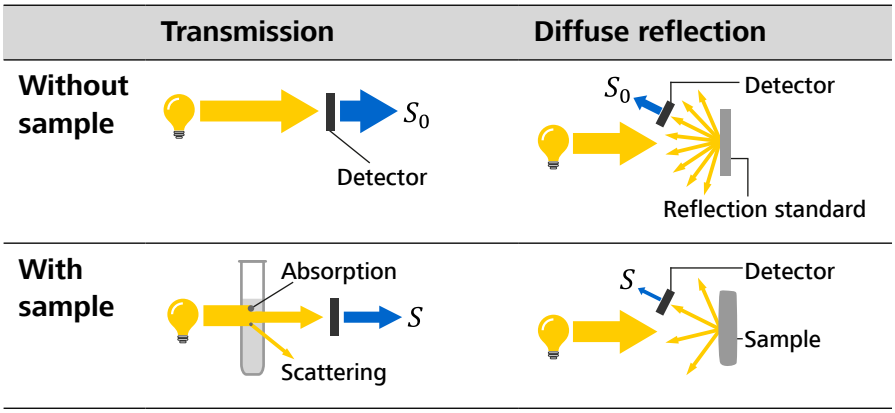
The measured photoelectric signal of a pixel is proportional to the mean value of the light intensity over the pixel area. Therefore $S_0/S = I_0/I$. The photoelectric signals can therefore be used to calculate the absorbance:

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

Transmission and reflection

In **transmission mode**, the light that passes through the sample is measured. S_0 is measured in the absence of the sample. S is measured from the light that has passed through the sample.

In **reflection mode**, the light reflected off the sample is measured. For the reference, a reflection standard is used instead of the sample. Ideally, the reflection standard reflects 100% of the light. Part of the reflected light is directed to the detector and provides the signal S_0. The signal S is measured in the same way, but with the sample that reflects the light.



The calculated absorbance A represents all the light that does not reach the detector. Therefore, A includes not only the light absorbed by the sample, but also:

- Light that does not reach the detector because it is scattered away from the detector.
- Light that is falsely scattered to the detector.

Absorption spectrum

The absorption spectrum is calculated on the basis of the reference scan (signal S_0) and the sample scan (signal S) using the above formula.

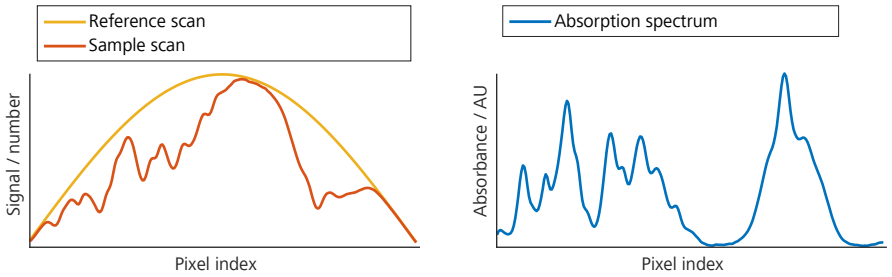


Figure 6 Reference scan and sample scan (left), and the calculated absorption spectrum (right), as a function of pixel index.

The above calculation assumes that the reference scan and the sample scan use the same optical paths or optical paths with similar optical

properties. In process environments, a multilevel referencing approach is used *(see "2060 The NIR", chapter 3.2.2, page 16)*.

From pixels to wavelengths

The pixel scale is converted into the wavelength scale. The instrument assigns each pixel to an exact wavelength, e.g.:

Pixel 6 ➔ Wavelength 1,009.4 nm

The exact wavelength for each pixel is determined by the wavelength calibration *(see "Wavelength calibration", chapter 3.1, page 13)*.

Wavelength scale conversion

The spectrum is transferred by interpolation to the standard wavelength scale:

1,000.0 nm, 1,000.5 nm, 1,001.0 nm, …

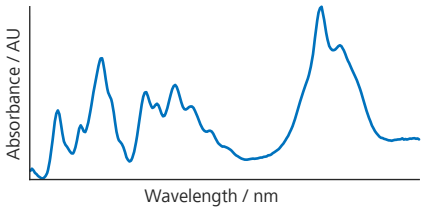


Figure 7 The absorption spectrum on the wavelength scale.

3 Instrument setup

The following steps ensure that identical spectra are obtained for a given sample with a given sample measurement setup, regardless of whether they were acquired at different times and on the same or on another instrument from the **OMNIS NIR Analyzer** product family, or a different instrument of the **2060 The NIR** type.

Both the x-axis and the y-axis of the spectra must be taken into account:

- **X-axis**: Wavelength calibration *(see "Wavelength calibration", chapter 3.1, page 13)*
- **Y-axis**: Reference standardization *(see "Reference standardization", chapter 3.2, page 14)*

Furthermore, it must be ensured that the instrument performance meets the requirements, i.e.:

- The **Instrument performance tests** must be successfully carried out before the instrument can be used to acquire spectra *(see "Instrument performance tests", chapter 3.3, page 24)*.

Standards

For the wavelength calibration and the Instrument performance tests, the instrument uses an internal, metrologically traceable **wavelength standard**.

Additionally, if the reflection mode is used, a **reflection standard** is needed for the reference standardization and the Instrument performance tests, which depends on the instrument type:

- **OMNIS NIR Analyzer**: Internal reflection standard
- **2060 The NIR**: External reflection standard

3.1 Wavelength calibration

The wavelength calibration standardizes the wavelength values, i.e., the x-axis of the spectra. It maps each pixel of the photosensor array to a wavelength.

The wavelength calibration uses an internal, metrologically traceable wavelength standard. The wavelength standard has an absorption spectrum with defined peaks and known peak positions.

The **CAL WL** command performs the following steps:

1. The absorption spectrum of the wavelength standard is acquired on the pixel scale, using an internal reference path.
2. In the acquired spectrum, the peak positions are identified with sub-pixel precision.
3. Polynomial regressions are performed using the measured peak positions on the pixel scale and the nominal peak positions of the wavelength standard.
4. The regression polynomial maps each pixel to its respective wavelength.

The regression coefficients are stored on the instrument:

- **OMNIS NIR Analyzer**: A set of regression coefficients is stored on the instrument for each sample presentation. This means that for the OMNIS NIR Analyzer Liquid/Solid, a wavelength calibration, and validation must be performed on both functional units.
- **2060 The NIR**: The regression coefficients are instrument-specific. The same set is used for all channels.

Validation of the wavelength calibration

A wavelength calibration must be validated after it is performed. The **VAL WL** command performs the following steps:

1. The absorption spectrum of the wavelength standard is acquired.
2. Validation of wavelengths:
    a. In the acquired spectrum, the peak positions are identified on the wavelength scale.
    b. The wavelength residuals between the measured peak positions and the known peak positions are calculated.
    c. For each peak, the wavelength residual must be within tolerance to pass the test.

3. Validation of bandwidths:
   a. The peak widths are determined in the acquired spectrum.
   b. The bandwidth residuals between the measured peak widths and the known peak widths are calculated.
   c. For each peak, the bandwidth residual must be within tolerance to pass the test.
4. The overall validation status is successful if all the above residuals are within tolerance.

Validation must be successfully carried out before the instrument can be used to acquire spectra.

3.2 Reference standardization

The reference standardization standardizes the absorbance values, i.e., the y-axis of the spectra.

Absorbance determination

Calculating the absorbance A of a sample requires the signals S_0 (reference scan) and S (sample scan) *(see "How light is converted into a spectrum", chapter 2.3, page 8)*:

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

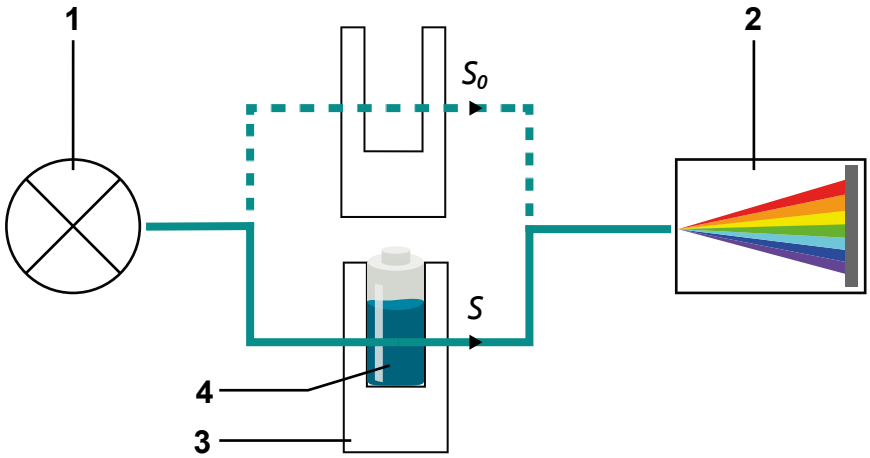


Figure 8 Optical path in transmission mode (as an example with a liquid sample presentation).

In *Figure 8*, the light passes from the light source (**1**) through a sample holder (**3**) to the detector (**2**).

The reference signal S_0 is measured without the sample, the signal S with the sample (**4**). Otherwise, the optical properties are identical for both optical paths and cause damping by the same percentage for both signals. This does not change the result of the above formula.

The equation relates S to S_0, the reference. The two signals are equally important. Any deviation in either of the two signals will result in a different absorbance value and ultimately a different spectrum.

Both S and S_0 are influenced by fluctuations in the instrument and by ambient conditions. To ensure that these influences cancel each other out, both signals should be measured within a short time span.

The implementation of this principle depends on the instrument type:

- **OMNIS NIR Analyzer**
  The absorption spectrum of the sample is calculated with the signals S and S_0 *(see "OMNIS NIR Analyzer", chapter 3.2.1, page 15)*.
- **2060 The NIR**
  In process environments, it is not practicable to use the same optical path for the S and S_0 measurements. Additional measures are therefore required *(see "2060 The NIR", chapter 3.2.2, page 16)*.

3.2.1 OMNIS NIR Analyzer

Reference standardization is accomplished by measuring the signals S_0 and S and by calculating the absorbance A.

Acquiring the spectrum of a sample

Before a functional unit can be used to acquire spectra, the Instrument performance tests must be successfully carried out on the functional unit *(see "Instrument performance tests", chapter 3.3, page 24)*.

1. The sample must be ready in the sample presentation.
2. The absorption spectrum will be calculated with the most recently acquired reference spectrum, S_0. To obtain a current value for S_0, the **MEAS REF SPEC** command can be executed.
   When using the solid sample presentation, the instrument automatically inserts a reflection standard into the optical path. This reflection standard does not require any correction of the S_0 signal.
3. The **MEAS SPEC** command measures the sample. This yields the S signal.
4. The software calculates A, the absorbance of the sample:

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

   S_0 corresponds thereby to the signal measured on the reference path and S to the signal measured via the sample.

3.2.2 2060 The NIR

Instruments of the **2060 The NIR** type need an external reference standardization.

External reference standardization

Measuring the signal S (with the sample) and S_0 (without the sample) repeatedly via optical paths with identical optical properties is time-consuming and susceptible to errors.

Two additional optical paths are therefore introduced *(see figure 9, page 16)*:

- An **internal reference** in the instrument. The internal reference path provides the S_{ref} signal, which can be easily measured.
- An additional external optical path, with the fibers connected to a **calibration fixture**. This optical path delivers the S_{fiber} signal.

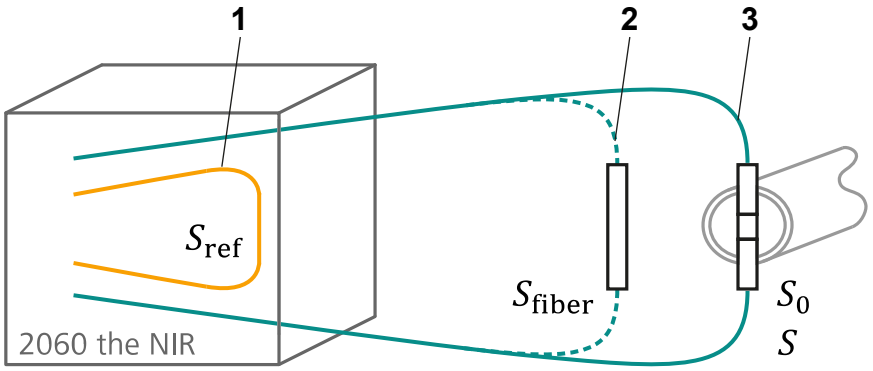


Figure 9 Optical paths as an example in transmission mode: Internal reference path (1), external fibers connected to a calibration fixture (2), and external fibers connected to the probe, with or without a sample (3). The optical paths 2 and 3 represent the same fiber optics, which are merely connected differently.

The calibration fixture fixes the fibers in place and thus forms the reference path (**2**). In transmission mode, the air serves as reference, transmitting 100% of the light. In reflection mode, the calibration fixture also takes up the reflection standard. To start with, an ideal reflection standard is assumed, reflecting 100% of the light.

The absorbance A of the sample is calculated from the S_0 and S signals. If the two additional signals, S_{ref} and S_{fiber}, are added to the numerator and to the denominator, then the result remains the same:

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

This equation can be converted into:

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

The 3 terms represent absorbances and can be denoted as:

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

Figure 10 illustrates how the S_{ref}, S_{fiber}, S_0 and S signals are measured.

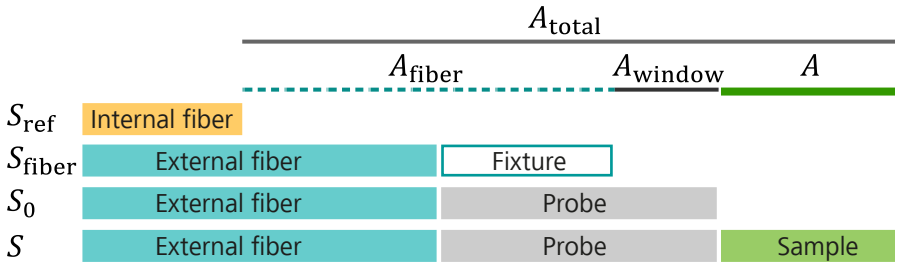


Figure 10 External reference standardization

A_{total} is the absorbance of the external fiber, the probe, and the sample, referenced to the internal fiber.

A_{fiber} is the absorbance of the external fiber plus the calibration fixture, referenced to the internal fiber.

A_{window} is the absorbance of the probe, minus the calibration fixture.

Eliminating environmental fluctuations

To determine the absorbance A of the sample, 3 absorbance values are measured. A is calculated from the 3 values as in the above equation:

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

This creates the possibility of determining the 3 absorbance values at different times. This way, fluctuations in the instrument or fluctuations in the ambient conditions can be eliminated easily for each of the 3 determinations:

- A_{total} is determined with each sample measurement. S_{ref} and S should be measured in this connection within a short time span to eliminate fluctuations.
- The **fiber optic correction**, A_{fiber}, can be determined less frequently. S_{ref} and S_{fiber} should be measured in this connection within a short time span to eliminate fluctuations.
- The **window correction**, A_{window}, can be determined less frequently, as a rule only once after installation. S_{fiber} and S_0 should be measured in this connection within a short time span to eliminate fluctuations.

When is a window correction needed?

If the calibration fixture fully represents the optical properties of the probe, then A_{window} equals 0. In this case, the window correction can be skipped.

Generally speaking, a window correction is required for the transmission mode. No window correction is required for the reflection mode. However, there are exceptions, as mentioned in the following table:

Measuring mode	Probe	Reference standardization
Transmission	Transmission pair	Fiber optic + Window
	Transmission probe	
	Immersion probe with single fibers	
Reflection	Reflection probe	Fiber optic
	Immersion probe with MicroBundle	Fiber optic + Window

To decide if a window correction is needed, each combination of calibration fixture and probe should be investigated separately. If the calibration fixture does *not* fully represent the optical properties of the probe, then a window correction will be needed.

Channels

An instrument of the **2060 The NIR** type provides multiple channels. Each channel can be connected to a different fiber and probe configuration. Reference standardization must therefore be performed separately for each channel.

All channels use the same internal reference path, i.e., the same S_{ref} signal. A multiplexer switches back and forth between the internal reference and the different measuring channels.

Performing a fiber optic correction

A fiber optic correction must be carried out after start-up or whenever the fiber configuration of a channel changes. A lamp replacement or an extreme change to the ambient conditions may also make restandardization advisable.

The procedure involves the use of a reference material:

- In reflection mode, the reference material is a reflection standard. A non-ideal reflection standard is assumed (e.g., 99%). The reflection standard has a known nominal absorption spectrum $A_{nominal}$.
- In transmission mode, the air serves as reference material. The nominal absorption spectrum is a zero line ($A_{nominal} = 0$), as it is assumed that the air absorbs no light.

Figure 11 illustrates the following procedure:

1. The external fibers must be connected to the calibration fixture.
   In reflection mode, the calibration fixture is combined with the reflection standard.
2. The **REF STD** command with the **Fiber optic** interface performs the following scans:
   a. An internal reference scan provides a value for S_{ref}.
   b. An external scan measures the external fibers, the calibration fixture, and the reference material. This yields the S_{raw} signal.
3. The software calculates A_{raw} (**Measured raw spectrum**):

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

   For this, A_{raw} corresponds to the absorbance of the external fibers, the calibration fixture, and the reference material, referenced to the internal optical path.
4. The nominal absorption spectrum of the reference material, $A_{nominal}$, is shown in the software as **Reference spectrum**. The reference spectrum must be subtracted from A_{raw} to get A_{fiber}:

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

   Here A_{fiber} corresponds to the absorbance of the fibers and the calibration fixture, as referenced to the internal optical path.
   Note: In transmission mode, $A_{nominal} = 0$ and $A_{fiber} = A_{raw}$.
   A_{fiber} stands for the fiber optic **Correction spectrum**.
5. A_{fiber} remains unchanged until a fiber optic correction for the respective channel is carried out again.

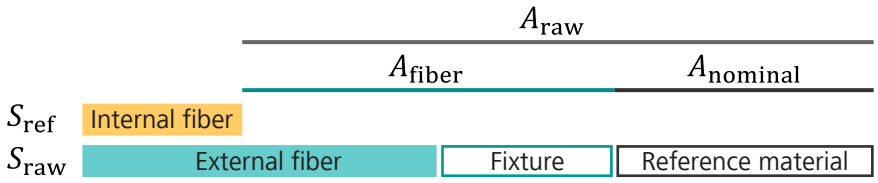


Figure 11 Fiber optic correction

Validating the fiber optic correction

The fiber optic correction must be validated using the same measuring parameters and the same calibration fixture.

The procedure involves the use of a reference material:

- In reflection mode, the reference material is a reflection standard. A non-ideal reflection standard is assumed (e.g., 99%). The reflection standard has a known nominal absorption spectrum $A_{nominal}$.
- In transmission mode, the air serves as reference material. The nominal absorption spectrum is a zero line ($A_{nominal} = 0$), as it is assumed that the air absorbs no light.

Figure 12 illustrates how validation residuals are determined:

1. The external fibers must be connected to the calibration fixture. In reflection mode, the calibration fixture is combined with the reflection standard.
2. The **VAL REF STD** command with the **Fiber optic** interface performs the following scans:
   a. An internal reference scan provides a value for S_{ref}.
   b. An external scan on the respective channel measures the external fibers, the calibration fixture, and the reference material. This yields the S_{raw} signal.
3. The software calculates A_{raw} (**Measured raw spectrum**):

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

4. A_{raw} is corrected by the fiber optic correction spectrum in order to eliminate the absorbance of the fibers and the calibration fixture:

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

   $A_{corrected}$ is shown in the software as **Measured corrected spectrum**.
5. Ideally, $A_{corrected}$ should be identical to the **Reference spectrum** $A_{nominal}$. The differences between them are calculated as **Validation residuals**:

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

   Note: In transmission mode, $A_{nominal} = 0$ und $A_{residual} = A_{corrected}$.

