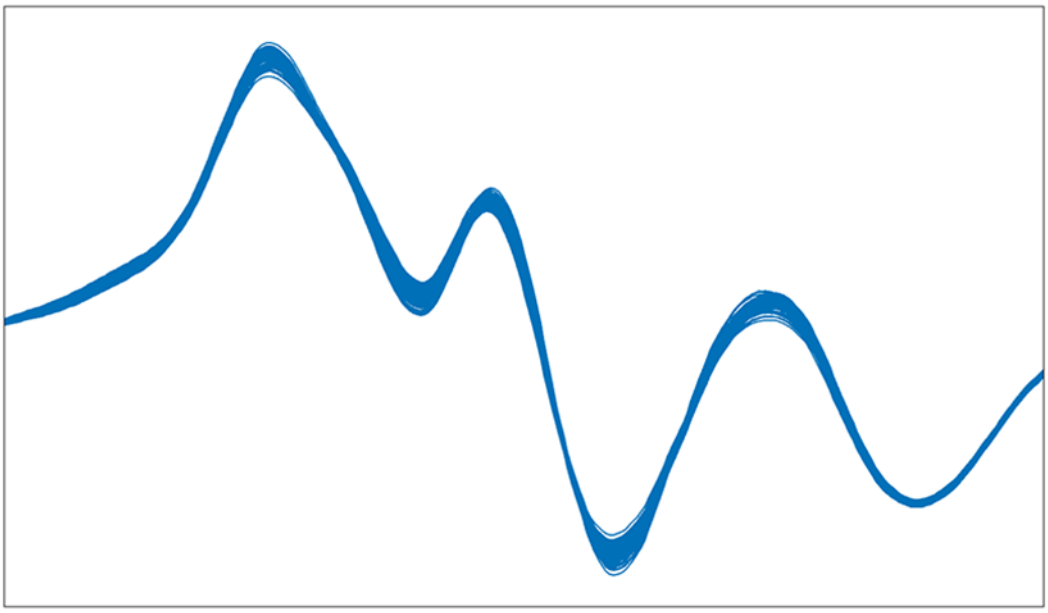
OMNIS NIR 理论



手册

8.0600.8101CN / 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 理论

手册

8.0600.8101CN / v9 /
2025-10-10

Technical Communication
Metrohm AG
CH-9100 Herisau

本文献为原件。

本文献经认真起草制定。但并不能完全排除会有错误存在。若有此类提示请联系上述地址。

免责条款

并非 Metrohm 造成的故障情况，例如不按规定储存、不按规定使用等，则不属于保修范围。擅自变更产品（比如改装或加装）会排除生产厂家对由此造成的损失及其后果的责任。要严格遵守 Metrohm 产品文档中的说明和注意事项。否则排除 Metrohm 的责任。

目录

1 概览

1.1 引言

近红外光谱（NIR 光谱）是一种适用于样品宽光谱的无损坏、快速并且无须试剂的分析方法。该方法可以同时分析多个参数并且既可以测定一个材料的物理属性也能够测定其化学属性。其中包括分析物浓度、密度、颗粒大小或固有粘稠度。

此外，NIR 光谱还可对未知样品（OMNIS Software 版本 4.0 以上）和校验样品（OMNIS Software 版本 4.2 以上）进行身份验证。

能够远距离无损测量样品在品控和工艺流程监控中具有至关重要的意义。

该手册描述了如何在 OMNIS Software 中应用记录、处理和分析近红外光谱的技术和算法。第 2 章中说明了如何将测量信号转换为吸收光谱。第 3 章所阐述的是仪器的校正。第 4 章说明的是能够预测分析参数（量化）或产品关联性（身份验证）的模型的开发。第 5 章所述为对未知情况的预测。第 6 章为包含不同算法说明的附录。

1.2 方案框架

所述步骤插入以下框架：

1. **校正、标准化和性能测试**
   确保仪器所采集的吸收光谱的可传输性和可靠性。
2. **开发模型**
   针对量化参数或身份验证样品预测开发模型。
   开发基于包含已知分析参数或已知产品关联性的样品。
3. **样品分析**
   对被分析的样品采集光谱。量化模型基于光谱提供量化预测，或对识别模型进行识别或验证样品。
4. **监控**
   监控模型和仪器用于确认系统适于进一步的分析。

1.3 文献说明

惯用图例

粗显大写字母代表矩阵，粗显小写字母代表向量。转置矩阵或向量通过一个上方小字符 t 标示，例如 $\mathbf{x}^t$。

标量通过小写字母标示。上尖角符号 (^) 标示预估（预测）的数值，例如 $\hat{y}$。横线代表平均值，例如 $\bar{y}$。

标量可能表示的是波长相关的变量，例如吸收率 A。标量书写方式可以与矢量书写方式交替使用。

1.4 详细信息

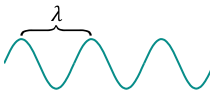
在以下页面上可以找到有关产品的附加信息：

- Metrohm 网站 *https://www.metrohm.com* – 产品系列概览、PDF 文档、附件说明以及应用信息。
- OMNIS Software 的帮助 *https://guide.metrohm.com* – 根据主题筛选的 OMNIS Software 信息。

2 近红外线光和光谱

2.1 光及其与材料的交替反应

光谱仪用于测量样品与光的相互作用。光可以以不同成都吸收或发散。交替反应取决于光的属性（特别是波长）以及材料的属性（特别是分子结构）。



波长

光是一种电磁辐射。光作为具有电场和磁场的波在空间中运动。波在空间和时间中传播。波相应地通过其波长 λ（例如以纳米 = 10^{-9} 米为单位）及其频率 (Hz) 标示。

相反，波长与波的频率成正比。越高频率波（每秒振动次数更多）的波长则越短。基于此关系，波既可以用波长（nm）也可以用（Hz）描述。

光可以将离散量子单元更换为光子。单个光子的能量 E 取决于其频率 f 或波长 λ:

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

h 相当于普朗克常数，c 相当于光速。

图 1 显示电磁辐射的不同区域。

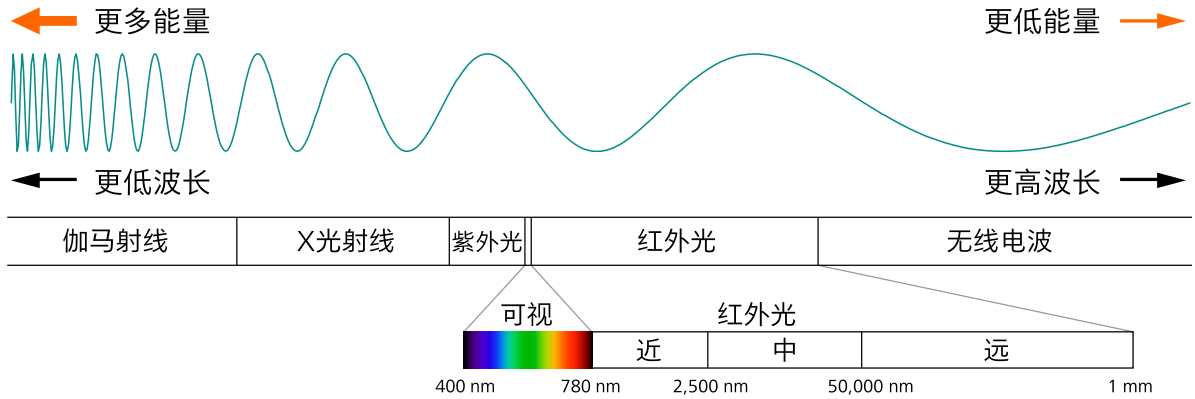


图 1 电磁辐射的区域。近红外光区域 (NIR) 位于可见光区域旁。NIR 涵盖 780 nm 至 2,500 nm 的波长。

辐射源

电磁辐射的不同区域具有不同的辐射源。NIR 区域中为 **热辐射** 源。一个红外摄像头用于识别例如人体，因为人体的温度比环境温度高。

强度

电磁波的振动幅度决定了光的强度。振动幅度越高，光的强度越高。对于可见光而言，光的强度以亮度感知。

光和材料的交替反应

接下来说明光通过分子吸收的过程。

光与材料相互作用的方式方法取决于电磁辐射的区域。例如如果可见光的能量被传输至分子，分子中的电子由一个较低的能级切换至一个较高的能级（电子跃迁）。红外区域内发生 **振动跃迁**。化学键、官能团和分子可能以不同方式振动，例如由于伸缩振动、变形振动或旋转振动。

分子只能吸收离散振动状态。环境温度条件下，大部分分子处于振动基态（0 级）。由振动基态跃迁到活跃态按以下模式命名：

振动跃迁 ***i* ➔ *j***	名称
0 ➔ 1	基态跃迁
0 ➔ 2	第一个谐波跃迁
0 ➔ 3	第二个谐波跃迁

振动状态越高，能级越高。对于由状态 i 跃迁到活跃态 j，分子必须吸收一定的跃迁能量 ΔE_{ij}。

光可以将能量更换为 $E = hf$，其中 f 是光的频率。如果光子能量 hf 等于跃迁能量 ΔE_{ij}，则光开始吸收。

所允许的振动状态取决于键强度和所参与原子的质量。因此，特定的键形式会有特定的跃迁能量和吸收的波长。

为了能吸收光，必须满足以下条件。振动跃迁的电荷分布推移必须确保分子的电偶极矩有变化。能量吸收的概率取决于偶极矩沿相关化学键变化程度的大小。

振动跃迁对于极性或非极性分子以及功能团都会影响偶极矩的变化。同源双核分子如 N_2 不吸收红外光。

所激发振动状态的时长有限。如果分子回到较低的振动状态，则能量将转换为热能。

NIR 光谱范围

基态跃迁的波长处于中红外区。近红外区涵盖谐波跃迁和合频带。*图 2* 显示了不同分子和官能团所吸收的波长带。

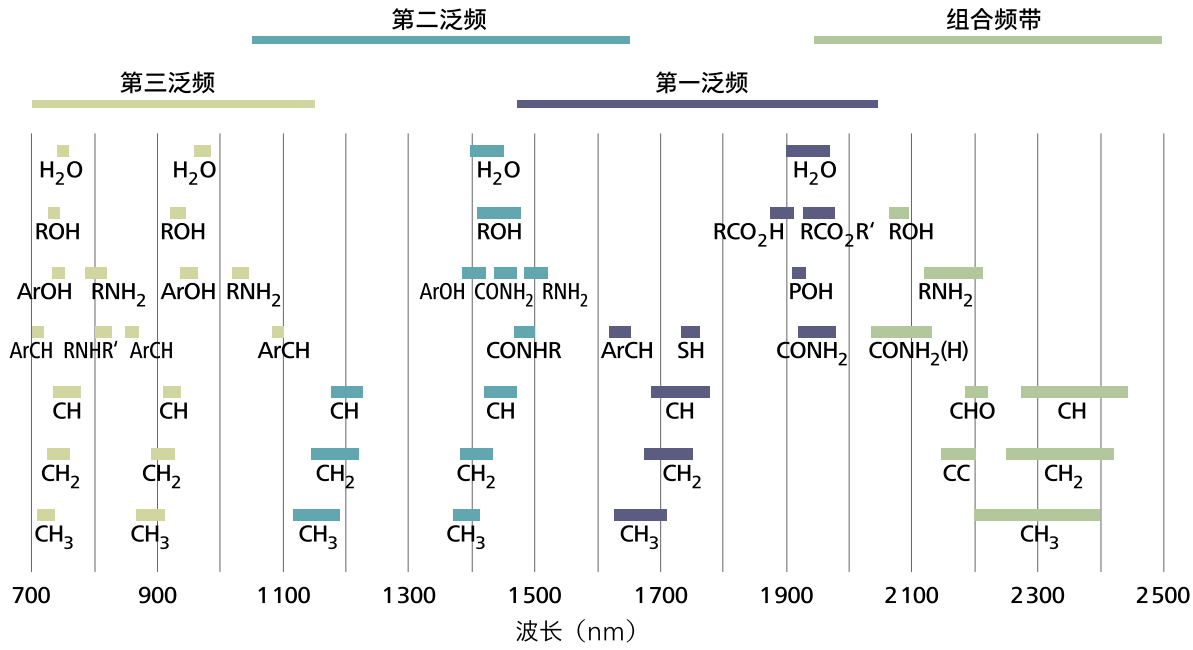


图 2　NIR 吸收带

基态跃迁是最有可能的跃迁，也最常出现。谐波跃迁的概率小一些。因此，基态跃迁比谐波跃迁吸收的光多。通常，吸收量会随着每个谐波而递减。因此，谐波适于强吸收分子。

两个或更多的基态振动可能会同时激发同一个光频率，即基态振动的组合频率。对应的吸收带被称为 **合频带**。有些合频带处于 NIR 区域，即 1,900 和 2,500 nm 之间。

2.2　数学基本原理

2.2.1　Beer-Lambert 定律

Beer-Lambert 定律描述了光通过均匀的样品时，吸收量与样品内物质属性的关联度。

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

其中，A 相当于吸光度，ε 为吸收器的摩尔消光系数 (L/mol/cm)，c 为吸收器的浓度 (mol/L) 以及 l 样品的液层厚度 (cm)。

摩尔消光系数 ε 是一个常数，表示一个物质的吸收量。摩尔消光系数对于特定波长 i 和特定物质 j 而言是一定的。混合物的总吸光度为混合物中所有物质的吸光度总和：

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

其中，A_i 为波长 i 的吸光度，N 是混合物中的物质数量，ε_{ij} 则是波长 i 和物质 j 的摩尔消光系数，c_j 为物质 j 的浓度。

Beer-Lambert 定律的前提条件是吸光度和浓度之间的线性关系以及吸光度与摩尔消光系数之间的线性关系。此线性关系适用于多种情况。

基于 Beer-Lambert 定律，光谱吸收量测量可以用于证明：

- 吸收器中浓度的变化。
  这是最常见的应用。
- 影响摩尔消光系数因素的变化。
  溶剂温度、粘稠度、pH 值或介电常数都可能影响摩尔消光系数。在某些情况下，可以将其用于光谱测量。

Beer-Lambert 定律不考虑散射。散射有时可被用于识别颗粒大小的变化。

2.2.2 线性回归

一个波长

最简单的情况是一个混合物中只有一个吸收器。根据 Beer-Lambert 定律，特定波长的吸光度与吸收器的浓度呈线性关系。

图 1 中每个点均代表一个吸收器已知浓度（x 轴）和吸光度测量值（y 轴）的样品。通过线性回归得出相应的回归线 **A**。

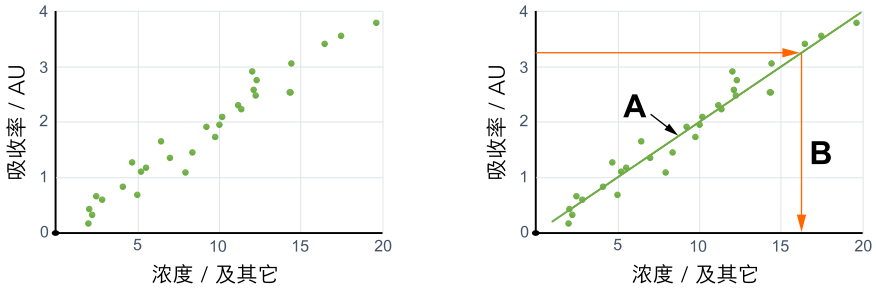


图 3　吸光度值和浓度之间的关系

对于吸收器未知浓度的样品，可以按照以下方式确定其浓度：

1. 测量指定波长时的吸光度。
2. 使用回归线确定浓度（**B**）。
   回归线是用于预测分析参数（例如吸收器浓度）的简单的量化模型。

但如果混合物中含有不同浓度的多个吸收器时，则该操作方法不适用。

多个波长

这种情况下，测量单一波长的吸收量就不适用了，而是要测量多个波长的吸收量，从而得到一个光谱。和上述情况一样，可以通过线性回归确定光谱与分析参数之间的关系。多个波长需要多元线性回归（MLR）。

即便混合物含有不同浓度的多个吸收器，多元线性回归也可以预测分析参数，这一点已得到证实 *（参见章节 6.1，第 69 页）*。

但对于多元线性回归而言，样品数目必须多于波长数目。其它方法也可用于光谱预测，例如 PCA 或 PLS。

2.3 光如何转变为光谱

一个光谱仪由一个光源和一个探测器单元组成。光源发射具有宽波长范围光谱的光，即多色光。光与样品相互作用。然后，光谱仪将剩余光的参数归纳为一个波长的函数。

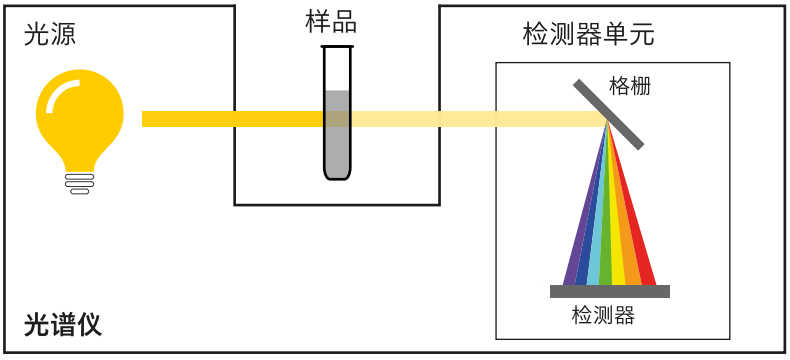


图 4　一个包含光源和检测器单元的光谱仪。

在光谱仪中，光通过格栅分解为不同的波长。检测器用于测量光，其中每个波长涉及行传感器中的一个元素（或像素）。

扫描 是测量所有像素。每个像素均产生一个光电信号，该信号与光强度成正比。信号可以通过与像素的对比方式展现。

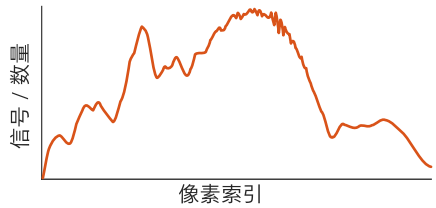


图 5　检测器信号的光谱作为像素函数。

积分时间

积分时间是检测器采集光的时间区间。较长积分时间可以提升信号。

积分时间过长会导致达到探测器饱和度以及信息丢失。过短的积分时间则会降低信号以及信噪比。

自动积分时间 能够确保最佳曝光，即最佳的信噪比，而不会达到饱和度。每次样品扫描以及每次参考扫描之前都会进行多次测量。积分时间的设置须确保高强度信号能够达到所识别区域的约 90%。

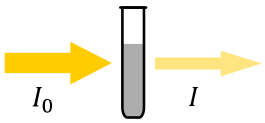
在进一步计算时将考虑积分时间的差别。

- **OMNIS NIR Analyzer**
  仪器始终自动设置积分时间。

- **2060 The NIR**
  积分时间可以手动或自动设置。
  手动积分时间可以缩短重复、类似测量的测量持续时间。为了防止积分时间过长而导致达到检测器饱和度，应在设置积分时间时考虑到留有足够的余量 *（参见"饱和度"，第 39 页）*。

吸光度

光谱测量用于确定一件样品吸收或散射的光量。样品通过光照射。检测器测量光源发出的光并在与样品交替作用后测量剩余的光。



吸光度 A 被定义为光照射与样品交替作用前光强度（I_0）和光照射与样品交替作用过后光强度的常用对数（I）：

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

吸光度为 1 意味着光的 10% 进入样品，而吸光度为 2 则代表 1% 的光进入了样品。

吸光度没有单位。

参考扫描和样品扫描

根据上述公式，计算吸收光谱需要两次扫描。扫描针对每个像素测量一个光电信号 S：

- **参考扫描** 测量光照射与样品交替作用前的信号 S_0。
- **样品扫描** 测量光照射与样品交替作用后的信号 S。

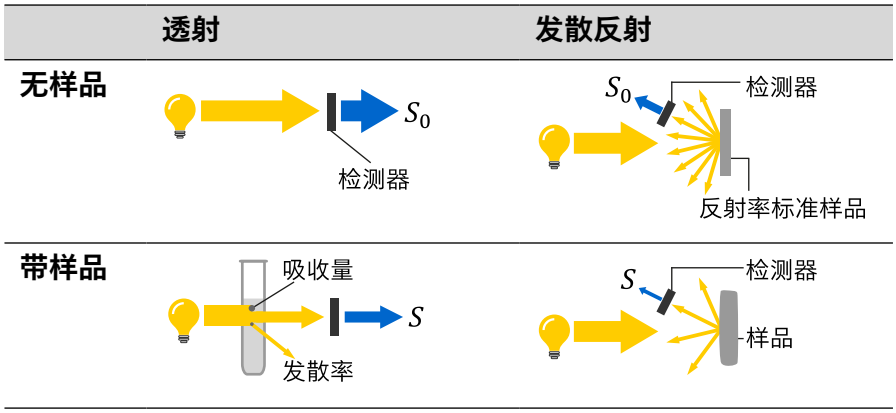
所测量的一个像素的光电信号在一个像素平面上与光强度的平均值成正比。因此 $S_0/S = I_0/I$。这样可以将光电信号用于计算吸光度：

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

透射和反射

透射模式用于测量穿过样品的光线。S_0 是在无样品的情况下测得的数值。S 是测量进入穿过样品的光线。

反射模式 是测量样品反射的光线。作为参考采用反射率标准样品替代样品。反射率标准样品能够在理想情况下反射 100% 的光。反射光的一部分被导向检测器并提供信号 S_0。信号 S 则以同样的方式测量，但采用的是反射光的样品。



所计算的吸光度 A 体现的是所有未达到检测器的光。因此，A 不仅包含样品吸收的光，还包括：

- 因检测器散射而未达到检测器的光。
- 被错误散射到检测器的光。

吸收光谱

吸收光谱根据参考扫描（信号 S_0）和参考扫描（S）依照上述公式计算。

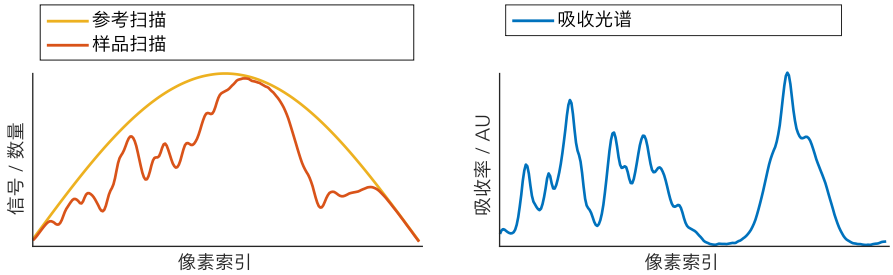


图 6　参考扫描和样品扫描（左）以及所计算出的吸收光谱（右）作为像素索引的函数。

上述计算的前提条件是参考扫描和样品扫描要使用相同的光路径，或具有相似光学属性的光路径。在过程环境中使用 1 个多级参考方案 *（参见章节3.2.2，第13 页）*。

