B&W TEK OEM 应用说明 101

用于拉曼定量分析的光谱预处理

技术说明

简介

拉曼光谱学因其可以无损测量、快速分析以及可以同时进行定性和定量分析的能力，已经成为制药和化学工业过程分析中日益普遍的技术。光谱预处理算法通常应用在定量光谱数据分析中，是为了在加强光谱特征的同时尽可能地减少与所讨论分析物无关的变异性。对于没有化学计量学专业背景的普通人来说，理解预处理步骤可能性以及知道如何正确应用它们可能就会令人怯步。本文的目的是通过实际应用的例子来讨论与拉曼光谱有关的主要预处理方案，并复习 B&W Tek 和 Metrohm 软件中可用的算法，以便读者能够自如地应用它们来建立拉曼定量的模型。

拉曼数据的光谱预处理

光谱预处理用于消除或尽量减少在光谱数据中的一些影响，这些影响与所研究系统相关的光谱变化并不直接相关。预处理还可用来提高细微光谱差异的区分能力，如小峰强度或光谱偏移。

让我们来探讨一些与拉曼数据较相关的光谱预处理步骤。

去除基线

基线去除或基线校正（在 Vision 软件中被称为去除多项式趋势）可用于去除拉曼数据中变化的背景，如荧光或有干扰的环境光，这在光谱仍然有清晰的拉曼峰时特别有用。有许多巧妙的数学方法可用来去除基线，普遍会用最小二乘法函数对多项式拟合来描述光谱的基线，然后从光谱中减去该函数。**图 1** 举例显示了在 BWSpec 软件中对碳黑粉末光谱进行基线校正。碳样品有着变化的背景，在进行进一步的光谱分析之前需要进行基线校正，例如计算 D 带和 G 带的强度比。

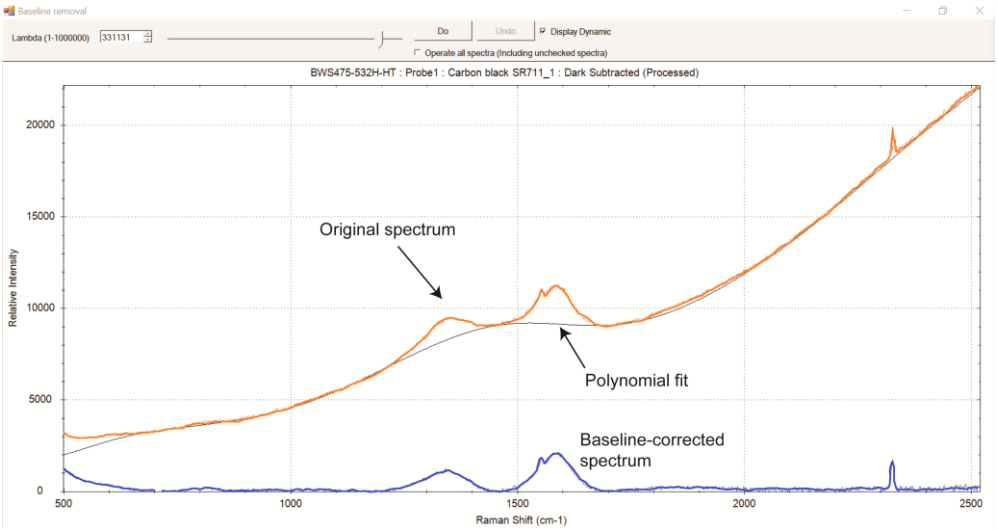

图 1. 基线校正后碳黑样品的拉曼光谱。操纵上方滑杆可改变多项式拟合（为了清晰显示，光谱已被人为地加粗）。

WSpec、BWIQ 和 Vision 软件中都有去除基线的算法。在对具有不同基线的大型数据集使用基线校正时应谨慎，因为一种拟合可能无法对数据集中的所有光谱进行优化。相反，推选对定量数据集使用求导来去除基线的影响。

求导

对拉曼和近红外数据使用求导也是常见的预处理步骤。求导应用在光谱数据中可以增强光谱特征和消除基线的影响。通常使用一阶和二阶求导，因为高阶求导会放大不必要的噪声。在 B&W Tek 的软件和 Vision 软件中有多种求导方法，但到目前为止，常用 Savitzky-Golay 求导法对拉曼数据进行处理。