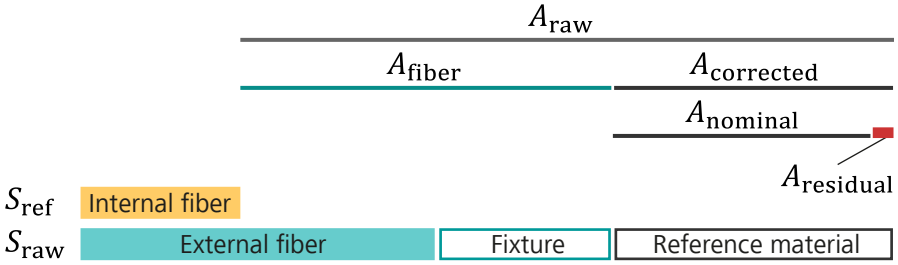


Figure 12 Residuals for the validation of the fiber optic correction

To examine the validation residuals, the wavelength range is subdivided into multiple segments. For each segment, the root mean square of the residuals over the wavelengths gives the **RMS residual** (unit: mAU):

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

Here, f corresponds to the number of wavelengths in the segment, and $A_{residual\ i}$ to the residual of wavelength i.

A predefined tolerance for A_{RMS} must be met for each segment. If the tolerance is met in all segments, then the overall validation is successful.

Validation must be successfully carried out before the instrument can be used to acquire spectra on the respective channel.

Performing a window correction

If a window correction is needed, then it must be carried out after start-up or whenever the probe or fiber configuration of a channel changes. Changes to the probe, such as contamination, may also make restandardization advisable.

The procedure involves the use of a reference material:

- In reflection mode, the reference material is a reflection standard. A non-ideal reflection standard is assumed (e.g., 99%). The reflection standard has a known nominal absorption spectrum A_{nominal}.
- In transmission mode, the air serves as reference material. The nominal absorption spectrum is a zero line ($A_{\text{nominal}} = 0$), as it is assumed that the air absorbs no light.

Figure 13 illustrates the following procedure:

1. To obtain a current value for A_{fiber}, a fiber optic correction should be performed as outlined above.
   Important: The fiber optic correction must be carried out before the window correction.
2. The external fibers must then be connected to the probe, with no sample present. If necessary, a reflection standard takes the place of the sample.
3. The **REF STD** command with the **Window** interface performs the following scans:
   a. An internal reference scan provides a value for S_{ref}.
   b. An external scan on the respective channel measures the external fibers, the probe, and the reference material. This yields the S_{probe} signal.
4. The absorbance A_{sample}, referenced to the internal optical path, is:
   $$A_{\text{probe}} = \log_{10} \frac{S_{\text{ref}}}{S_{\text{probe}}}$$
5. The software calculates A_{raw} (**Measured raw spectrum**):
   $$A_{\text{raw}} = A_{\text{probe}} - A_{\text{fiber}}$$
   Here A_{raw} corresponds to the absorbance of the probe and the reference material, as referenced to the calibration fixture.

6. The nominal absorption spectrum of the reference material, $A_{nominal}$, is shown in the software as **Reference spectrum**. The reference spectrum must be subtracted from A_{raw} to get A_{window}:

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

   Here, A_{window} corresponds to the absorbance of the probe, as referenced to the calibration fixture.
   Note: In transmission mode, $A_{nominal} = 0$ and $A_{window} = A_{raw}$.
   A_{window} stands for the window **Correction spectrum**.
7. A_{window} remains unchanged until a window correction for the respective channel is carried out again.

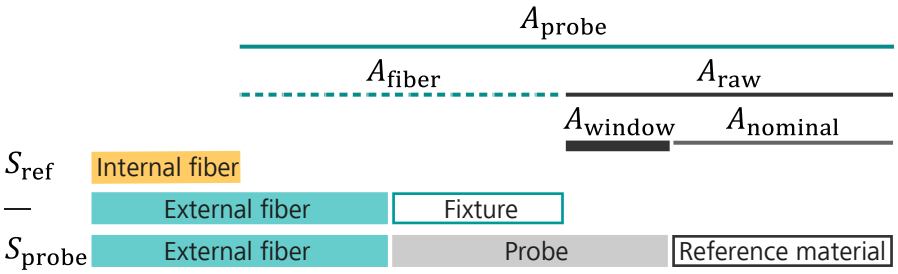


Figure 13 Window correction

Validating the window correction

The window correction must be validated using the same measuring parameters and the same calibration fixture.

The procedure involves the use of a reference material:

- In reflection mode, the reference material is a reflection standard. A non-ideal reflection standard is assumed (e.g., 99%). The reflection standard has a known nominal absorption spectrum $A_{nominal}$.
- In transmission mode, the air serves as reference material. The nominal absorption spectrum is a zero line ($A_{nominal} = 0$), as it is assumed that the air absorbs no light.

Figure 14 illustrates how validation residuals are determined:

1. The external fibers must be connected to the probe, with no sample present. If necessary, a reflection standard takes the place of the sample.
2. The **VAL REF STD** command with the **Window** interface performs the following scans:
   a. An internal reference scan provides a value for S_{ref}.
   b. An external scan on the respective channel measures the external fibers, the probe, and the reference material. This yields the S_{probe} signal.
3. The absorbance A_{sample}, referenced to the internal optical path, is:

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

4. The software calculates A_{raw} (**Measured raw spectrum**):
   $A_{raw} = A_{probe} - A_{fiber}$
5. The subtraction of the window correction spectrum from A_{raw} eliminates the absorbance differences between the calibration fixture and the probe:
   $A_{corrected} = A_{raw} - A_{window}$
   $A_{corrected}$ is shown in the software as **Measured corrected spectrum**.
6. Ideally, $A_{corrected}$ should be identical to the **Reference spectrum** $A_{nominal}$. The differences between them are calculated as **Validation residuals**:
   $A_{residual} = A_{corrected} - A_{nominal}$
   Note: In transmission mode, $A_{nominal}$ = 0 und $A_{residual} = A_{corrected}$.

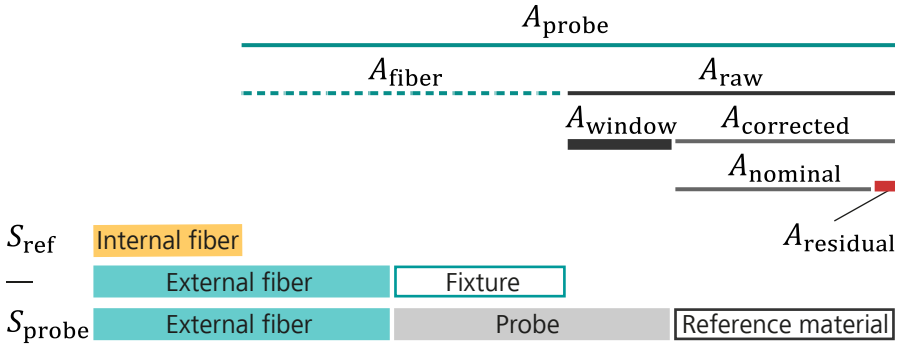


Figure 14 Residuals for the validation of the window correction

To examine the validation residuals, the wavelength range is subdivided into multiple segments. For each segment, the root mean square of the residuals over the wavelengths gives the **RMS noise** (unit: mAU):

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

Here, *f* corresponds to the number of wavelengths in the segment, and $A_{residual\ i}$ to the residual of wavelength *i*.

A predefined tolerance for A_{RMS} must be met for each segment. If the tolerance is met in all segments, then the overall validation is successful.

Acquiring the spectrum of a sample

The Instrument performance tests must be successfully carried out on the respective channel *(see "Instrument performance tests", chapter 3.3, page 24)* before the instrument can be used to acquire spectra.

The procedure for acquiring the spectrum of a sample is illustrated in *Figure 15*:

1. The external fibers must be connected to the probe. The sample must be present.
2. The absorption spectrum will be calculated with the most recently acquired reference spectrum, S_{ref}. To obtain a current value for S_{ref}, execute the **MEAS REF SPEC** command.
3. The **MEAS SPEC** command measures the sample, including the probe and the fibers. This yields the S signal.
4. The software calculates A_{total}, the absorbance of the sample, including the probe and the fibers, in reference to the internal optical path:

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

5. The absorbance of the sample is then calculated as described above, using the fiber optic correction spectrum A_{fiber} and the window correction spectrum A_{window} of the respective channel:

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

   A stands for the spectrum of the sample.

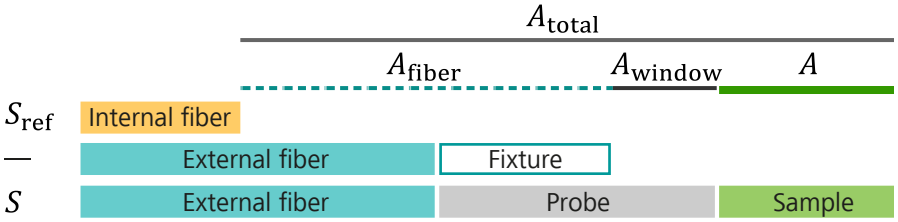


Figure 15 Acquiring the spectrum of a sample

3.3 Instrument performance tests

The Instrument performance tests can be performed via internal and via external optical paths.

- **OMNIS NIR Analyzer**
  - **Internal Instrument performance tests** (obligatory): The internal tests use the reference path of the respective sample presentation. The tests check the wavelengths and signal noise. Before running the tests, the wavelength calibration for the respective sample presentation must be successfully carried out and validated.
    The internal tests must be successfully carried out before the instrument can be used to acquire spectra on the respective sample presentation.
  - **External Instrument performance tests** (optional): The external tests support the validation according to pharmacopoeia such as USP <856>, Ph.Eur 2.2.40, and JP 2.27. Wavelengths, signal noises and photometric linearity are checked *(see "External Instrument performance tests (OMNIS NIR Analyzer)", chapter 3.3.1, page 28)*.
- **2060 The NIR**
  The tests can use the internal optical path or one of the external optical paths. The tests check the wavelengths and signal noise. Before running the tests, the wavelength calibration and the external reference standardization on the respective channel must be successfully carried out and validated.
  The Instrument performance tests must be successfully carried out before the instrument can be used to acquire spectra on the respective channel. Predefined tolerances must be met. The permissible tolerances depend on the fiber configuration of the respective channel, which is listed in the instrument-specific data (measuring mode, fiber type, and fiber length).

Wavelength test

The wavelength test investigates wavelength accuracy and wavelength precision. For this purpose, a wavelength standard is used that has an absorption spectrum with defined peaks and known peak positions:

- **Internal**: The absorption spectrum of the internal, metrologically traceable wavelength standard is determined via the internal optical path:

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

  A_{WL} corresponds thereby to the absorbance of the internal wavelength standard, S_{ref} to the signal measured on the internal reference path, and $S_{\mathrm{ref,WL}}$ to the signal measured on the internal reference path with the internal wavelength standard.
- **External** for instruments of the **OMNIS NIR Analyzer** product family: The external wavelength test is optional *(see "External wavelength test", page 28)*.

- **External** for instruments of the **2060 The NIR** type:
  The external fibers must be connected to the calibration fixture for single fibers and to the reflection standard for MicroBundle.
  The absorption spectrum of the internal, metrologically traceable wavelength standard is determined via an external optical path:

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

  A_{WL} corresponds thereby to the absorbance of the internal wavelength standard, S_{ref} to the signal measured on the internal reference path, $S_{\mathrm{fiber,WL}}$ to the signal measured on the optical path of the respective channel, with the fibers connected to the calibration fixture or to the reflection standard, respectively, and with the internal wavelength standard inserted into the optical path, and A_{fiber} to the fiber optic correction spectrum from the reference standardization.

A_{WL} also includes the absorbance of the reflection standard, which is irrelevant for the following calculations. Ideally, the peak positions of A_{WL} would be identical with the nominal peak positions of the wavelength standard.

The wavelength accuracy and the wavelength precision are checked as follows:

1. A number of absorption spectra of the wavelength standard is acquired as described above (A_{WL}).
2. The peak positions are identified in the acquired spectra.
3. For each peak position, the following statistics are calculated, based on the acquired spectra:
   a. Mean value (unit: nm)
   b. Standard deviation (unit: nm)
4. **Accuracy**: For each peak, the difference between the mean peak position and the nominal peak position must be within the predefined tolerance.
   ➨ **Note**: The nominal peak positions may vary slightly from test to test. This is because the peak positions are temperature-corrected. A similar temperature correction is made during wavelength calibration. This helps to ensure that measurements at a given temperature deliver comparable results on all instruments.
5. **Precision**: For each peak, the standard deviation must be within the predefined tolerance.
6. If the tolerances for all peaks are met, then the overall status of the wavelength test is successful.

Noise test

The signal noise can be checked internally or externally:

- **Internal**: The noise is determined as an absorbance of the internal optical path, referenced to the absorbance from another measurement on the same optical path:

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

  $S_{\text{ref},1}$ and $S_{\text{ref},2}$ correspond thereby to the signals measured on the internal reference path.
- **External** for instruments of the **OMNIS NIR Analyzer** product family: The external noise test is optional *(see "External noise tests", page 28)*.
- **External** for instruments of the **2060 The NIR** type:
  The external fibers must be connected to the calibration fixture for single fibers and to the reflection standard for MicroBundle.
  The noise is determined as the difference between the measured absorption spectrum and the nominal absorption spectrum:

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

  S_{ref} corresponds thereby to the signal measured on the internal reference path, S_{fiber} to the signal measured on the external optical path, whereby the fibers are connected to the calibration fixture or the reflection standard, respectively, A_{fiber} to the fiber optic correction spectrum from the reference standardization, and A_{nominal} to the nominal absorption spectrum of the reflection standard.
  Note: In transmission mode, $A_{\text{nominal}} = 0$.

Ideally, $A_{\text{noise}} = 0$.

The noise test performs the following steps:

1. A number of noise spectra is acquired as described above (A_{noise}).
2. The noise spectra are subdivided into different wavelength segments.
3. 3 quantities are calculated for each noise spectrum and each segment:
   a. Photometric noise (unit: mAU)
   b. Peak-to-peak noise (unit: mAU)
   c. Baseline bias of the noise (unit: mAU)
4. The mean value of the acquired noise spectra is calculated for each of the 3 quantities in every segment.
5. If all mean values are within the predefined tolerances, then the overall status of the noise test is successful.

3.3.1 External Instrument performance tests (OMNIS NIR Analyzer)

Instruments of the **OMNIS NIR Analyzer** product family can be validated according to pharmacopoeia such as USP <856>, Ph.Eur 2.2.40, and JP 2.27 (from OMNIS Software version 4.4). These tests require external, metrologically traceable reference standards. The reference standards have individual nominal absorption spectra measured with a reference instrument at room temperature.

External wavelength test

The wavelength accuracy and the wavelength precision are checked as follows:

1. The external wavelength standard (transmission or reflection) must be brought into position.
2. A number of absorption spectra of the external wavelength standard is acquired:

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

   A_{WL} corresponds thereby to the absorbance of the external wavelength standard, S_{ref} to the signal measured on the internal reference path, and S_{WL} to the signal measured via the external wavelength standard.
3. The peak positions are identified in the acquired spectra.
4. For each peak position, the following statistics are calculated, based on the acquired spectra:
   a. Mean value (unit: nm)
   b. Standard deviation (unit: nm)
5. **Accuracy**: For each peak, the difference between the mean peak position and the nominal peak position must be within the predefined tolerance.
6. **Precision**: For each peak, the standard deviation must be within the predefined tolerance.
7. If the tolerances for all peaks are met, then the overall status of the wavelength test is successful.

External noise tests

The signal noise is checked once at low photon flux (low flux) and once at high photon flux (high flux):

1. The external reference standard (transmission or reflection) for the low flux test or the high flux test must be brought into position.

2. A series of noise spectra is acquired as the difference between measured absorption spectra and the nominal absorption spectrum of the reference standard:

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

   S_{ref} corresponds thereby to the signal measured on the internal reference path, S_{ND} to the signal measured via the external reference standard, and A_{nominal} to the nominal spectrum of the reference standard.
3. The noise spectra are subdivided into different wavelength segments.
4. 3 quantities are calculated for each noise spectrum and each segment:
   a. Photometric noise (unit: mAU)
   b. Peak-to-peak noise (unit: mAU)
   c. Baseline bias of the noise (unit: mAU)
5. The mean value of the acquired noise spectra is calculated for each of the 3 quantities in every segment.
6. If all mean values are within the predefined tolerances, then the overall status of the noise test is successful.

Photometric linearity

The aim of this test is to verify a linear relationship between the reflectance (or transmittance) and the measured absorbance across the entire wavelength range:

1. Absorbance spectra of 5 reference standards with different reflectance (or transmittance) are acquired.
2. The linear relationship between reflectance (or transmittance) and the measured absorbance is ensured by a linear regression at multiple wavelengths.
3. If the slopes and y-intercepts of all regression lines are within the predefined tolerances, then the overall status of the test is successful.

4 Model development

A distinction is made between the following types of models:

- A **Quantification model** describes how a parameter of interest (e.g., water content) depends on the acquired spectra of the samples.
- An **Identification model** (from OMNIS Software version 4.0) classifies the samples into different products (e.g., different varieties of coffee beans), based on the acquired spectra.
  A product designates a specific chemical substance or a specific chemical substance with specific physical properties (e.g., particle size).

A spectrum of an unknown sample is acquired for the purpose of analyzing that sample. Depending on the application, the spectrum is used as follows:

- Quantification: Based on the spectrum, a quantification model creates a prediction, e.g., for the water content of the sample.
- Identification: Based on the spectrum, an identification model identifies the sample, e.g., as Arabica coffee.
- Verification (starting with OMNIS Software version 4.2): Based on the spectrum, an identification model verifies, for example, whether or not the sample is Arabica coffee.

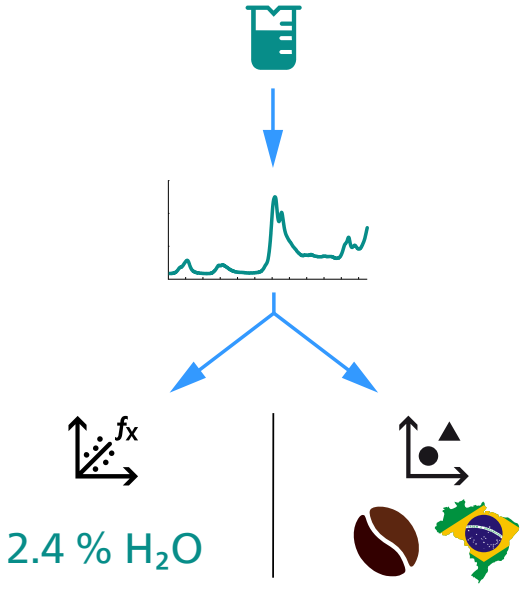


Figure 16 Quantification (bottom left) and identification (bottom right).

Developing a model and analyzing samples involves the following steps:

1. **Sampling**
   Physical samples are collected and processed:
   a. A spectrum is acquired for each sample.
   b. For quantification, a reference value for the parameter of interest (e.g., moisture content) is measured with a reference method (e.g., titration). The reference scan must be accurate and precise.
   c. The product membership of the sample must be known for the identification.
2. **Model development**
   Model development takes place in an iterative process that is comprised of the following steps:
   a. Splitting of the dataset into a calibration dataset, a validation dataset, and an outlier dataset.
   b. Application of appropriate data preprocessing and wavelength ranges to the spectra.
   c. Calculation of a model based on the calibration dataset.
   d. The validation of the model ensures that the model fulfills the requirements. The validation relies primarily on the validation dataset, which was not used in the development of the model. During quantification, the model predicts the parameter of interest for the spectra in the validation dataset. The calculated values are then compared with the known reference values. The model assigns the spectra to different products for the identification. The calculated product memberships are compared with the respective true product memberships.
3. **Monitoring**
   Monitoring the model ensures that the predictive ability does not decrease over time. Any changes in the process or the samples require a revalidation.

4.1 Physical samples

Physical samples are collected and analyzed at the start. Proper sampling is a prerequisite for developing a robust model. There are several things to take into account.

Range of variations

The samples should include any typical sample variations expected in the future. Concentrations of all chemical components and particle sizes should at least cover the expected range of variation.

The samples should cover a reasonable variation of conditions and a reasonable time span. All variations should be taken into account, for instance process fluctuations, seasonal fluctuations or fluctuations in ambient conditions.

The samples should be evenly distributed over the range of variations. For quantification, this includes the range of reference values. For instance, if the range of reference values is 1% to 10%, then the samples should be evenly distributed between 1% and 10%.

Calibration and validation samples

As a rule, 2 sample sets are used:

- Calibration set: Samples used for model development.
- Validation set: Samples used to validate the model.

Both sets should cover the expected variations. Preferably, independent samples should be collected for the validation set in order to test the robustness of the model. Scenarios such as different operators, different suppliers or different instruments, if applicable, should be taken into account.