从像素到波长

像素标度转换为波长标度。仪器为每个像素分配一个准确的波长，例如：

像素 6 ➔ 波长 1,009.4 nm

每个像素准确的波长通过波长校正确定 *（参见章节3.1，第11 页）*。

转换波长标度

光谱通过内推法转换为标准波长标度：

1,000.0 nm，1,000.5 nm，1,001.0 nm，…

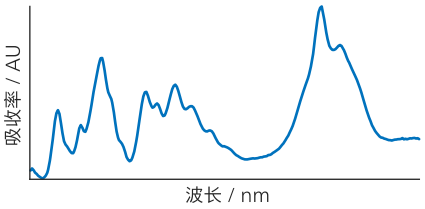


图 7　吸收光谱转换为波长标度。

3 仪器设置

以下步骤确保了对于具有特定样品测量架构的样品能够采用一致的光谱，无论何时、使用 **OMNIS NIR Analyzer** 产品系列相同或不同的仪器、还是使用 **2060 The NIR** 型的仪器记录光谱。

光谱的 x 轴和 y 轴都必须予以考虑：

- **x 轴**：波长校正 *（参见章节 3.1，第 11 页）*
- **y 轴**：参比标准化 *（参见章节 3.2，第 12 页）*

此外，必须确保仪器性能符合要求，即：

- **仪器性能测试** 必须在能够使用仪器记录光谱之前成功执行 *（参见章节 3.3，第 21 页）*。

标准溶液

对于波长校正和 仪器性能测试，仪器使用一个内部可追踪的计量学 **波长标准版**。

使用反射模式时，需要一个用于参比标准化的 **反射率标准样品** 和根据仪器产品类型而不同的 仪器性能测试：

- **OMNIS NIR Analyzer**：内部反射率标准样品
- **2060 The NIR**：外部反射率标准样品

3.1 波长校正

波长校正对波长值进行标准化，例如光谱的 x 轴。波长校正为光电传感器数组的每一个像素分配一个波长。

波长校正使用 1 个内部可追踪的计量学波长标准版。波长标准版具有 1 个已定义波峰和已知最高位置的吸收光谱。

CAL WL 命令执行以下步骤：

1. 波长标准版的吸收光谱在使用内部参照路径的情况下在像素标度上记录。
2. 记录的光谱中，最高位置以亚像素精度识别。
3. 在使用测得的像素标度最高位置和波长标准版标称最高位置的情况下执行多项式回归。
4. 回归多项式为每个像素分配其相应的波长。

回归系数保存在仪器上：

- **OMNIS NIR Analyzer**：对于每个样品展示在仪器上保存一个回归系数集。这也就是说，对于 OMNIS NIR Analyzer Liquid/Solid 必须在两个功能单元上执行波长校正和校验。
- **2060 The NIR**：回归系数因仪器型号而异。所有通道均使用相同的系数集。

波长校正的校验

执行波长校正后必须对其进行校验。**VAL WL** 命令执行以下步骤：

1. 波长标准版的吸收光谱将被记录。
2. 波长的校验结果：
   a. 记录的光谱中识别有波长标度上的最高位置。
   b. 测量的最高位置和已知最高位置之间的波长残差将被计算。
   c. 对于每个波峰而言，波长残差必须处于允差范围内，才能通过测试。
3. 波动幅度的校验：
   a. 在记录的光谱中确定有波峰宽度。
   b. 测量的波峰宽度和已知的波峰宽度之间的波峰宽度残差将被计算得出。
   c. 对于每个波峰而言，带宽残差必须处于允差范围内，才能通过测试。
4. 如果以上所述的残差均在允差范围内，则校验的整体情况合格。

校验必须在能够使用仪器记录光谱之前成功执行。

3.2 参比标准化

参比标准化用于指定吸光度值，即光谱的 y 轴。

吸光度的测定

一件样品吸光度 A 的计算需要信号 S_0（参考扫描）和S（样品扫描）*（参见章节 2.3，第 7 页）*：

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

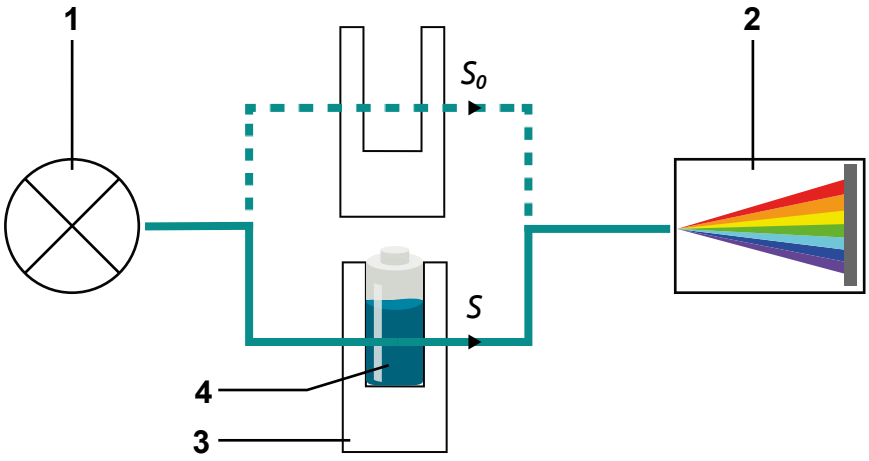


图 8　透射模式中的光路径（液体样品展示作为示例）。

图 8 中，光源（**1**）发射的光穿透样品支架到达（**3**）检测器（**2**）。

参考信号 S_0 在不带样品的情况下测量，信号 S 的测量则使用样品（**4**）。否则，两个光路径的光学属性相同，并对两种信号造成相同百分比的缓冲。这不会改变上述公式的结果。

该方程将 S 与参照 S_0 关联。两个信号同样重要。两个信号中的一个出现偏差将导致吸光度不同，最终导致得到的是不同的光谱。

S 和 S_0 受到仪器和环境条件变动的影响。为了确保这些影响相互抵消，应将两个信号在很短的时间间隔内测量。

原理的执行取决于仪器产品类型：

- **OMNIS NIR Analyzer**
  样品的吸收光谱通过信号 S 和 S_0 计算得出 *（参见章节3.2.1，第13 页）*。
- **2060 The NIR**
  在过程环境下不可对 S 和 S_0 的测量使用相同的光路径。因此需要进一步的措施 *（参见章节3.2.2，第13 页）*。

3.2.1 OMNIS NIR Analyzer

参比标准化通过测量信号 S_0 和 S 并计算吸光度 A 实现。

记录一个样品的光谱

在能够使用功能单元记录光谱之前必须对功能单元成功执行 仪器性能测试 *（参见章节3.3，第21 页）*。

1. 样品必须在样品展示中准备就绪。
2. 吸收光谱通过之前记录的参考光谱 S_0 计算而出。为了能够得到当前数值 S_0，可以执行 **MEAS REF SPEC** 命令。
   使用固体物质样品展示时，仪器自动将反射率标准样品插入光路径。该反射率标准样品不需要修正 S_0 信号。
3. 通过 **MEAS SPEC** 命令测量样品。执行该命令将得到信号 S。
4. 软件计算 A，即样品的吸光度：

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

   其中，S_0 对应在参考路径上测得的信号，S 对应通过样品测得的信号。

3.2.2 2060 The NIR

2060 The NIR 类型的仪器需要一个外部参比标准化。

外部参比标准化

在相同光学属性的光路径上进行信号 S（带样品）和信号 S_0（无样品）的重复测量既费时由容易出错。

因此将导入两个更多的照射过程 *（参见图9，第14 页）*：

- 仪器中的 **内部参比**。内部参考路径能够提供可以以简单方式测得的信号 S_{ref}。

- 光纤束与 **校正装置** 关联的另一个外部光路径。该光路径能够提供信号 S_{fiber}。

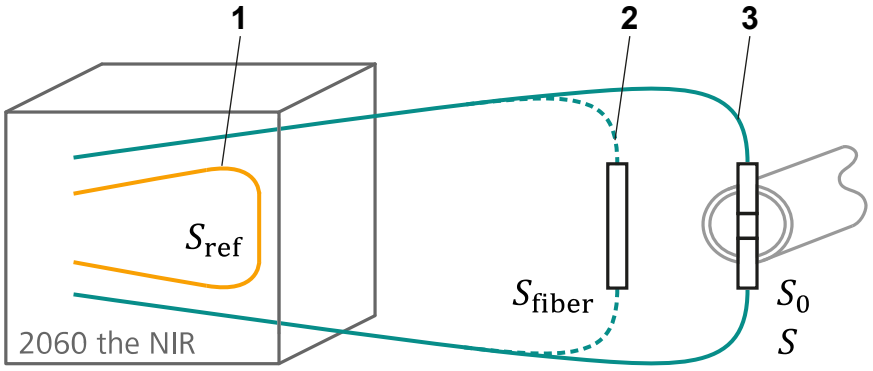


图 9　作为投射模式下光路径的示例：内部参考路径 (1)，外部光纤束与一个校正装置 (2) 以及外部光纤束与带或不带样品的探针关联 (3)。光路径 2 和 3 表示仅以不同方式关联的相同光纤束。

校正装置固定光纤束从而构成参考路径（**2**）。在透射模式下，空气用作参考，传送 100% 的光。在反射模式下，校正装置也记录反射率标准样品。首先，设定理想的反射率标准样品能够反射 100% 的光。

样品的吸光度 A 由信号 S_0 和 S 计算得出。如果将两个附加信号 S_{ref} 和 S_{fiber} 添加至分子和分母，则结果保持不变：

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

该方程可转换为：

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

3 个光谱项用于描述吸光度值并且可以按如下方式定义说明：

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

图 10 展示了信号 S_{ref}、S_{fiber}、S_0 和 S 如何测量。

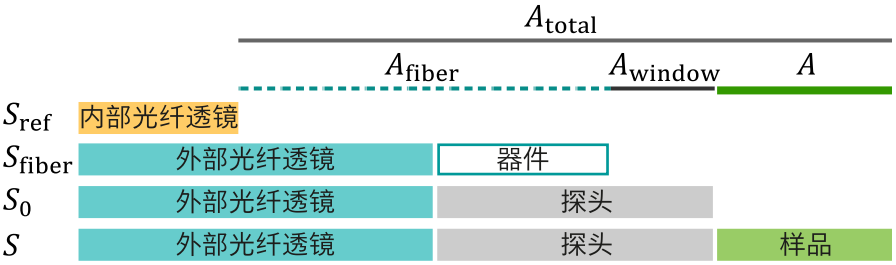


图 10　外部参比标准化

A_{total} 是外部光纤束、探针和样品基于内部光纤束的吸光度。

A_{fiber} 是外部光纤束加上校正装置基于内部光纤束的吸光度。

A_{window} 是探针减去校正装置的吸光度。

消除环境波动

为了确定样品的吸光度 A，需要测量 3 个吸光度值。根据上述方程由 3 个数值计算出 A:

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

由此可以测定 3 个不同时点的吸光度值。通过这种方式可以轻松消除 3 次中每次测定的仪器波动或环境条件波动：

- A_{total} 通过每次样品测量测定。其中，S_{ref} 和 S 需要在短时间间隔内测量，以便消除波动。
- **玻璃纤维校正** A_{fiber} 不需要多次测定。其中，S_{ref} 和 S_{fiber} 需要在段时间间隔内测量，以便消除波动。
- **窗口校正** A_{window} 不需要多次测定，通常安装后测定一次即可。其中，S_{fiber} 和 S_0 需要在段时间间隔内测量，以便消除波动。

什么时候需要窗口校正?

如果校正装置完全反应探针的光学属性，则 A_{window} 等于 0。在这种情况下，窗口校正可以跳转。

对于透射模式，通常要求进行窗口校正。对于反射模式则不需要窗口校正。但也有例外情况，详情参阅下表格：

测量模式	探针	参比标准化
透射	透射对	玻璃纤维 + 窗口
	透射探针	
	透射反射探针	
反射	反射探针	玻璃纤维
	带 MicroBundle 的透射反射探针	玻璃纤维 + 窗口

为了确定是否需要进行窗口校正，需要对每个校正装置和探针组合进行一一的分析。如果校正装置*不*能完全反应探针的光学属性，则需要进行窗口校正。

通道

2060 The NIR 类型的仪器提供多个通道。每个通道都可以关联其它光纤束和探针配置。因此，对于每个通道都必须分别执行参比标准化。

所有通道均使用相同的内部参考路径，例如相同的信号 S_{ref}。一个多路复用器在内部参考和不同测量通道之间来回切换。

执行玻璃纤维校正

首次投入运行后或当一个通道的光纤束配置变更时，必须执行玻璃纤维校正。更换灯或环境条件极端变化也建议重新进行标准化。

该步骤使用一个参考材料：

- 在反射模式中，参考材料是反射率标准样品。设定反射率标准样品为非理想状况（例如 99 %）。反射率标准样品具有已知的标称吸收光谱 $A_{nominal}$。
- 在透射模式下，空气用作参考。标称吸收光谱是谱带基线（$A_{nominal} = 0$），因为设定条件空气不吸收光。

图 11 展示了以下步骤：

1. 外部光纤束必须连接至校正装置。
   在反射模式下，校正装置与反射率标准样品结合。
2. 命令 **REF STD** 通过 **玻璃纤维** 接口执行以下扫描：
   a. 内部参考扫描提供一个 S_{ref} 数值。
   b. 外部扫描测量外部光纤束、校正装置和参考材料。从而给出信号 S_{raw}。
3. 软件计算 A_{raw}（**测得的原始光谱**）：
   $$A_{raw} = \log_{10} \frac{S_{ref}}{S_{raw}}$$
   其中，A_{raw} 为外部光纤束、校正装置和参考材料基于内部光路径的吸光度。
4. 参考材料的标称吸收光谱 $A_{nominal}$ 在软件中作为 **参考光谱** 显示。参考光谱必须从 A_{raw} 减去，从而得到 A_{fiber}：
   $$A_{fiber} = A_{raw} - A_{nominal}$$
   其中，A_{fiber} 相当于光纤束、校正装置基于内部光路径的吸光度。
   注意：在透射模式下，$A_{nominal} = 0$ 并且 $A_{fiber} = A_{raw}$。
   A_{fiber} 代表玻璃纤维 **修正光谱**。
5. A_{fiber} 保持不变，直至对相关通道执行了玻璃纤维校正。

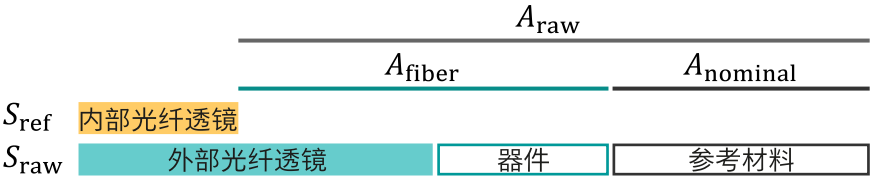


图 11　玻璃纤维校正

校验玻璃纤维校正

玻璃纤维校正必须通过相同的测量参数和相同的校正装置校验。

该步骤使用一个参考材料：

- 在反射模式中，参考材料是反射率标准样品。设定反射率标准样品为非理想状况（例如 99 %）。反射率标准样品具有已知的标称吸收光谱 $A_{nominal}$。

- 在透射模式下，空气用作参考。标称吸收光谱是谱带基线（$A_{nominal} = 0$），因为设定条件空气不吸收光。

图 12 展示了如何测定校验残差：

1. 外部光纤束必须连接至校正装置。
   在反射模式下，校正装置与反射率标准样品结合。
2. 命令 **VAL REF STD** 通过 **玻璃纤维** 接口执行以下扫描：
   a. 内部参考扫描提供一个 S_{ref} 数值。
   b. 相关通道的外部扫描测量外部光纤束、校正装置和参考材料。从而给出信号 S_{raw}。
3. 软件计算 A_{raw}（**测得的原始光谱**）：

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

4. A_{raw} 通过玻璃纤维修正光谱修正，以便消除光纤束和校正装置的吸光度：

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

   $A_{corrected}$ 在软件中作为 **测得的修正光谱** 显示。
5. 在理想状况下，$A_{corrected}$ 应与 **参考光谱** $A_{nominal}$ 一致。二者之间的差值作为 **校验残差** 计算得出：

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

   注意：在透射模式下，$A_{nominal} = 0$ 并且 $A_{residual} = A_{corrected}$。

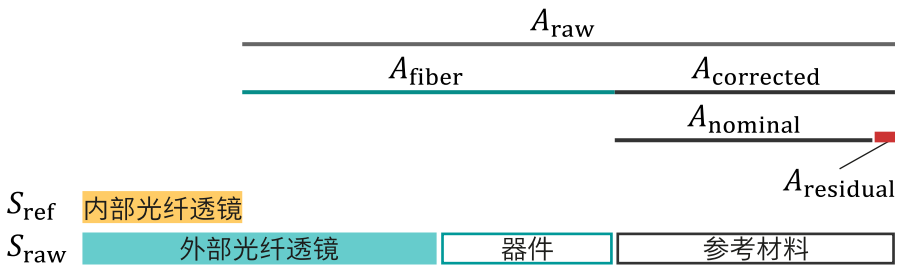


图 12　校验和玻璃纤维校正的残差

为了分析校验残差，波长范围被分为多个分段。对于每个分段，整个波长范围的残差平方平均值得出 **RMS 残差**（单位：mAU）：

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

其中，f 相当于分段中的波长数目，$A_{residual\ i}$ 相当于波长 i 的残差。

每个分段必须确保 A_{RMS} 的预定义允差。如果所有分段的允差都得以确保则代表整体校验成功。

校验必须在能够使用仪器在每个通道上记录光谱之前成功执行。

进行窗口校正

如果必须进行窗口校正，则首次投入运行后或每次更改探针或一个通道的光纤束配置时都要执行。探针如果发生变化（例如沾污）也建议重新标准化。

该步骤使用一个参考材料：

- 在反射模式中，参考材料是反射率标准样品。设定反射率标准样品为非理想状况（例如 99 %）。反射率标准样品具有已知的标称吸收光谱 $A_{nominal}$。
- 在透射模式下，空气用作参考。标称吸收光谱是谱带基线（$A_{nominal} = 0$），因为设定条件空气不吸收光。

图 13 展示了以下步骤：

1. 为了得到 A_{fiber} 的当前数值，应按上述方式执行玻璃纤维校正。
   重要：玻璃纤维校正必须在窗口校正前完成。
2. 外部光纤束必须在无当前样品的情况下连接至探针。必要时，让反射率标准样品占位样品。
3. 命令 **REF STD** 通过 **窗口** 接口执行以下扫描：
   a. 内部参考扫描提供一个 S_{ref} 数值。
   b. 相关通道的外部扫描测量外部光纤束、探针和参考材料。从而得到信号 S_{probe}。
4. 基于内部光路径的吸光度 A_{probe} 为：

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

5. 软件计算 A_{raw}（**测得的原始光谱**）：

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

   其中，A_{raw} 为探针和参考材料基于校正装置的吸光度。
6. 参考材料的标称吸收光谱 $A_{nominal}$ 在软件中作为 **参考光谱** 显示。参考光谱必须从 A_{raw} 中减去，从而得到 A_{window}：

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

   其中，A_{window} 为探针基于校正装置的吸光度。
   注意：在透射模式下，$A_{nominal} = 0$ 并且 $A_{window} = A_{raw}$。
   A_{window} 代表窗口 **修正光谱**。
7. A_{window} 保持不变，直至再次执行相关通道的窗口校正。

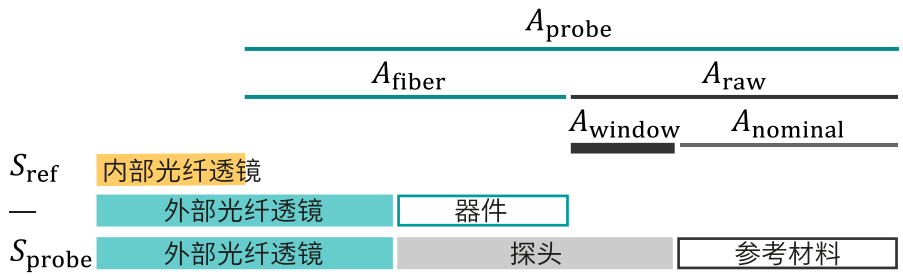


图 13　窗口校正

窗口校正的校验

窗口校正必须通过相同的测量参数和相同的校正装置校验。

该步骤使用一个参考材料：

- 在反射模式中，参考材料是反射率标准样品。设定反射率标准样品为非理想状况（例如 99 %）。反射率标准样品具有已知的标称吸收光谱 $A_{nominal}$。
- 在透射模式下，空气用作参考。标称吸收光谱是谱带基线（$A_{nominal} = 0$），因为设定条件空气不吸收光。

图 14 展示了如何测定校验残差：

1. 内部光纤束必须在无当前样品的情况下连接至探针。必要时，让反射率标准样品占位样品。
2. 命令 **VAL REF STD** 通过 **窗口** 接口执行以下扫描：
   a. 内部参考扫描提供一个 S_{ref} 数值。
   b. 相关通道的外部扫描测量外部光纤束、探针和参考材料。从而得到信号 S_{probe}。
3. 基于内部光路径的吸光度 A_{probe} 为：

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

4. 软件计算 A_{raw}（**测得的原始光谱**）：

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

5. 通过从 A_{raw} 减去窗口校正光谱消除校正装置和探针之间的吸光度差值：

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

   $A_{corrected}$ 在软件中作为 **测得的修正光谱** 显示。
6. 在理想状况下，$A_{corrected}$ 应与 **参考光谱** $A_{nominal}$ 一致。二者之间的差值作为 **校验残差** 计算得出：