Savitzky-Golay 求导法是通过对一段小区间数据点进行多项式拟合，然后该函数的求导是根据中心点附近的数据段计算的。该段通常被称为 "窗口 "尺寸。通常会使用大的窗口尺寸，因为小的窗口尺寸会产生更多的噪声数据，会对拉曼位移的微小变化更加敏感。**图 2** 显示对同一组数据，取两个不同窗口尺寸进行 Savitzky-Golay 一阶求导后的对比；具有较大窗口尺寸的光谱比具有较小窗口尺寸的光谱显示出更小的噪声。

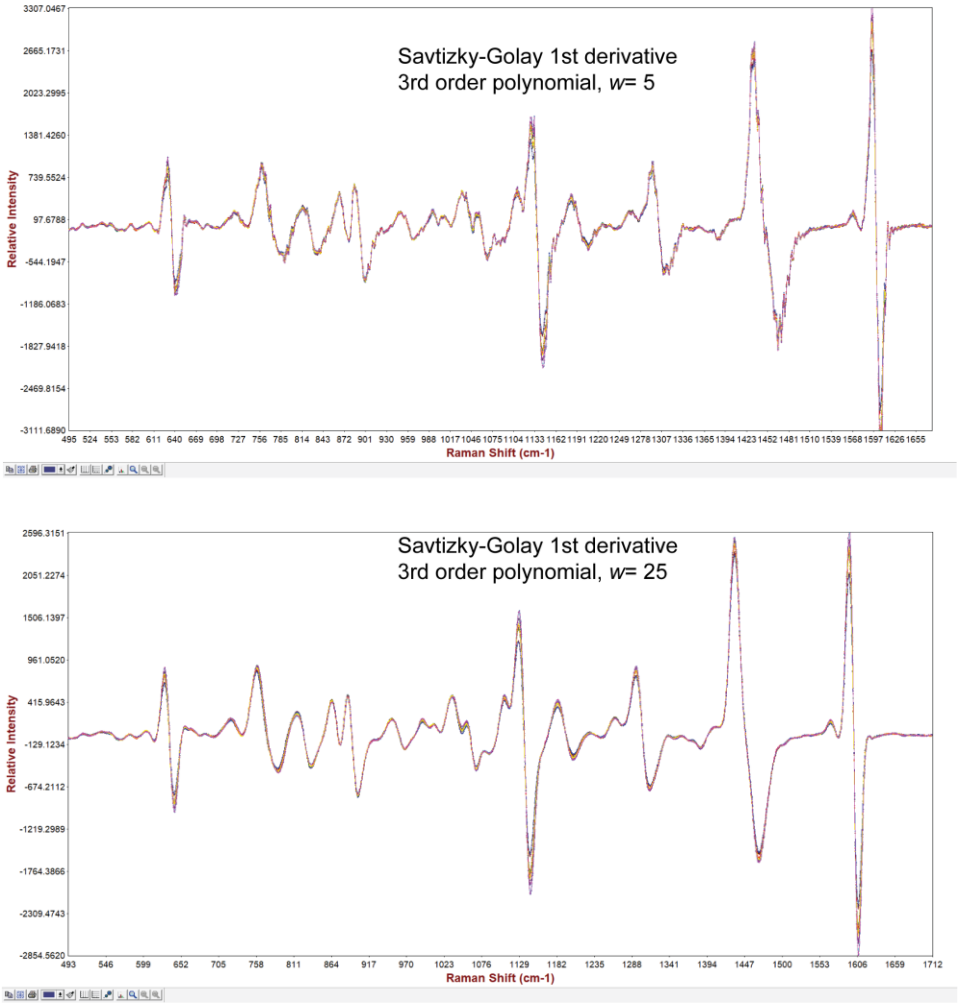

图 2.（上图）用 SG 一阶求导处理的数据，窗口=5；（下图）用 SG 一阶求导处理的数据，窗口=25。该光谱显示的是杀虫剂中的乳化剂样品。

区域选择

可以用特定的光谱区域建立模型，以排除那些信息量小或不相关的变化区域；这会使模型更简单，潜在变量也更少。在 BWIQ 和 Vision 软件中，可以手动选择光谱区域。在通常情况下，选择整个指纹区域（约 200-1800cm^{-1}）就足以建立拉曼模型。经验丰富的光谱分析人员可以选择更多的定制区域，但在使用多元回归法时，一个常见的错误是选择的区域太窄，该区域只包含了与感兴趣分析物对应的特征。