If it is not practical to collect independent samples for calibration and validation, an alternative is to split the spectra of the available samples into a calibration dataset and a validation dataset. To ensure that the sets are still independent of one another, the automatic splitting algorithm should be used.

Once a model has been developed, the use of explicit outlier samples can be taken into account for the purpose of checking the outlier detection.

All samples should be treated in the same way. The same method for acquiring the spectra should be used, with the same hardware configuration and the same measuring parameters. The same reference method should be used for quantification, with the same measuring parameters.

Number of samples

The more variations which must be covered with respect to conditions, chemical components or particle sizes, the more samples that will be required.

Quantification: For the statistical analysis to work properly, a minimum number of around 50 samples is required, with at least 20–25 samples each in the calibration set and the validation set.

Identification and verification: For each product, the samples must cover the expected variations. The products can have different numbers of samples, with the minimum number being 3.

An initial model without a validation set can be developed with at least 10 to 20 samples (depending on the number of variations). If the cross-validation (for quantification models) or internal validation (for identification models) indicates that an adequate model can be built, then additional calibration samples and validation samples must be collected for development of the final model.

Replicates

Sometimes there are very few samples in a particular range of conditions or reference values. To compensate for that, one may be tempted to replicate these samples. Doing so would however be problematic. If duplicates of a sample are in both the calibration set and the validation set, then the figures of merit will be misleading (too optimistic). Duplicates in the same set are also to be avoided.

Reference method (quantification)

During quantification, a reference method is used to measure the reference values. The **standard error of laboratory (SEL)** for the reference method used plays a significant role in the development of a quantification model. The SEL is the standard deviation of the differences between the measurements of duplicate samples.

The SEL is often the largest error contributing to the standard error of prediction (SEP) for the NIR method *(see "SEP – Standard error of prediction", page 66)*. The SEL should not exceed 0.7 times, preferably 0.5 times, the SEP required. The range of reference values should be at least 3 times (preferably 5 times) that of the SEL.

One way to reduce the SEL is to perform repeated reference scans on each sample. The average of the measured values should be defined as the reference value for the sample. The same number of reference scans should be carried out for each sample. The figures of merit will be expressed relative to a defined number of repeated reference scans. Varying numbers of repeated reference scans would lead to incorrect estimates of the figures of merit and should be avoided.

Sample temperature

Sample temperature significantly affects spectra acquired from liquids containing water or other hydrogen bonds. Spectra of other polar liquids may also be affected, as well as spectra of solids containing water, moisture, or solvents. Such samples should be measured at a defined temperature.

Outliers

Some samples may be identified later as outliers. An outlier is a sample that deviates from the majority of the samples for one reason or another. To avoid a negative influence on the model, samples flagged as outliers are not included in the calculation of the model.

There are different types of outliers:

- **Spectral outliers**
  If the acquired spectrum of a sample differs from most other spectra, then the sample may be detected as a spectral outlier *(see "Spectral outliers", chapter 4.3.3, page 48)*.

- **Reference value outliers (quantification)**
  During quantification, individual reference values may exhibit anomalies and may be detected later as reference value outliers *(see "Reference value outliers (quantification)", chapter 4.3.4, page 54)*.
  The OMNIS Model Developer (OMD) detects the respective samples as outliers, based on outlier detection according to ASTM D8321-22 *(see "OMNIS Model Developer (OMD)", chapter 4.4.3, page 67)*.

4.2 Principal Component Analysis (PCA)

The spectroscopic data of the calibration samples contains a large number of variables (wavelengths). The variables are highly correlated. The data is therefore highly redundant. Latent variable models such as PCA and PLS are used to deal with this kind of data.

Principal component analysis (**PCA**) focuses on the spectra, without taking reference values into account.

The OMNIS Software uses PCA during the model development for the automatic dataset splitting and the spectral outlier detection.

Preliminary steps

The following preliminary steps are required:

1. **Data preprocessing**: The OMNIS Software applies the specified data preprocessing to the spectra *(see "Data preprocessing", chapter 4.3.1, page 38)*.
2. **Wavelength ranges**: The OMNIS Software applies the defined wavelength selection to the spectra *(see "Wavelength ranges", chapter 4.3.2, page 46)*.
3. **Mean centering**: For each wavelength, the mean absorbance value is calculated and subtracted from the respective value in each spectrum.

The first principal component

After the preliminary steps, PCA rearranges the information in the spectral data and separates the relevant data from the noise. To achieve this, PCA transforms the wavelength variables into a new variables space called **principal components** (**PC**).

PCA transforms the relevant information from a multitude of wavelength variables into only a few principal components. To illustrate the concept with a simple example, suppose that instead of thousands, there are only 2 wavelength variables, and that these 2 variables are reduced to 1 principal component.

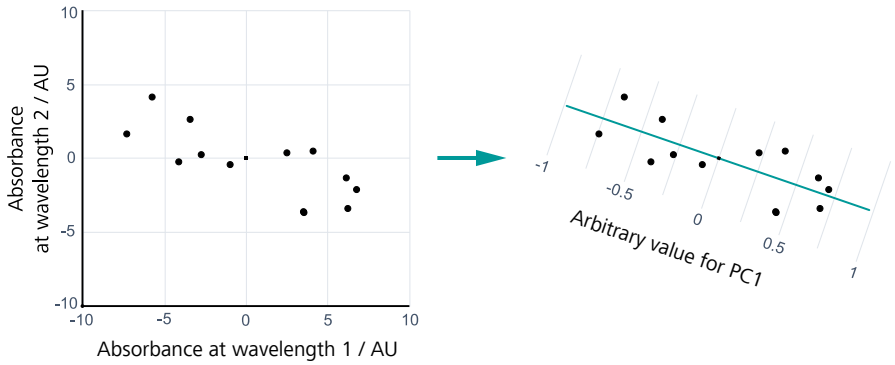


Figure 17 Points representing spectra in a 2-dimensional wavelength space (left). The same points shown in a 1-dimensional principal component space (right).

In *Figure 17* on the left, the horizontal and vertical axes form the original wavelength space with 2 variables. Thus, each point represents a spectrum with only 2 wavelengths. The mean value of all wavelength values forms the zero point.

On the right, the principal component PC1 is the direction through the data that explains the maximum variance. In this example, PC1 is the only variable in the principal component space. As a result, the 2 original variables are reduced to 1.

Scores and residuals

Figure 18 shows the magnitudes that characterize a spectrum *i*:

- The distance s_i from the center, measured in the principal component space. In the example with only 1 principal component, s_i is measured in the direction of PC1. The distance s_i is called the **score** of spectrum *i*.
- The offset e_i from the principal component space to the spectrum. The distance e_i is called the **residual** of spectrum *i*.

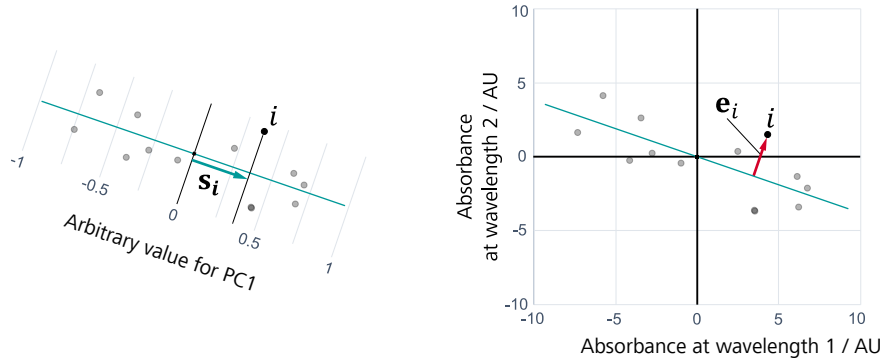


Figure 18 Spectrum i with score (left) and residual (right).

The score s_i is measured in the principal component space. The residual e_i is measured in the original wavelength space.

Transformation to multiple principal components

Usually more than one principal component is required for an adequate description of the spectroscopic data.

In *Figure 19* there are 3 original variables x_1, x_2, x_3. Each point represents a spectrum with 3 wavelengths.

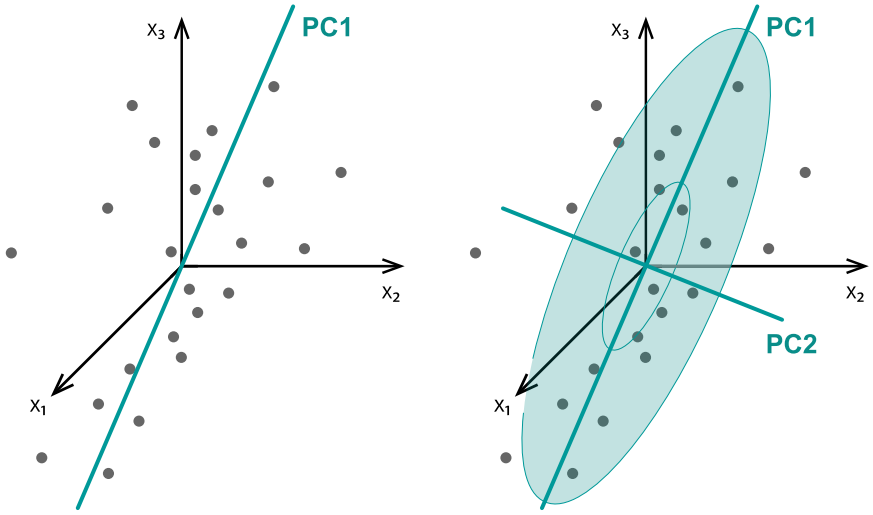


Figure 19 3 original variables are reduced to 1 principal component (left) or 2 principal components (right). PC1 and PC2 form a 2-dimensional principal component space.

The first principal component PC1 is once again the direction through the data that explains the maximum amount of variance.

The second principal component PC2 is the direction through the data that explains the maximum amount of remaining variance. The same is true for all subsequent principal components, each one describes the respective maximum amount of remaining variance. The first few principal components thus account for most of the variance in the data, while others contain mainly noise and can be discarded. In this way, the number of variables can be reduced.

As an essential characteristic of PCA is that all principal components are **orthogonal** (at right angles) to one another. Therefore, the scores are uncorrelated.

Mahalanobis distance

As seen above, the score of a spectrum *i* is measured in the principal component space, while the residual is measured in the original wavelength space.

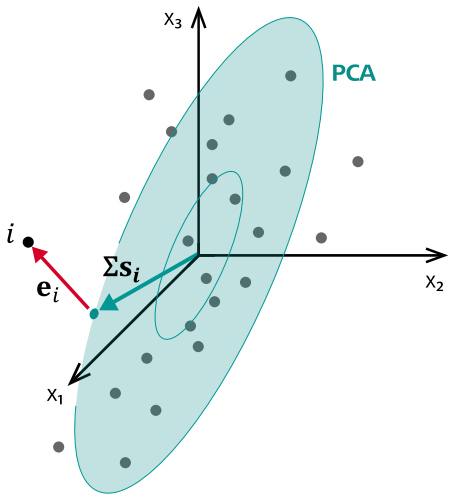
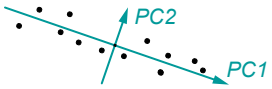


Figure 20 Score and residual of spectrum i. The green point is the orthogonal projection of point i (representing spectrum i) onto the principal component space.

In *Figure 20*, the score vector $\mathbf{\Sigma s}_i$ represents the absolute distance (Euclidean distance) from the center of the PCA model to the orthogonal projection of the spectrum onto the principal component space.



In the example, the Euclidean distances of the spectra are more spread out in the direction of PC1 than in the direction of PC2. The spreading can be measured as **variance**. The variance in PC1 is higher than in PC2.

The normalized score vector $\mathbf{s}_i$ represents a normalized distance, referred to as the **Mahalanobis distance**. The Mahalanobis distance accounts for the different variance in different principal component directions. Each direction is given the same weight. Therefore, a small Euclidean distance in a low variance direction can count as much as a large Euclidean distance in a higher variance direction.

Transforming spectra with numerous wavelengths

The same concepts apply when transforming spectra with a multitude of wavelength variables into principal components. In *Figure 21*, each spectrum is represented by a curve (left) and a point (right).

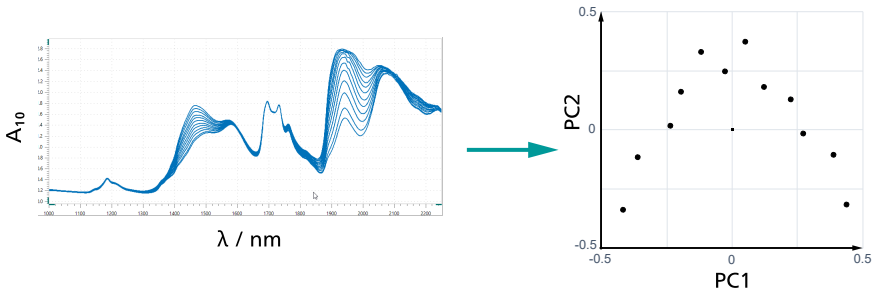


Figure 21 Transformation of spectral data into a principal component space. On the right, scores are expressed in arbitrary units.

The figure on the right shows the first 2 principal components, PC1 and PC2. Subsequent principal components PC3, PC4, etc. can also be visualized in the same fashion.

A PCA model uses a fixed number of principal components. The more principal components, the more relevant spectral variations the model explains. At the same time, however, the model also captures more irrelevant spectral variations (noise). A balanced compromise is required.

When the OMNIS Software performs a principal component analysis, the number of principal components is selected so that the explained variance is at least 95%.

PCA algorithm

There are several ways to transform the original data into a principal component space. The OMNIS Software performs a singular value decomposition *(see "PCA algorithm", chapter 6.2, page 88)*.

4.3 Data preparation

4.3.1 Data preprocessing

Spectroscopic models rely on the relationship between absorbance values and either the parameter of interest (quantification) or the product membership (identification, verification). **Parameterization** of the spectra ensures that the spectra are in good shape to express this relationship. The aim is to eliminate the irrelevant variance without losing the useful information. Artifacts and nonlinearities are corrected. If done properly, parameterization increases the accuracy and robustness of the model, as well as the repeatability and reproducibility of the predictions.

The parameterization is applied to the calibration dataset, the validation dataset and the outlier dataset, as well as to all future unknown samples analyzed with the same model.

Data preprocessing is the first parameterization step. All data preprocessing is performed in the order specified. The relevant wavelength

ranges can be specified in a second parameterization step *(see "Wavelength ranges", chapter 4.3.2, page 46)*.

Noise reduction

Spectra can contain different kinds of random fluctuations around the signal. Examples are high-frequency noise associated with the instrument's detector and electronic circuits, or low-frequency noise induced by instrument drift during the scanning measurements.

The spectrometer provides a spectrum that is averaged from a number of individual measurements. This significantly reduces high-frequency noise. Further noise reduction can be achieved with a smoothing filter. These filters are based on the idea that the noise is high-frequency and the signal is low-frequency. They approximate the signal by the neighboring absorbance values and reduce the noise by averaging.

Scatter correction

Scattering refers to the change in direction of light caused by interaction with the sample. Stray light that does not reach the detector results in baseline fluctuations of the spectra.

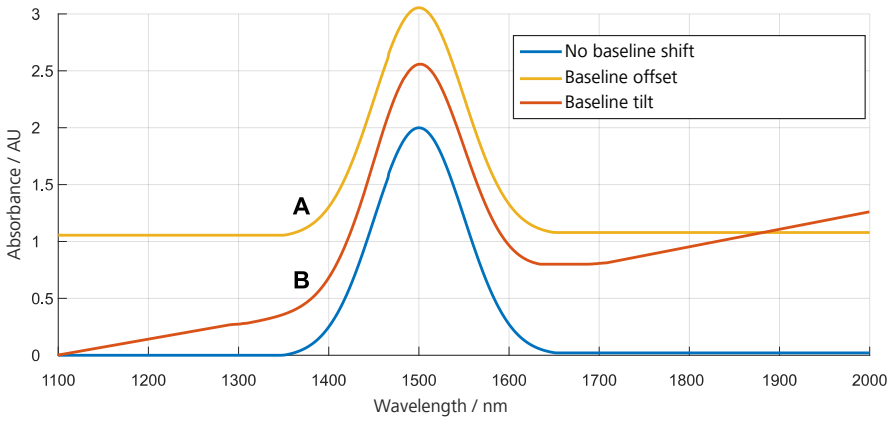


Figure 22 The main types of baseline shift are baseline offset and baseline tilt.

Various types of baseline shifts can be distinguished:

- A constant additive factor, resulting in a **baseline offset** (spectrum **A**).
- A wavelength-dependent multiplicative factor, resulting in a **baseline tilt** (spectrum **B**).
- A second-order wavelength-dependent multiplicative factor, resulting in a **quadratic baseline tilt**.
  Higher order baseline tilts can also occur.
- Absorbance-dependent multiplicative factors, resulting in **amplification**. Amplifications, however, are irrelevant.

Scattering is most evident in solid samples. The resulting baseline shifts can be used to detect changes in particle size or other physical variations.

If the interest is in chemical variations, however, baseline shifts should be minimized by appropriate preprocessing.

Data preprocessing

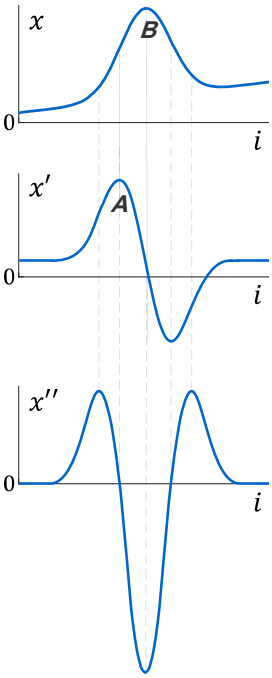
All subsequent data preprocessing is applied to each case to an individual spectrum. No other spectra are involved in the calculations.

Data preprocessing changes the signal values. The numbers on the y-axis no longer have any meaning.

The applied data preprocessing involves linear transformations. Therefore the Beer-Lambert law is still applicable.

4.3.1.1 Derivatives

Derivatives can be performed using the Gap-Segment filter or the Savitzky-Golay filter.



The derivative of a spectrum describes the slope or steepness of the curve at any given point. The slope is the rate of change of the initial spectrum.

In the spectrum, x_i is the absorbance for the wavelength i. The first derivative x_i' represents the slope of the spectrum at wavelength i. The first derivative has a maximum (**A**) where the initial spectrum is steepest. The first derivative equals 0 where the initial spectrum exhibits a peak (**B**).

The first derivative removes baseline offsets and transforms baseline tilts into baseline offsets.

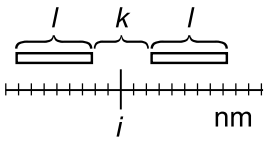
The second derivative x_i'' corresponds to the slope of the first derivative for the wavelength i. Positive peaks in the original spectrum (**B**) become negative peaks and vice versa.

The second derivative removes baseline offsets and baseline tilts from the original spectrum.

Caution is required if the original spectrum contains a significant level of noise. Each derivative worsens the signal-to-noise ratio considerably. For this reason, derivatives are combined with a smoothing function in the Gap-Segment filter or the Savitzky-Golay filter.

4.3.1.2 Gap-Segment

The Gap-Segment filter smoothes the spectrum. Optionally, the Gap-Segment filter performs a first or second derivative. The calculation depends on whether derivatives are used:



- **Derivative order 0**: For each wavelength i, the Gap-Segment filter calculates the mean value from 2 segments with segment size l, e.g., 10 nm. The 2 segments are separated by a distance of the size k, e.g., 5 nm.

- **Derivative order 1**: For the first derivative, the mean values of the 2 segments are calculated separately. The difference between the two mean values is then calculated.
- **Derivative order 2**: The second derivative can be calculated in the same way from the first derivative.

At the beginning and at the end of the spectrum, $l + k/2$ wavelengths are calculated using zero values for segment wavelengths that fall outside the spectrum.

At the beginning and at the end of the spectrum, zero values are used for segment wavelengths that fall outside the spectrum.

Smoothing may be accompanied by a slight shift of the peaks and a certain amount of bias.

Parameter Settings

Greater smoothing is achieved with:

- a lower derivative order,
- a larger segment size,
- a larger gap size.

Excessive smoothing results in the loss of relevant variance, reducing the predictive ability of the model.