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

   注意：在透射模式下，$A_{nominal} = 0$ 并且 $A_{residual} = A_{corrected}$。

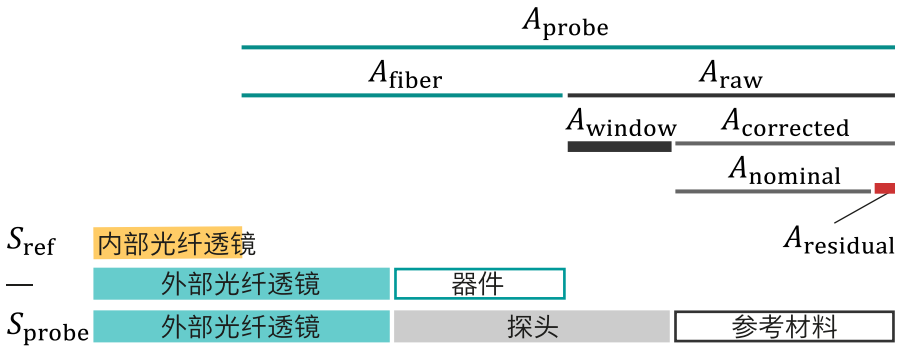


图 14　窗口校正的校验残差

为了分析校验残差，波长范围被分为多个分段。对于每个分段，波长范围的残差平方平均值得出 **RMS 干扰**（单位：mAU）：

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

其中，f 相当于分段中的波长数目，$A_{\mathrm{residual}\,i}$ 相当于波长 i 的残差。

每个分段必须确保 A_{RMS} 的预定义允差。如果所有分段的允差都得以确保则代表整体校验成功。

记录一个样品的光谱

在能够使用仪器记录光谱之前必须在每个通道上成功执行 仪器性能测试 *（参见章节 3.3，第 21 页）*。

记录样品光谱的步骤参见 *图 15*：

1. 外部光纤束必须连接至探针。必须存在样品。
2. 吸收光谱通过最近一次记录的参考光谱 S_{ref} 计算得出。为了得到 S_{ref} 的当前数值，可以执行 **MEAS REF SPEC** 命令。
3. 命令 **MEAS SPEC** 测量样品包括探针和光纤束。执行该命令将得到信号 S。
4. 软件计算 A_{total}、样品、包括探针和光纤束基于内部光路径的吸光度：

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

5. 然后，样品的吸光度如上所述在使用相关通道的玻璃纤维校正光谱 A_{fiber} 和窗口校正光谱 A_{window} 的情况下计算得出：

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

   A 代表样品的光谱。

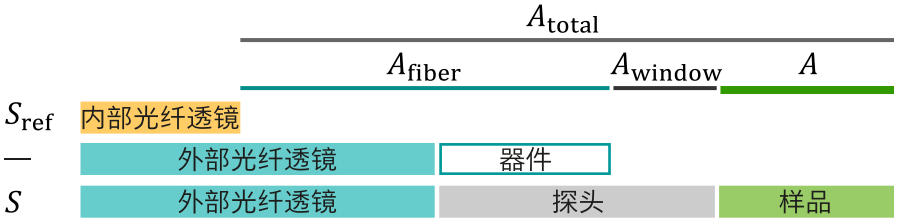


图 15　记录 1 个样品的光谱

3.3 仪器性能测试

仪器性能测试 可通过内部和外部光路径进行。

- **OMNIS NIR Analyzer**
  - **内部 仪器性能测试**（强制性）：内部测试使用相应样品展示的参考路径。这些测试检查波长和信号干扰。
    测试前必须成功执行和校验相关样品展示的波长校正。
    在能够使用仪器记录相关样品展示的光谱之前必须成功执行内部测试。
  - **外部 仪器性能测试**（可选）：外部测试支持根据例如 USP <856>、Ph.Eur 2.2.40 和 JP 2.27 等药典进行校验。需要检查波长、信号干扰和光度线性 *（参见章节3.3.1，第23页）*。
- **2060 The NIR**
  测试可以使用内部光路径或外部光路径。这些测试检查波长和信号干扰。
  测试前必须成功执行和校验相关通道的波长校正和外部参比标准化。
  仪器性能测试 必须在能够使用仪器记录相关通道的光谱之前成功执行。必须遵守预定义的允差。允许的允差取决于仪器技术数据中注明的相关通道的光纤束配置（测量模式、纤维类型和纤维长度）。

波长测试

波长测试将对波长准确度和波长精度进行分析。为此，要使用 1 个波长标准版，它具有已定义波峰和已知最高位置的吸收光谱：

- **内部**：内部可追踪的计量学波长标准版的吸收光谱可通过内部光路径进行测定：

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

  其中，A_{WL} 对应内部波长标准版的吸光度，S_{ref} 对应内部参考路径上测得的信号，$S_{\mathrm{ref,WL}}$ 则对应用内部波长标准版在内部参考路径上测得的信号。
- **外部**用于 **OMNIS NIR Analyzer** 产品系列的仪器：外部波长测试是可选的 *（参见"外部波长测试"，第23 页）*。

- **外部**用于 **2060 The NIR** 类型的仪器：
  对于单纤维，外部光纤束必须连接至校正装置，MicroBundle 则必须连接至反射率标准样品。
  内部可追踪的计量学波长标准版的吸收光谱可通过外部光路径进行测定：

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

  其中，A_{WL} 对应内部波长标准版的吸光度，S_{ref} 对应内部参考路径上测得的信号，$S_{\mathrm{fiber,WL}}$ 则对应相关通道光路径上测得的信号，其中纤维与校正装置及反射率标准样品连接并且内部反射率标准样品装入光路径中，A_{fiber} 对应参比标准化的玻璃纤维校正光谱。

A_{WL} 还包含对于以下计算无关紧要的反射率标准样品吸光度。在理想情况下，A_{WL} 的最高位置与波长标准版的标称最高位置相同。

如下检查波长准确度和波长精度：

1. 如上所述记录一系列的波长标准版吸收光谱（A_{WL}）。
2. 记录的光谱中将对最高位置进行识别。
3. 对于每个最高位置将计算出记录光谱的以下统计数据：
   a. 平均值（单位：nm）
   b. 标准偏差（单位：nm）
4. **精确度**：对于每个波峰而言，最高位置平均值和最高位置标称值之间的差值必须位于预定义的允差范围内。
   ➡ **注意**：每次测试的标称最高位置可能会略有偏差。原因是最高位置是经过温度修正的。相似的温度校正在波长校正中进行。该校正能够帮助在特定温度下对所有仪器的测量都能得到相近的结果。
5. **精度**：对于每个波峰而言，标准偏差必须位于预定义允差范围内。
6. 如果所有波峰的允差都得以确保，则波长测试的整体情况合格。

干扰测试

信号干扰可在内部或外部检查：

- **内部**：干扰作为内部光路径的吸光度测定，基于同一光路径上的其它测量的吸光度：

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

  其中，$S_{\mathrm{ref,1}}$ 和 $S_{\mathrm{ref,2}}$ 对应内部参考路径上测得的信号。
- **外部**用于 **OMNIS NIR Analyzer** 产品系列的仪器：外部干扰测试是可选的 *（参见"外部干扰测试"，第 24 页）*。

- **外部**用于 **2060 The NIR** 类型的仪器：
  对于单纤维，外部光纤束必须连接至校正装置，MicroBundle 则必须连接至反射率标准样品。
  以测量吸收光谱和标称吸收光谱之间差值的形式测定干扰：

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

  其中，S_{ref} 对应内部参考路径上测得的信号，S_{fiber} 对应外部光路径上测得信号，其中纤维与校正装置和反射率标准样品连接，A_{fiber} 对应参比标准化的玻璃纤维校正光谱，A_{nominal} 则对应反射率标准样品的标称吸收光谱。
  注意：在透射模式下，$A_{\text{nominal}} = 0$。

在理想情况下，$A_{\text{noise}} = 0$。

干扰测试执行以下步骤：

1. 如上所述记录一系列的干扰光谱（A_{noise}）。
2. 干扰光谱分为不同的波长分段。
3. 为每个干扰光谱和每个分段计算 3 个变量：
   a. 光度噪声干扰（单位：mAU）
   b. 峰对峰干扰（单位：mAU）
   c. 干扰的基线偏差（单位：mAU）
4. 对于每个分段中的 3 个变量中的每一个，都要计算所记录干扰光谱的平均值。
5. 如果所有平均值均位于预定义的允差范围内，则干扰测试的整体情况合格。

3.3.1 外部 仪器性能测试（OMNIS NIR Analyzer）

OMNIS NIR Analyzer 产品系列的仪器可根据例如 USP <856>、Ph.Eur 2.2.40 和 JP 2.27 等药典进行验证（从 OMNIS Software 版本 4.4 起）。这些测试需要外部可追踪的计量学参考标准样品。参考标准样品具有单独的、在环境温度下用参考仪器测得的标称吸收光谱。

外部波长测试

如下检查波长准确度和波长精度：

1. 必须将外部波长标准版（透射或反射）移入正确位置。
2. 记录一系列的外部波长标准版吸收光谱：

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

   其中，A_{WL} 对应外部波长标准版的吸光度，S_{ref} 对应内部参考路径上测得的信号，S_{WL} 对应通过外部波长标准版测得的信号。
3. 记录的光谱中将对最高位置进行识别。
4. 对于每个最高位置将计算出记录光谱的以下统计数据：
   a. 平均值（单位：nm）
   b. 标准偏差（单位：nm）

5. **精确度**：对于每个波峰而言，最高位置平均值和最高位置标称值之间的差值必须位于预定义的允差范围内。
6. **精度**：对于每个波峰而言，标准偏差必须位于预定义允差范围内。
7. 如果所有波峰的允差都得以确保，则波长测试的整体情况合格。

外部干扰测试

信号干扰要在低光子流（低通量）和高光子流（高通量）条件下各测试一次：

1. 必须将低通量测试或高通量测试的外部参考标准样品（透射或反射）移入正确位置。
2. 以测量吸收光谱和参考标准样品标称吸收光谱之间差值的形式记录一系列的干扰光谱：

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

   其中，S_{ref} 对应在内部参考路径上测得的信号，S_{ND} 对应通过外部参考标准样品测得的信号，A_{nominal} 对应参考标准样品的标称光谱。
3. 干扰光谱分为不同的波长分段。
4. 为每个干扰光谱和每个分段计算 3 个变量：
   a. 光度噪声干扰（单位：mAU）
   b. 峰对峰干扰（单位：mAU）
   c. 干扰的基线偏差（单位：mAU）
5. 对于每个分段中的 3 个变量中的每一个，都要计算所记录干扰光谱的平均值。
6. 如果所有平均值均位于预定义的允差范围内，则干扰测试的整体情况合格。

光度线性

该测试的目的是在整个波长范围内证明反射率（或透射率）与测得的吸光度之间存在线性关系：

1. 记录 5 个具有不同反射率（或透射率）的参考标准样品的吸收光谱。
2. 反射率（或透射率）与测得吸光度之间的线性关系通过多个波长下的线性回归确保。
3. 如果所有回归线的斜率和 y-轴截距都在预定义的允差之内，则测试的整体情况合格。

4 开发模型

分为以下类型的方法段：

- **量化模型** 说明了分析参数（例如 水份含量）与记录的样品光谱的依赖性。
- **识别模型**（OMNIS Software 版本 4.0 以上）根据记录的光谱将样品分类为不同产品（例如 不同的咖啡豆品种）。
  一个产品代表一个特定的化学物质或一个具有特定物理属性（例如 颗粒大小）的特定的化学物质。

对于一个未知样品的分析将记录样品的光谱。根据应用的不同按照以下方法使用光谱：

- 量化：量化模型基于光谱生成一个预测，例如针对样品的水份含量。
- 身份验证：识别模型基于光谱识别样品，例如作为阿拉比卡咖啡。
- 校验（OMNIS Software 版本 4.2 以上）：例如，识别模型根据光谱校验样品是否为阿拉比卡咖啡。

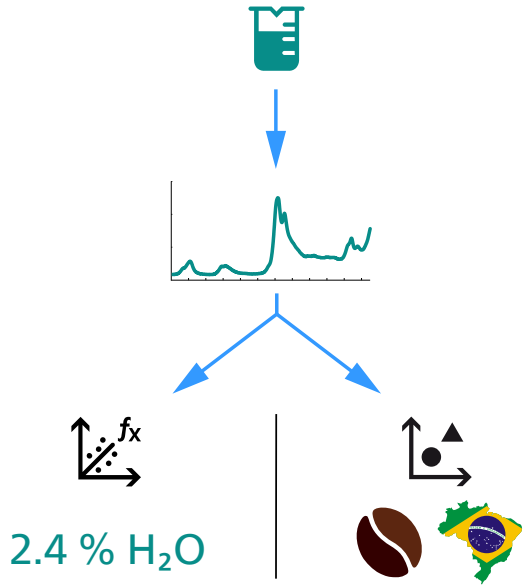


图 16　量化（左下）和身份验证（右下）。

开发模型和分析样品包含以下步骤：

1. **采样**
   采样和处理物理样品：
   a. 对于每件样品都将记录光谱。
   b. 对于量化将通过参考方法段（例如 滴定）测量用于分析参数的参考值（例如 水份含量）。参考测量必须准确和精确。
   c. 对于身份验证必须已知样品的产品关联性。
2. **开发模型**
   开发模型采用迭代方法，包含以下步骤：
   a. 将数据组划分成校准数据集、校验数据集和离群数据集。
   b. 对光谱应用合适的数据预处理和波长范围。
   c. 基于校准数据集计算模型。
   d. 模型的校验能够确保模型满足要求。校验首先基于开发模型时未使用的校验数据集。
      模型在量化时预测校验数据集中的光谱的分析参数。然后，将计算值与已知参考值进行比较。
      对于身份验证，模型将光谱归类为不同的产品。预测的产品关联性与相关的实际产品关联性比较。
3. **监控**
   模型的监控确保了预测力不随着时间降低。如果过程或样品发生变化，则需要再校验。

4.1 样品

首先，对物理样品进行采集和分析。正确的采样是开发可靠模型的前提条件。其中，需要注意多个方面。

变异宽度

样品应考虑典型、未来预计的样品变异。所有化学组分的浓度和颗粒大小应至少覆盖预期的变异宽度。

样品应覆盖合适的条件变异和合适的时间间隔。应考虑所有变异，例如过程波动、季节性波动或环境条件波动。

样品应在整个变异宽度均匀分布。对于量化，变异宽度还包括参考值的范围。如果参考值的范围在 1% 到 10% 之间，则样品应在 1% 和 10% 之间平均分布。

校正和校验样品

通常使用 2 个样品组：

- 校正组：用于模型开发的样品。
- 校验组：用于校验模型的样品。

两组都必须覆盖所期待的变异。对于校验组应首选采集不相关的样品，以便检验模型的耐用性。应考虑例如不同操作人员、不同供应商或不同仪器等场景。

如果对于校正和校验采集不相关样品不可行，也可以将可用样品的光谱划分为校准数据集和校验数据集。为了确保样品组的不相关性，应采用自动划分算法。

一旦模型已开发，则可以考虑使用明确的离群值样品，以便检验离群值识别性能。

所有样品应以相同方式处理。对于光谱记录应采用具有相同应将配置和相同测量参数的相同方法。对于量化应采用具有相同测量参数的相同参考方法。

样品数目

需要考虑的条件、化学组分或颗粒大小变异越大，越需要更多的样品。

量化：为了确保统计分析正常工作，至少需要 50 件样品，其中校正组和校验组分别至少必须包含 20 到 25 件样品。

身份验证和校验：对于每件产品，样品都必须覆盖所期待的变异。产品的样品数目可能不同，最小数量为 3 件。

通过至少 10 到 20 件样品（根据变异的数量而定）可以开发第一个无校验组的模型。如果交叉验证（对于量化模型）或内部校验（对于识别模型）表明可以创建一个合适的模型，必须采集更多的校正和校验样品以开发最终的模型。

复制品

有时特定条件或参考值范围内只有非常少的样品。为了加以补偿，可以尝试复制样品。但会存在问题。如果校正组和校验组中都存在一个样品的复制品，则模型品质指标可能会引起误导（过于理想）。因此须避免相同的样品组中存在复制品。

参考方法段（量化）

量化采用参考方法段测量参考值。所采用的参考方法段的**实验室的标准误差（SEL）** 在量化模型的开发中的作用非常重要。SEL 是复制样品测量之间差值的标准偏差。

SEL 通常是造成 NIR 方法段产生预测标准误差（SEP）的最大错误 *（参见"SEP – 预测标准误差"，第 55 页）*。SEL 应不超过所要求 SEP 的 0.7 倍，首选不超过 0.5 被。参考值的范围应至少为 SEL 的 3 倍，最好为 5 倍。

对每件样品重复执行参考测量可以降低 SEL。测量值的平均值应作为样品的参考值确定。对于每件样品应执行一定数量的参考测量。模型品质指标将相对根据特定数量的重复参考测量表示。不同数量的重复参考测量将会导致错误的模型品质指标预测值，因此需要避免。

样品温度

样品温度对包含水份或其它氢键的液体光谱起到重大影响。其它极性液体的光谱可能如同包含固体物质、水份、湿度或溶剂的光谱受到同样的影响。此类样品应在设定的温度下测量。

离群值

有些样品必要时之后作为离群值识别。离群值是处于某种原因与大部分样品不同的样品。为了防止对模型的负面影响，识别为离群值的样品不纳入模型的计算中。

离群值的类型包括：

- **光谱离群值**
  如果采集的样品光谱与大部分其它光谱不同，则该样品被识别为光谱离群值 *（参见章节 4.3.3，第 40 页）*。
- **离群值参考值（量化）**
  量化时，单个参考值可能包含异常并在之后被识别为离群值参考值 *（参见章节 4.3.4，第 45 页）*。
  OMNIS Model Developer (OMD) 基于根据 ASTM D8321-22 的离群值识别方法识别相应的样品作为离群值 *（参见章节 4.4.3，第 55 页）*。

4.2 主要成分分析（PCA）

校正样品的光谱数据包含很多变量（波长）。变量之间相互关联程度很大。因此，数据过多。为了能够绕过此类数据，需要使用诸如 PCA 和 PLS 这样的潜变量模型。

主要成分分析（PCA，英文：*principal component analysis*）聚焦于光谱，而不考虑参考值。

在开发模型过程中，OMNIS Software 将 PCA 用于数据组的自动划分和光谱离群值的识别。

准备步骤

需要以下准备步骤：

1. **数据预处理**：OMNIS Software 在光谱上应用所选的数据预处理 *（参见章节 4.3.1，第 32 页）*。
2. **波长范围**：OMNIS Software 在光谱上应用所确定的波长选项 *（参见章节 4.3.2，第 38 页）*。
3. **平均值集中**：对于每个波长计算平均吸光度值并从每个光谱相关的数值中减去。

第一主要成分

准备步骤后，PCA 将光谱数据中的信息重新整理并将重要的数据排除干扰。为此，PCA 将波长变量转换为新的变量空间，即所谓的 **主要成分（PC**，英文：Principal Components）。

PCA 将大量波长变量中的重要信息转变为少数的主要成分。为了能够用简单的例子描述该方案，设定只有 2 个波长变量而非上千个，并且这 2 个变量简化为 1 个主要成分。

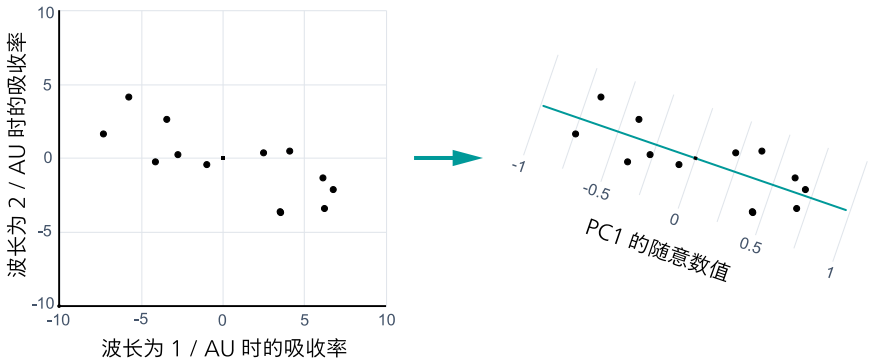


图 17　在 2 维波长空间内表示光谱的点位（左）。在 1 维主要成分空间内的相同点位（右）。

左 *图 17* 中，水平和垂直轴构成了包含 2 个变量的初始波长空间。这样以来，每个点位均表示一个仅包含 2 个波长的光谱。所有波长值的平均值构成谱带基点。

右侧为穿过数据的方向，说明了最大方差，即主要成分 PC1。在该例子中，PC1 是主要成分空间中唯一的变量。因此，2 个初始变量将简化为 1。

得分和残差

图 18 显示了确定光谱 *i* 属性的规模：

- 由主要成分空间中测得的中间的间距 s_i。仅 1 个主要成分的示例中，s_i 在 PC1 方向中测得。间距 s_i 被称为光谱 *i* 的 **得分**。
- 从主要成分空间到光谱的偏置值 e_i。间距 e_i 被称作光谱 *i* 的 **残差**。

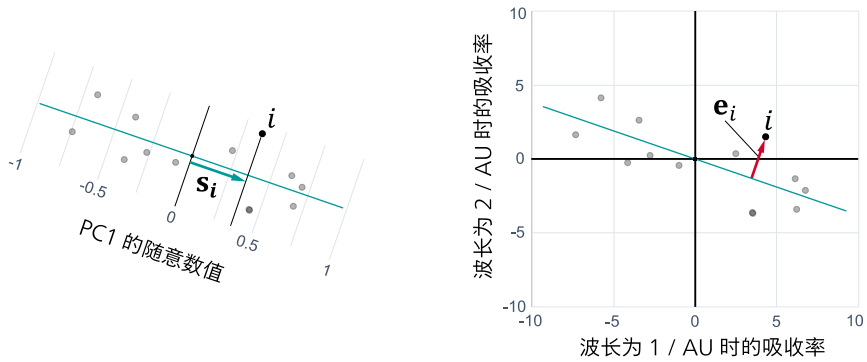


图 18　带得分（左）和残差（右）的光谱 i。

得分 s_i 在主要成分空间中测得。残差 e_i 将在初始波长空间内测得。

转变为多个主要成分

通常对于光谱数据的合适说明需要一个以上的主要成分。

图 19 中有 3 个初始变量 x_1、x_2、x_3。每个点位代表一个包含 3 个波长的光谱。

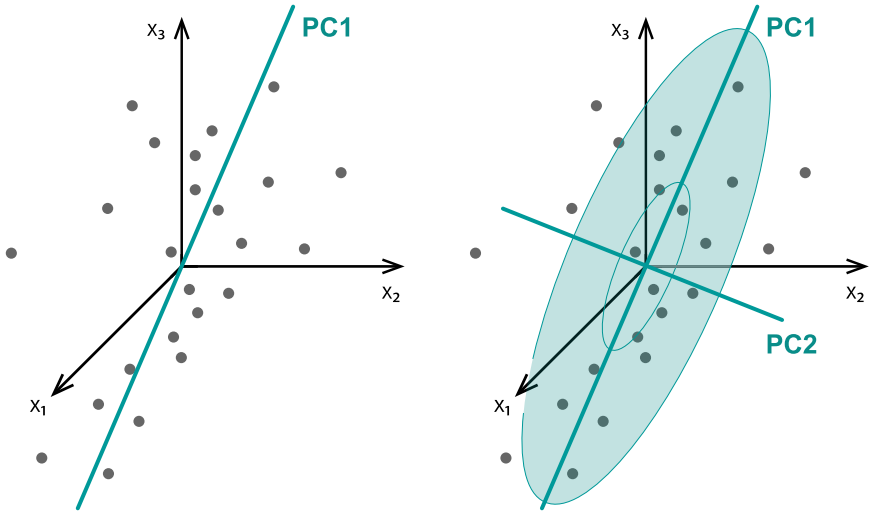


图 19　3 个初始变量被简化为 1 个主要成分（左）或 2 个主要成分（右）。PC1 和 PC2 构成了一个 2 维主要成分空间。

第一个主要成分 PC1 是穿过数据的说明了最大方差的方向。

第二个主要成分 PC2 是穿过数据的说明了最大剩余方差的方向。这一点对于说明最大剩余方差的以下主要成分同样适用。因此，第一对主要成分占了数据中最大的方差部分，而其它数据主要包含干扰并且可以弃用。通过这种方式可以简化变量的数量。