而分析物浓度的确定需要对分析物和参照物（含特征值外所有数据）进行量化，所以如果只包括分析物的特征，模型就缺乏参照物，并且可能变得不稳定。如果还使用了归一化步骤，这种不稳定的情况就会更明显。

举个例子，考虑一个简单的偏最小二乘法（PLS）模型来量化苯腈与环己烷的混合物。从 6 个样品的 12 个光谱组成得到该模型，苯甲腈的浓度从 10%到 35% v/v 不等。如**图 3** 所示，苯甲腈在 2232 cm^{-1} 处有一个较强的峰，这是由氰基-CN 拉伸强度引起的。

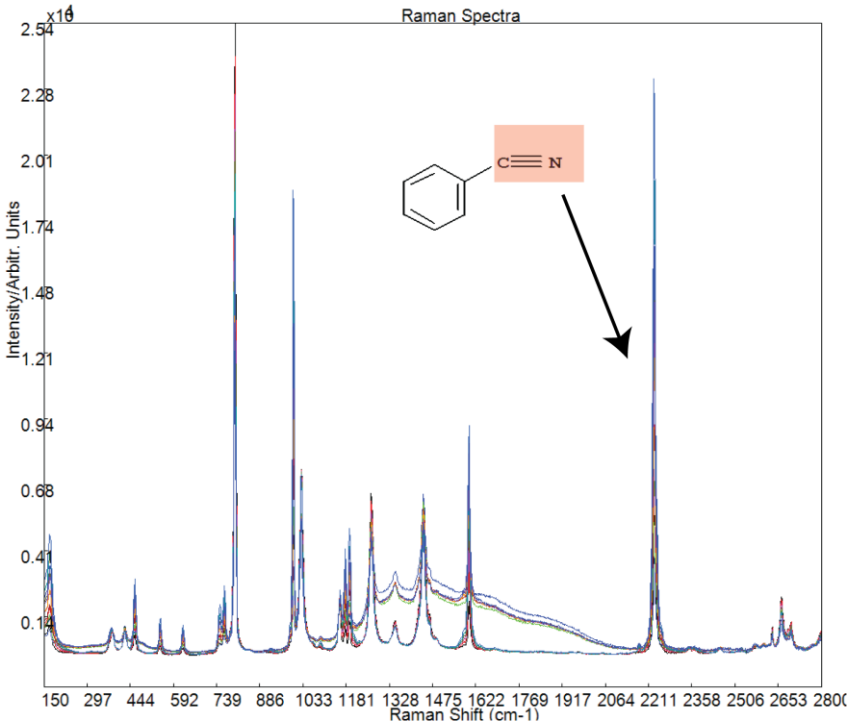

图 3. 苯甲腈和环己烷混合物的原始光谱。

图 4 对比显示对数据使用两种创建的 PLS 模型后对应的预测与测量图。如果选择 300-2300 cm^{-1} 的区域（**图 4a**），可以得到一个具有良好线性的模型（均方根误差 RMSE 为 0.21%）。另一方面，如果选择 2150-2300 cm^{-1} 的狭窄区域（**图 4b**），该模型的线性度就较差，均方根误差 RMSE 则高达 1.64%。后一个模型的性能差原因是缺乏参照物，导致归一化后所有光谱的光谱强度相同。