4.3.1.3 Savitzky-Golay

Like the Gap-Segment filter, the Savitzky-Golay filter smoothes the spectrum and optionally performs a first or second derivative. However, the Savitzky-Golay filter uses a different smoothing technique.

For each wavelength *i*, the Savitzky-Golay filter fits a low-degree polynomial in the range of the respective wavelength. The value of the polynomial at wavelength *i* is the smoothed value. If a derivative is carried out, then the value of the derivative is used.

A weighted sum of neighboring values calculates everything at once:

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

k corresponds here to filter width, c_j to convolution coefficients that depend on the derivative order, polynomial degree and filter width and can be looked up in tables, and x_{i+j} to the absorbance values of the initial spectrum at wavelength *i+j*.

At the beginning and at the end of the spectrum, extrapolated values are used for filter wavelengths that fall outside the spectrum (horizontal extrapolation).

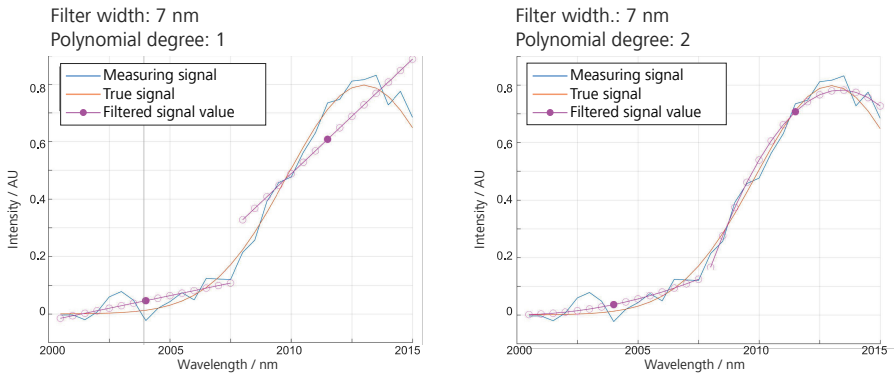


Figure 23 Savitzky-Golay filtering with different polynomial degrees

Figure 23 illustrates the Savitzky-Golay filtering. The filter width is 7 nm. The polynomials for the wavelengths 2004 nm and 2011.5 nm are shown. The new values are the filled points or, if derivatives are used, their derivative. All other wavelengths are treated in the same way.

The filter width defines the wavelength range into which each polynomial is fitted. The convolution is weighted such that the influence of the absorbance values decreases towards both sides of the respective wavelength.

Parameter Settings

Greater smoothing is achieved with:

- a lower derivative order,
- a larger filter width,
- a lower polynomial degree.

Excessive smoothing results in the loss of relevant variance, reducing the predictive ability of the model.

4.3.1.4 SNV – standard normal variate

SNV normalizes an individual spectrum to variance 1 and mean value 0. SNV normalizes the absorbance values x_i for each wavelength i within a defined wavelength range as follows:

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

Here, m corresponds to the mean value and s to the standard deviation of all absorbance values within the defined wavelength range.

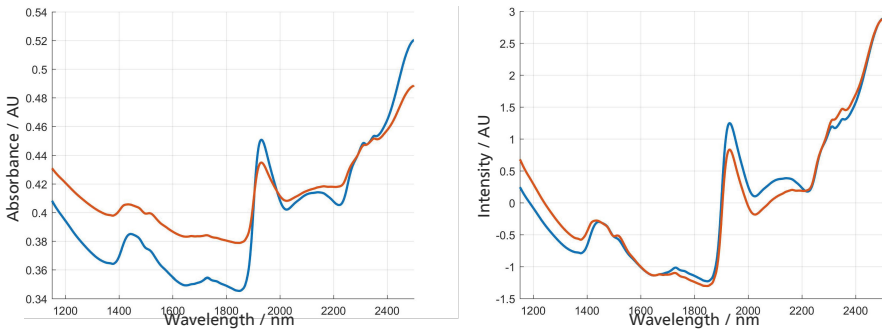


Figure 24 Absorption spectra (left) and SNV-treated spectra (right).

Normalization removes the variance between spectra. This is useful if the variance results from properties that are not of interest, e.g., varying pathlengths in granular or powdery samples or in turbid media.

Note: If a derivative is used after SNV, then some of the removed variance may reappear. Derivatives should therefore be applied before SNV. If SNV must be performed in exceptional cases prior to the derivative, then the following order may be taken into account: Detrend, SNV, derivative. However, the typical sequence is: Derivative, SNV, detrend *(see "Sequence for multiple data preprocessing steps", chapter 4.3.1.7, page 46)*.

Parameters

- **Wavelength ranges**
  If artifacts affect certain wavelength ranges (e.g., due to saturation or strong noise), then these ranges can be excluded.
  Only the defined wavelength ranges are taken into account when calculating the mean value and standard deviation. The subsequent standardization is carried out both in the defined wavelength ranges and in the intermediate areas. For excluded wavelengths at the beginning of the spectrum, the normalized value of the adjacent start wavelength is adopted. For excluded wavelengths at the end of the spectrum, the normalized value of the adjacent end wavelength is adopted.
  If necessary, excluded wavelengths can also be excluded for the model calculation *(see "Wavelength ranges", chapter 4.3.2, page 46)*.
  Note: Starting with OMNIS Software version 4.6, multiple wavelength ranges can be defined.

4.3.1.5 Detrend

Detrend fits a second-order polynomial to the entire spectrum with the aid of the method of least squares. Detrend then subtracts the polynomial from the spectrum.

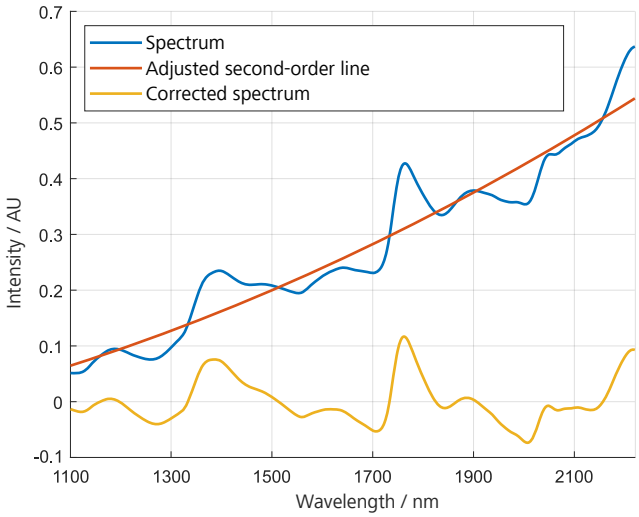


Figure 25 Detrend transforms the blue spectrum into the yellow spectrum.

Detrend reduces wavelength-dependent scattering effects, up to quadratic baseline tilts.

The figure above shows a spectrum (blue) in which the trend dominates. If this dominating trend is similar for all spectra, then detrend may work well. In other cases, detrend tends to remove useful variations. In these cases, derivatives are probably a better choice.

Since an individual polynomial is fitted to each spectrum, additional interfering variance may arise. As a rule, SNV is applied before detrend. This gives more robust estimates of the polynomial coefficients.

Parameters

- **Wavelength ranges**
  If artifacts affect certain wavelength ranges (e.g., due to saturation or strong noise), then these ranges can be excluded.
  A polynomial is fitted to all intensity values of the defined wavelength ranges. The polynomial is then subtracted from the spectrum in all defined wavelength ranges. The intensity value is set to zero for all excluded wavelengths.
  If necessary, excluded wavelengths can also be excluded for the model calculation *(see "Wavelength ranges", chapter 4.3.2, page 46)*.
  Note: Starting with OMNIS Software version 4.6, multiple wavelength ranges can be defined.

4.3.1.6 Overview of data preprocessing

Preprocessing	Purpose	Positive effects	Negative effects
Gap-Segment	Smoothing Greater smoothing is achieved with a lower derivative order, a larger segment size or a larger gap size.	▪ Reduces high-frequency noise.	▪ Excessive smoothing results in the loss of relevant variance.
Derivatives within Gap-Segment	Baseline correction	▪ First derivative: Removes baseline offsets. ▪ Second derivative: Removes baseline offsets and baseline tilts.	▪ Increases the noise. ▪ Changes the appearance of the spectrum.
Savitzky-Golay	Smoothing Greater smoothing is achieved with a lower derivative order, a larger filter width or a lower polynomial degree.	▪ Reduces high-frequency noise.	▪ Excessive smoothing results in the loss of relevant variance.
Derivatives within Savitzky-Golay	Baseline correction	▪ First derivative: Removes baseline offsets. ▪ Second derivative: Removes baseline offsets and baseline tilts.	▪ Increases the noise. ▪ Changes the appearance of the spectrum.
SNV – standard normal variate	Scatter correction *	▪ Removes baseline offsets.	▪ Relevant variance may be removed.
Detrend	Scatter correction *	▪ Removes baseline offsets. ▪ Removes baseline tilts and quadratic baseline tilts.	▪ Relevant variance may be removed. ▪ Irrelevant variance may be created.

* Note: Wavelength ranges containing artifacts (e.g., saturation or high noise) should be excluded.

4.3.1.7 Sequence for multiple data preprocessing steps

If multiple data preprocessing steps are used, then the sequence can be decisive. The basic rule is the following.

Preferably, Gap-Segment or Savitzky-Golay should be applied prior to SNV, and SNV before detrend.

Example with a first derivative and SNV: The first derivative transforms baseline tilts into baseline offsets. A subsequent SNV removes these offsets. If the sequence is reversed, then the SNV does not change the baseline tilts. The subsequent first derivative transforms them into baseline offsets. The offsets would remain.

Example with a second derivative and SNV: Baseline offsets and baseline tilts are removed in all cases. However, applying the second derivative and SNV in the proper sequence enables the removal of quadratic baseline tilts. The second derivative transforms quadratic baseline tilts into baseline offsets. A subsequent SNV removes these offsets. If the sequence is reversed, then the SNV does not change the quadratic baseline tilt. The subsequent second derivative transforms it into baseline offsets. The offsets would remain.

4.3.2 Wavelength ranges

The second parameterization step is carried out after data preprocessing *(see "Data preprocessing", chapter 4.3.1, page 38)*: The selection of wavelength ranges enables the exclusion of ranges that are not suitable for the purpose. In particular, noisy or saturated wavelength ranges might deteriorate the subsequent calculations and should be excluded.

Noise

Noise occurs at high absorbance values, where only a small amount of light reaches the detector. The following figure shows noisy ranges.

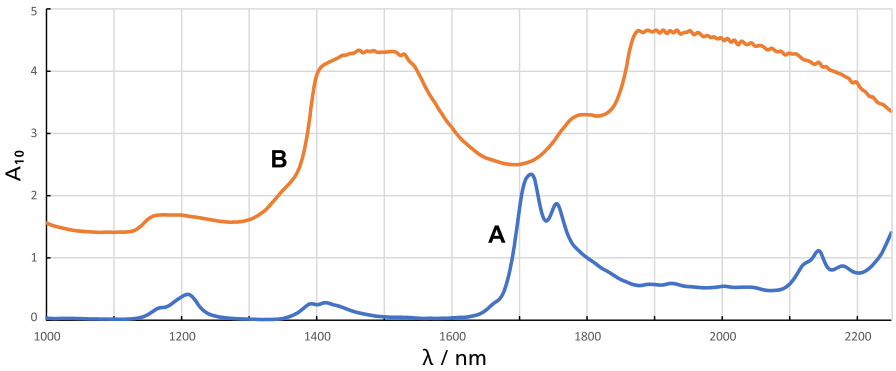


Figure 26 Example with noisy wavelength ranges.

Spectrum **A** has normal shapes of the peak. Spectrum **B** has 2 noisy regions: One from 1,400 to 1,550 nm and one above 1,870 nm. The

noisy regions are very noisy and do not resemble a bell curve or a combination thereof.

Saturation

Saturation of the detector can occur if a large amount of light reaches the detector, i.e., at low absorbance values.

- **OMNIS NIR Analyzer**
  The integration time is always set automatically. This prevents saturation and minimizes the noise *(see "Integration time", page 8)*.
- **2060 The NIR**
  If automatic integration time is activated, then no saturation will occur. If manual integration time is activated, then excessively long integration times can lead to saturation of the detector. Saturated ranges appear at low absorbance values, but are not always easy to detect visually. Therefore, the manual integration time should be set with sufficient leeway *(see "Integration time", page 8)*.

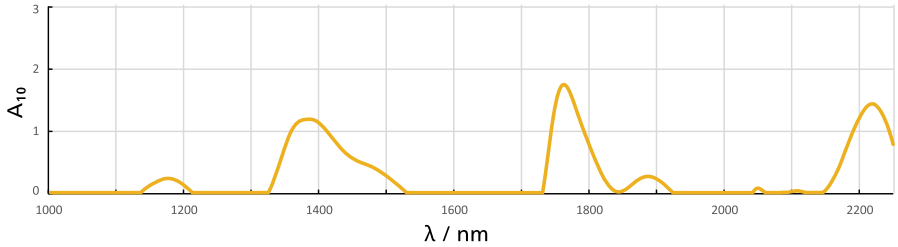


Figure 27 Example with saturated wavelength ranges.

Other reasons

There are other reasons to include or exclude wavelength ranges. The selection can be based on knowledge of the parameter of interest and its respective absorption bands *(see "Light and its interaction with matter", chapter 2.1, page 3)*. However, it must be taken into account that – depending on the type of preprocessing – the relevant information may be shifted to other wavelength ranges.

If the data preprocessing has introduced anomalies at the start and at the end of the spectra, then those wavelengths can be excluded.

Variations in chemical components or fluctuations in ambient conditions can affect certain wavelength ranges. Excluding these wavelength ranges may improve the robustness of the model.

Caution must be exercised when excluding well-formed wavelength ranges. Ranges that seemingly have no information may actually provide hidden and important information. They may be useful for detecting outliers or when processing interfering absorption bands. In fact, interfering absorption bands are the main reason for performing multivariate measurements *(see "Linear regression example", chapter 6.1, page 84)*.

4.3.3 Spectral outliers

A spectrum that differs from most other spectra is called a spectral outlier.

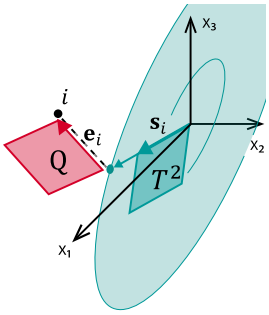
Outliers should be carefully investigated. An outlier may distort the model, if for instance a contaminated sample or a measurement error is the cause. In this case, the outlier should not be taken into account when calculating the model.

On the other hand, an outlier may represent properties that are not well covered by the other spectra. In this case, the outlier will actually improve the model. If the outlier appears to represent a valid sample, a check should be carried out to verify that the calibration samples are evenly distributed across the range of variations.

Hotellings T^2 and Q residuals are suitable measures for detecting outliers.

Hotellings T² and Q residuals

When transforming spectroscopic data into a principal component space, the spectra can be characterized by their scores and residuals *(see "Principal Component Analysis (PCA)", chapter 4.2, page 34)*. The same is true for a transformation into a latent variables space *(see "PLS regression", chapter 4.4.1, page 57)*.



Example: A 3-dimensional wavelength space (x_1, x_2, x_3) is transformed into a 2-dimensional space (green). The point *i* represents the spectrum *i* and is projected from the 3-dimentional space into the 2-dimentional space. This results in the following:

- A score vector $\Sigma\mathbf{s}_i$ within the 2-dimentional space, or its normalized score vector $\mathbf{s}_i$, respectively, which represents the Mahalanobis distance.
- A residual $\mathbf{e}_i$ within the 3-dimensional space.

The following magnitudes can be derived from $\mathbf{s}_i$ and $\mathbf{e}_i$*(see "Hotelling's T² and Q residuals", chapter 6.4, page 91)*:

- **Hotellings *T²***, or simply ***T²***, is the squared Mahalanobis distance, i.e., the squared normalized distance from the model center to the orthogonal projection of the spectrum onto the principle component space or latent variables space, respectively.
  If all scores of a spectrum correspond to the mean value, then T^2 is 0 and the spectrum is in the center of the model. The model fits best close to the center.
  Far from the center, the model may fit poorly. The T^2 values are high. A high T^2 value indicates an extreme spectrum, e.g., representing a sample with an extreme composition of chemical components.

- The **Q residual** is derived as the squared residual, i.e., the squared orthogonal distance from the spectrum to the principle component space or to the latent variables space, respectively.
  Q residuals reveal the variations that are not explained by the model. A high Q residual indicates the spectrum may not fit the model, e.g., if the measured sample contains a different substance.

Detecting spectral outliers

The detection of spectral outliers identifies spectra that deviate from the population.

1. Parameterization is taken into account as follows:
   a. From OMNIS Software version 4.2: The user decides whether the parameterization (data preprocessing and wavelength selection) is applied or not.
      Subsequent changes to the parameterization have no influence on dataset splitting.
   b. From OMNIS Software version 3.3 to OMNIS Software version 4.1: The user decides whether or not data preprocessing is taken into account. The wavelength selection and subsequent changes to the data preprocessing have no influence on dataset splitting.
   c. Up to OMNIS Software version 3.2: Data preprocessing is taken into account as specified at the time of outlier detection. The wavelength selection and subsequent changes to the data preprocessing have no influence on dataset splitting.
2. Spectral outlier detection is based on the PCA model of all mean-centered spectra *(see "Principal Component Analysis (PCA)", chapter 4.2, page 34)*. The number of principal components is chosen in such a way that the explained variance is at least 95%.
3. The Hotellings T^2 and Q residual values of the spectra are used for the detection of spectral outliers. The algorithm assesses whether the Hotellings T^2 or the Q residual of the spectrum being tested is the result of random variation or systematic variation.
   A description of the algorithm can be found in the appendix *(see "Spectral outliers – Algorithm", chapter 6.5, page 92)*.

4.3.3.1 Influence plot

The influence plot shows basic properties of the spectra and helps to analyze spectral outliers.

The basis for the influence plot is a PCA model *(see "Principal Component Analysis (PCA)", chapter 4.2, page 34)* or a PLS model:

- **Quantification**: The influence plot is optionally based on **PCA** or **PLS**. Similar to PCA, PLS regression reduces the spectroscopic data to fewer variables. PLS also takes the reference values into account.
  The principal components of PCA are called **latent variables** in PLS. *(see "PLS regression", chapter 4.4.1, page 57)*
- **Identification**: A **PCA**-based influence plot is available (OMNIS Software version 4.3 or higher).

Types of spectral outliers

The influence plot visualizes the Hotellings T^2 and Q residual values for each spectrum *(see "Hotellings T² and Q residuals", page 48)*. Hotellings T^2 and Q residuals reveal different types of spectral outliers:

- **Hotellings T^2 outlier**, also called leverage outlier: A large T^2 shows that the projection of the spectrum onto the principle component space (PCA), or latent variables space (PLS), is far from the model center.
- **Q residuals outlier**: A large Q residual means the spectrum is poorly described by the model.

Figure 28 shows multiple spectra in different views:

- Influence plot on the left: The Q residuals describe the variations not explained by the model, whereas Hotellings T^2 takes into account the variations within the model itself.
  The dashed lines show the **critical values**, or **limit values**, for the specified significance level *(see "Spectral outliers – Algorithm", chapter 6.5, page 92)*.
  The higher the significance level, the lower the limit values, so that more points can lie outside the limit values.
- On the right: Example of an original space with 3 variables x_1, x_2, x_3, which is transformed as an example into a 2-dimensional latent variables space.
  For points A to D, the orthogonal distance to the plane is shown (dashed lines), as well as the modeled point in the latent variables space (green points).