PCA 的主要特征是所有主要成分相互 **正交**（右侧角度）。因此，得分不关联。

马氏距离

如上所示，光谱 *i* 的得分在主要成分空间中测得，而允差在初始波长空间中测得。

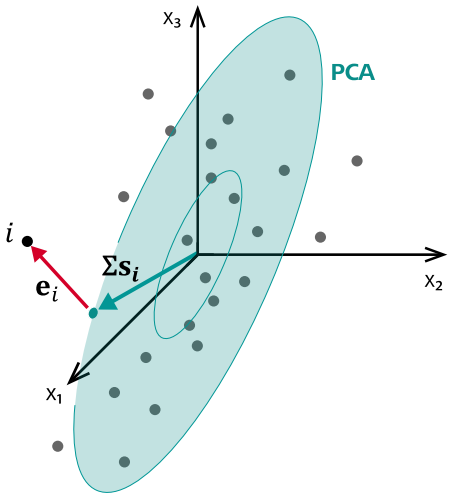


图 20　光谱的得分和允差。绿色点位为主要成分空间上点位 i 的正交投影（表示光谱 i）。

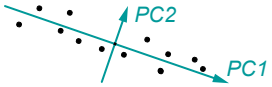


图 20 中，得分向量 Σs_i 表示从 PCA 模型中点到光谱在主要成分空间上正交投影的绝对距离（欧式距离）。

本例中，PC1 方向的光谱的欧氏距离比 PC2 方向上的距离更大。分布范围可以作为 **方差** 测量。PC1 内的方差大于 PC2 内的方差。

正常化的得分矢量 s_i 代表正常化的距离、即所谓的 **马氏距离**。马氏距离考虑了不同主要成分方向内的不同方差。每个方向均得到了相同的加权。因此，低方差方向的小欧氏距离与高方差方向的大欧氏距离具有同样的意义。

转换多个波长的光谱

相同方案适用于将具有大量波长变量的光谱转换为主要成分。在 *图 21* 中通过一条曲线（左）和一个点位（右）显示了每个光谱。

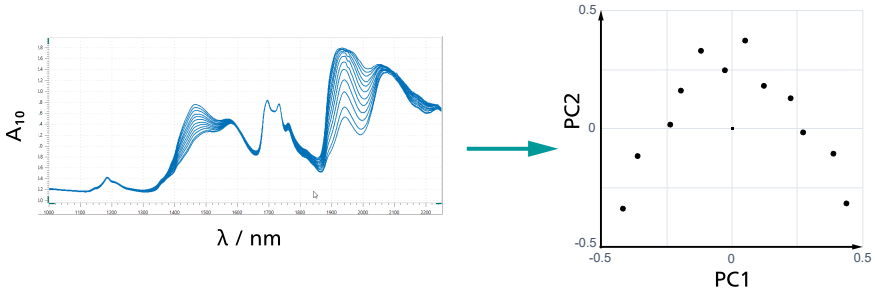


图 21　将光谱数据转换为一个主要成分空间。右侧的得分以任意的单位表现。

右图展示了前 2 个主要成分 PC1 和 PC2。同样方式可以展示主要成分 PC3、PC4 等。

PCA 模型采用固定数量的主要成分。主要成分越多，说明模型的光谱变异越多。但同时，模型采集更多的非关联光谱变异（干扰）。需要一个均衡的妥协。

如果 OMNIS Software 执行主要成分分析，则主要成分的数量选择须考虑到所设定的方差至少应为 95 %。

PCA 算法

原始数据转换为主要成分空间有多种方式。OMNIS Software 执行一个奇异值分解 *（参见章节 6.2，第 72 页）*。

4.3 数据预处理

4.3.1 数据预处理

光谱模型基于吸光度值与感兴趣的参数（量化）或产品关联性（身份验证、校验）之间的关系。光谱的**参数化**确保了光谱能够充分表达该关系。目标是在不朽是有用信息的情况下消除非关联方差。伪迹和非线性将被修正。正确执行参数化设置能够提高模型的精确度和耐用性以及预测的重复性和重现性。

参数化设置将被应用于校准数据集、校验数据集和离群数据集以及所有今后使用相同模型分析的未知样品。

参数化设置的第一个步骤是 **数据预处理**。数据预处理以设定顺序进行。参数化设置的第二个步骤可以确定相关的波长范围 *（参见章节 4.3.2，第 38 页）*。

干扰消除

光谱可能包含信号周围不同形式的随机波动。例如因检测器和仪器的电路引起的高频干扰或扫描测量期间仪器漂移引起的低频干扰。

光谱仪提供由一系列单次测量确定的光谱。因此，高频干扰将大幅减少。进一步的干扰消除可以通过一个平滑滤波器实现。该滤波器基于干扰为高频、信号为低频的想法。滤波器通过相邻的吸光度值对信号进行近似分解并通过构建平均值减少干扰。

散射修正

散射是因样品与光交替作用产生的方向变化。未到达检测器的散射光将会导致光谱内的基线波动。

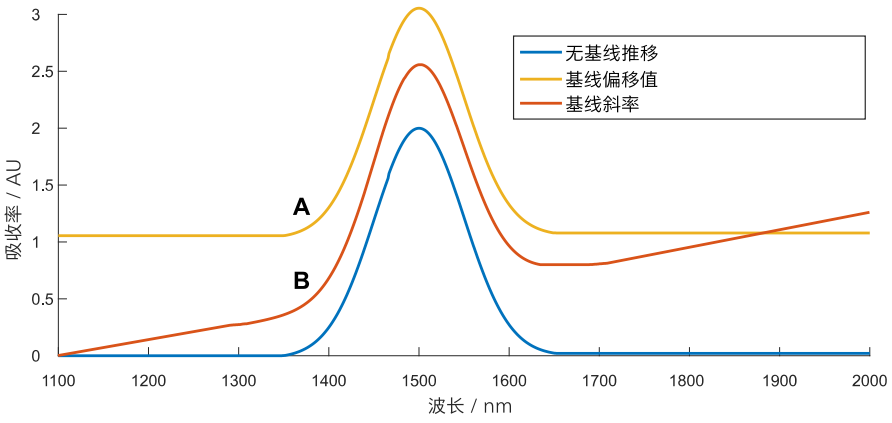


图 22　基线推移的最重要形式是基线的偏置和基线的倾斜。

可以区分不同形式的基线推移：

- 一个导致 **基线偏置** 的恒定加法因子（光谱 **A**）。
- 一个导致 **基线倾斜** 的与波长相关的乘法因子（光谱 **B**）。
- 一个导致 **基线平方倾斜** 的与波长相关的二阶乘法因子。
  也可能出现更高阶数的基线倾斜。

- 导致 **数量增大** 的吸光度相关的乘法因子。但数量增大无关紧要。

散射对于固体样品而言是最明显的。由此得到的基线推移可以用于识别颗粒大小的变化或其它物理变异。

但如果关乎化学变异，则应通过合适的预处理将基线推移降低到最小程度。

数据预处理

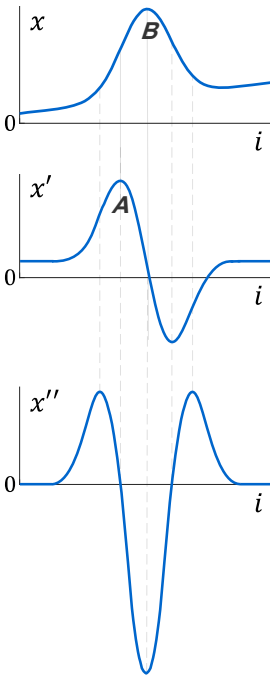
所有以下数据预处理均相应应用于单个光谱。计算不纳入更多光谱。

数据预处理不改变信号值。y 轴上的数字不再有意义。

所应用的数据预处理为线性变换。因此，Beer-Lambert 定律依旧有效。

4.3.1.1 导数

导数可以通过 Gap-Segment 滤波器或 Savitzky-Golay 滤波器求得。



光谱的导数描述了每个点位的曲线斜率或倾斜度。斜率是初始光谱的变化率。

光谱中，x_i 是波长为 *i* 时的吸光度。一阶导数 x_i' 表示波长 *i* 的光谱斜率。初始光谱最大斜率的未知，一阶导数的值最大（**A**）。初始光谱的波峰（**B**）位置，一阶导数等于 0。

一阶导数能够消除基线的偏置并将基线倾斜转换为基线偏置。

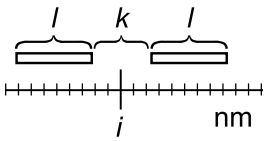
二阶导数 x_i'' 为波长 *i* 的一阶导数的斜率。原始光谱的正波峰（**B**）将变为负波峰，而负波峰则变为正波峰。

二阶导数能够消除基线偏置以及原始光谱的基线倾斜。

对于原始光谱存在大量干扰的情况须小心谨慎。每次求导数都会对信噪比产生严重的负面影响。出于此原因，求导数还须结合 Gap-Segment 滤波器或 Savitzky-Golay 滤波器中的平滑功能。

4.3.1.2 Gap-Segment

Gap-Segment 滤波器能够平滑光谱。可以选择 Gap-Segment 滤波器执行求一阶或二阶导数。计算取决于是否使用求导数功能：



- **导数阶次 0**：Gap-Segment 滤波器对每个波长 *i* 计算由分段尺寸为 *l*（例如 10 nm）的 2 个分段计算平均值。2 个分段以分段尺寸 *k*（例如 5 nm）的距离间隔。

- **导数阶次 1**：对一阶导数分别计算 2 个分段的平均值。从而构成了两个平均值的差值。
- **导数阶次 2**：二阶导数可以按一阶导数的相同方式计算。

在光谱的前端和后端处计算 *l* + *k*/2 波长，其中对光谱以外的分段波长使用零值。

光谱的前端和后端处对光谱以外的分段波长使用零值。

使用平滑功能可能会造成给波峰略微偏移并产生一定程度的扭曲。

参数设定

更强的平滑通过以下方式实现：

- 降低导数阶次，
- 加大分段尺寸，
- 加大分段间距。

过度平滑会造成重要的方差损失，从而减弱模型的预测力。

4.3.1.3 Savitzky-Golay

如果 Gap-Segment 滤波器一样，Savitzky-Golay 滤波器平滑光谱并选择执行一阶或二阶导数的方法。Savitzky-Golay 滤波器使用过另一种平滑方法。

Savitzky-Golay 滤波器对每个波长 *i* 在相关波长范围内拟合一个低阶多项式。波长为 *i* 的多项式数值为平滑数值。如果求导数，则使用求导数值。

相邻数值的加权求和可以一次性计算出所有参数：

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

其中，*k* 为滤波器宽度，c_j 为卷积系数（取决于导数阶次、多项函数阶数和滤波器宽度并可在表格中查找），x_{i+j} 为波长 *i*+*j* 的初始光谱吸光度值。

在光谱的前端和后端对处于光谱以外的滤波器波长使用外插数值（水平外推法）。

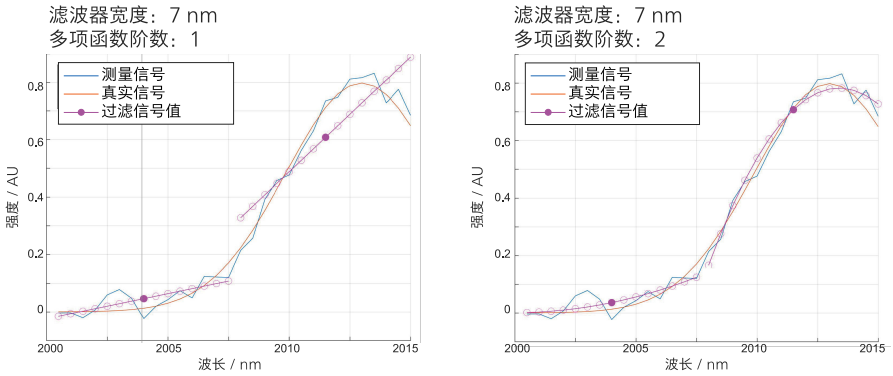


图 23　使用不同多项函数阶数的 Savitzky-Golay 滤波

图 23 中显示 Savitzky-Golay 滤波。滤波器宽度为 7 nm。显示的是波长 2,004 nm 和 2,011.5 nm 的多项式。新数值为填充的点位或在使用导数时为其求导数值。其它所有波长将以相同方式处理。

滤波器宽度定义每个多项式中插入的波长范围。卷积加权须确保吸光度值的影响根据相关波长的两侧而减弱。

参数设定

更强的平滑通过以下方式实现：

- 降低导数阶次，
- 加大滤波器宽度，
- 降低多项函数阶数。

过度平滑会造成重要的方差损失，从而减弱模型的预测力。

4.3.1.4　SNV - 标准正态变量

SNV 将单一光谱归一化为方差 1，将平均值标准化为 0。SNV 将按照以下方式在确定波长范围内对每个波长 *i* 的吸光度值 x_i 进行归一化：

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

其中，*m* 为平均值，*s* 为所确定波长范围内所有吸光度值的标准偏差。

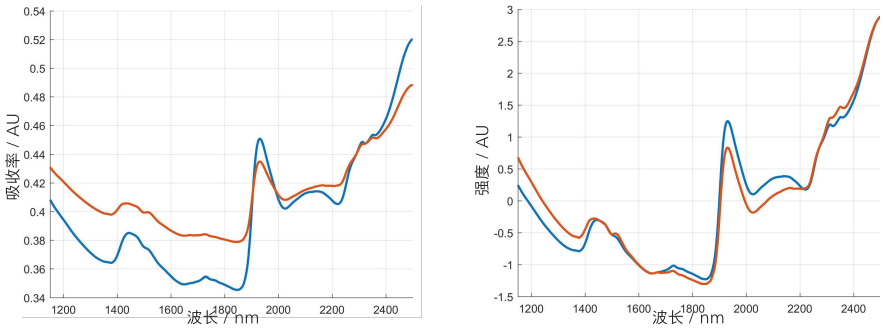


图 24　吸收光谱（左）和经过 SNV 处理的光谱（右）。

通过归一化消除光谱之间的方差。这对于由不感兴趣的属性（例如颗粒型或粉末状样品或浑浊介质中不同的层厚度）得出的方差很有意义。

注意：如果在 SNV 后求导数，则已消除的方差可能会部分再次出现。因此，必须在 SNV 之前求导数。如果在例外情况下必须在 SNV 之前求导数，可以考虑以下顺序：detrend、DNV、导数。通常的顺序为：导数、SNV、detrend *（参见章节 4.3.1.7，第 38 页）*。

参数

- **波长范围**
  如果伪迹对某些波长范围产生不良影响（例如 由于达到饱和度或强噪音），则可以排除这些范围。
  在计算平均值和标准偏差时，仅考虑定义的波长范围。随后的归一化将在定义的波长范围及中间区间内进行。对于光谱起始处被排除的波长，将采用相邻起始波长的标准化数值。对于光谱结束处被排除的波长，将采用相邻结束波长的标准化数值。
  如有需要，也可在计算模型时排除已排除的波长 *（参见章节 4.3.2，第 38 页）*。
  注意：从 OMNIS Software 版本 4.6 起，可以定义多个波长范围。

4.3.1.5 Detrend

Detrend 根据最小平方的方法对光谱拟合一个二阶多项式。然后，detrend 从光谱减去多项式。

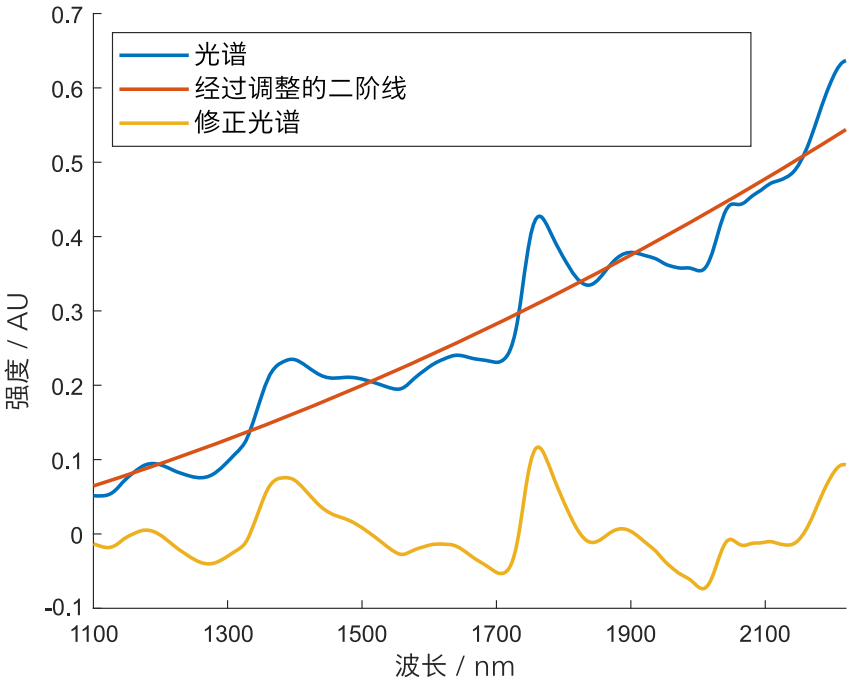


图 25　detrend 将蓝色光谱转换为黄色光谱。

detrend 减弱波长相关的散射效应直至基线的平方倾斜。

上图显示了趋势中占统治地位的光谱（蓝色）。如果该占统治地位的趋势对所有光谱而言相近，则表明 detrend 功能正常。在其它情况下，detrend 倾向于删除有用的变异。在某些情况下，求导数可能时更好的选择。

因为对每个光谱均拟合了各自的多项式，则可能会出现干扰性的方差。通常，SNV 在 detrend 之前应用。这样会得到更可靠的多项式系数的预测值。

参数

- **波长范围**
  如果伪迹对某些波长范围产生不良影响（例如 由于达到饱和度或强噪音），则可以排除这些范围。
  按所有定义波长范围的强度值调整多项式。随后，在所有定义的波长范围内，从光谱中减去该多项式。所有被排除波长的强度值将被置零。
  如有需要，也可在计算模型时排除已排除的波长 *（参见章节 4.3.2，第 38 页）*。
  注意：从 OMNIS Software 版本 4.6 起，可以定义多个波长范围。

4.3.1.6 数据预处理一览

预处理	目的	正效应	负效应
Gap-Segment	平滑 更强的平滑通过一个更低的导数阶次、更大的分段尺寸或更大的分段距离达到。	▪ 减弱高频干扰。	▪ 过度平滑会导致重要的方差丢失。
通过 Gap-Segment 求导数	基线修正	▪ 一阶导数：消除基线偏置。 ▪ 二阶导数：消除基线偏置和基线倾斜。	▪ 加强干扰。 ▪ 改变光谱外观。
Savitzky-Golay	平滑 更强的平滑通过降低导数阶次、加大录波器宽度或降低多项函数阶数达到。	▪ 减弱高频干扰。	▪ 过度平滑会导致重要的方差丢失。
Savitzky-Golay 中求导数	基线修正	▪ 一阶导数：消除基线偏置。 ▪ 二阶导数：消除基线偏置和基线倾斜。	▪ 加强干扰。 ▪ 改变光谱外观。
SNV – 标准正态变量	散射修正 *	▪ 消除基线偏置。	▪ 在必要时删除重要的方差。

预处理	目的	正效应	负效应
Detrend	散射修正 *	▪ 消除基线偏置。 ▪ 消除基线的倾斜和基线的平方倾斜。	▪ 在必要时删除重要的方差。 ▪ 可能会出现无关紧要的方差。

* 注意：应排除具有伪迹的波长范围（例如达到饱和度或强干扰）。

4.3.1.7 多个数据预处理步骤时的顺序

使用多个数据预处理步骤时，顺序可能是至关重要的。基本规则如下。

首选 Gap-Segment 或 Savitzky-Golay 在 SNV 之前应用，SNV 首选在 detrend 之前应用。

求一阶导数和 SNV 的示例：一阶导数将极限倾斜转变为基线偏置。接下来的 SNV 消除该偏置。如果顺序相反，则 SNV 不会更改基线倾斜。接下来的求一阶导数会将其变换为基线偏置。这种情况下，偏置将被保留。

求二阶导数和 SNV 的示例：基线的偏置和基线的倾斜将确定被消除。以正确的顺序应用二阶导数和 SNV 能够消除基线的平方倾斜。二阶导数能够将基线的平方倾斜转变为基线的偏置。接下来的 SNV 消除该偏置。如果顺序相反，则 SNV 不会改变基线的平方倾斜。接下来的二阶导数会将其转转变为基线偏置。这种情况下，偏置将被保留。

4.3.2 波长范围

数据预处理 *（参见章节 4.3.1，第 32 页）* 后执行参数化设置的第二个步骤：选择波长范围能够排除不适用于目的的范围。特别是受干扰或饱和的波长范围可能会对后续计算产生不良影响，因此应被排除。