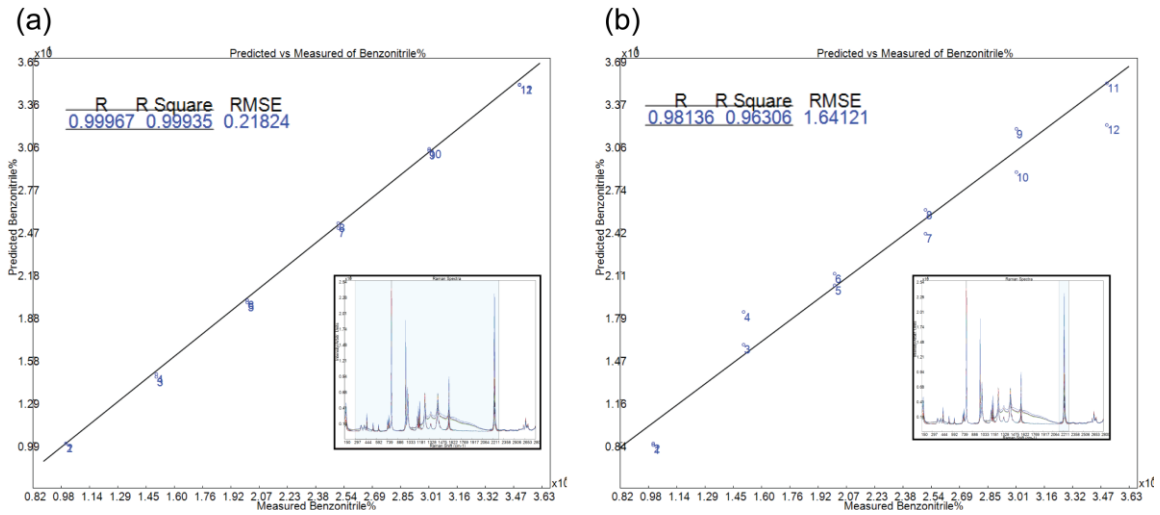

图 4.（a）选择 300-2300 cm^{-1} 的 PLS 模型和（b）选择 2150-2300 cm^{-1} 的 PLS 模型。

归一化

定量拉曼的模型很大程度会受到整体光谱强度波动的影响。强度波动可能来自许多因素，如光谱仪的光通量的漂移、激发功

率的不稳定、光程长度的差异和采样点位置的物理差异。尽管在短期内这些波动是可以避免的，但从长期来看它们是不可避免的。光谱归一化可以有效地消除整体强度变化的影响，因此是搭建稳健回归模型的一个必要步骤。

在光谱学中，有许多不同的数学方法可用于归一化。标准正态变量变换（SNV）和多元散射校正（MSC）是振动光谱学中常见的两种归一化算法，在 BWIQ 和 Vision 软件中可使用。与 MSC 相比，光谱学家倾向于 SNV，因为 MSC 是基于整个数据集的平均值的散射校正，而 SNV 是基于单个样品光谱的标准差，它不依赖于整个数据集。

图 5 显示的光谱为含有不同量葡萄糖和乳酸的水溶液。将使用 SNV 的数据与基线不断变化的非归一化数据进行比较（插图）。在归一化之前需完成区域选择，这样就不用考虑被排除的区域了。

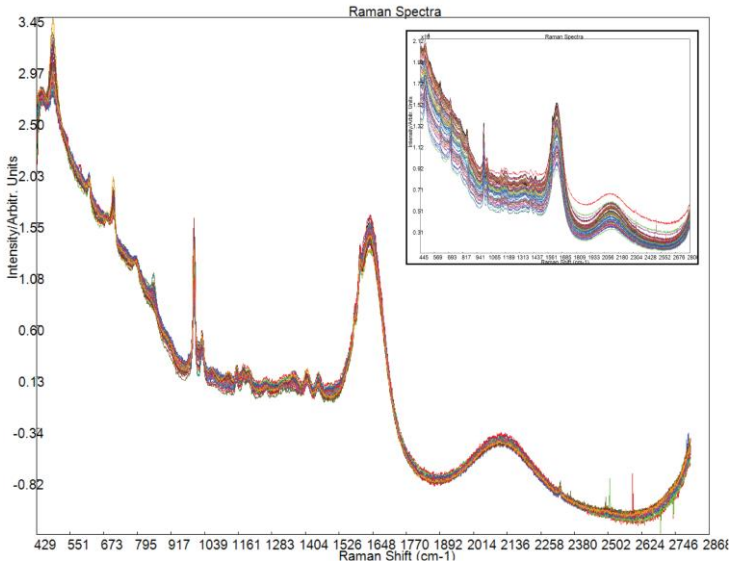

图 5. 使用 SNV 后葡萄糖和乳酸的水溶液的拉曼数据集(插图显示的是原始非归一化的数据)