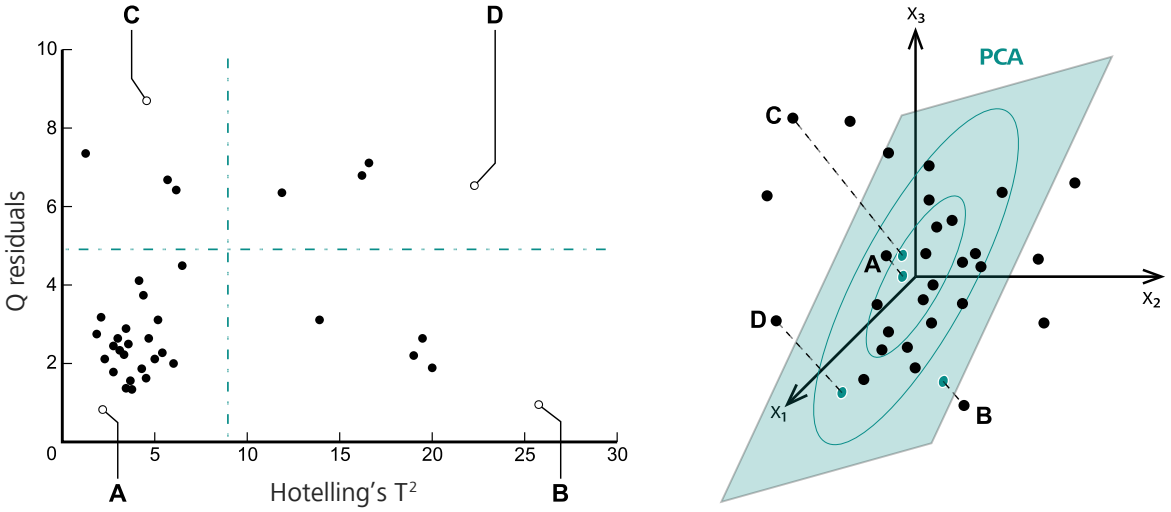


Figure 28 Influence plot (left), original space and latent variables space (right). Each spectrum is represented by a point in the left figure and a point in the right figure.

Both views highlight 4 points with different characteristics:

- Spectrum A has low scores and low residuals. It is close to the model center and well explained by the model.
- Spectrum B is a Hotellings T^2 outlier. It is off the mean, but well explained by the model.
- Spectrum C is a Q residuals outlier. It is off the mean and poorly explained by the model.
- Spectrum D is both a Hotellings T^2 outlier and a Q residuals outlier. It is off the mean and partially explained by the model.

The influence plot shows how different spectra influence the model. Since all latent variables go through the center, spectra near the center (e.g., spectrum A) have little chance to change the directions of the latent variables. They have no leverage. The further the distance from the center, the greater the leverage and the potential to influence the model. Some spectra actually succeed in pulling the model in their direction (spectrum B), while others do so only to a certain extent (spectrum D) or not at all (spectrum C).

Compared to a model based on all spectra, calculating a model without spectrum B is likely to change the model more than without spectrum D, and even more so than without spectrum C. Spectrum B will probably influence the quantification model strongly, for either better or worse. Deciding whether or not to delete a potential outlier in the lower right quadrant of the influence plot requires special care.

Ideally, the model should capture the variance of a large number of spectra. It is not desirable for the model to be characterized by only a few spectra. In the figure above, a few spectra have large distances from the

center and also from most of the other spectra. This is suspicious. The model is governed by few spectra. These are potential outliers and should be investigated. Furthermore, it should be ensured that the samples are evenly distributed over the range of variation.

PCA and PLS influence plot

The PCA influence plot depends on the spectra only. The PLS influence plot depends on the spectra and the reference values.

The following table shows how different settings affect the PCA influence plot and the PLS influence plot.

	PCA influence plot	PLS influence plot
Spectra	The underlying PCA model is based on all spectra in the calibration dataset, in the validation dataset and in the outlier dataset.	The underlying PLS model is based on all spectra in the calibration dataset. Based on this PLS model, the T^2 and Q residual values of the spectra in all 3 datasets are calculated and displayed in the plot.
Parameterization	Takes into account the selected data preprocessing and wavelength ranges. Note: The outlier detection is based on PCA and takes into account the parameterization according to the user setting and OMNIS Software version *(see "Detecting spectral outliers", page 49)*.	Takes into account the selected data preprocessing and wavelength ranges. Note: The outlier assessment in the prediction is based on PLS and takes into account data preprocessing and wavelength ranges.
Number of variables	Uses the number of principal components that achieves an explained variance of at least **95%**.	Uses the currently selected number of latent variables.
Significance level and critical values	Uses the currently selected significance level to calculate and visualize the critical values (dashed lines). Identification: If the last dataset splitting was carried out without determining outliers, then the influence plot will use a significance level of 5%. Note: Increasing the significance level results in decreased critical values, leading to more outliers during model development.	Uses the currently selected significance level to calculate and visualize the critical values (dashed lines). Note: Increasing the significance level results in decreased critical values, leading to more outliers during the prediction.

	PCA influence plot	PLS influence plot
Reference values (quantification)	Reference values have no impact on the PCA model. However, any spectrum can have an associated reference value outlier and thus be flagged as an outlier.	The reference values affect the PLS model and thus the PLS influence plot. Furthermore, any spectrum can have an associated reference value outlier and thus be flagged as an outlier.

Analyzing outliers

When analyzing potential outliers, the following factors are to be taken into account:

- A Hotellings T^2 outlier indicates a sample with an extreme composition of the chemical components compared to the other samples.
- A Q residuals outlier may indicate, for example, a contaminated sample or an error during spectra acquisition.

The potential outliers should be examined carefully. True outliers should be removed from the spectra list. Valid samples should be retained. In case the dataset is then split again, the outlier detection may find potential outliers that were not found in the first run. One possible reason is that the new PCA model requires fewer principal components to reach the 95% explained variance. If the newly found outliers prove to be valid samples, then they should be retained. In this case, the automatic splitting can be repeated without the outlier detection.

4.3.3.2 Score plot

The basis for the score plot is a PCA model or a PLS model:

- **Quantification**: The score plot (OMNIS Software version 3.0 or higher) is based on **PLS** *(see "PLS regression", chapter 4.4.1, page 57)*.
- **Identification**: The score plot (OMNIS Software version 4.3 or higher) is based on **PCA** *(see "Principal Component Analysis (PCA)", chapter 4.2, page 34)*.

Each spectrum has a score value for each principal component or latent variable. In the score plot, each spectrum is represented by a dot. The x-axis shows, for example, the score for the first latent variable, the y-axis shows, for example, the score for the second latent variable. Likewise, any pair of latent variables can be displayed.

Since for each wavelength variable the absorbance values have been mean-centered, the scores of each latent variable will also be mean-centered as well. A point near the center of the score plot (0/0) represents a mean spectrum with regard to the two latent variables displayed. Points close to one another represent similar spectra, while points further apart

represent dissimilar spectra with respect to the two latent variables displayed.

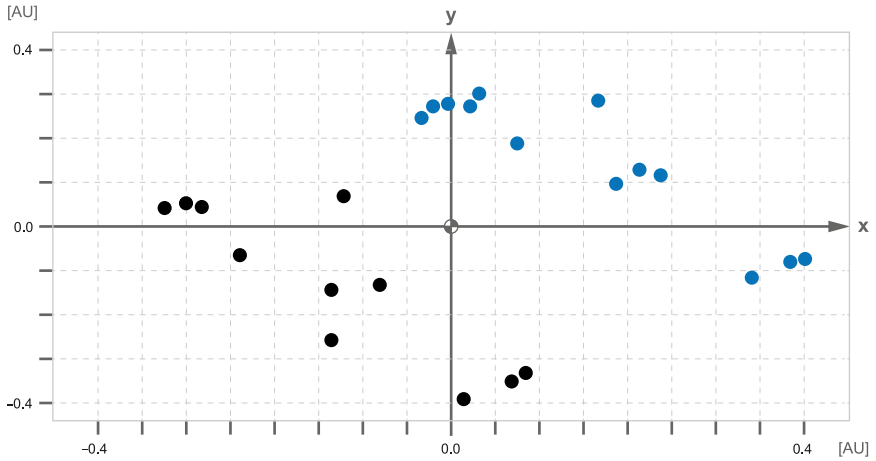


Figure 29 Score plot for the latent variable 1 (x-axis) and the latent variable 2 (y-axis). AU = arbitrary unit.

Figure 29 shows an example of 2 datasets measured at different conditions. The scores are normalized, each latent variable is assigned the same weighting.

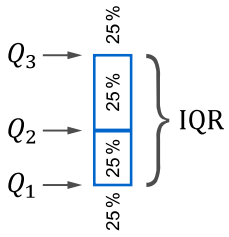
The scores of all principal components or latent variables of a spectrum can be combined into a single value (Hotellings T^2), which is displayed on the x-axis of the influence plot.

4.3.4 Reference value outliers (quantification)

In quantification models, reference value outliers are also determined in addition to the spectral outliers. Reference value outliers show anomalies in the reference value.

Typically, reference value outliers are incorrectly transmitted numbers, e.g., 143 instead of 14.3 or 15.9 instead of 51.9. The outlier detection identifies such transmission or transcription errors based on an empirical approach. Only obvious errors are flagged for further investigation.

Boxplots



Reference value outliers are found using a method based on boxplots. A **boxplot** sorts the reference values in ascending order. Quartiles subdivide the dataset into 4 parts. Each part contains 25% of the reference values.

The first quartile Q_1 separates the 25% lowest values from the rest. Q_2 is the median and separates the 50% lowest values from the rest. The third quartile Q_3 separates the 75% lowest values from the rest. A vertical rectangle represents the middle 50% of the data, the interquartile range (IQR).

Data that falls outside the IQR box by a specific amount is considered to represent potential outliers and can be shown as small circles. The lower and upper cutoff values for outliers are often defined as 1.5 times the IQR:

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

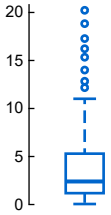
In doing so, Q_1 corresponds to the first quartile, Q_3 to the third quartile and IQR to the interquartile range ($Q_3 - Q_1$).

To complete the boxplot, whiskers above and below the box reach out to the most remote points that are not flagged as potential outliers.

Adjustment for skewed distributions

The standard boxplot assumes an approximately symmetrical distribution of the data. As a rule, many regular reference values are flagged as potential outliers in connection with skewed distributions. For this reason, the OMNIS Software uses an adjusted version of the boxplot that takes the skewness of the distribution into account.

In the following example, the distribution is skewed towards the higher reference values. This yields 8 outliers with high values. After adjusting the outlier cutoff values, only 2 of them remain. On the other side, a new outlier appears with a low value.

Skewed distribution	Adjusted outlier cutoff values
	

A description of the adjustment can be found in the appendix *(see "Reference value outliers – Algorithm", chapter 6.6, page 95)*.

4.3.5 Dataset splitting

The dataset consists of spectra and reference values (quantification) or spectra and associated product names (identification, verification). The dataset is used for the development and validation of the model. The dataset should therefore be split into a calibration dataset, a validation dataset, and an outlier dataset. The splitting can be done manually or automatically.

Assuming there are enough spectra available, 20% to 30% of the total data for the validation dataset can be used, for example.

Automatic splitting algorithm

The automatic dataset splitting is done in a way that ensures that the calibration dataset and the validation dataset are representative of the

population and independent of one another. The objective is to divide the data into two sets which will cover approximately the same region in the PCA space and exhibit similar statistical properties. The algorithm used is a slightly modified duplex algorithm from R. D. Snee, *Validation of Regression Models: Methods and Examples*, Technometrics Volume 19, No. 4 (Nov. 1977), pages 415–428.

Replicates

Replicates should not be in the dataset. The algorithm does not remove replicates or pseudoreplicates, nor does it force replicates of a given sample to be added to the same dataset.

1. Parameterization is taken into account as follows:
   a. From OMNIS Software version 4.2: The user decides whether the parameterization (data preprocessing and wavelength selection) is applied or not.
      Subsequent changes to the parameterization have no influence on dataset splitting.
   b. From OMNIS Software version 3.3 to OMNIS Software version 4.1: The user decides whether or not data preprocessing is taken into account. The wavelength selection and subsequent changes to the data preprocessing have no influence on dataset splitting.
   c. Up to OMNIS Software version 3.2: Data preprocessing is taken into account as specified at the time of outlier detection. The wavelength selection and subsequent changes to the data preprocessing have no influence on dataset splitting.
2. The splitting is based on the PCA model of all mean-centered spectra. The number of principal components is chosen in such a way that the explained variance is at least 95%.
3. The distances between all possible pairs of spectra are calculated from the PCA scores.
4. The 2 spectra that are furthest apart from one another are assigned to the calibration dataset.
5. Of the remaining spectra, the 2 furthest apart from one another are assigned to the validation dataset.
6. Of the remaining spectra, the spectrum that is furthest apart from the spectra already in the calibration dataset is assigned to the calibration dataset.
7. Of the remaining spectra, the spectrum that is furthest apart from the spectra already in the validation dataset is assigned to the validation dataset.
8. The alternation continues until one of the datasets has reached its intended size. The remaining spectra are assigned to the other dataset.

4.4 Quantification

4.4.1 PLS regression

The OMNIS Software uses PLS regression to calculate quantification models.

Similar to PCA, **partial least squares regression** (**PLS regression**) reduces the spectroscopic data to a few variables. However:

- The principal components of PCA are called **latent variables** in PLS. The directions of the latent variables and the principal components are usually similar, but not identical.
- In addition to the spectra, PLS regression also takes the **reference values** into account. Therefore, fewer latent variables are needed to achieve sufficient correlation with the reference values, resulting in less noise.

PLS extracts the underlying **latent variables** from the extensive, highly redundant spectroscopic data. Latent variables are also referred to as hidden variables, as they are not measured directly. The latent variables explain as much of the data variation as possible, while at the same time modeling the reference values well.

Preliminary steps

The preparatory steps for PLS regression are similar to those for PCA:

1. **Parameterization**: The OMNIS Software applies the specified data preprocessing and specified wavelength selection to the spectra.
2. **Mean centering**: For each wavelength, the mean absorbance value is calculated and subtracted from the respective value in each spectrum.
   The reference values are also mean-centered.

Transformation to latent variables

After the preliminary steps, PLS regression transforms the spectral data into a latent variables space, taking the reference values into account. *Figure 30* shows the first 2 latent variables, LV1 and LV2. The latent variables LV3, LV4, etc. can be visualized in the same way.

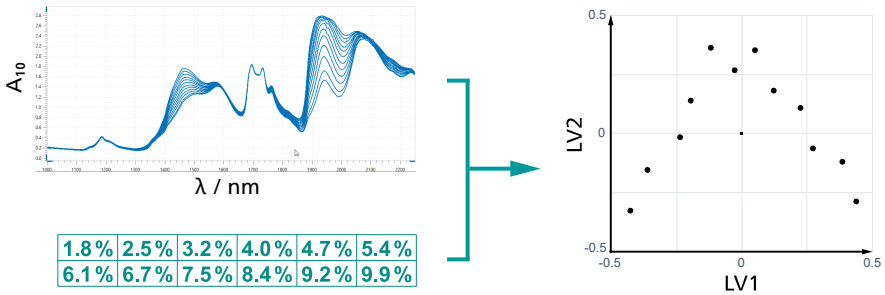
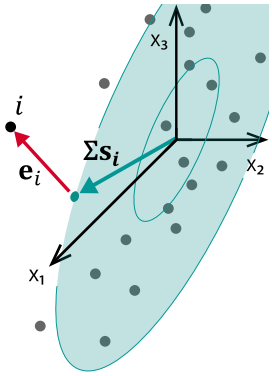


Figure 30 Transformation of spectra and reference values into a latent variables space. On the right, scores are expressed in arbitrary units.

The first latent variable LV1 best explains the variance in the spectral data and also exhibits the highest possible correlation with the reference values. All subsequent latent variables LV2, LV3, etc. best explain the remaining variance and also exhibit the greatest possible correlation with the reference values. Thus, the first few latent variables explain most of the variance and maximize the correlation, while others contain mainly noise and can be discarded.

Scores and residuals



PLS has similar sizes as PCA:

- **Scores**: Scores are measured in the latent variables space. The orthogonal projection of the sample *i* onto each latent variable direction gives the score vector $\Sigma \mathbf{s}_i$, representing the Euclidean distance from the center.
  The **Mahalanobis distance** $\mathbf{s}_i$ is the normalized score vector, which assigns the same weight to each direction.
- **Residuals**: The residual vector $\mathbf{e}_i$ is the offset between the sample *i* and the latent variables space, measured in the original wavelength space.

PLS algorithm

The PLS algorithm maximizes the covariance between the spectra and the reference values *(see "PLS algorithm", chapter 6.3, page 90)*.

4.4.1.1 Number of latent variables

The choice of the number of latent variables in the quantification model is fundamental for the predictive ability of the model. If the number of latent variables is too low, then relevant spectral variations are not captured. This is called **underfitting** and leads to less accurate predictions.

If the number of latent variables is too high, the calibration samples will be modeled too well. The model captures irrelevant spectral variations (noise). This is called **overfitting** and results in fluctuating, unstable, and less accurate predictions of unknown samples.

Finding the optimal number of latent variables means finding a balance between the following objectives:

- The SEP is sufficiently close to its minimum value.
  Without a validation dataset, SECV is used.
- The quantification model should use as few latent variables as possible.
  In case of doubt, the lower number should be used.
- The correlation plot for the validation dataset is close to its optimum.
  Ideally, the slope is close to 1, the intercept is close to 0 and the data points exhibit minimal scatter.
  In the absence of a validation dataset, the cross-validation values are used *(see "Cross-validation", page 60)*.

It is to be noted that the figures of merit are only estimates based on the available calibration samples and validation samples. In general, a quantification model based on fewer latent variables will be more robust.

4.4.1.2 Loading plot

The loading plot in the OMNIS Software is based on the PLS model *(see "PLS algorithm", chapter 6.3, page 90)*.

PLS loadings show how the original wavelength variables (including parameterization) contribute to the formation of each latent variable. It is irrelevant whether the loadings are positive or negative.

The loadings are computed in such a way that the first latent variable captures the variance which is most predictive for the reference parameter. All subsequent latent variables capture the remaining variance that is most predictive for the reference parameter. Thus, PLS loadings differing greatly from 0 indicate that the respective wavelengths are well suited to model the reference parameter.

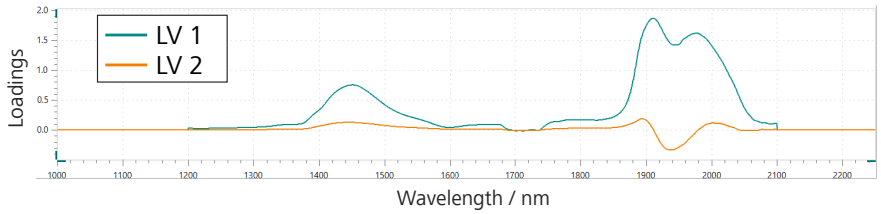


Figure 31 Loading plot for the latent variables LV1 and LV2.

In *Figure 31*, the wavelength selection was limited to the range from 1,200 nm to 2,100 nm. Therefore, no loadings occur outside this range.

4.4.2 Validation of quantification models

During validation, a check is made to determine whether the model fulfills the requirements regarding its performance and robustness. To accomplish this, the expected prediction error must be estimated as realistically as possible.

As a rule, a quantification model is validated as follows:

1. Starting with a limited number of samples and without a validation dataset, a model is developed and tested using cross-validation (see below).
2. With a greater number of samples, the dataset is split into a calibration dataset, which is used to develop a model, and a validation dataset, which is used to validate the model.
3. Finally, samples for the validation dataset are collected and measured on a different day, possibly by a different person and with a different instrument.

Cross-validation

Cross-validation relies exclusively on the calibration dataset. It results in an estimate for each calibration sample, using a temporary model created without the respective calibration sample.

Cross-validation uses a multi-round procedure in one of the following ways:

- **Leave-One-Out**
  In each round, the leave-one-out cross-validation (LOO-CV) puts aside 1 sample while a model is created using all the remaining samples. This model then predicts the parameter of interest for the sample put aside. This prediction serves as an estimate for the sample.
  The cycle continues until each sample has been put aside once.
- **K-fold**
  The k-fold cross-validation subdivides the calibration dataset into *k* blocks of as equal a size as possible.
  In each round, 1 block is put aside while a model is created using the remaining blocks. This model then predicts the parameter of interest for the samples put aside.
  The cycle continues until each block has been put aside once.
  The blocks can be selected in several ways:
  - **Fixed Blocks (DUPLEX)** (from OMNIS Software version 3.2): The blocks are selected reproducibly, based on a Duplex algorithm. Each prediction serves directly as an estimate for the respective sample.
  - **Random**: The blocks are selected at random. The above procedure is repeated several times. The calibration dataset is subdivided into *k* blocks differently each time. Each sample will ultimately have multiple estimates. Their mean value serves as an estimate for the respective sample.

In general, Leave-One-Out is preferred. For large calibration datasets, the k-fold cross-validation can be used to save computation time. A typical value for *k* is 5.