干扰

干扰在高吸光度值、并且仅少量光到达检测器时出现。下图显示了受干扰的范围。

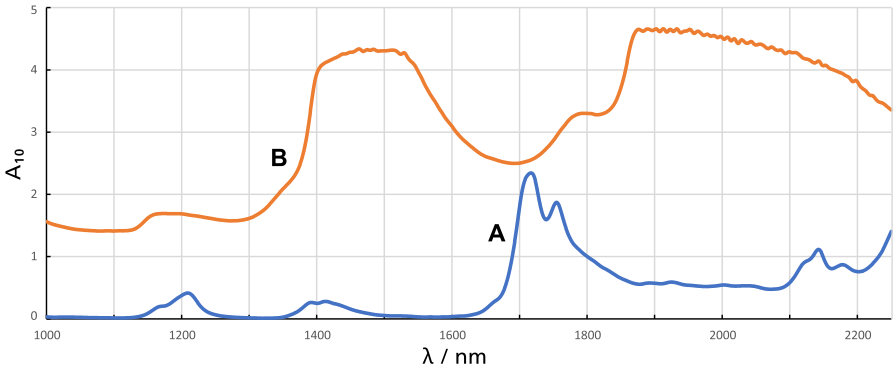


图 26 受干扰波长范围的示例。

光谱 **A** 包含正常的波峰型。光谱 **B** 包含 2 个受干扰范围：1,400 至 1,550 nm 和 1,870 nm 以上区域。受干扰的范围为强干扰区域并且与钟形曲线或其结合体不相近。

饱和度

如果大量光到达检测器，即吸光度值低的情况下，可能会出现检测器达到饱和度。

- **OMNIS NIR Analyzer**
  仪器始终自动设置积分时间。这样便防止了饱和度出现并将干扰最小化 *（参见"积分时间"，第 7 页）*。
- **2060 The NIR**
  如果激活了自动积分时间，则不会出现达到饱和度的情况。
  如果激活了手动积分时间，则过长积分时间可能导致检测器达到饱和度。饱和的区域在低吸光度值时出现，但在视觉上并不容易识别。因此，手动积分时间的设置应留有余地 *（参见"积分时间"，第 7 页）*。

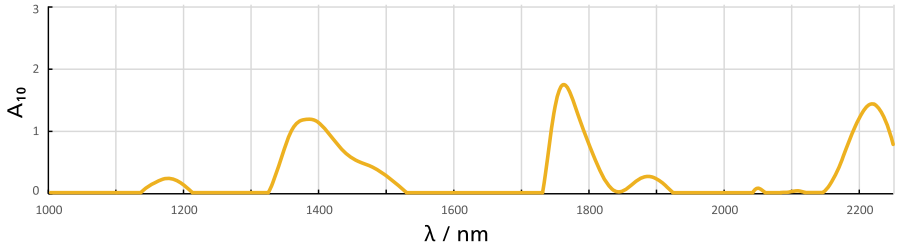


图 27　饱和波长范围的示例。

更多原因

另外还有纳入或排除波长范围的原因。选择可以基于分析参数及其相关吸收带 *（参见章节 2.1，第 3 页）*。但还需考虑相关信息可以根据预处理的形式必要时推移至其它波长范围。

如果在数据预处理时在光谱前后端引入了异常，可以排除相应的波长。

化学组分的变异或环境条件的波动可能会影响特定的波长范围。该波长范围的排除可以改善模型的耐用性。

排除形状良好的波长范围时须小心谨慎。看上去不包含信息的范围实际上可以提供隐藏和重要的信息。这些信息对于识别离群值或在处理干扰吸收带时有帮助。实际上，干扰吸收带是执行多变量测量的主因 *（参见章节 6.1，第 69 页）*。

4.3.3 光谱离群值

与大部分其它光谱不同的光谱被称为光谱离群值。

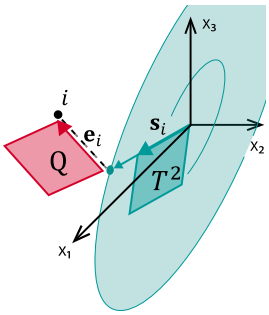
应仔细检查离群值。当原因是污染的样品或测量误差时，离群值可能会让模型扭曲。在这种情况下，计算模型时不考虑离群值。

否则，离群值代表其它光谱不充分覆盖的属性。在这种情况下，离群值甚至可以改进模型。如果离群值看上去可能是有效的样品，则应在变异宽度上均匀分配校正样品。

识别离群值的合适措施为 Hotellings T^2 和 Q 检验残差。

Hotelling T² 和 Q 检验残差

将光谱数据转换为主要成分空间时，光谱可以通过得分和残差描述特征 *（参见章节 4.2，第 28 页）*。这一点对于转变为潜变量空间同样适用 *（参见章节 4.4.1，第 47 页）*。



示例：一个 3 维波长空间（x_1、x_2、x_3）被转变为一个 2 维空间（绿色）。点位 *i* 代表光谱 *i* 并由 3 维空间投影为 2 维空间。由此可得出：

- 2 维空间及其归一化得分向量 $\mathbf{s}_i$ 内、代表马氏距离的的得分向量 $\Sigma \mathbf{s}_i$。
- 3 维空间内的残差 $\mathbf{e}_i$。

由 $\mathbf{s}_i$ 和 $\mathbf{e}_i$ 可以导出以下尺寸 *（参见章节 6.4，第 75 页）*：

- **Hotellings *T*²** 或简而言之 ***T*²** 是平方马氏距离，即在主要成分空间或潜变量空间上从模型终点到光谱正交投影的平方归一化的距离。
  如果一个光谱的所有得分相当于平均值，T^2 等于 0 并且光谱位于模型中点。模型在中点附近最佳。
  模型在远离中点可能表现不佳。T^2 数值较大。较大 T^2 数值表示极端光谱，例如一个具有极端化学组分成分的样品。
- **Q 残差** 是平方残差，即从光谱到主要成分空间或潜变量空间的平方正交距离。
  Q 检验残差显示了无法通过模型说明的变异。高 Q 残差表示光谱可能与模型不匹配，例如 当测得的样品包含其它物质。

识别光谱离群值

识别光谱离群值能够识别与总体不同的光谱。

1. 按如下方式考虑参数化：
   a. OMNIS Software 版本 4.2 以上：用户决定是否应用参数化（数据预处理和波长选择）。
   以后对参数化的更改对数据组分划没有影响。
   b. OMNIS Software 版本 3.3 以上至 OMNIS Software 版本 4.1：用户决定是否考虑数据预处理。波长选择和以后对数据预处理的更改对数据组分划没有影响。
   c. OMNIS Software 3.2 版本以下：按照识别离群值时确定的方式考虑数据预处理。波长选择和以后对数据预处理的更改对数据组分划没有影响。
2. 光谱离群值的识别基于平均值集中光谱的 PCA 模型 *（参见章节 4.2，第 28 页）*。主要成分的数量选择须考虑到所设定的方差至少应为 95 %。
3. 光谱离群值的识别采用用于 Hotellings T^2 和 Q 检验残差的数值。
算法评估所检验光谱的 Hotellings T^2 或 Q 残差是一个随机或系统性变异的结果。
算法的说明请参阅附录 *（参见章节 6.5，第 76 页）*。

4.3.3.1 影响图

影响图显示光谱的基本属性，有助于分析光谱离群值。

影响图的基础是 PCA 模型 *（参见章节 4.2，第 28 页）* 或 PLS 模型：

- **量化**：影响图可选择基于 **PCA** 或 **PLS**。
如同 PCA 一样，PLS 回归会将分光数据简化为少数变量。PLS 在这时也考虑参考值。
PCA 的主要成分在 PLS 中被称为**潜变量**。
（参见章节 4.4.1，第 47 页）
- **身份验证**：提供基于 **PCA** 的影响图（OMNIS Software 版本 4.3 以上）。

光谱离群值的形式

影响图显示每个光谱的 Hotelling T^2 和 Q 检验残差数值 *（参见 "Hotelling T^2 和 Q 检验残差"，第 40 页）*。Hotelling T^2 和 Q 检验残差可以识别不同类型的光谱离群值：

- **Hotelling T^2 离群值**也称为 Hebelarm 离群值（英文：Leverage outlier）：较高的 T^2 值表示光谱在主要成分空间 (PCA) 或潜变量空间(PLS) 的投影距离模型中点较远。
- **Q 检验残差离群值**：较高的 Q 检验残差表示模型对光谱的描述不佳。

图 28 以不同视图展示了多个光谱：

- 左影响图：Q 检验残差描述了通过模型说明的变异，而 Hotellings T^2 则考虑了模型中的变异。
  虚线显示了所设定显著性水平的**危险数值**或**极限值** *（参见章节 6.5，第 76 页）*。
  显著性水平越高，极限值越小，因此可能有更多的点超出极限值。
- 右图：具有 x_1、x_2、x_3 3 个变量的代表性原始空间，作为示例转换为具有潜变量的 2 维空间。
  对于点位 A 到 D 展示了到层级（虚线）的正交距离以及潜变量空间中的模型化点位（绿色点位）。

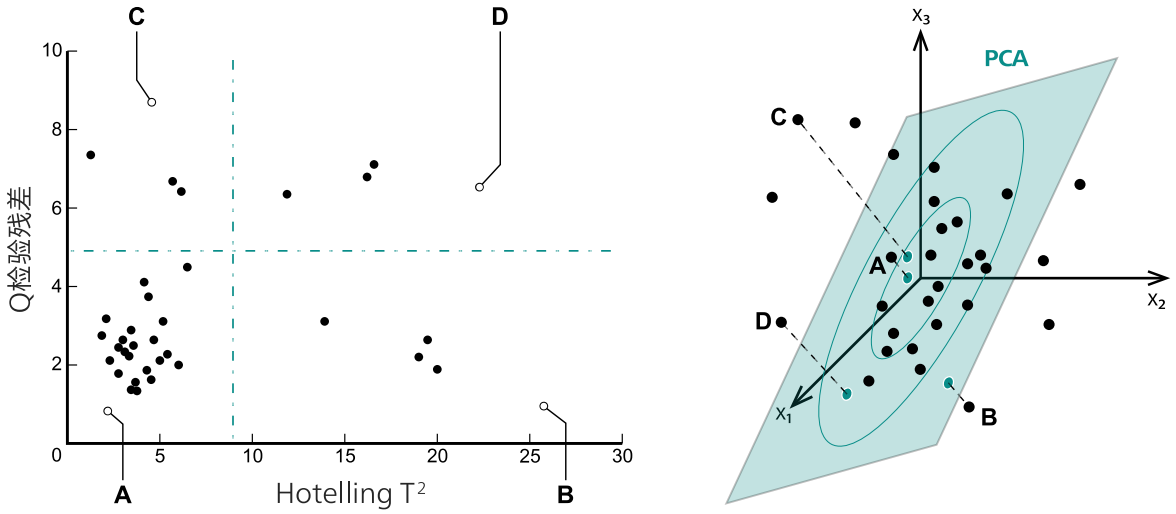


图 28　影响图（左）、原始空间和潜变量空间（右）。每个光谱分别通过一个点位在左右图中展示。

两个视图中均突出显示了具有不同特征的 4 个点位：

- 光谱 A 的得分和残差都低。其位于模型中点附近，模型描述良好。
- 光谱 B 是 Hotelling T^2 离群值。它位于中点外，但模型描述也良好。
- 光谱 C 是 Q 检验残差离群值。它位于中点外，模型描述不佳。
- 光谱 D 既是一个 Hotellings T^2 离群值，也是一个 Q 检验残差离群值。它位于中点外，模型只能部分描述。

影响图展示了不同光谱如何影响模型。因为所有潜变量均穿过中点，所以中点附近的光谱（例如光谱 A）几乎没有机会改变潜变量的方向。这些光谱没有杠杆值。距离中点的距离越大，杠杆值和模型影响的可能性越大。有些光谱实际上能够将模型向其方向拉动（光谱 B），而其它光谱只能在某种程度上实现拉动（光谱 D）或完全无法拉动（光谱 C）。

基于所有光谱对比一个模型，无光谱 B 模型的计算相对无光谱 D 模型的计算改变程度更大，相对无光谱 C 模型的计算改变程度还要更大。

光谱 B 可能会很大程度影响量化模型 — 可能是好的也可能是不好的影响。是否应删除影响图右下平方中可能的离群值需要小心谨慎。

理想情况下，模型将采集大量光谱的方差。模型仅覆盖少数光谱并无意义。上图中少数光谱与中点和大部分其它光谱的距离较大。这很可疑。模型会受到有些光谱的影响。涉及应检查的可能的离群值。此外，应确保样品在整个变异宽度均匀分布。

PCA 和 PLS 影响图

PCA 影响图仅取决于光谱。PLS 影响图则取决于光谱和参考值。

以下表格显示了不同设置如何影响 PCA 影响图和 PLS 影响图。

	PCA 影响图	PLS 影响图
光谱	以此为准的 PCA 模型基于校准数据集、校验数据集、离群数据集中的所有光谱。	以此为准的 PLS 模型基于校准数据集中的所有光谱。 基于此 PLS 模型，计算所有 3 个数据组中光谱的 T^2 和 Q 残差数值并在影响图中展示。
参数化设置	考虑所选的数据预处理和波长范围。 注意：离群值识别基于 PCA，并根据用户设置和 OMNIS Software 版本考虑了参数化 *（参见"识别光谱离群值"，第 40 页）*。	考虑所选的数据预处理和波长范围。 注意：预测的离群值评估基于 PLS 并考虑了数据预处理和波长范围。
变量数量	使用达到至少 **95%** 声明方差的主要成分数量。	使用当前所选的潜变量数量。
显著性水平和危险数值	使用当前所选的显著性水平用于计算和显示危险数值（虚线）。 身份验证：如果最近一次完成的数据组分划是在没有确定离群值的情况下进行的，影响图会使用 5 % 的显著性水平。 注意：提高显著性水平会导致更低的危险数值，从而会在模型开发中出现更多的离群值。	使用当前所选的显著性水平用于计算和显示危险数值（虚线）。 注意：提高显著性水平会导致更低的危险数值，从而会在预测时产生更多的离群值。
参考值（量化）	参考值对 PCA 模型没有影响。 但是，每个光谱都可能具有相关的离群值参考值，从而作为离群值标示。	参考值影响 PLS 模型，从而影响 PLS 影响图。 此外，每个光谱都具有相关的离群值参考值，从而作为离群值标示。

分析离群值

分析可能的离群值须考虑以下因素：

- Hotellings T^2 离群值通过与其它样品对比的极端化学组分成分指向一件样品。
- Q 检验残差离群值可能会指向例如结合的样品或光谱记录错误。

应仔细检查可能的离群值。真正的离群值应从光谱表删除。因保留有效的样品。如果数据组重新划分，则离群值识别在必要时会找到在第一次运行时未找到的可能的离群值。可能的原因是新的 PCA 模型需要较少的主要成分便可以达到声明的 95 % 的方差。如果新找到的离群值被证明为有效样品，则应将其保留。在这种情况下可以重复自动划分，无离群值识别。

4.3.3.2 得分图

得分图的基础是 PCA 模型或 PLS 模型：

- **量化**：得分图（OMNIS Software 版本 3.0 以上）基于 **PLS** *（参见章节 4.4.1，第 47 页）*。
- **身份验证**：得分图（OMNIS Software 版本 4.3 以上）基于 **PCA** *（参见章节 4.2，第 28 页）*。

每个光谱都有一个供每个主要成分或潜变量使用的得分值。在得分图中，每个光谱都由一个点来表示。x 轴显示例如第一个潜变量的得分，y 轴则代表例如第二个潜变量的得分。同样，也可以显示每一个潜变量对。

因为每个波长变量的吸光度值均已经过平均值集中，所以每个潜变量的得分也会被平均值集中。临近得分图中点附近的点位（0/0）表示基于两个所显示潜变量的平均值光谱。相邻点位附近的类似光谱、相距较远的点位代表基于两个显示潜变量的不相似光谱。

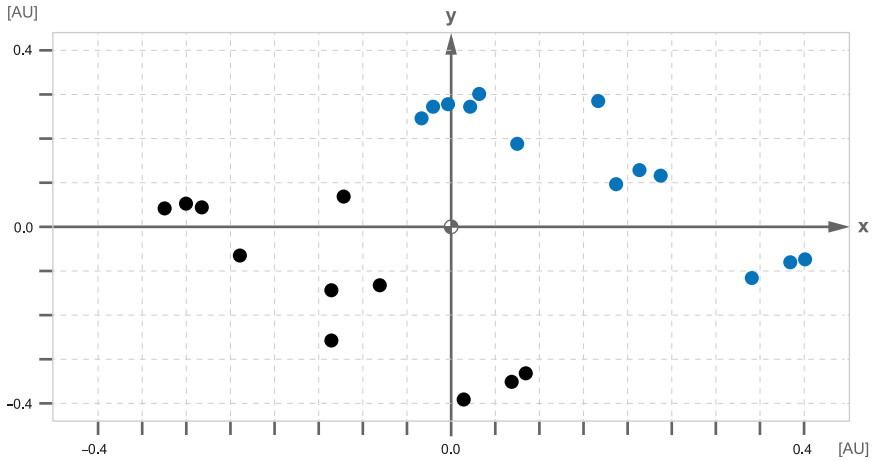


图 29　潜变量 1（x 轴）和潜变量 2（y 轴）的得分图。AU = 任意单位。

图 29 显示了已用于不同条件的 2 个数据组的示例。这些得分已被标准化，每个潜变量均包含相同的加权。

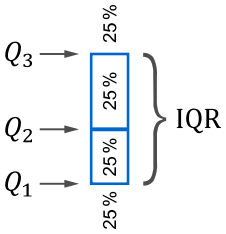
一个光谱的所有主要成分或潜变量的得分可以归纳为一个单一值（Hotelling T^2），该单一值显示于影响图的 x 轴上。

4.3.4 离群值参考值（量化）

量化模型中，除了光谱离群值以外还会确定离群值参考值。离群值参考值显示了参考值中的异常。

通常离群值参考值是错误传送的数字，如 143 而非 14.3，或 15.9 而非 51.9。离群值识别会基于经验原则确定传送或转录错误。仅明显的错误才会在后续检查中被标示。

箱线图



离群值参考值通过一个基于箱线图的方法确定。**箱线图** 以升序排列参考值。四分位数将数据组分为 4 部分。每个部分包含 25% 的参考值。

第一个四分位数 Q_1 将 25% 的最小数值与其余数值分离。Q_2 为中位数，将 50% 的最小数值与其余数值分离。第三个四分位数 Q_3 将 75% 最小数值与其余数值分离。垂直矩形标示数据的中间 50%，即四分位间距（IQR，英文：interquartile range）。

IQR 箱以外、特定值周围的数据被视作可莪能的离群值并且可以作为小圆圈展示。离群值的上极限值和下极限值通常被定义为 IQR 的 1.5 倍：

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

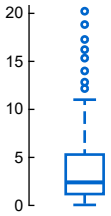
其中，Q_1 是第一个四分数，Q_3 是第三个四分数，IQR 是四分位间距 $(Q_3 - Q_1)$。

为了完善箱线图，箱上下的两条触须延伸至未标示未可能的离群值的最远处的点位。

对倾斜分布的调整

通常的箱线图设定数据接近系统性的分布。倾斜分布时，通常许多正常的参考值会标为可能的离群值。出于此原因，OMNIS Software 采用考虑分布倾斜度的、经过调整的箱线图版本。

在下例中，分布被推至了更高的参考值。由此得出了具有高数值的 8 个离群值。调整离群值极限值后，仅保留其中 2 个。另一侧则出现新的低数值离群值。

倾斜分布	经过调整的离群值极限值
	

您可在附录中找到调整的相关说明*（参见章节 6.6，第 77 页）*。

4.3.5 数据组分划

数据组由光谱和参考值构成（量化）构成，或由光谱和相关的产品名称构成（身份验证、校验）。数据组用于开发和校验模型。因此，应将数据组划分成校准数据集、校验数据集和离群数据集。划分可以手动或自动设置。

设定有足够的光谱可用的条件下，可以将总体数据的 20% 到 30% 用于校验数据集。

自动划分算法

自动数据划分时将确保校准数据集和校验数据集对于总体而言具有代表性并且相互不关联。目标是将数据划分为两组，接近覆盖 PCA 空间中的相同区域并具有相近的统计属性。所使用的算法是一个略微改动的 Duplex 算法，参阅 R. D. Snee 编著的 *Validation of Regression Models: Methods and Examples*, Technometrics Band 19, No. 4 (Nov. 1977), 第 415–428 页。

复制品

数据组内不得有复制品。算法不会移除复制品或假重复，不强制要求将特定样品的复制品插入相同的数据组。

1. 按如下方式考虑参数化：
   a. OMNIS Software 版本 4.2 以上：用户决定是否应用参数化（数据预处理和波长选择）。
      以后对参数化的更改对数据组分划没有影响。
   b. OMNIS Software 版本 3.3 以上至 OMNIS Software 版本 4.1：用户决定是否考虑数据预处理。波长选择和以后对数据预处理的更改对数据组分划没有影响。
   c. OMNIS Software 3.2 版本以下：按照识别离群值时确定的方式考虑数据预处理。波长选择和以后对数据预处理的更改对数据组分划没有影响。
2. 划分基于所有平均值集中光谱的 PCA 模型。主要成分的数量选择须考虑到所设定的方差至少应为 95 %。
3. 由 PCA 得分计算所有可能的光谱对之间的距离。
4. 相距最远的 2 个光谱将被分配给校准数据集。
5. 剩余的光谱中，2 个相距最远的光谱将比分配给校验数据集。
6. 剩余的光谱中，校准数据集中已包含的光谱中最远处的光谱将被分配给校准数据集。
7. 剩余的光谱中，校验数据集中已包含的光谱中最远处的光谱将被分配给校验数据集。
8. 切换将继续至数据组中的一个已经达到其预定的大小。剩余的光谱将被归类为其它数据组。

4.4 量化

4.4.1 PLS 回归

OMNIS Software 使用 PLS 回归用于计算量化模型。

类似于 PCA，**部分最小平方回归**（**PLS 回归**，英文：*partial least squares regression*）将光谱数据简化为少数变量。但是：

- PCA 的主要成分在 PLS 中被称为 **潜变量**。潜变量的方向和主要成分通常类似但不相同。
- 除了光谱以外，PLS 回归也考虑 **参考值**。因此，需要少数潜变量达到与参考值的关联性，导致小幅干扰。

PLS 从大量的、高度冗余的光谱数据中减去相关的 **潜变量**。潜变量也被称作隐藏的变量，因为非直接测量。潜变量说明了尽可能大部分的数据变异并且同时对参考值的模型化良好。

准备步骤

PLS 回归的准备步骤与 PCA 的步骤类似：

1. **参数化设置**：OMNIS Software 在光谱上应用所选的数据预处理并将确定的波长选择应用于光谱。
2. **平均值集中**：对于每个波长计算平均吸光度值并从每个光谱相关的数值中减去。
   参考值也同样经过平均值集中。

转换为潜变量

根据准备步骤，PLS 回归在考虑参考值的情况下将光谱数据转换为一个潜变量空间。*图 30* 显示了前 2 个潜变量 LV1 和 LV2。同样方式可以展示潜变量 LV3、LV4 等。

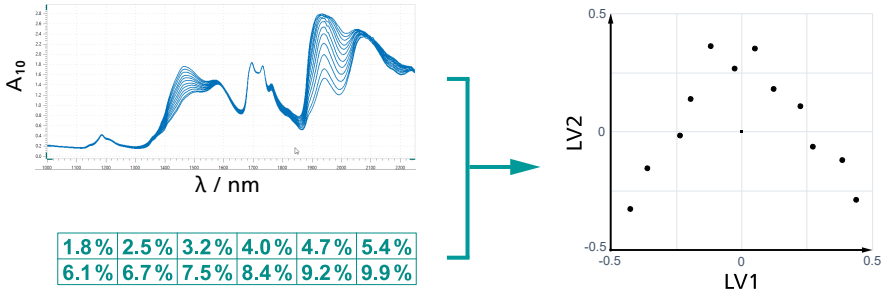
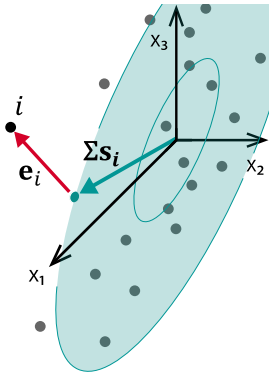


图 30　将光谱和参考值转换为一个潜变量空间。右侧的得分以任意的单位表现。

第一个潜变量 LV1 以最佳方式说明了光谱数据中的方差并包含带参考值的最大可能的关联性。后续所有潜变量 LV2、LV3 等以最佳方式说明了剩余的方差并包含带参考值的最大可能的关联性。因此，第一对潜变量占了方差中最大的部分并最大化关联，而其它数据主要包含干扰并且可以弃用。



得分和残差

PLS 如同 PCA 具有相近的大小：

- **得分**：得分将在潜变量空间内测得。样品 *i* 对潜变量每个方向的正交投影得到了得分向量 $\Sigma\mathbf{s}_i$，标示距离中点的欧氏距离。
  马氏距离 $\mathbf{s}_i$ 是赋予每个方向相同加权的标准化得分向量。
- **残差**：残差向量 $\mathbf{e}_i$ 是样品 *i* 和潜变量空间之间的偏置，在原始波长空间中测得。

PLS 算法

PLS 算法最大化光谱和参考值之间的协方差 *（参见章节 6.3，第 74 页）*。

4.4.1.1 潜变量数量

量化模型中潜变量的数量的选择对于模型的预测力具有关键意义。如果潜变量数量过低，则无法识别相关光谱变量。这种情况被称为**调整不足**并会导致预测准确度下降。

如果潜变量的数量过高，校正样品将被充分模型化。模型采集非关联光谱变异（干扰）。这种情况被称为 **调整过度** 并会导致未知样品预测波动、不稳定以及准确度下降。

为了找到最佳潜变量数量，必须达到以下目标之间的平衡：

- SEP 足够贴近其最小值。
  SECV 在无校验数据集的情况下使用。
- 量化模型应使用尽可能少的潜变量。如有怀疑，应使用更低的数量。
- 校验数据集的相关图接近其最佳情况。在理想情况下，斜率接近 1，y-轴截距接近 0，数据点为包含最小发散度。
  交叉验证的数值在无校验数据集的情况下使用 *（参见"交叉验证"，第 49 页）*。

注意模型品质指标仅涉及基于可用校正样品和校验样品的预测值。一般而言，基于少数潜变量的量化模型更耐用。

4.4.1.2 载荷图

OMNIS Software 中的载荷图基于 PLS 模型 *（参见章节 6.3，第 74 页）*。

PLS 载荷显示原始的波长变量（包括参数化设置）是如何帮助架构每个潜变量的。载荷的正负无关紧要。

载荷的计算须确保第一个潜变量采取对于参考参数最具说服力的方差。所有后续潜变量均获取对于参考参数最具说服力的剩余方差。与

0 存在严重偏差的 PLS 载荷表明相应的波长适于参考参数 d 额模型化。

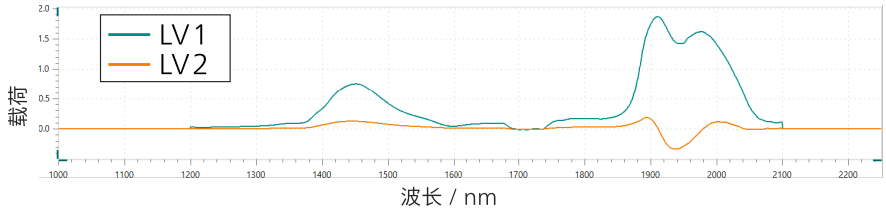


图 31　载荷图 LV1 和 LV2 的载荷图。

图 31 中波长选择被限制在 1,200 nm 至 2,100 nm 的范围内。因此在该区域以外不会出现载荷。

4.4.2　量化模型的校验

校验时检查模型是否满足对其性能高和耐用性的要求。为此必须尽可能实际地评估预期的预测错误。

量化模型通常按如下方式校验：

1. 由一个受限的样品数目和无校验数据集的情况下开发一个模型并通过交叉验证测试（见下）。
2. 加大样品数目会将数据组划分为 1 个用于开发模型的校准数据集和 1 个用于校验模型的校验数据集。
3. 最后，在另一天采集校验数据集的样品，并在必要时由另一个人用另一台仪器进行测量。

交叉验证

交叉验证完全基于校准数据集。交叉验证能够在使用一个在无相关校正样品情况下生成的临时模型时得到一个针对每个校正样品的预测值。

交叉验证中使用一个基于以下方式之一的多次舍入方法：

- **留一法**
  在每一次舍入中，留一法交叉验证都会返回 1 件样品，同时由剩余的样品创建一个模型。该模型预测返回样品的分析参数。该预测用作样品的预测值。
  循环将继续，直至每件样品都被返回。

- **K-fold**
  K-fold 交叉验证法是将校准数据集分为尽可能相同大小的 *k* 组块。每次舍入都会返回 1 个组块，同时由剩余的组块创建一个模型。该模型针对返回的样品预测分析参数。
  循环将继续直至返回每个组块。
  组块通过不同方式选择：
  - **Fixed Blocks (DUPLEX)**（OMNIS Software 版本 3.2 以上）：组块基于 Duplex 算法以可重现的方式选择。每次预测均直接用于相关样品的预测值。
  - **Random**：组块将以随机方式选择。上述操作流程将被多次重复。校准数据集每次都会以不同方式划分为 *k* 组块。最后，每件样品都会得出多个预测值。其平均值用作相关样品的预测值。

通常首选留一法。对于较大的校准数据集，可以采用 K-fold 交叉验证法，以便节省运算时间。*k* 的典型数值为 5。

4.4.2.1 相关图

相关图展示了参考值和计算值之间的相关性。它大致显示了量化模型的评估。

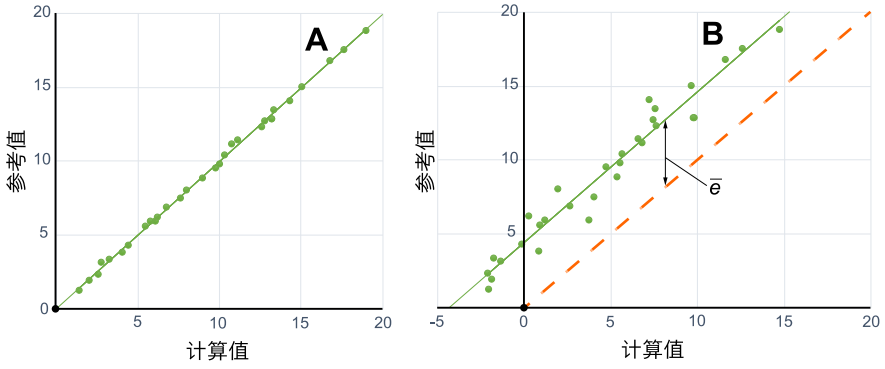
可按以下方式确定计算值：

- 校验数据集和离群数据集：通过量化模型预测
- 校准数据集：交叉验证的预测值

在相关图中通过一个点位展示每件样品。x 轴为样品的计算值，y 轴为参考值。回归线显示了变量之间的系统关联性。理想情况下，回归线的斜率为 1，y-轴截距为 0，所有点位位于直线上。这也意味着，每件样品的计算值都相当于参考值。

基于与理想情况的偏差可以区分系统性错误和随机错误。回归线的位置显示系统性错误。回归线的点位距离表示随机错误。

以下相关图 **A** 展现了一个新的相关性。其它图标展示了不同类型的错误，对此将在以下内容中说明。



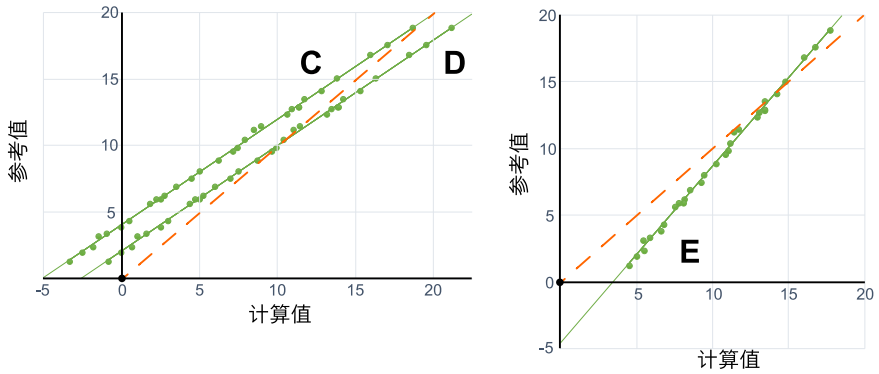


图 32　相关图。每个点位都代表一见样品并展现其参考值及计算值。虚线为理想情况的 45° 直线。

系统性错误

系统性错误是指总是出现、对于特定应用可能重复出现的错误。系统性错误可以修正。系统性错误通过偏差 $\bar{e}$ 和回归线的斜率 b 量化：

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

如果斜率等于 1 并且偏差等于 0，则不存在系统性错误。

偏差 是指参考值与计算值之间的平均错误：

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

其中，n 为样品数目，e_i 为第 i 个样品的错误，y_i 为第 i 个样品的参考值，$\hat{y}_i$ 是第 i 个阳光品的计算值，$\bar{y}$ 是参考值的平均值，$\bar{\hat{y}}$ 是计算值的平均值。

直线 **B** 和 **C** 具有正偏差，直线 **E** 具有负偏差。

相反正负号的错误相互抵消。因此，直线 **D** 的偏差接近 0。

回归线的**斜率**：

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

此处，$s_{\hat{y}y}$ 对应参考值和计算值之间的协方差，$s_{\hat{y}}^2$ 对应计算值的方差。

斜率可以被视作属性相关的错误：

- $b > 1$ (直线 **E**)：计算值越大，导致偏差的错误越大（正值越大）。
- $b < 1$ (直线 **C** 和 **D**)：计算值越大，导致偏差的而错误越小（负值越小）。
- $b = 1$ (直线 **A** 和 **B**)：导致偏差的错误不变。

带 y 轴的回归线的 **y-轴截距** 为 $\bar{y} - b\bar{\hat{y}}$。

斜率和 y-轴截距通过将参考值作为关联变量（y 轴）、将计算值作为关联变量（x 轴）进行计算。

随机错误

如果所有点位直接位于回归线上，则无随机错误。点位越分散，随机错误越高。

相关图 **B** 中的随机错误大于其它相关图中的随机错误。

显示错误类型

上图中的直线显示了以下错误类型：

	系统性错误			随机错误
直线	**偏差**	**斜率**	**y-轴截距**	
A	~ 0	~ 1	~ 0	小
B	> 0	~ 1	> 0	大
C	> 0	< 1	> 0	小
D	~ 0	< 1	> 0	小
E	< 0	> 1	< 0	小

4.4.2.2 模型品质指标

模型品质指标代表参考值与预测值一致。计算值通过量化模型确定。

R^2 – 测量系数

测量系数 R^2（英文：coefficient of determination）测量量化模型的调整质量。对于特定的数据组，该参数为通过量化模型确定的参考值变异比例：

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

其中，回归平方和 SS_{reg}（预测值的方差，即 声明的方差），总平方和 SS_{tot}（参考值方差），残差平方和 SS_{res}（残方差），即 未声明的方差），y_i 为第 *i* 个样品的参考值，$\hat{y}_i$ 为第 *i* 个样品的预测值，$\bar{y}$ 为参考值的平均值。

R^2 数值是为 1 的分数。R^2 等于 1 表示预测值与参考值完美吻合。R^2 为 0.9 表示参考值方差的 90 % 可以通过预测值说明，10 % 无法实现。

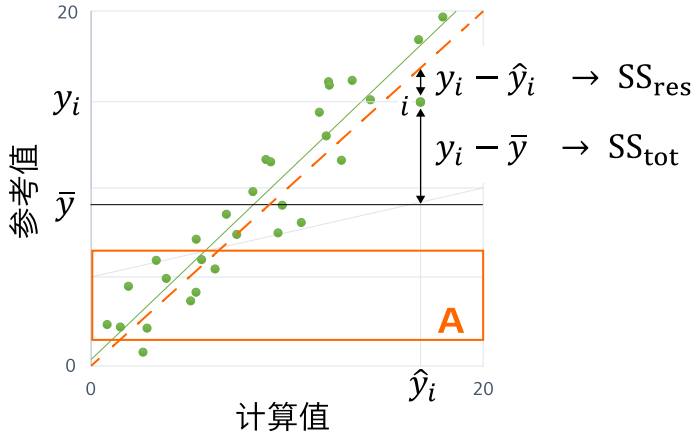


图 33　用于计算 R² 的组分。虚线为 45° 直线。

上图通过一件样品 *i* 及其由 PLS 回归得出的允差展示相关图。残余变异包含在 SS_{res} 中，参考值方差包含在 SS_{tot} 中。

高 R² 值是一个可用量化模型或准确预测的保障。R² 的大小直接取决于参考值变异。

小范围参考值的回归（范围 A）具有大致相同的残余变异，但参考值变异更小。得出的 R² 数值更小。

高 R² 值的原因是参考值范围出现不切实际多大的情况。另一方面，加工工艺的数据可能包含受限的值范围，从而导致 R² 数值更小。对于预测力的评估应采用标准错误。

绝对 R² 数值须小心谨慎地对待。更具说服力的是每个附加潜变量产生的变化程度 *（参见章节 4.4.1.1，第 48 页）*。

根据计算采用哪些数值的不同，得出的 R² 数值不同：

- R²C（OMNIS Software 中不显示）：通过校准数据集中光谱的预测值计算。
- **R²CV**：通过校准数据集中光谱的交叉验证的预测值计算 *（参见“交叉验证”，第 49 页）*。
- **R²P**：通过校验数据集中光谱的预测值计算。

OMNIS Software 采用 Pearson 抽样相关系数 $r_{y,\hat{y}}$ 的平方用于计算：

交叉验证的测量系数：

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

其中，y_i 为第 *i* 个样品的参考值，$\bar{y}$ 为参考值的平均值，$\hat{y}_{\mathrm{cv}i}$ 为第 *i* 个样品交叉验证的预测值，$\bar{\hat{y}}_{\mathrm{cv}}$ 为交叉验证预测值的平均值，*n* 为校准数据集中的样品数量。注意校准数据集中的每件样品都应具有一个交叉验证预测值。

预测的测量系数：

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

其中，v_i 为第 *i* 个校正样品的参考值，$\bar{v}$ 为参考值的平均值，$\hat{v}_i$ 为第 *i* 个校正样品的预测值，$\bar{\hat{v}}$ 为预测值的平均值，v 为校正样品的数量。

SEC - 校正的标准误差

校正的标准误差 (SEC) 基于校准数据集。SEC 可以被视为理论最佳预测准确度的预测值。SEC 是偏最小二乘回归（PLS）残差的标准偏差：

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

其中，**e** 对应包含在校准数据集中、不通过模型描述的所有参考值变异，*n* 是校正样品数量，*k* 为潜变量数量。分母 *n*-*k*–1 是残差向量 **e** 的自由度数量。

换句话说：SEC 是校准数据集中参考值和预测值之间差值的标准偏差：

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

其中，y_i 相当于第 *i* 个校正样品的参考值，$\hat{y}_i$ 为第 *i* 个校正样品的预测值，*n* 对应校正样品的数量，*k* 是潜变量的数量。

SEC 有时也被称为 RMSEC。SEC 包含随机错误和系统性错误（斜率和偏差）。

SECV – 交叉验证标准误差

交叉验证标准误差（SECV） 基于校准数据集。SECV 基于校准数据集和交叉验证法预判预测力 *（参见“交叉验证”，第 49 页）*。SECV 可以用于第一次模型评估或用于确定潜变量的最佳数量。

SECV 是校准数据集中样品的交叉验证的参考值和预测值之间的标准偏差。

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

其中，y_i 相当于第 *i* 个样品的参考值，$\hat{y}_{cv_i}$ 对应第 *i* 个样品交叉验证的预测值，*n* 为校准数据集中样品的数量。

SECV 包含所有错误：随机错误和系统性错误（斜率和偏差）。其它原则采用分离的数值用于偏差修正的 SECV 以及用于被称为 RMSECV 的未修正 SECV。低偏差时，这些数值相近。

SEP – 预测标准误差

预测标准误差 (SEP) 基于校验数据集。因此，SEP 提供预测准确度的实际预测值。

SEP 是校验数据集中样品的参考值和预测值之间差值的标准偏差：

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

其中，v_i 为第 *i* 个校正样品的参考值，$\hat{v}_i$ 对应第 *i* 个校正样品的预测值，v 为校正样品的数值。

SEP 包含所有错误：随机错误和系统性错误（斜率和偏差）。其它原则采用分离的数值用于偏差修正的 SEP 以及用于被称为 RMSEP 的未修正 SEP。低偏差时，这些数值相近。

模型品质指标的说明

模型品质指标为预测值。例：SEP 数值是标准偏差的 *预测值* – 基于当前样品 – 也具有自己的标准偏差。校正样品的数量越高，预测值越可靠。

SEP 应可以与 SECV 和 SEC 对比。过高的差异代表可能存在过度调整 *（参见章节 4.4.1.1，第 48 页）*。原则是差异不应超过 20%。

标准错误也应对比参考方法段实验室标准错误（SEL）看待。NIR 方法段的精度可接受的条件是 SEP 大于 SEL 1.4 至 2.0 倍。更大的 SEP 在满足要求的情况下可以接受。

参考方法段的 SEL 下的 SEC 或 SECV 表明过度调整。

对于上述与 SEL 的关系，设定 SEL 基于每件样品重复参考测量的正确数量 *（参见"参考方法段（量化）"，第 27 页）*。

4.4.3 OMNIS Model Developer（OMD）

量化模型的研发要求严格、耗费时间并且需要一定的专业知识。OMNIS Model Developer（OMD，OMNIS Software 版本 4.0 以上）能够实现自动开发并提供经过充分优化的量化模型。

工作原理

正确的采样是前提条件 *（参见章节 4.1，第 26 页）*。一个由光谱和相应参考值构成的数据组用作 OMD 的输入。

OMD 通过 5% 显著性水平和附录中列举的算法在不考虑数据预处理的情况下确定光谱离群值 *（参见"模型开发时识别光谱离群值"，第 76 页）*。

数据组的划分取决于离群值识别后保留的光谱数量：

光谱数量	交叉验证法	校验数据集
> 99	K-fold（5 个组块，DUPLEX）	光谱的 25%
30–99	K-fold（5 个组块，DUPLEX）	—
< 30	留一法	—

划分后将根据 ASTM D8321-22 在确定校准数据集中的附加离群值。

模型的评估和排列一句不同的特征值确定。OMD 优化数据预处理、波长选择以及潜变量数量，目标是达到过度调整的风险和调整不足之间的平衡。

结果

OMD 的结果是一个根据预测力排序的模型的列表。**预测力** 基于模型复杂度、模型品质指标和数据组大小计算。

列表标有色码，以便简化首选模型的选择：

- 绿色：预测力良好。
  如果样品数目足够多，则模型对于所有相同类型的未知样品而言效果都不错。模型品质指标能够得到对于今后错误可靠的预测值。
- 黄色：预测力中等。
  如果样品数目足够大，模型预计将良好运行。模型品质指标对于今后的样品而言可能是最佳的。推荐进行特别的校验。
- 红色：预测力不足。
  模型具有严重的缺点。不应使用。

相同颜色的模型根据有利于低预判预测错误和低潜变量数量之间平衡的信息标准排序。

优化参数化设置

无需自动创建整个模型，也可以只优化参数化。当前设置（例如，数据组分划、交叉验证法）保持不变，但不会影响优化。

4.4.4 斜率/y 轴截距校正

斜率/y 轴截距校正能够实现应用量化模型时的系统性错误（偏差、斜率）修正。

校准数据集中存在系统性错误的可能原因：

- 量化模型中存在系统性错误。例如未识别的离群值或样品数目不足。
- 光谱测量方法中存在系统性错误。
- 参考测量方法中存在系统性错误。

如果校验数据集中出现系统性错误，则可能存在其它原因：

- 光谱测量方法的变化，例如仪器的变化。

- 参考测量方法的变化，例如新的实验室、新的装备。
- 样品的变化，例如搬运、储存或运输。

偏差纠正、特别是斜率/y 轴截距校正应小心谨慎应用。

如果系统性错误不明显，则不应应用修正。如果错误明显，则应彻底检查。错误的原因因尽可能排除。如果错误出于合理原因无法排除，则可以应用偏差纠正和斜率/y 轴截距校正。