4.4.2.1 Correlation plot

The correlation plot visualizes the correlation between reference values and calculated values. It enables an assessment of the quantification model at a glance.

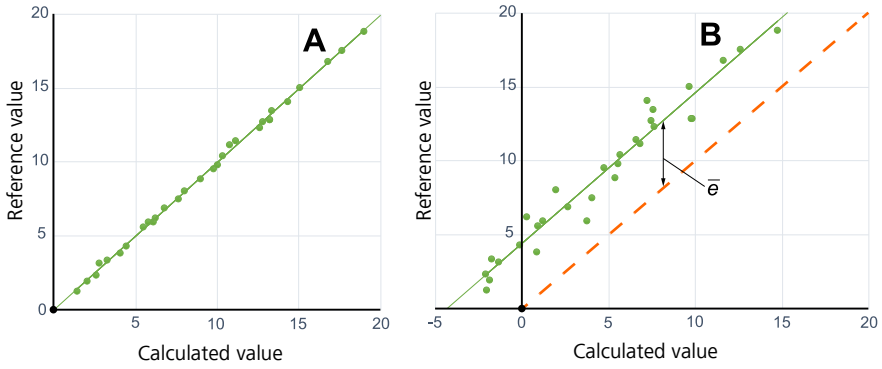
The calculated values are determined as follows:

- Validation dataset and outlier dataset: Prediction using the quantification model
- Calibration dataset: estimates of the cross-validation

In the correlation plot, each sample is represented by a point. The calculated value of the sample can be found on the x-axis, the reference value on the y-axis. A regression line reveals the systematic relationship between the variables. Ideally, the regression line has a slope of 1 and a y-intercept of 0, and all points are located on the lines. This means that for each sample, the calculated value corresponds to the reference value.

Systematic errors and random errors can be distinguished, based on the deviations from the ideal case. The position of the regression line reveals the systematic errors. The distances of the points from the regression line indicate the random errors.

The following correlation plot **A** shows a good correlation. The other plots show different types of errors, which are explained below.



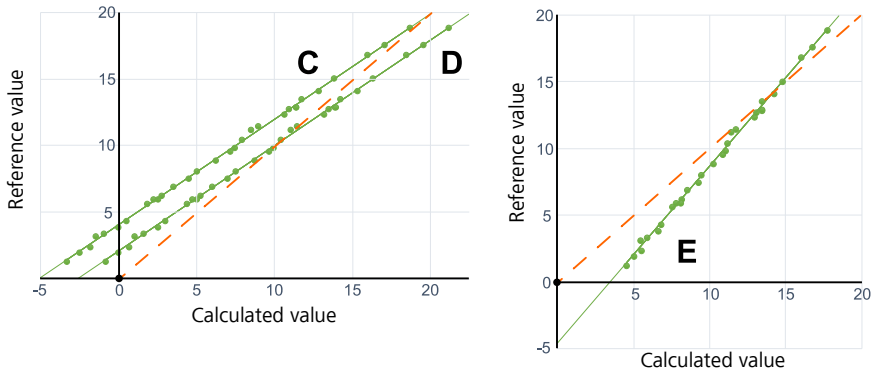


Figure 32 Correlation plots. Each point represents a sample and shows its reference value and its calculated value. The dashed line indicates the ideal 45° line.

Systematic errors

Systematic errors are errors that always occur and which are repeatable for a given application. Systematic errors can be corrected. They are quantified through the bias $\bar{e}$ and the slope b of the regression line:

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

If the slope equals 1 and the bias equals 0, then there are no systematic errors.

The **bias** is the average error between the reference values and the calculated values:

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

Here, n corresponds to the number of samples, e_i to the error of the ith sample, y_i to the reference value of the ith sample, $\hat{y}_i$ to the calculated value of the ith sample, $\bar{y}$ to the mean value of the reference values, and $\bar{\hat{y}}$ to the mean value of the calculated values.

The lines **B** and **C** have a positive bias, line **E** has a negative bias.

Errors with opposite algebraic signs cancel each other out. Therefore, the bias of line **D** is approximately 0.

The **slope** of the regression line is:

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

$s_{\hat{y}y}$ corresponds thereby to the covariance between the reference values and the calculated values and $s_{\hat{y}}^2$ stands for the variance of the calculated values.

The slope can be seen as a property-dependent error:

- $b > 1$: (Line **E**) The higher the calculated value, the higher (more positive) the error that contributes to the bias.
- $b < 1$: (Lines **C** and **D**): The higher the calculated value, the smaller (more negative) the error that contributes to the bias.
- $b = 1$ (Lines **A** and **B**): The error that contributes to the bias is constant.

The **intercept** of the regression lines with the y-axis is $\bar{y} - b\bar{\hat{y}}$.

The slope and the y-intercept are calculated with the reference values as the dependent variable (y-axis) and the calculated values as the independent variable (x-axis).

Random errors

If all points lie directly on the regression line, then there are no random errors. The more scattered the points, the higher the random errors.

In correlation plot **B**, the random errors are greater than in the other correlation plots.

Visualization of error types

The lines in the above figures show the following error types:

	Systematic errors			Random errors
Line	**Bias**	**Slope**	**y-intercept**	
A	~ 0	~ 1	~ 0	small
B	> 0	~ 1	> 0	large
C	> 0	< 1	> 0	small
D	~ 0	< 1	> 0	small
E	< 0	> 1	< 0	small

4.4.2.2 Figures of merit

Figures of merit express the agreement between reference values and calculated values in numbers. The calculated values are determined by the quantification model.

R^2 – Coefficient of determination

The **coefficient of determination R^2** measures the goodness of fit of the quantification model. For a given dataset, this is the proportion of the reference value variation that is explained by the quantification model:

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

Here, SS_{reg} corresponds to the regression sum of squares (calculated values variance, i. e. explained variance), SS_{tot} to the total sum of squares (reference value variance), SS_{res} to the residual sum of squares (residual variance, i. e. unexplained variance), y_i to the reference value of the *i*th sample, $\hat{y}_i$ to the calculated value of the *i*th sample, and $\bar{y}$ to the mean value of the reference values.

The R² value is a fraction of 1. An R² of 1 means that the calculated values fit the reference values perfectly. An R² of 0.9 indicates that 90% of the reference value variance is explained by the calculated values, and 10% is not explained.

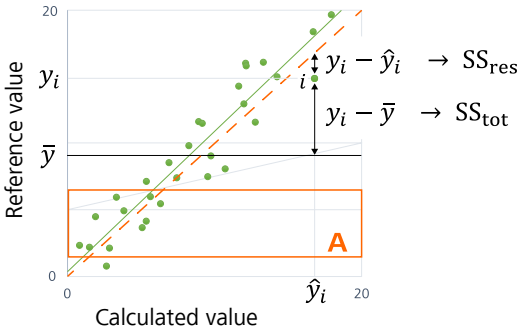


Figure 33 The components to calculate R². The dashed line is the 45 degree line.

The above figure shows a correlation plot with a sample *i* and its residual from PLS regression. The residual variation is contained in SS_{res} and the reference value variation in SS_{tot}.

A high R² value does not assure a useful quantification model or accurate predictions. The magnitude of R² depends directly on the reference value variation.

A regression with a smaller range of reference values (range **A**) has approximately the same residual variation, but the reference value variation is smaller. The resulting R² value is lower.

Thus, the reason for a high R² could be that the reference value range is unrealistically large. On the other hand, data from a manufacturing process may for instance exhibit a limited value range, thus resulting in a lower R² value. The standard errors should be applied to evaluate predictive capability.

The absolute R² value should be viewed with caution. The degree of change with each additional latent variable is more meaningful *(see "Number of latent variables", chapter 4.4.1.1, page 58)*.

Different R² values are generated, depending on the values used for calculation:

- R²C (not shown in the OMNIS Software): Calculated with the calculated values of the spectra in the calibration dataset.

- **R²CV**: Calculated with the cross-validation estimates of the spectra in the calibration dataset *(see "Cross-validation", page 60)*.
- **R²P**: Calculated with the calculated values of the spectra in the validation dataset.

The OMNIS Software uses the square of the Pearson sample correlation coefficient for the calculation $r_{y,\hat{y}}$:

Coefficient of determination of cross-validation:

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

Here y_i corresponds to the reference value of the *i*th sample, $\bar{y}$ to the mean of the reference values, $\hat{y}_{\mathrm{cv}_i}$ to the cross-validation estimate of the *i*th sample, $\bar{\hat{y}}_{\mathrm{cv}}$ to the mean value of the cross-validation estimates, and *n* to the number of samples in the calibration dataset. Note that each sample in the calibration dataset has exactly one cross-validation estimate.

Coefficient of determination of prediction:

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

Here v_i corresponds to the reference value of the *i*th validation sample, $\bar{v}$ to the mean value of the reference values, $\hat{v}_i$ to the calculated value of the *i*th validation sample, $\bar{\hat{v}}$ to the mean value of the calculated values, and v to the number of validation samples.

SEC – Standard error of calibration

The **standard error of calibration (SEC)** is based on the calibration dataset. The SEC can be seen as an estimate for the theoretically best prediction accuracy. SEC is the standard deviation of the residuals of the partial least squares regression (PLS):

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

Here **e** corresponds to the residual vector containing all reference value variations in the calibration dataset that are not described by the model, *n* to the number of calibration samples, and *k* to the number of latent variables. The denominator *n*–*k*–1 is the number of degrees of freedom of the residual vector **e**.

In other words: The SEC is the standard deviation of the differences between reference values and calculated values for the samples in the calibration dataset:

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

Here y_i corresponds to the reference value of the *i*th calibration sample, $\hat{y}_i$ to the calculated value of the *i*th calibration sample, *n* to the number of calibration samples, and *k* to the number of latent variables.

The SEC is sometimes called RMSEC. The SEC contains the random errors and the systematic errors (slope and bias).

SECV – standard error of cross-validation

The **standard error of cross-validation (SECV)** is based on the calibration dataset. The SECV estimates the prediction accuracy based on the calibration dataset and a cross-validation method *(see "Cross-validation", page 60)*. The SECV can be used for an initial model assessment or to determine the optimal number of latent variables.

The SECV is the standard deviation of the differences between reference values and cross-validation estimates of the samples in the calibration dataset.

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

Here y_i corresponds to the reference value of the *i*th sample, $\hat{y}_{\mathrm{cv}_i}$ to the cross-validation estimate of the *i*th sample, and *n* to the number of samples in the calibration dataset.

The SECV contains all errors: The random errors and the systematic errors (slope and bias).

Other approaches use separate values for a bias-corrected SECV and for the uncorrected SECV, which is then called RMSECV. If the bias is low, then these values will be similar.

SEP – Standard error of prediction

The **standard error of prediction (SEP)** is based on the validation dataset. Therefore, the SEP provides the most realistic estimate of the prediction accuracy.

The SEP is the standard deviation of the differences between reference values and calculated values for the samples in the validation dataset:

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

Here v_i corresponds to the reference value of the *i*th validation sample, $\hat{v}$ to the calculated value of the *i*th validation sample, and v to the number of validation samples.

The SEP contains all errors: the random errors and the systematic errors (slope and bias).

Other approaches use separate values for a bias-corrected SEP and for the uncorrected SEP, which is then called RMSEP. If the bias is low, then these values will be similar.

Interpretation of the figures of merit

The figures of merit are estimates. Example: The SEP value is an *estimate* of a standard deviation – based on the samples at hand – and also has its own standard deviation. The higher the number of validation samples, the more trustworthy is the estimate.

The SEP should be comparable to the SECV and the SEC. Excessively large differences may indicate overfitting *(see "Number of latent variables", chapter 4.4.1.1, page 58)*. As a rule of thumb, the differences should not exceed 20%.

The standard errors should also be taken into account in relation to the standard error of laboratory (SEL) for the reference method. The precision of the NIR method is acceptable if the SEP is 1.4 to 2.0 times higher than SEL. A larger SEP can nonetheless be accepted if it meets the requirements.

An SEC or SECV lower than the SEL for the reference method indicates overfitting.

The above relationships with the SEL assume that the SEL is based on the correct number of repeated reference measurements as performed for each sample *(see "Reference method (quantification)", page 33)*.

4.4.3 OMNIS Model Developer (OMD)

Developing a quantification model is challenging, time-consuming, and requires a certain level of expertise. The OMNIS Model Developer (OMD, starting with OMNIS Software version 4.0) automates the development and delivers well-optimized quantification models.

Mode of operation

Proper sampling is a prerequisite *(see "Physical samples", chapter 4.1, page 31)*. A dataset consisting of spectra and corresponding reference values serves as input for the OMD.

The OMD determines spectral outliers with a significance level of 5% and the algorithm listed in the appendix without taking the data preprocessing into account *(see "Spectral outlier detection during model development", page 93)*.

The splitting of the dataset depends on the number of spectra remaining after outlier detection:

Number of spectra	Cross-validation method	Validation dataset
> 99	K-fold (5 blocks, DUPLEX)	25% of the spectra
30–99	K-fold (5 blocks, DUPLEX)	—
< 30	Leave-One-Out	—

After splitting, additional outliers are detected in the calibration dataset according to ASTM D8321-22.

The models are assessed and classified on the basis of various key figures. The OMD optimizes data preprocessing, wavelength selection, and the number of latent variables, for which a balance is sought between the risk of overfitting and the risk of underfitting.

Result

The result of the OMD is a list of models sorted according to their predictive ability. **Predictive ability** is calculated on the basis of model complexity, figures of merit, and dataset size.

The list is color-coded to aid in the selection of the preferred model:

- Green: Good predictive ability.
  If the number of samples is sufficiently large, then the model will perform well on all unknown samples of the same type. The figures of merit enable a reliable estimate of future errors.
- Yellow: Medium predictive ability.
  If the number of samples is sufficiently large, then the model is expected to perform well. The figures of merit may be too optimistic for future samples. Separate validation is recommended.
- Red: Insufficient predictive ability.
  The model has serious drawbacks. It should not be used.

Models of the same color are sorted according to an information criterion that favors a balanced trade-off between low estimated prediction error and low number of latent variables.

Optimizing parameterization

Instead of creating the entire model automatically, the parameterization can also be optimized by itself. The current settings (e.g., dataset splitting, cross-validation method) remain unchanged, but have no influence on the optimization.

4.4.4 Slope/y-intercept correction

The slope/y-intercept correction enables the correction of systematic errors (bias, slope) in the application of a quantification model.

If systematic errors occur with the calibration dataset, possible causes are:

- Systematic errors in the quantification model. Unrecognized outliers or an insufficient number of samples, for example.
- Systematic errors in the spectroscopic measurement process.
- Systematic errors in the reference measurement process.

If systematic errors occur with a validation dataset, then additional causes may include:

- Changes in the spectroscopic measurement procedure, e.g., of the instrument.
- Changes in the reference measurement procedure, e.g., new laboratory, new equipment.
- Changes in the samples, e.g., during handling, storage or transport.

The bias correction and even more the slope/y-intercept correction should be used with caution.

If the systematic errors are not significant, then no correction should be applied. If the errors are significant, then they should be thoroughly investigated. If possible, the cause of the errors should be eliminated. If there is a legitimate reason that the errors cannot be eliminated, then a bias correction or a slope/y-intercept correction may be applied.

A reliable estimate of the bias needs at least 20 samples. A reliable estimate of the slope needs at least 30 samples.

Bias correction

The following correlation plot shows a bias correction. The slope of the original regression line **F** remains unchanged. After the correction (regression line **G**), the positive errors and the negative errors cancel each other out.

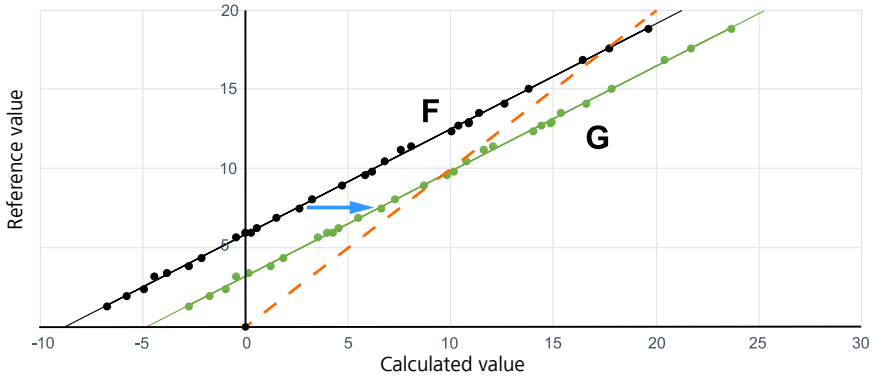


Figure 34 Bias correction

Slope/y-intercept correction

The following correlation plot shows a slope/y-intercept correction. The original regression line **H** is corrected by the slope and the y-intercept. As a result, both the slope and the bias are corrected (regression line **K**).

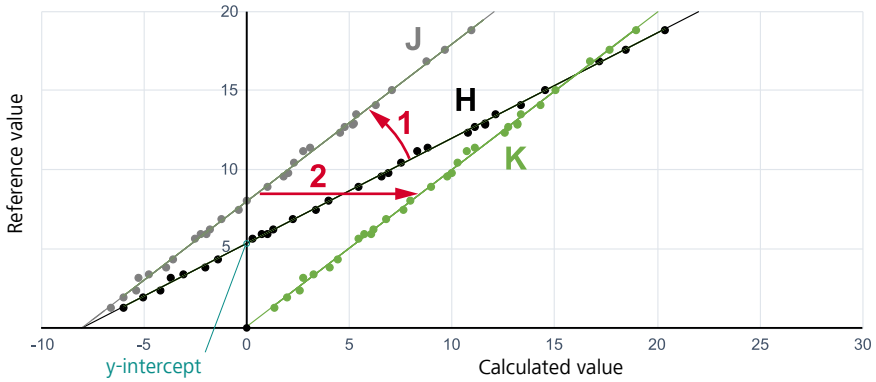


Figure 35 Slope/y-intercept correction

SEP

The samples in the correction dataset form the basis for the slope/y-intercept correction. Based on these samples, the OMNIS Software calculates the following **standard errors of prediction (SEP)**. The denominators take into account the corresponding degrees of freedom:

Type of correction	SEP
Uncorrected	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n}}$
Bias correction	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-1}}$
Slope/y-intercept correction	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-2}}$

Here, v_i corresponds to the reference value of the *i*th sample in the correction dataset, $\hat{v}_i$ corresponds to the predicted value of the *i*th sample in the correction dataset, and *n* corresponds to the number of samples in the correction dataset.

The SEP contains all errors: The random errors and the systematic errors (slope and bias).

4.5 Identification and verification

4.5.1 Support Vector Machine (SVM)

Identification models (from OMNIS Software version 4.0) use Support Vector Machines (SVM) for classification between different products. A support vector machine is a supervised machine learning algorithm. Based on the calibration samples, it learns how to assign new samples to a product.

In the interest of simplicity, classification between 2 products is described below. The concept can be extended, the final model classifies between any number of products.

Linear classification

Figure 36 (left) shows input data with 2 variables.

In the interest of simplicity, the parameterized spectra are shown in a 2-dimensional variables space. Each point represents a spectrum, the color indicates the product membership.

The products are linearly separable. The SVM algorithm creates a hyperplane between the products (figure on the right).

Hyperplanes are a generalization of planes from 3-dimensional space to spaces of any dimension. The dimension of a hyperplane is one less than the dimension of the space surrounding it. A hyperplane in a 2-dimensional space is a line.

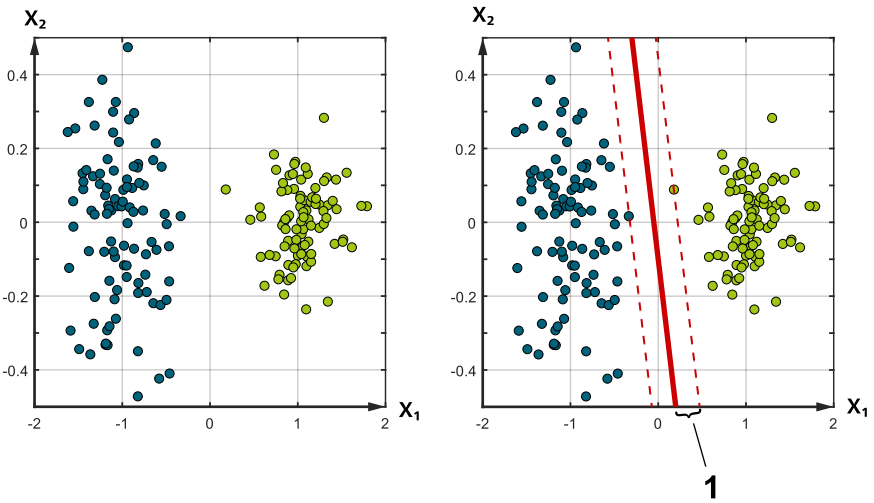


Figure 36 Input data (on the left) and the hyperplane created by the SVM (solid red line on the right). The values are expressed in arbitrary units.