得到可靠的预测值需要至少 20 个样品。得到可靠的斜率预测值需要至少 30 个样品。

偏差纠正

以下相关图展示了偏差纠正。原始回归线 **F** 的斜率保持不变。纠正（回归线 **G**）后，正和负错误相互抵消。

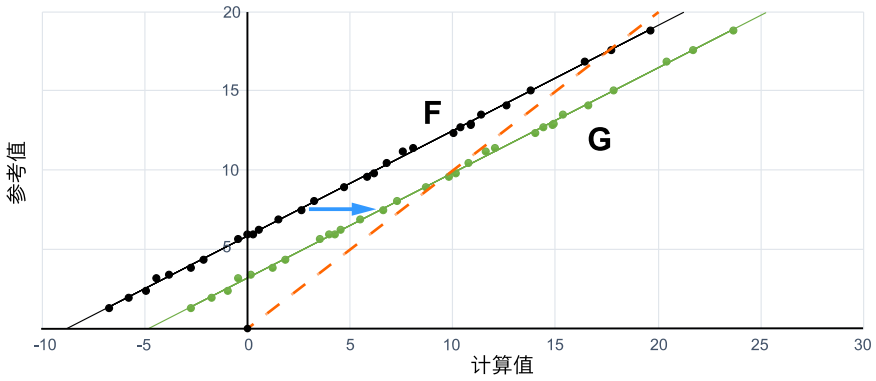


图 34　偏差纠正

斜率/y 轴截距校正

以下相关图展示了斜率/y 轴截距校正。原始的回归线 **H** 围绕斜率和 y-轴截距纠正。从而纠正斜率和偏差（回归线 **K**）。

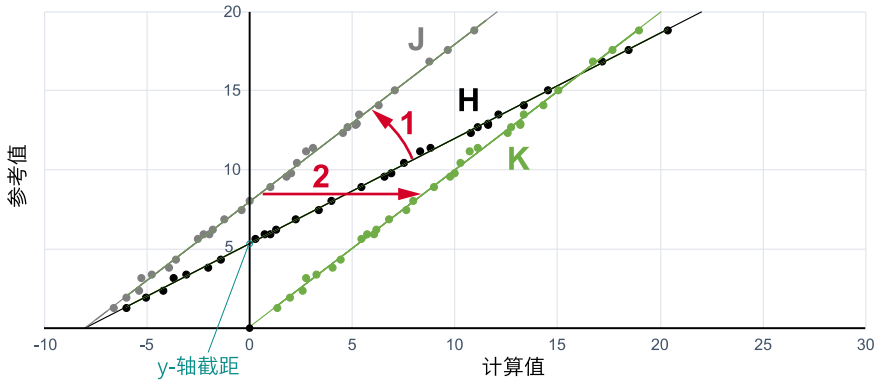


图 35　斜率/y 轴截距校正

SEP

纠正数据集中的样品是斜率/y 轴截距校正的基础。OMNIS Software 基于这些样品计算以下**预测标准误差（SEP）**。各分母项均考虑相应的自由度：

校正类型	SEP
未校正	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n}}$
偏差纠正	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-1}}$
斜率/y 轴截距校正	$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(v_i - \hat{v}_i)^2}{n-2}}$

其中，v_i 表示纠正数据集中第 *i* 个样品的参考值，$\hat{v}_i$ 表示纠正数据集中第 *i* 个样品的预测值，*n* 表示纠正数据集中的样品数量。

SEP 包含所有错误：随机错误和系统性错误（斜率和偏差）。

4.5 身份验证和校验

4.5.1 Support Vector Machine (SVM)

识别模型（OMNIS Software 版本 4.0 以上）采用 Support Vector Machines (SVM) 用于不同产品之间的分类。Support Vector Machine 是一种受监控的机器学习算法。该算法根据校正样品将新的样品分类为一件产品。

为了方便理解，接下来对 2 种产品之间的分类进行说明。方案可以扩展，最终的模型在任意数量的产品之间分类。

线性分类

图 36 (左) 展示了 2 个变量的输入数据。

为了便于理解，我们用一个 2 维变量空间展示参数化设置的光谱。每个点位代表一个光谱，颜色代表产品关联性。

产品可线性分离。SVM 算法在产品之间创建一个超平面（右图）。

超平面是将 3 维空间的平面统一维任意维度的空间。一个超平面的维度比围绕其的空间小一个维度。一个 2 维空间内的超平面是一条直线。

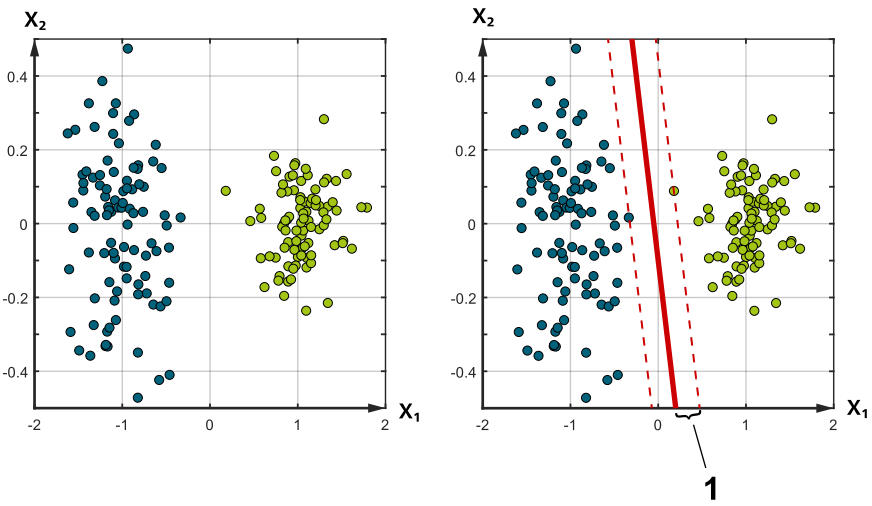


图 36　输入数据（左）和由 SVM 创建的超平面（右侧红色直线）。数值以任意单位标注。

SVM 算法将每一侧的超平面和附近点位之间的间距 (**1**) 最大化。新的光谱根据其所处超平面哪一侧可以在相同的空间上构建并归类为一个产品。

对于超平面的定义，SVM 算法仅考虑距离对面产品的点位最近的点位。这些点位或向量支持构建超平面并被称作支持向量。

如果点位无法线性分离 — 例如因离群值 — 可以确定一条线性分类超平面。在这种情况下，优化算法能够找到加大超平面与每一侧支持向量之间的间距和确保所有点位位于正确的超平面一侧之间找到平衡妥协。正则化参数检查平衡妥协效果以及超平面的最终位置。

非线性分类

图 37 (左) 中产品不可线性分离。产品的分离需要一个非线性分类因素。

一个线性或非线性的核函数（Kernel function）将数据转换为一个高维度特征空间。变换的执行须确保特征空间内的数据可以通过一个超平面线性分离。

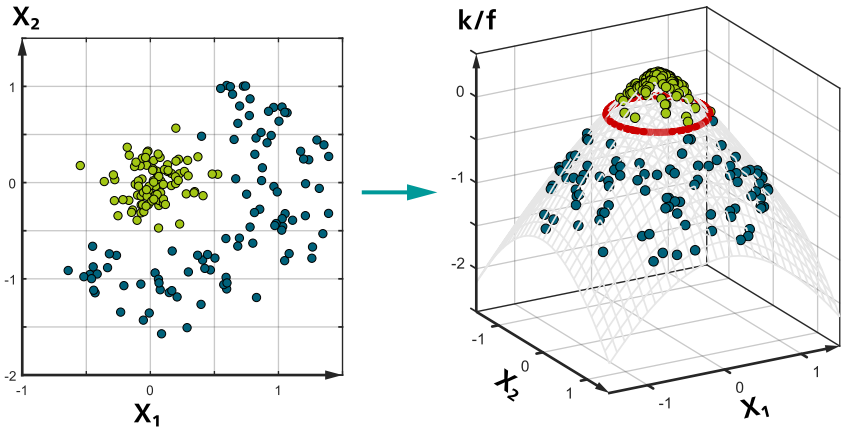


图 37　非线性分离的产品（左）。一个附加维度 k/f（核特征）可以简化分离（右）。数值以任意单位标注。

图中，2 维空间的数据被转换为 3 维空间。一个产品的点位将通过原始平面提升，而另一件产品的点位将向下推移。产品之间的线性超平面是与原始平面非常相近的 2 维平面。在 2 维空间内观察，决策边界为一条非线性线，在这种情况下为红色圆形线。

超平面在这里同样用作分类器。一个新的样品可以根据其光谱位于超平面的哪一侧归类为一件产品。

OMNIS Software 使用一个带径向基础函数的核将数据转换为特征空间。核采用一个控制非线性度的标度参数。

参数选择

对于控制超平面位置的正则化参数、控制非线性度的正则化参数，必须选择合适的数值。

两个参数对 Support Vector Machine 的普遍化能力产生影响，即对该算法将校准光谱普遍化为新未知光谱效果的影响。如果标度参数能够实现高度非线性，超平面可能针对校正光谱调整的程度过大（过度调整）。如果标度参数只能实现低程度的非线性，超平面可能对校正光谱的调整程度不足（调整不足）。

采用**栅格查找**确保一般化效果良好。算法通过尽可能少的尝试找到最佳参数组合：

1. 采用一套预定义的参数组合。
2. 对于有些所选的参数组合，SVM 学习将具有最大精确度的不同产品的校正光谱进行分类的规则。
3. 对于每个分类规则而言，根据交叉验证预测产品归类效果如何。
4. 基于交叉验证的结果选择更多参数组合。
   SVM 重新学习相应的分类规则，交叉验证法预测分类准确度。
   若干次重复后确定最终的参数组合。
5. SVM 通过最终的参数组合和所有校正光谱评估最终分类规则。

概率

识别模型会根据分类规则计算某个特定样品否属于某一特定产品的概率。概率取决于与此产品总体的距离以及模型参数。

如此计算得出的概率可用于控制样品与产品的分配 *（参见"分配样品（从 OMNIS Software 版本 4.4 起）"，第 61 页）*。

为了尽量减少假阳性身份验证，从 OMNIS Software 版本 4.4 起会额外地为每个产品训练 1 个定性模型 *（参见章节 4.6，第 64 页）*。

4.5.2 样品产品所属性的预测

为了对特定的样品进行身份验证，识别模型会为每个产品提供 1 个概率 *（参见"概率"，第 61 页）*。

每个概率都是介于 0% 和 100% 之间的单独数值。产品的数值总和非 100%。
这些数值需相互关联进行观察，从而可对不同产品进行比较。

分配样品（从 OMNIS Software 版本 4.4 起）

评估通过可设置的**概率阈值**和各个产品的定性进行：

1. 对于概率高于概率阈值的每个产品，要用相应的定性模型对样品进行定性。如果定性失败，则将相应产品的概率置零。
2. 用步骤 1 中修正的概率进行评估：
   a. 如果没有概率高于概率阈值，则身份验证失败（身份验证状态 **未校验**）。
   b. 如果只有唯一一个概率高于概率阈值，则样品识别成功并将其分配给相应产品（身份验证状态 **被识别**）。
   c. 如果有多个概率高于概率阈值，则预测含糊不清且身份验证失败（身份验证状态 **含糊不清**）。

表格 1 展示了不同概率阈值的评估示例。在本示例中，产品 A 和 C 的定性均成功通过。

表格 1　不同概率阈值的评估示例（从 OMNIS Software 版本 4.4 起）

概率	概率阈值	定性	识别结果
产品 A：87 % 产品 B：71 % 产品 C：68 % 产品 D：30 %	90 %	–	未校验
	80 %	产品 A：成功 → 87 %	产品 A
	70 %	产品 A：成功 → 87 % 产品 B：失败 → 0 %	产品 A
	60 %	产品 A：成功 → 87 % 产品 B：失败 → 0 % 产品 C：成功 → 68 %	含糊不清

分配样品（OMNIS Software 版本 4.0 至 4.3）

通过可设置的**概率阈值**评估样品的概率：

- 如果没有概率高于概率阈值，则身份验证失败（身份验证状态 **未校验**）。
- 如果只有唯一一个概率高于概率阈值，则样品识别成功并将其分配给相应产品（身份验证状态 **被识别**）。
- 如果有多个概率高于概率阈值，则预测含糊不清且身份验证失败（身份验证状态 **含糊不清**）。

表格 2 展示了不同概率阈值的评估示例。

表格 2　不同概率阈值的评估示例（OMNIS Software 版本 4.0 至 4.3）

概率	概率阈值	识别结果
产品 A：87 %	90 %	未校验
产品 B：72 %	80 %	产品 A
产品 C：68 %	70 %	含糊不清

4.5.3　识别模型的校验

识别模型通常按照如下方式校验：

1. 由一个受限样品数目和无校验数据集的情况下开发一个模型并通过校准数据集中的样品测试。
2. 样品数目足够时，数据组可以自动划分为校准数据集和校验数据集。校验数据集中的样品不用于开发模型。
3. 最后，在另一天采集外部校验数据集的样品，并在必要时由另一个人用另一台仪器进行测量。

但对于每件样品将预测的产品关联性与相关实际的产品关联性对比。如果二者相符，预测正确（= 成功），否则错误（= 失败）。

校验

OMNIS Software 显示用于识别模型的以下参数，这些参数给出了模型的效果。理想情况下，所有特征值为 100%。

成功 %（整体） 测量 *精确度*，模型的整体正确性。百分比数值回答了问题：多少样品可以正确识别模型？

$$成功\ \%（整体）= \frac{正确分类}{所有分类}$$

_成功 %（整体）_ 为每件样品赋予了相同的权重。因此，具有更多样品的产品比具有少数样品的产品对百分比值的影响更大。

成功 % 的类似数值对于每件产品可用。

改善身份验证

以下操作可能有助于改善模型：

- **调整概率阈值**
  - 如果许多预测都含糊不清，或者出现许多 0.0 % 概率，则可以提高概率阈值。
  - 如果有许多样品因未达到概率阈值而无法识别，则可降低概率阈值。
- **调整参数化设置**
  确定更合适的数据预处理和波长选择。
- **使用模型层级**
  模型层级能够实现识别模型的分级构建。
  示例：具有 4 件不同产品的识别模型在没有其它辅助条件的情况下无法区分诸如果糖和葡萄糖等类似产品。如果将果糖和葡萄糖归纳在"糖"的产品组中，则模型能够在糖与其它两种产品之间很好地区分。如果样品被识别为糖，则另一个模型会接受果糖和葡萄糖之间的区分。因为该模型更具特性，因此模型可以更好地区分类似的产品。

未识别的样品

如果错误地无法识别样品：

- 检查样品和样品处理是否异常。
- 检查并在必要时降低概率阈值。
- 检查数据预处理和波长选择。
- 在得分图中检查未识别样品的光谱：
  - 如果光谱尚未包含在模型中：将这些光谱增加至相应产品的校验数据集。
  - 在得分图中将待检查光谱的得分与校准数据集中的光谱得分进行比较。

  如果样品并非异常值，但其变异在校准数据集中未得到充分体现，则应相应扩展校准数据集。

4.6 定性

定性模型（从 OMNIS Software 版本 4.4 起）可将一组样品与其他样品区分开。这些模型比如适用于将可用样品（正样品）与不可用样品（负样品）区分开。

4.6.1 定性模型的计算

定性模型的计算与识别模型的计算类似 *（参见章节 4.5.1，第 58 页）*。但是，定性的校准数据集只包含唯一一种样品（正样品）。

Support Vector Machine（SVM）可将输入数据转换到一个更高维的空间中。正则化参数规定超平面的位置，而缩放参数决定非线性度。网格搜索确定合适的投影方式。这种投影的决策边界构成了定性模型的基础。

4.6.2 定性模型的校验

操作方法

定性模型通常逐步开发和校验。在此，要逐渐增加样品的数量：

1. 正校验数据集：
   a. 在正样品数量有限时，暂不创建正校验数据集。
   b. 只要有足够数量的正样品可用，就可通过自动数据组分划创建 1 个正校验数据集。

   负校验数据集：
   a. 将采集的负样品分配给负校验数据集。
   b. 自动确定负光谱可用于识别光谱离群值并将其分配给负校验数据集 *（参见章节 6.5，第 76 页）*。
2. 最后，在另一天采集正和负校验数据集的样品，并在必要时由另一个人用另一台仪器进行测量。

正和负校验数据集中的样品不用于计算模型。

校验

定性模型为每个样品确定 1 个结果（有效或无效）。对于校准数据集和正校验数据集中的样品，预期会得到正结果。对于负校验数据集中的样品，预期会得到负结果。如果结果与预期一致，则预测正确（= 成功），否则错误（= 失败）。

OMNIS Software 为定性模型显示以下值：

*成功 %（整体）*测量模型的整体正确性。

$$\textit{成功 \%（整体）} = \frac{\textit{正确的预测}}{\textit{所有预测}}$$

与*成功 %* 类似的数字可用于每个数据组。理想情况下，所有特征值为 100%。

优化定性

更合适的参数化（数据预处理和波长选择）可优化定性模型。

未通过定性的样品

如果样品被错误地判定为未通过定性：

- 检查样品和样品处理是否异常。
- 检查数据预处理和波长选择。
- 在得分图中检查未通过定性样品的光谱：
  - 如果光谱尚未包含在模型中：将这些光谱增加至正校验数据集。
  - 在得分图中将待检查光谱的得分与校准数据集中的光谱得分进行比较。

  如果样品并非异常值，但其变异在校准数据集中未得到充分体现，则应相应扩展校准数据集。

5 预测

5.1 量化

对于使用未知分析参数预测一件样品：

1. 记录样品的光谱。
2. 量化模型采用与校准数据集中光谱相同的数据预处理和波长范围。
3. 模型根据获得的光谱预测分析参数。

计算量化模型时，校正光谱和参考值已经过平均值集中。上述操作流程当然也考虑了这一点。

5.1.1 离群值和结果监视

在预测量化的感兴趣的参数时，可以识别不同类型的离群值并监测结果：

离群值/监控		原因（示例）
光谱离群值	Hotelling T^2	化学组分的浓度超出校正样品的浓度范围。
	Q 检验残差	该样品包含校正样品中不存在的组成部分。
最近的离群值（可选，OMNIS Software 版本 4.2）		样品发生样品变异，这些变异在该组合中未在校正样品上表示出来。
结果监视（可选）		感兴趣的参数的结果值超出客户选择的验收标准。

光谱离群值

光谱离群值按照以下方式确定：

1. 软件基于量化模型（PLS 模型）计算光谱的 Hotelling T^2 的数值以及 Q 检验残差。
2. 如果光谱的 T^2 或 Q 检验残差大于模型计算的相应临界值，根据所应用的模型应将样品标识为离群值 *（参见"预测的离群值评估（量化）"，第 77 页）*。

在 OMNIS Software 中，T^2 和 Q 检验残差作为变量提供。它们可以与量化模型 PLS 影响图中的数值比较。虚线展示了临界数值。

Nearest Neighbor 离群值

（OMNIS Software 版本 4.2 以上）

在理想情况下，校正样品涵盖样品变异的所有可能组合。实际上，有些组合出现得比较频繁，有些则根本不出现。因此，校正样品在潜在变量空间中分布不均匀。在某些区域有许多校正样品，但它们之间存在间隙。

如果未知样品的光谱落在校正样品之间的间隙中，则预测结果可能无效或不准确。为了检测这种情况，计算从未知样品 *i* 到每个校正样品 *u* 的间距 *D*：

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

其中 $\mathbf{s}_i$ 对应于未知样品 *i* 的得分，而 $\mathbf{s}_u$ 对应于校正样品 *u* 的得分。得分是标准化的并且是正交的。

最小间距是指与最近的校正样品的间距，并称为 **Nearest Neighbor Distance** (NND)：

如果 NND 值超过特定 NND 极限值，则将未知样品称为 Nearest Neighbor 离群值。

按以下方式确定 NND 极限值：

1. 确定每个校正样品的 NND 值。此数值相当于到其余校正样品的最近距离。
2. 所有校正样品的最大 NND 值是 NND 极限值。

在 OMNIS Software 中，未知样品的 NND 值和 NND 极限值作为变量提供。

结果监视

在 OMNIS Software 中，结果监视可用于定义预测结果范围的警告界限和干预界限。可以选择定义违反定义的限值时触发的操作。

5.2 身份验证和校验

在身份验证或校验样品时按照以下方式操作：

1. 记录样品谱图。
2. 识别模型采用与校准数据集中光谱相同的数据预处理和相同的波长选择。
3. 识别模型对光谱进行评估 *（参见章节 4.5.2，第 61 页）*。
4. 如果识别模型是模型层级的一部分，并且另一个识别模型与确定的产品相关联，则执行该模型。如果一个或多个量化模型与确定的产品相关联，则执行这些模型。
5. 显示识别结果或校验结果，如果是模型层级，可能还会显示量化结果。

身份验证状态

- 已识别
  身份验证成功。

- 含糊不清
  多个产品超过概率阈值。身份验证失败。
- 未识别
  无产品超过概率阈值。身份验证失败。

校验状态

- 成功
  样品识别成功，结果与预期产品一致。
- 失败
  校验失败。

5.3 定性

样品的定性方法如下：

1. 记录样品谱图。
2. 定性模型采用与校准数据集中光谱相同的数据预处理和相同的波长选择。
3. 模型根据获得的光谱定性样品。
4. 显示定性结果。

定性状态

- 成功
- 失败

6 附录

6.1 线性回归示例

变换指标线性回归

最简单的情况是一个混合物中只有一个吸收器和光谱只有一个波峰。具有不同吸收器浓度的样品包含具有不同吸光度值的波峰 *（参见章节 2.2.1，第 5 页）*。

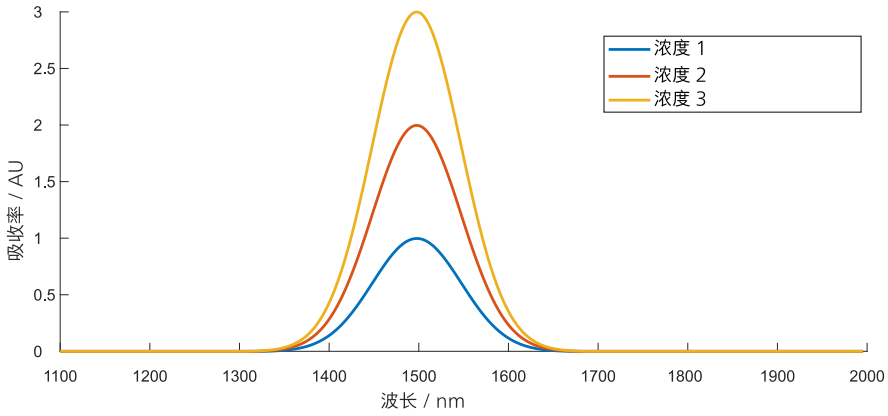


图 38　3 件具有一个 1,500 nm 波峰样品的模型化数据。

1,500 nm 时 3 个测得的吸光度值可以相对吸收器的浓度展现。

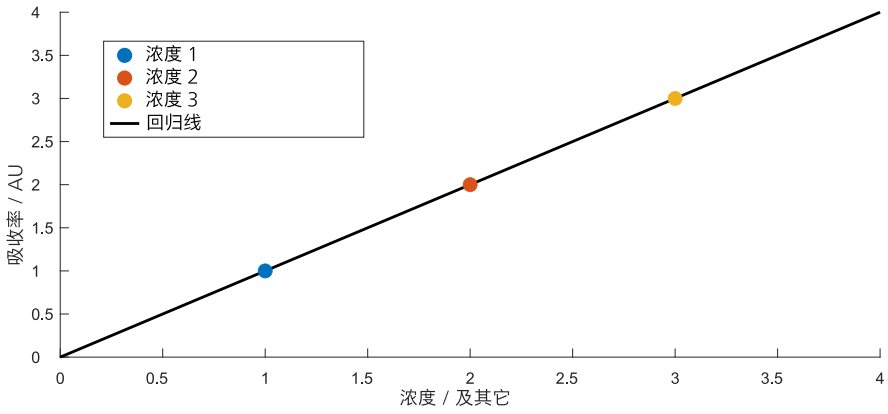


图 39　吸光度值和浓度之间的关系。

根据 Beer-Lambert 定律，吸光度值和浓度之间存在线性关系。通过一组 3 件样品可以执行线性回归，从而构成一条回归线。

因为只有 1 个变量（1 个波长），所以是指标变换线性回归。该回归可用作量化模型。但对于未知浓度的样品，在 1,500 nm 的波长条件下测量吸光度 A。由回归线得出吸收器的相应浓度 c：

$$c = bA$$

系数 b 是恒定的并且与回归线的斜率相同。

须注意所有样品都必须包含具有相同摩尔消光系数的相同的吸收器。此外，所有吸光度测量均必须在相同层厚度的情况下进行。

指标变换线性回归

实际的混合物包含一个以上的吸收器。记录的光谱为所有吸收光谱的总和 （参见章节 2.2.1，第 5 页）。

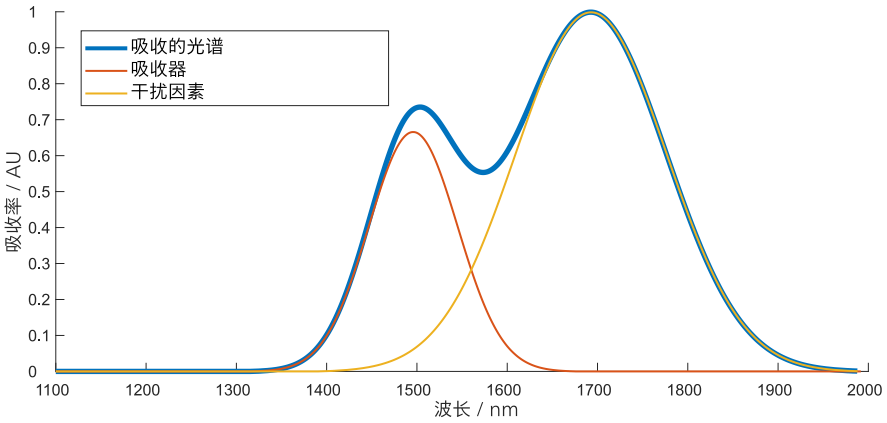


图 40　具有 2 个组分的模型化数据。吸收器（红线）应被量化。

记录的光谱（蓝线）为纯吸收光谱和重叠干扰光谱的和。波长 1,500 nm 的条件下，所测得的吸光度值不仅包含吸收器的吸光度，还包含干扰器的吸光度。分析参数无法通过 1,500 nm 条件下的唯一一次测量进行量化。无法得知是否存在干扰器以及测量是否可靠。

如果在 2 个波长（例如 1,500 nm 和 1,700 nm）条件下测量会发生什么？波长 1 测得的吸光度 A_1 是纯吸收器信号 A_1^{a}（角标 a = 吸收器）和纯干扰信号 A_1^{f}（角标 f = 干扰器）的和。这一点同样适用于波长 2 测得的吸光度，A_2：

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

其中，$\varepsilon_1^{\mathrm{a}}$ 和 $\varepsilon_1^{\mathrm{f}}$ 为吸收器和干扰器波长 1 的摩尔消光系数，c_{a} 和 c_{f} 为吸收器和干扰器的浓度。

上述方程中，层厚度 l 从 Beer-Lambert 中排除。这让后续的代数计算更加简单。层厚度当然必须对于所有样品保持相同。方程中的吸光度因此为每厘米的吸光度。

方程可以以矩阵形式说明：

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

因此：

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

求解吸收器浓度：

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

由此得出吸收器的浓度在存在干扰器的情况下也可以计算，方法是测量两个波长的吸光度并将每个吸光度乘以一个常数。

该常数为摩尔消光系数并且可以在表格中查找。但实际上从不会发生。实际上，浓度通过校正步骤和借助 PLS 回归这样的指标变换线性回归求解线性方程来确定。因此，常数被称作回归常数 b_1 和 b_2：

$$c = b_1 A_1 + b_2 A_2$$

2 个以上吸收器

如上所示，如果是 1 个吸收器，则测定 1 个波长的吸光度即可。如果是 2 个吸收器，则测定 2 个波长的吸光度即可。

这种方式可以归纳。多个吸收器需要测定不同的波长 i 的多个吸光度值 A_i。始终为线性关系：

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

开发量化模型

在上述方程可以预测未知样品中的浓度之前，必须测定系数 b_1、b_2 等。为此需要一个校正步骤。需要测量多个不同分析参数浓度的样品。

与之后用于 PCA 和 PLS 的项一致，可以将 c 用 y 替代并将 A 用 x 替代。然后，上述用于每件校正样品的方程如下：

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

其中，n 为样品数目，f 为波长的数量，y_1 为通过参考方法段（例如滴定）测得的样品 1 的参考值，$x_{1,1}$ 为波长 1 时测得的样品 1 的吸光度，$\{x_{1,1}\ldots x_{1,f}\}$ 为在波长 f 时测得的样品 1 的光谱。$b_1\ldots b_n$ 对应回归系数，$e_1\ldots e_n$ 相当于展示回归系数模型化测量数据效果的错误项。

更紧凑的矩阵形状得到如下结果：

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

X 被定义为矩阵 $f \times n$。$\mathbf{X}^\mathbf{t}$ 是转置矩阵 **X**，即行和列相互转置，从而得到上面的矩阵 $n \times f$。预测向量 **p** 相当于上述回归系数 **b**。

回归向量 **p** 能够实现对一个新样品基于其光谱预测分析参数 **x**。计算值 $\hat{y}$ 为：

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

指标变换线性回归的任务是测定得到最小错误项的回归系数。但多元线性回归（MLR）比波长数目要求的校正样品数目更多。另一个障碍是变量之间的高相关性。

其它方法也可用于光谱预测。通过 PCA 可以大大降低数据量并完全消除相关性。通过一次 PLS 回归将另行考虑样品的参考值。

6.2 PCA 算法

主要成分分析（**PCA**，英文：*principal component analysis*）将用于自动数据划分和识别光谱离群值 *（参见章节 4.2，第 28 页）*。

对于参数化设置和平均值集中光谱的主要成分分析，OMNIS Software 执行 **奇异值分解** (**SVD**，英文：*singular value decomposition*）。奇异值分解将原始的数据矩阵 **X**（校准数据集的光谱）分解为 3 个矩阵并计算 n 个主要成分：

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

其中，**X** 对应原始光谱数据（一个矩阵 $f \times n$ 带 f 个波长和 n 件样品），**L** 为载荷（一个矩阵 $f \times n$），Σ 是奇异值（一个矩阵 $n \times n$），**S** 代表得分（一个矩阵 $n \times n$）。方程可以以图形形式展现 *（参见图 41，第 73 页）*。

载荷 **L** 将波长空间的轴投影到主要成分空间的轴上。**L** 的列向量是主轴或主方向。

得分 **S** 是原始光谱在主轴上的投影。**S** 的每个行向量包含特定的光谱。

Σ 为奇异值的对角矩阵 σ_i。奇异值描述了通过每个主要成分声明的方差。通过约定对矩阵进行排列，使得 $\sigma_1 > \sigma_2 > \ldots > \sigma_n$。

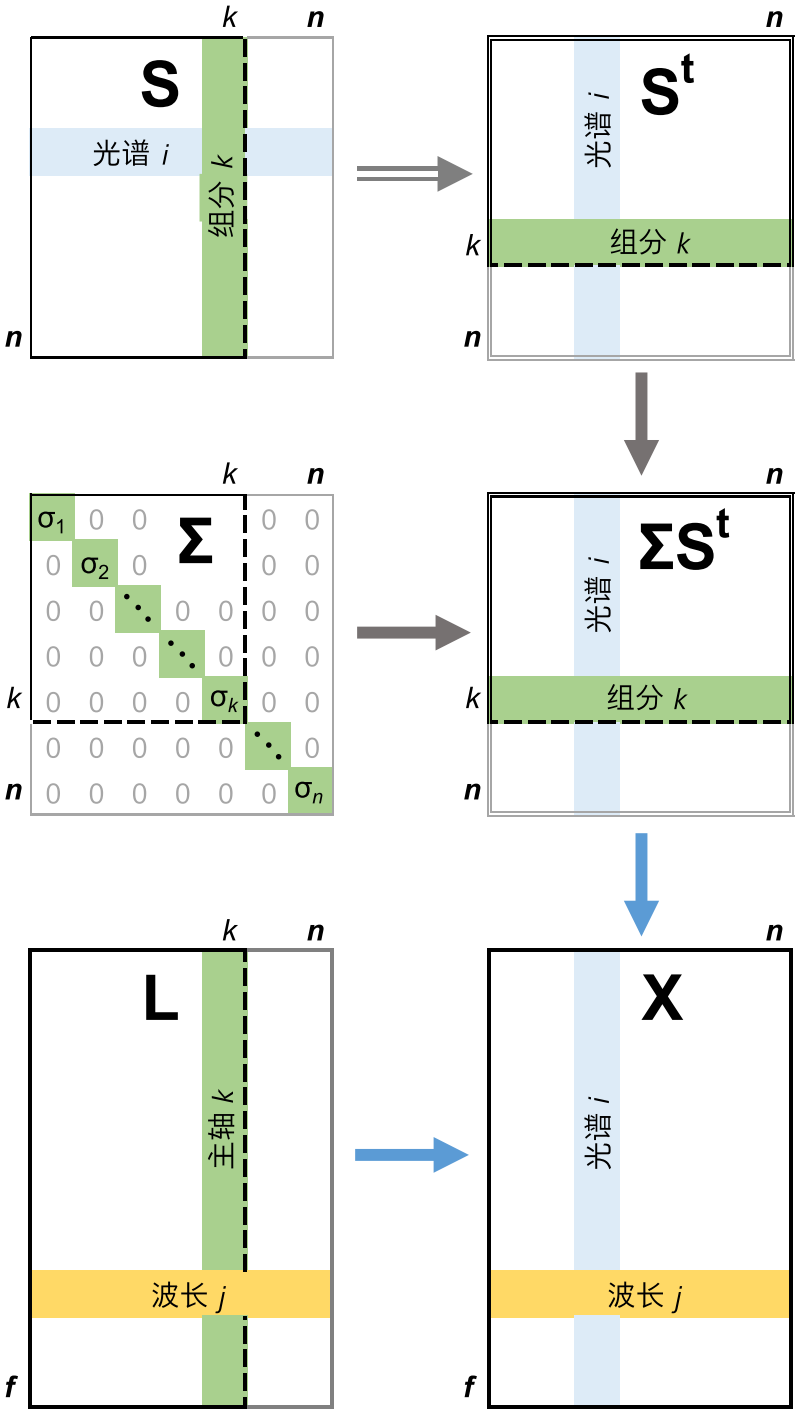


图 41　图形形式的奇异值分解方程。虚线展示了一个带 k 主要成分的模型的缩减矩阵。从 n-k 流出的主要成分信息流入残余矩阵（一个 f × n 矩阵，未描述）。

残余矩阵

一个 PCA 模型仅使用所计算的 *n* 主要成分的第一个对。*图 41* 中展示了 *n* 个主要成分中的前 *k* 个。原始数据 **X** 可以划分为由模型描述的数据以及模型未描述的数据：

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

其中，**X** 代表原始的光谱数据（一个矩阵 $f \times n$），$\mathbf{L_a}$（一个矩阵 $f \times k$）对应 **L** 的前 *k* 列，$\mathbf{\Sigma}_a$（一个对角矩阵 $k \times k$）是前 *k* 个奇异值，$\mathbf{S_a}$（一个矩阵 $n \times k$ 包含 *k* 个主要组分）为 **S** 的前 *k* 列，**E** 是参与数列（$f \times n$），包含无法由模型描述的所有光谱变异在 **X** 中。

通常为 $k \ll n \ll f$。例如 $k = 3$ 个主要成分，$n = 100$ 个样品，$f = 2{,}500$ 波长。

残余矩阵 **E** 的每列 $\mathbf{e}_i$ 显示光谱 *i* 与 PCA 空间的正交距离，称为 **残差**。模型使用的主要成分越多，残差越小。

6.3 PLS 算法

偏最小二乘回归（**PLS 回归**，英文：*partial least squares regression*）用于计算量化模型 *（参见章节 4.4.1，第 47 页）*。

PLS 引用两个数据组块：

- 参数化设置和平均值集中的矩阵 **X**（光谱）。
- 平均值集中的 **y** 向量（参考值）。

PLS 将矩阵 **X** 分解为 2 个矩阵：

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

其中，**X**（一个矩阵 $f \times n$，*f* 个波长和 *n* 件样品）相当于经过预处理并且平均值集中的光谱，**L** 对应载荷（一个矩阵 $f \times k$，*k* 个潜变量），**S** 为得分（一个矩阵 $n \times k$），**Z** 是残余矩阵（一个矩阵 $f \times n$），其中包含无法由模型描述的所有光谱变异在 **X** 中。

PCA 中，得分矩阵 **S** 说明了 **X** 的方差，而 PLS 中，得分矩阵 **S** 说明了 **X** 和 **y** 之间的协方差。PLS 通过得分将说明的协方差最大化。这就意味着，得分不仅能够以最佳方式说明 **X** 的方差，而且具有与参考值最大程度的相关性。

为了最大化 **X** 和 **y** 之间的协方差，PLS 算法交换 **X** 和 **y** 之间的数据。因此，**X** 和 **y** 融合为一个唯一的系统。其中，得分 **S** 相对参考值 **y** 回归，以获得回归系数 **b**：

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

其中，**e** 是包含在 **y** 中无法由模型描述的所有参考值变异的残差向量。

预测

由回归系数 **b** 可以确定预测向量 **p**。对于一件新样品的分析参数 $\hat{y}$ 的预测将采用预测向量 **p** 和经过预处理及平均值集中的光谱 **x**：

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

OMNIS Software **通过** SIMPLS **算法和唯一一组参考值**（PLS-1）**执行** PLS **算法**。

6.4 Hotellings T² 和 Q 检验残差

Hotellings T^2 和 Q 检验残差通过一个 PCA 或 PLS 模型体现光谱的特征。这样能帮助对可能的离群值进行身份验证 *（参见"Hotelling T² 和 Q 检验残差"，第 40 页）*。

Hotelling T²

马氏距离是光谱与模型中点偏差大小的标度。距离将被标准化。每个主要成分或潜变量均包含相同的权重。

设定光谱或得分正常分布，平方的码数距离 MD² 则符合 Hotellings T^2 分布：

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

光谱 *i* 的平方马氏距离是前 *k* 个主要成分或潜变量的平方和：

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

其中，$\mathbf{s_i}$ 是缩减的得分矩阵 **S** 的第 *i* 行，$s_{i,a}$ 对于光谱 *i* 和主要成分 *a*（或潜变量）的标准化得分，*k* 为所使用的主要成分或潜变量的数量。

MD² 可被称为 T^2。$T^2 = 0$ 的光谱投影在平均值集中的模型的中点，即所有得分位于中央。光谱在超平面上的投影距离中点越远，T^2 数值越大。模型在中点附近最佳。模型在远离中点可能表现不佳。

可以确定一个识别离群值、用于假设检验的 **显著性水平**（例如 5%）*（参见章节 6.5，第 76 页）*。

Q 残差

光谱 *i* 的 Q 残差是 PCA 或 PLS 模型距离光谱的平方距离：

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

其中，$\mathbf{e_i}$ 相当于残余矩阵 **E** 的第 *i* 列，$e_{i,j}$ 对应光谱 *i* 的残差，*f* 是波长的数量。

Q 检验残差显示了无法通过模型说明的变异。Q 残差表示光谱可能与模型不一致，例如 当测得的样品包含另一种物质时。

可以确定一个识别离群值、用于假设检验的 **显著性水平**（例如 5%）*（参见章节 6.5，第 76 页）*。

6.5 光谱离群值 - 算法

识别光谱离群值能够识别与总体不同的光谱 *（参见"识别光谱离群值"，第 40 页）*。算法评估被检验光谱的 Hotellings T^2 或 Q 残差的数值是否为一个随机或系统性变异的结果。

模型开发时识别光谱离群值

1. 按如下方式考虑参数化：
   a. OMNIS Software 版本 4.2 以上：用户决定是否应用参数化（数据预处理和波长选择）。
   以后对参数化的更改对数据组分划没有影响。
   b. OMNIS Software 版本 3.3 以上至 OMNIS Software 版本 4.1：用户决定是否考虑数据预处理。波长选择和以后对数据预处理的更改对数据组分划没有影响。
   c. OMNIS Software 版本 3.2 以下：按照识别离群值时确定的方式考虑数据预处理。波长选择和以后对数据预处理的更改对数据组分划没有影响。
2. 光谱离群值的识别基于光谱表中所有平均值集中光谱的 PCA 模型*（参见章节 4.2，第 28 页）*。须测试的光谱同样被记录在 PCA 模型中。主要成分的数量选择须考虑到所设定的方差至少应为 95%。
3. **Hotelling T^2离群值**：
   基于 Hotellings T^2 数值 *（参见"Hotelling T²"，第 75 页）* 执行假设检验，依据 H. Hotelling, *The Generalization of Student's Ratio*, The Annals of Mathematical Statistics Band 2, No. 3 (Aug. 1931), P. 360–378。
   a. 零假设是指须分析的光谱的 T^2 数值符合 PCA 模型的 T^2 数值。如果零假设是真实的，则 T^2 符合 Hotellings T^2 分布。分布可以作为标度化的 F 分布展现：
   $$T^2 \sim \frac{k(n-1)}{n-k} F_{k,n-k}$$
   其中 k 是主要成分的数量，n 是光谱的数量，$F_{k,n-k}$ 是 F 分布（包含 k 和 n-k 参数）。
   b. 可设置的显著性水平能够操控零假设为真时被拒绝的概率（称为 1 类型错误）。标准值为 5 %。
   c. 基于该分布和显著性水平计算 T^2 的危险数值。
   d. 如果光谱的 T^2 数值大于危险数值，零假设将被拒绝，光谱将被标示为可能的离群值。

4. **Q 检验残差离群值**
   基于 Q 检验残差 *（参见"Q 残差"，第 75 页）* 执行假设检验，依据 J. E. Jackson and G. S. Mudholkar, *Control Procedures for Residuals Associated With Principal Component Analysis*, Technometrics Band 21, No. 3 (Aug. 1979), P. 341–349。
   a. 零假设是指须分析的光谱的 Q 检验残差符合 PCA 模型的 Q 检验残差。如果零假设为真，则可以创建一个接近正常分布的 Q 检验残差测试统计。
   b. 可设置的显著性水平能够操控零假设为真时被拒绝的概率（称为 1 类型错误）。标准值为 5 %。
   c. 基于该测试统计和显著性水平计算 Q 的危险数值。
   d. 如果光谱的 Q 检验残差大于危险数值，零假设将被拒绝，光谱将被标示为可能的离群值。

预测的离群值评估（量化）

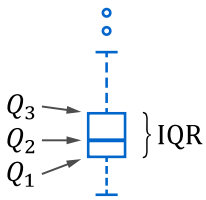
量化时，预测可使用离群值评估：

1. 准备步骤：
   a. 量化模型采用上述相同的算法。但基础是包含校准数据集的所有平均值集中光谱的量化模型（PLS 模型）。数据预处理、波长范围和潜变量的数量将如量化模型中所示予以考虑。
   b. 量化模型基于确定的显著性水平计算 T^2 和 Q（PLS 影响图中的虚线）的危险数值。危险数值将保留在量化模型中。
2. 预测时，光谱的 T^2 和 Q 数值基于量化模型计算。
3. 如果光谱的 T^2 或 Q 数值大于相关的临界数值，样品根据所应用的量化模型被视为可能的离群值。

6.6 离群值参考值 - 算法

箱线图能够识别参考值的离群值 *（参见章节 4.3.4，第 45 页）*。

为了考虑分布的倾斜，离群值极限是将通过以下计算调整。



Medcouple (MC) 测量参考值的倾斜。计算由箱线图的中位数，Q_2。函数通过所有参考值的上半部分和下半部分中的对（英文：*couples*）计算。结果的中位数为 Medcouple：

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

其中，Q_2 是定义箱线图中线的第二个四分位，y_i、y_j 是一对参考值。

Medcouple 始终在 –1 和 1 之间。对称分布时，MC = 0。MC > 0 的倾斜分布被扭曲为更高的参考值，MC < 0 将扭曲为更低的参考值。

调整的离群值极限值 的计算取决于分布被推向哪一侧：

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

对称分布 (MC = 0) 时，极限值和箱之间的距离为 1.5 IQR。

对于不同倾斜度的不同分布，幂函数能够实现准确而稳固的离群值识别，参阅 M. Hubert und E. Vandervieren, *An adjusted boxplot for skewed distributions*, Computational Statistics & Data Analysis Band 52, No. 12 (Aug. 2008), P. 5186–5201。

所期待的被标记离群值的百分比约为 1% 并且与正常分布的标准箱线图的百分比相近。注意：该百分比与显著性水平无关。