The SVM algorithm maximizes the margin (**1**) from the hyperplane to the nearest point on each side. New spectra can be mapped into the same space and assigned to a product, depending on which side of the hyperplane they fall on.

To define the hyperplane, the SVM algorithm takes into account only the points that are closest to the points of the opposite product. These points, or vectors, support the formation of the hyperplane and are called support vectors.

If the points are not linearly separable – for instance due to an outlier – a linearly classifying hyperplane can nonetheless still be determined. In this case, an optimization algorithm finds a trade-off between increasing the margin from the hyperplane to the support vectors on each side and ensuring that all points are on the correct side of the hyperplane. A regularization parameter controls the trade-off and thus the final position of the hyperplane.

Nonlinear classification

The products are not linearly separable in *Figure 37 (left)*. A nonlinear classifier is needed to separate the products.

A linear or nonlinear kernel function transforms the data into a higher-dimensional feature space. The transformation is done in such a way that the data in the feature space can be linearly separated by a hyperplane.

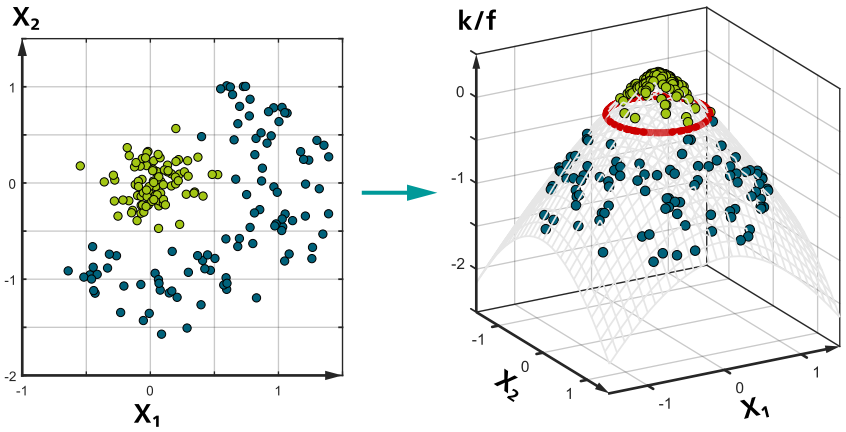


Figure 37 Products not linearly separable (on the left). An additional dimension k/f (kernel feature) facilitates the separation (on the right). The values are expressed in arbitrary units.

In the figure, the data is transformed from 2-dimensional space to 3-dimensional space. The points of one product are raised above the original plane, while the points of the other product are shifted below. The linear hyperplane between the products is a 2-dimensional plane very similar to the original plane. Viewed in 2 dimensions, the decision boundary is a nonlinear line, in this case the red circular line.

Here, too, the hyperplane acts as a classifier. A new sample can be assigned to a product, depending on which side of the hyperplane its spectrum falls on.

The OMNIS Software uses a radial basis function kernel to transform the data into the feature space. The kernel uses a scaling parameter that controls the degree of nonlinearity.

Parameter selection

Suitable values must be chosen for the regularization parameter, which controls the position of the hyperplane, and for the scaling parameter, which controls the degree of nonlinearity.

Both parameters affect the generalization ability of the support vector machine, i.e., how well it can generalize from the calibration spectra to new, unseen spectra. For instance, if the scaling parameter enables a high degree of nonlinearity, then the hyperplane may be excessively fitted to the calibration spectra (overfitting). If the scaling parameter enables only a low degree of nonlinearity, then the hyperplane may not be sufficiently fitted to the calibration spectra (underfitting).

To achieve good generalization, **grid search** is used. With as few attempts as possible, the algorithm finds the best parameter combination:

1. A set of predefined parameter combinations is used.
2. For some selected parameter combinations, the SVM learns a classification rule that distinguishes between the calibration spectra of the different products with maximum accuracy.
3. A cross-validation estimates how successful the assignment of the product is for each classification rule.
4. Additional parameter combinations are selected, based on the cross-validation results.
   The SVM learns once again the corresponding classification rules, and a cross-validation method estimates the classification accuracies.
   After a few iterations, the final parameter combination is determined.
5. With the final parameter combination and with all calibration spectra, the SVM evaluates the final classification rule.

Probabilities

Based on the classification rule, the identification model calculates the probability of whether a particular sample belongs to a particular product or not. The probability depends on the distances to the populations of the products and on the model parameters.

The probabilities calculated in this way make it possible to control the assignment of the samples to the products *(see "Assignment of a sample (from OMNIS Software version 4.4)", page 74)*.

In order to minimize false positive identifications, a qualification model is also trained for each product, starting with OMNIS Software version 4.4 *(see "Qualification", chapter 4.6, page 77)*.

4.5.2 Prediction of the product membership of a sample

The identification model provides a probability per product for the identification of a particular sample *(see "Probabilities", page 73)*.

Each probability is an individual value between 0% and 100%. The values do not add up to 100% across the products.
The values are to be considered relative to each other, which makes it possible to compare the different products.

Assignment of a sample (from OMNIS Software version 4.4)

The evaluation is carried out with the help of an adjustable **probability threshold** and with qualifications for individual products:

1. For each product whose probability is above the probability threshold, a qualification of the sample is performed using the corresponding qualification model. If the qualification fails, then the probability for the corresponding product is set to zero.
2. Evaluation with the modified probabilities from step 1:
    a. If no probability is above the probability threshold, then the identification fails (identification status **Unidentified**).
    b. If a single probability is above the probability threshold, then the sample is successfully identified and assigned to the corresponding product (identification status **Identified**).
    c. If multiple probabilities are above the probability threshold, then the prediction is ambiguous and the identification has failed (identification status **Ambiguous**).

Table 1 shows an evaluation example with different probability thresholds. In this example, the qualifications for products A and C are both successful.

Table 1 Evaluation example with different probability thresholds (from OMNIS Software version 4.4)

Probability	Probability threshold	Qualification	Identification result
Product A: 87% Product B: 71% Product C: 68% Product D: 30%	90%	–	Unidentified
	80%	Product A: Successful → 87%	Product A
	70%	Product A: Successful → 87% Product B: Failed → 0%	Product A
	60%	Product A: Successful → 87% Product B: Failed → 0% Product C: Successful → 68%	Ambiguous

Assignment of a sample (OMNIS Software version 4.0 to 4.3)

The probabilities of the sample are evaluated with the help of an adjustable **probability threshold**:

- If no probability is above the probability threshold, then the identification fails (identification status **Unidentified**).
- If a single probability is above the probability threshold, then the sample is successfully identified and assigned to the corresponding product (identification status **Identified**).
- If multiple probabilities are above the probability threshold, then the prediction is ambiguous and the identification has failed (identification status **Ambiguous**).

Table 2 shows an evaluation example with different probability thresholds.

Table 2 Evaluation example with different probability thresholds (OMNIS Software version 4.0 to 4.3)

Probability	Probability threshold	Identification result
Product A: 87% Product B: 72% Product C: 68%	90%	Unidentified
	80%	Product A
	70%	Ambiguous

4.5.3 Validation of identification models

As a rule, an identification model is validated as follows:

1. Starting with a limited number of samples and without a validation dataset, a model is developed and tested with the samples in the calibration dataset.

2. When there is a sufficient number of samples, the dataset can be split into a calibration dataset and a validation dataset. The samples in the validation dataset are not used for the development of the model.
3. Finally, samples for an external validation dataset are collected and measured on a different day, possibly by a different person and with a different instrument.

The calculated product membership is compared with the respective true product membership for each sample. If they match, the prediction is correct (= successful), otherwise incorrect (= failed).

Validation

The OMNIS Software shows the following magnitudes for identification models, which indicate how well the models perform. Ideally, all figures are 100%.

Successful % (Total) measures the *Accuracy*, the overall correctness of a model. The percentage answers the question: How many of the samples can the model identify correctly?

$$\textit{Successful \% (Total)} = \frac{\textit{correct classifications}}{\textit{all classifications}}$$

Successful % (Total) gives equal weight to each sample. Therefore, products with more samples have a greater impact on the percentage than do products with fewer samples.

Similar *Successful %* numbers are available for each product.

Improving the identification

The following actions can help improve the model:

- **Adjust probability threshold**
  - If many predictions are ambiguous or many 0.0 % probabilities occur, then the probability threshold can be raised.
  - If many samples are not identified because the probability threshold was not reached, then the probability threshold can be decreased.
- **Adjust the parameterization**
  Determine more suitable data preprocessing and wavelength selection.

- **Using model hierarchies**
  A model hierarchy enables a hierarchical structuring of identification models.
  Example: An identification model with 4 different products cannot easily distinguish between the similar products fructose and glucose. If fructose and glucose are combined into a product group 'sugar', then the model is able to distinguish between sugar and the other two products. If a sample is identified as sugar, then another model takes over and distinguishes between fructose and glucose. This model, being more specialized, can more easily distinguish between the similar products.

Non-identified samples

If samples are falsely not identified:

- Check samples and sample processing for anomalies.
- Check probability threshold, reduce if necessary.
- Check data preprocessing and wavelength selection.
- Check the spectra of the unidentified samples in the score plot:
  - If the spectra are not already included in the model: Add the spectra to the validation dataset of the respective product.
  - In the score plot, compare the scores of the spectra to be checked with the scores of the spectra in the calibration dataset.

  If the samples are not outliers, but their variations are underrepresented in the calibration dataset, then the calibration dataset should be extended accordingly.

4.6 Qualification

Qualification models (from OMNIS Software version 4.4) distinguish a group of samples from other samples. These models are suitable, for example, for distinguishing usable samples (positive samples) from unusable samples (negative samples).

4.6.1 Calculation of qualification models

The calculation of a qualification model is proceeds in a similar fashion to that of the identification model *(see "Support Vector Machine (SVM)", chapter 4.5.1, page 71)*. However, the calibration dataset for the qualification contains only a single type of sample (positive samples).

A support vector machine (SVM) transforms the input data into a higher-dimensional space. A regularization parameter defines the position of the hyperplane, while a scaling parameter determines the degree of nonlinearity. A grid search determines a suitable projection. The decision boundary for this projection forms the basis for the qualification model.

4.6.2 Validation of qualification models

Procedure

A qualification model is usually developed and validated step by step. The number of samples is gradually increased:

1. Positive validation dataset:
   a. If the number of positive samples is limited, then no positive validation dataset will be created initially.
   b. As soon as a sufficient number of positive samples is available, a positive validation dataset is created with the help of automatic dataset splitting.

   Negative validation dataset:
   a. The negative samples collected are assigned to the negative validation dataset.
   b. The automatic determination of negative spectra can be used to detect spectral outliers and assign them to the negative validation dataset *(see "Spectral outliers – Algorithm", chapter 6.5, page 92)*.
2. Finally, samples for the positive and negative validation dataset are collected and measured on a different day, possibly by a different person and with a different instrument.

The samples in the positive and negative validation dataset are not used for the calculation of the model.

Validation

The qualification model determines a result (positive or negative) for each sample. A positive result is expected for the samples in the calibration dataset and in the positive validation dataset. A negative result is expected for the samples in the negative validation dataset. If the result corresponds to the expectation, then the prediction is correct (= successful), otherwise it is incorrect (= failed).

The OMNIS Software displays the following values for qualification models:

Successful % (Total) measures the overall correctness of a model.

$$\textit{Successful \% (Total)} = \frac{\textit{Correct predictions}}{\textit{All predictions}}$$

Similar *Successful %* numbers are available for each dataset. Ideally, all figures are 100%.

Improving the qualification

A more suitable parameterization (data preprocessing and wavelength selection) can improve the qualification model.

Non-qualified samples

If samples are falsely not qualified:

- Check samples and sample processing for anomalies.
- Check data preprocessing and wavelength selection.
- Check the spectra of the unqualified samples in the score plot:
  - If the spectra are not already included in the model: Add the spectra to the positive validation dataset.
  - In the score plot, compare the scores of the spectra to be checked with the scores of the spectra in the calibration dataset.

  If the samples are not outliers, but their variations are underrepresented in the calibration dataset, then the calibration dataset should be extended accordingly.

5 Prediction

5.1 Quantification

When predicting a sample with an unknown parameter of interest, the following procedure applies:

1. The spectrum of the sample is acquired.
2. The quantification model applies the same data preprocessing and wavelength ranges as for the spectra in the calibration dataset.
3. Based on the resulting spectrum, the model predicts the parameter of interest.

The calibration spectra and the reference values were mean-centered when the quantification model was calculated. Naturally, this is accounted for in the above procedure.

5.1.1 Outliers and result monitoring

When predicting a quantitative parameter of interest, different types of outliers can be recognized and the results can be monitored:

Outlier/monitoring		Cause (example)
Spectral outliers	Hotellings T^2	The concentration of a chemical component is outside the concentration range of the calibration samples.
	Q residuals	The sample contains parts that are not present in the calibration samples.
Nearest Neighbor Outlier (optional, starting with OMNIS Software version 4.2)		The sample exhibits sample variations that are not represented in this combination in the calibration samples.
Result monitoring (optional)		The result value of the parameter of interest is outside the acceptance criteria selected by the customer.

Spectral outliers

Spectral outliers are determined as follows:

1. The software calculates the values for Hotellings T^2 and Q residuals for the spectrum on the basis of the quantification model (PLS model).

2. If either the T^2 or Q residual value of the spectrum is larger than the corresponding critical value calculated by the model, then the sample is identified as an outlier in relation to the applied model *(see "Outlier assessment during prediction (quantification)", page 94)*.

The T^2 and Q residual values are available as variables in the OMNIS Software. They can be compared to the values in the PLS influence plot of the quantification model. The dashed lines indicate the critical values.

Nearest Neighbor Outlier

(from OMNIS Software version 4.2)

Ideally, the calibration samples cover all possible combinations of sample variations. In reality, some combinations occur more frequently, others not at all. Accordingly, the calibration samples are unevenly distributed in the latent variable space. In some areas there are many calibration samples, in between there are gaps.

If the spectrum of an unknown sample falls into a gap between the calibration samples, then the prediction result may be invalid or inaccurate. To recognize such cases, the distance *D* from the unknown sample *i* to each calibration sample *u* is calculated:

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

Here $\mathbf{s}_i$ corresponds to the scores of the unknown sample *i* and $\mathbf{s}_u$ to the scores of the calibration sample *u*. The scores are normalized and orthogonal.

The smallest distance is the distance to the nearest calibration sample and is referred to as the **Nearest Neighbor Distance** (NND):

If the NND value exceeds a certain NND limit value, then the unknown sample is referred to as a Nearest Neighbor Outlier.

The NND limit value is determined as follows:

1. An NND value is determined for each calibration sample. This value corresponds to the distance to the nearest of the remaining calibration samples.
2. The maximum NND value of all calibration samples is the NND limit value.

The NND value of the unknown sample and the NND limit value are available as variables in the OMNIS Software.

Result monitoring

In the OMNIS Software, result monitoring can be used to define warning limits and intervention limits for the range of the prediction results. As an

option, actions can be defined that are triggered when the defined limits are violated.

5.2 Identification and verification

The following procedure is used to identify or verify a sample:

1. The sample spectrum is recorded.
2. The identification model applies the same data preprocessing and the same wavelength selection as for the spectra in the calibration dataset.
3. The identification model evaluates the spectrum *(see "Prediction of the product membership of a sample", chapter 4.5.2, page 74)*.
4. If the identification model is part of a model hierarchy and if another identification model is linked to the determined product, then this model is executed. If one or more quantification models are linked to the determined product, then these models are executed.
5. The identification result or the verification result is displayed; in the case of a model hierarchy, quantification results may also be displayed.

Identification status

- Identified
  The identification is successful.
- Ambiguous
  Several products exceed the probability threshold. The identification failed.
- Not identified
  No product exceeds the probability threshold. The identification failed.

Verification status

- Successful
  The sample was successfully identified and the result corresponds to the expected product.
- Failed
  The verification failed.

5.3 Qualification

The following procedure is used for the qualification of a sample:

1. The sample spectrum is recorded.
2. The qualification model applies the same data preprocessing and the same wavelength selection as for the spectra in the calibration dataset.
3. Based on the resulting spectrum, the model qualifies the sample.
4. The qualification result is displayed.

Qualification status

- Successful
- Failed

6 Appendix

6.1 Linear regression example

Univariate linear regression

In the simplest case, a mixture has one only absorber and the spectrum has only one peak. Samples with different concentrations of the absorber exhibit peaks with different absorbance values *(see "Beer-Lambert law", chapter 2.2.1, page 6)*.

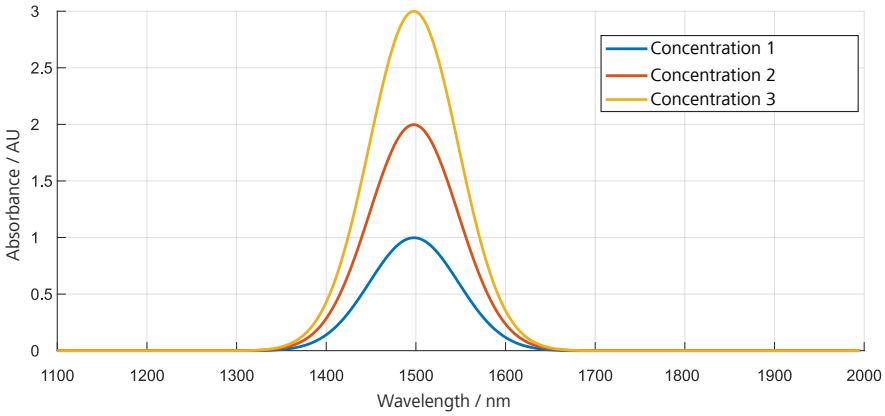


Figure 38 Modeled data with 3 samples with a peak at 1,500 nm.

The 3 measured absorbance values at 1,500 nm can be plotted against the concentration of the absorber.

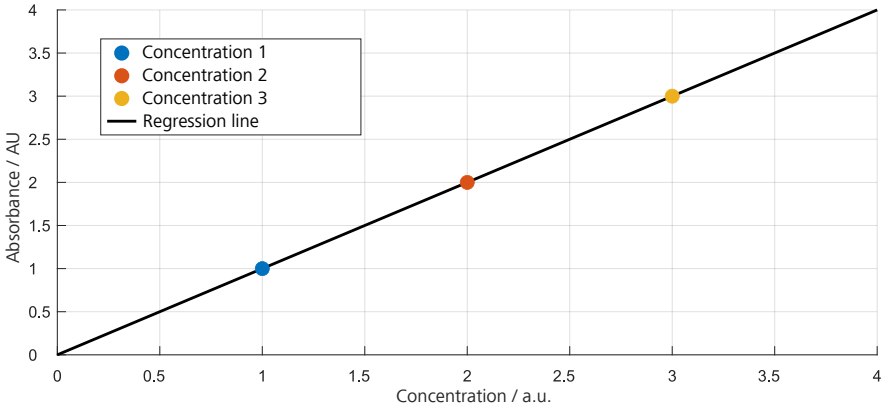


Figure 39 Relationship between absorbance values and concentrations.

In accordance with the Beer-Lambert law, the relationship between the absorbance values and the concentrations is linear. A linear regression can be performed with the set of 3 samples that yields a regression line as shown in the figure.

Since there is only 1 variable (1 wavelength), it is a univariate linear regression. This regression can be used as a quantification model. For a sample with unknown concentration, the absorbance A is measured at 1,500 nm. The regression line then reveals the corresponding concentration c of the absorber:

$$c = bA$$

The coefficient b is constant and identical to the slope of the regression line.

Note that all samples must contain the same absorber with the same molar absorptivity. Furthermore, all absorption measurements must be performed with identical pathlengths.

Multivariate linear regression

Real life mixtures contain more than one absorber. The acquired spectrum is the sum of all absorber spectra *(see "Beer-Lambert law", chapter 2.2.1, page 6)*.

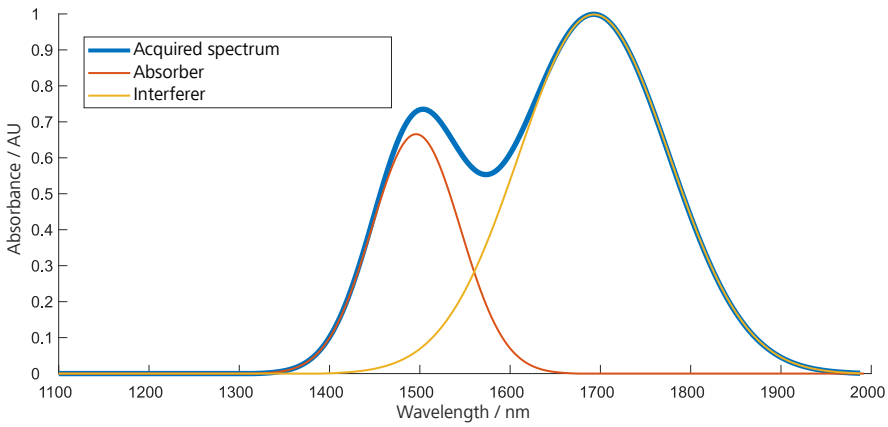


Figure 40 Modeled data with 2 components. The absorber (red line) is to be quantified.

The acquired spectrum (thick blue line) is the sum of the pure absorber spectrum and an overlapping interferer spectrum. At 1,500 nm, the measured absorbance value consists not only of the absorbance from the absorber, but also of the absorbance from the interferer. It is not possible to quantify the parameter of interest with a single measurement at 1,500 nm. Neither is it possible to know whether an interferer was present and whether the measurement is reliable.

What happens when measuring at 2 wavelengths, e.g., at 1,500 nm and 1,700 nm? The measured absorbance at wavelength 1, A_1, is the sum of the pure signal of the absorber, A_1^a (index a = absorber), and the pure signal of the interferer, A_1^f (index f = interferer). The same holds true for the measured absorbance at wavelength 2, A_2:

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

Here, $\varepsilon_1{}^{\mathrm{a}}$ and $\varepsilon_1{}^{\mathrm{f}}$ correspond to the molar absorptivities at wavelength 1 for the absorber and the interferer, respectively, and c_{a} and c_{f} to the concentrations for the absorber and the interferer.

In the above equations, the pathlength l from the Beer-Lambert law is excluded. This makes the algebra later a bit easier. The pathlength must of course be the same for all samples. The absorbances in the equations are therefore absorbances per cm.

The equations can be written in matrix form as follows:

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

Therefore:

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

Solving for the concentration of the absorber results in:

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

The result of this is that, even in the presence of an interferer, the concentration of the absorber can still be calculated – by measuring the absorbance at two wavelengths and multiplying each absorbance by a constant.

The constants are related to the molar absorptivities and could be found by consulting tables. In reality, however, this is never done. Instead, they are found via a calibration step and by solving the linear system of equations using a multivariate linear regression, such as PLS regression. The constants are therefore referred to as regression coefficients, named b_1 and b_2:

$$c = b_1 A_1 + b_2 A_2$$

More than 2 absorbers

As shown above, with 1 absorber it is sufficient to determine the absorbance at 1 wavelength. With 2 absorbers it is sufficient to determine the absorbance at 2 wavelengths.

This can be generalized. More absorbers require more absorbance values A_i at different wavelengths i. The relationship is still linear:

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

Developing a quantification model

Before the above equation can predict the concentration in unknown samples, the coefficients b_1, b_2 etc. must first be determined. This involves a calibration step. Several samples with different concentrations of the parameter of interest are measured.

To be consistent with the terminology used later for PCA and PLS, *c* can be replaced by *y*, and *A* can be replaced by *x*. When written out for each calibration sample, the above equation then appears as follows:

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

Here, *n* corresponds to the number of samples, *f* to the number of wavelengths, y_1 to the reference value for sample 1, as measured with the reference method (e.g., titration), $x_{1,1}$ to the measured absorbance of sample 1 at wavelength 1, and $\{x_{1,1} \ldots x_{1,f}\}$ to the spectrum of sample 1, as measured at *f* wavelengths. $b_1 \ldots b_n$ correspond to the regression coefficients, and $e_1 \ldots e_n$ to error terms which reveal how well the regression coefficients model the measured data.

In more compact matrix form, this is:

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

X is defined as matrix $f \times n$. $\mathbf{X^t}$ is the transposed matrix **X**, i.e., the rows and columns are switched to obtain the above matrix $n \times f$. The prediction vector **p** corresponds to the above regression coefficients **b**.

The prediction vector **p** enables the prediction of the parameter of interest for a new sample, based on its spectrum **x**. The calculated value $\hat{y}$ is:

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

The task of the multivariate linear regression is to determine regression coefficients that produce minimized error terms. A multiple linear regression (MLR) would require a number of calibration samples greater than the number of wavelengths. An additional obstacle is the high correlation between the variables.

For spectroscopic predictions, other methods can be used. PCA can greatly reduce the amount of data and completely eliminate correlation. A PLS regression also takes the reference values of the samples into account.

6.2 PCA algorithm

Principal component analysis (**PCA**) is used for the automatic dataset splitting and the spectral outlier detection *(see "Principal Component Analysis (PCA)", chapter 4.2, page 34)*.

For the principal component analysis of the parameterized and mean-centered spectra, the OMNIS Software performs a **singular value decomposition** (**SVD**). It decomposes the original data matrix **X** (the spectra of the calibration dataset) into 3 matrices and calculates n principal components:

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

Here **X** corresponds to the original spectral data (an $f \times n$ matrix with f wavelengths and n samples), **L** to the loadings (an $f \times n$ matrix), Σ to the singular values (an $n \times n$ matrix), and **S** to the scores (an $n \times n$ matrix). The equation can be shown in graphical form *(see figure 41, page 89)*.

The loadings **L** project the axes of the wavelength space onto the axes of the principal component space. The column vectors of **L** are the principal axes or principal directions.

The scores **S** are the projections of the original spectra onto the principal axes. Each row vector of **S** contains a specific spectrum.

Σ is a diagonal matrix of the singular values σ_i. The singular values describe the variance explained by each principal component. By convention, they are sorted such that $\sigma_1 > \sigma_2 > \ldots > \sigma_n$.

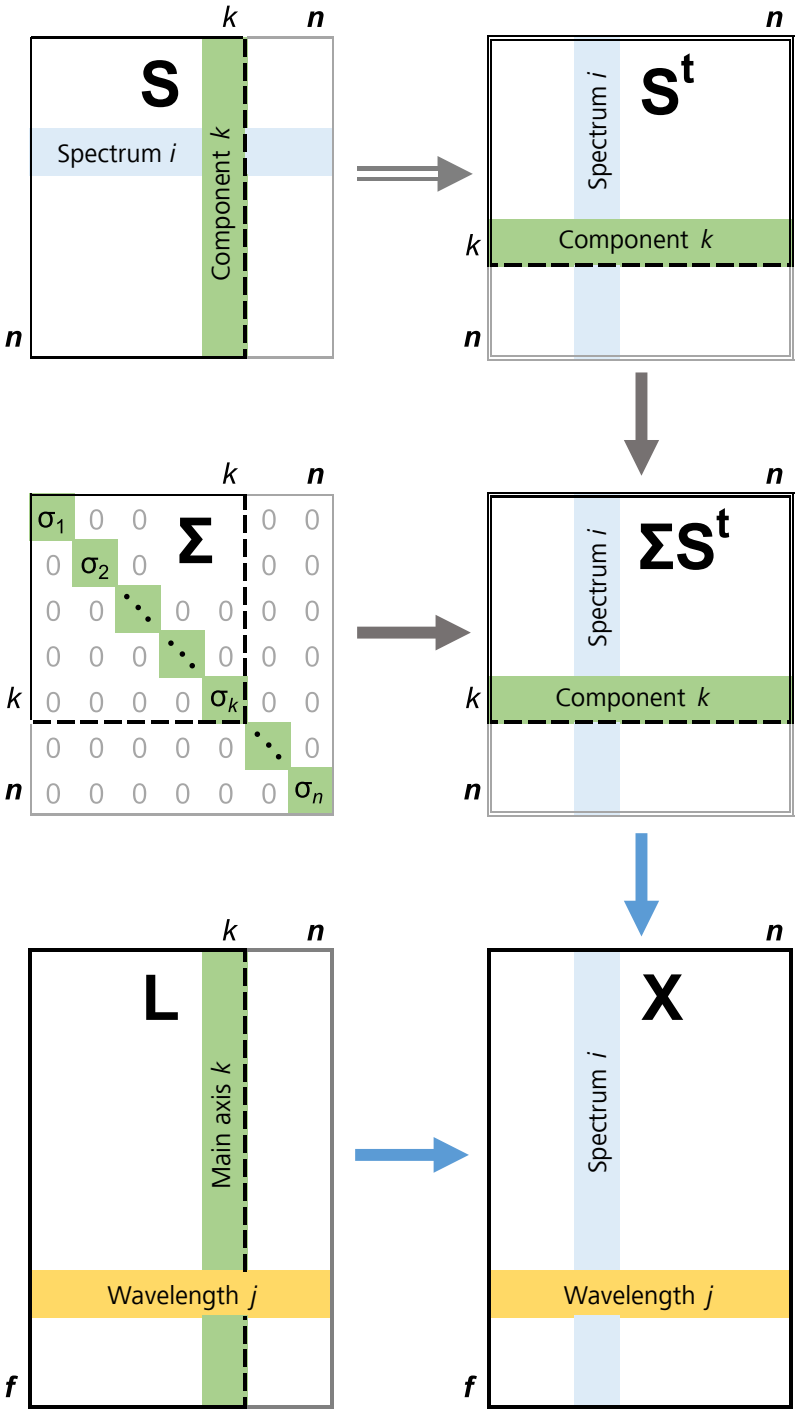


Figure 41 *Equation of the singular value decomposition in graphical form. The dashed lines show the truncated matrices for a model with k principal components. The information in the n–k omitted principal components flows into the residual matrix (an f × n matrix, not shown).*

Residual matrix

A PCA model uses only the first few of the calculated *n* principal components. In *Figure 41*, it is the first *k* of *n* principal components. The original data **X** can be subdivided into data described by the model and data not described by the model:

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

Here **X** corresponds to the original spectral data (an $f \times n$ matrix), $\mathbf{L_a}$ (an $f \times k$ matrix) to the first *k* columns of **L**, $\mathbf{\Sigma}_a$ (a $k \times k$ diagonal matrix) to the first *k* singular values, $\mathbf{S_a}$ (an $n \times k$ matrix with *k* principal components) to the first *k* columns of **S**, and **E** to the residual matrix (an $f \times n$ matrix) containing all spectral variations in **X** that cannot be described by the model.

As a rule, $k \ll n \ll f$. For example: $k = 3$ principal components, $n = 100$ samples, and $f = 2{,}500$ wavelengths.

Each column $\mathbf{e_i}$ of the residual matrix **E** shows the orthogonal distance of the spectrum *i* to the PCA space, called **residual**. The more principal components that the model uses, the smaller the residual will be.

6.3 PLS algorithm

The **partial least squares regression** (**PLS regression**) is used to calculate the quantification model *(see "PLS regression", chapter 4.4.1, page 57)*.

PLS includes two blocks of data:

- The parameterized and mean-centered **X** matrix (the spectra).
- The mean-centered **y**-vector (the reference values).

PLS decomposes the matrix **X** into 2 matrices:

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

Here **X** (an $f \times n$ matrix with *f* wavelengths and *n* samples) corresponds to the preprocessed and mean-centered spectra, **L** to the loadings (an $f \times k$ matrix with *k* latent variables), **S** to the scores (an $n \times k$ matrix), and **Z** to the residual matrix (an $f \times n$ matrix) containing all spectral variations in **X** that cannot be described by the model.

Whereas with PCA, the score matrix **S** explains the variance of **X**, with PLS the score matrix **S** explains covariance between **X** and **y**. PLS maximizes the covariance explained by the scores. That means the scores not only best explain the variance in **X**, but also exhibit the greatest possible correlation with the reference values.

To maximize the covariance between **X** and **y**, the PLS algorithm swaps data between **X** and **y**. **X** and **y** therefore merge into a single, integrated

system. In the process, the scores **S** are regressed against the reference values **y** in order to obtain the regression coefficients **b**:

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

For this, **e** is the residual vector containing all reference value variations in **y** that cannot be described by the model.

Prediction

The prediction vector **p** can be determined from the regression coefficients **b**. The prediction vector **p** and the preprocessed and mean-centered spectrum **x** are used to predict the parameter of interest $\hat{y}$ of a new sample:

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

The OMNIS Software implements PLS with the SIMPLS algorithm and with a single set of reference values (PLS-1).

6.4 Hotelling's T² and Q residuals

Hotelling's *T*² and Q residuals characterize the spectra in a PCA or PLS model. In particular, they help to identify possible outliers *(see "Hotellings T² and Q residuals", page 48)*.

Hotelling's T²

The Mahalanobis distance is a measure for the magnitude a spectrum deviates from the model center. The distance is normalized. Each principal component, or latent variable, is assigned the same weight.

Assuming that the spectra or the scores, respectively, are normally distributed, the squared Mahalanobis distances, MD², follow a Hotelling's *T*² distribution:

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

The squared Mahalanobis distance for spectrum *i* is the sum of squares of the normalized scores for the first *k* principal components or latent variables:

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

Here $\mathbf{s_i}$ corresponds to the *i*th line of the truncated score matrix **S**, $s_{i,a}$ to the normalized score for spectrum *i* and principal component (or latent variable) *a*, and *k* to the number of principal components, or latent variables used.

MD² can be referred to simply as T^2. A spectrum with $T^2 = 0$ projects onto the center of the mean-centered model, i.e., all scores are at the mean. The further a spectrum's projection onto the hyperplane moves away from the center, the higher the T^2 value will be. The model fits best close to the center. Far from the center, the model may fit poorly.

A **significance level**, e.g., of 5%, can be defined for a hypothesis test that detects outliers *(see "Spectral outliers – Algorithm", chapter 6.5, page 92)*.

Q-residual

The Q residual for the spectrum *i* is the squared distance from the PCA model or PLS model to the spectrum:

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

Here $\mathbf{e_i}$ corresponds to the *i*th column of the residual matrix **E**, $e_{i,j}$ to the residual for spectrum *i*, and *f* to the number of wavelengths.

Q residuals reveal the variations that are not explained by the model. A high Q residual indicates that the spectrum may be inconsistent with the model, e.g., if the measured sample contains a different substance.

A **significance level**, e.g., of 5%, can be defined for a hypothesis test that detects outliers *(see "Spectral outliers – Algorithm", chapter 6.5, page 92)*.

6.5 Spectral outliers – Algorithm

The detection of spectral outliers identifies spectra that deviate from the population *(see "Detecting spectral outliers", page 49)*. The algorithm assesses whether the values for Hotellings T^2 or for the Q residual of the spectrum being tested are the result of random variation or systematic variation.

Spectral outlier detection during model development

1. Parameterization is taken into account as follows:
   a. From OMNIS Software version 4.2: The user decides whether the parameterization (data preprocessing and wavelength selection) is applied or not.
      Subsequent changes to the parameterization have no influence on dataset splitting.
   b. From OMNIS Software version 3.3 to OMNIS Software version 4.1: The user decides whether or not data preprocessing is taken into account. The wavelength selection and subsequent changes to the data preprocessing have no influence on dataset splitting.
   c. Up to OMNIS Software version 3.2: Data preprocessing is taken into account as specified at the time of outlier detection. The wavelength selection and subsequent changes to the data preprocessing have no influence on dataset splitting.
2. The spectral outlier detection is based on the PCA model of all mean-centered spectra in the spectra list *(see "Principal Component Analysis (PCA)", chapter 4.2, page 34)*. The spectrum being tested is also included in the PCA model. The number of principal components is chosen in such a way that the explained variance is at least 95%.
3. **Hotellings T² outlier**:
   Based on the Hotellings T^2 values *(see "Hotelling's T²", page 91)*, a hypothesis test is performed according to H. Hotelling, *The Generalization of Student's Ratio*, The Annals of Mathematical Statistics Volume 2, No. 3 (Aug. 1931), pages 360–378.
   a. The null hypothesis is that the T^2 value of the spectrum to be examined fits the T^2 values of the PCA model. If the null hypothesis is true, then T^2 follows a Hotellings T^2 distribution. The distribution can be expressed as a scaled F-distribution:
   $$T^2 \sim \frac{k(n-1)}{n-k} F_{k,n-k}$$
   Where k is the number of principal components, n is the number of spectra, and $F_{k,n-k}$ is the F-distribution with parameters k and $n–k$.
   b. The adjustable significance level controls the probability with which the null hypothesis is rejected in the event that it is true (known as a type I error). The default value is 5%.
   c. Based on this distribution and the significance level, a critical value for T^2 is calculated.
   d. If the T^2 value for the spectrum is larger than the critical value, then the null hypothesis is rejected and the spectrum is flagged as a potential outlier.

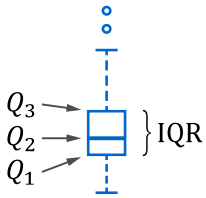
4. **Q residuals outlier**
   Based on the Q residuals *(see "Q-residual", page 92)*, a hypothesis test is performed according to J. E. Jackson and G. S. Mudholkar, *Control Procedures for Residuals Associated With Principal Component Analysis*, Technometrics Volume 21, No. 3 (Aug. 1979), pages 341–349.
   a. The null hypothesis is that the Q residual of the spectrum to be examined fits the Q residuals of the PCA model. If the null hypothesis is true, then an approximately normally distributed Q residual test statistic can be constructed.
   b. The adjustable significance level controls the probability with which the null hypothesis is rejected in the event that it is true (known as a type I error). The default value is 5%.
   c. Based on this test statistic and on the significance level, a critical value for Q is calculated.
   d. If the Q residual value for a spectrum is larger than the critical value, then the null hypothesis is rejected and the spectrum is flagged as a potential outlier.

Outlier assessment during prediction (quantification)

With quantification, an outlier assessment is available during prediction:

1. Preliminary steps:
   a. The quantification model uses the same algorithm as above. However, the base is the quantification model (PLS model) with all mean-centered spectra of the calibration dataset. Data preprocessing, wavelength ranges, and the number of latent variables are taken into account as specified in the quantification model.
   b. Based on the specified significance level, the quantification model calculates the critical values for T^2 and Q (dashed lines in the PLS influence plot). The critical values are recorded in the quantification model.
2. During prediction, the T^2 and Q values of the spectrum are calculated on the basis of the quantification model.
3. If either the T^2 or Q value of the spectrum is larger than the respective critical value, then the sample is considered to be a potential outlier in relation to the applied quantification model.

6.6 Reference value outliers – Algorithm



Boxplots enable the detection of outliers in the reference values *(see "Reference value outliers (quantification)", chapter 4.3.4, page 54)*.

To take into account the skewness of the distribution, the outlier cutoff values are adjusted with the following calculations.

The **medcouple** (MC) measures the skewness of the reference values. The calculation starts with the median of the boxplot, Q_2. A function is calculated with all possible *couples* of the upper half and the lower half of the reference values. The median of the results is the medcouple:

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

Here Q_2 corresponds to the second quartile that the center line defines in the boxplot, and y_i, y_j to a pair of reference values.

The medcouple is always between –1 und 1. For a symmetrical distribution, MC = 0. A skewed distribution with MC > 0 is skewed towards the higher reference values, with MC < 0 towards the lower reference values.

The calculation of the **adjusted outlier cutoff values** depends on the side to which the distribution is skewed:

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

In the case of a symmetric distribution (MC = 0), the distances between the cutoff values and the box are 1.5 IQR.

The exponential function enables accurate and robust outlier detection for various distributions with varying skewness, as empirically shown by M. Hubert und E. Vandervieren, *An adjusted boxplot for skewed distributions*, Computational Statistics & Data Analysis Volume 52, No. 12 (Aug. 2008), pages 5186–5201.

The expected percentage of flagged outliers is around 1% and is quite similar to the percentage of the standard boxplot for the normal distribution. Note: This percentage is independent of the significance level.