中心化处理

中心化处理是从将每个光谱减去数据集的平均光谱。这是得到基于 PLS 和 PCA 模型的必要步骤，因为这两种技术都是分析数据集的方差。BWIQ 软件中有一个单独的步骤可进行中心化处理，因此在预处理步骤中明确已包含，而 Vision 软件中的中心化处理为已暗含训练光谱。

实际应用例子

我们用上一节学到的信息来检查一个实际应用的例子。这个数据是用一个透射拉曼装置收集得到的。样品是一组 3.0mm 厚的药片，其中含有低剂量的对乙酰氨基酚（又称扑热息痛，APAP），以及纤维素、甘露醇、交联羧甲纤维素和硬脂酸镁等辅料。对乙酰氨基酚的浓度范围为 0-1.5%（w/w），目标浓度 0.5%，对应的目标剂量为每片~300mg 的药片中含有乙酰氨基酚 1.5mg。为了建立一个能预测新样品的模型，使用 3 秒的积分时间和 10 个光谱平均数来收集校准光谱。图 6 显示了原始数据；除了暗减法和相对强度校正外，没有应用其他预处理。光谱被导入 BWIQ 进行处理并建立 PLS 模型。

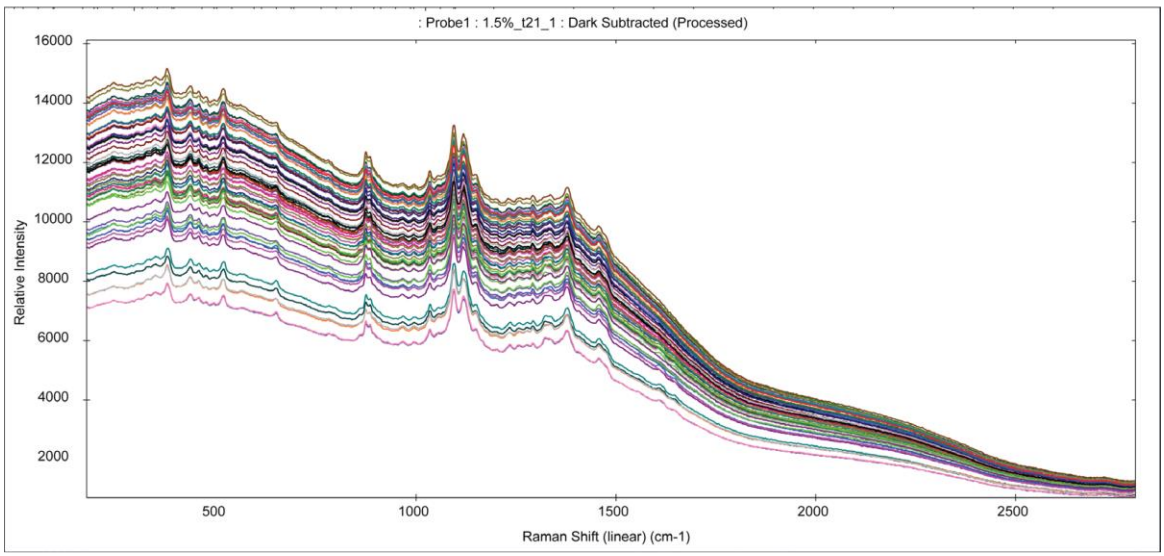

图 6. 含有 0-1.5% w/w 对乙酰氨基酚药片的光谱，并且图谱为原始光谱

在收集数据后，将样品光谱与组成样品的各个纯成分的光谱进行比较是非常有用的。**图 7** 显示并比较了含有 1.5%的对乙酰氨基酚的样品与纯对乙酰氨基酚、纤维素和甘露醇的光谱（交联羧甲纤维素和硬脂酸镁的特征峰太宽或太弱，无法直观地区分）。样品光谱中标记为绿色的峰是由样品中的对乙酰氨基酚贡献的。

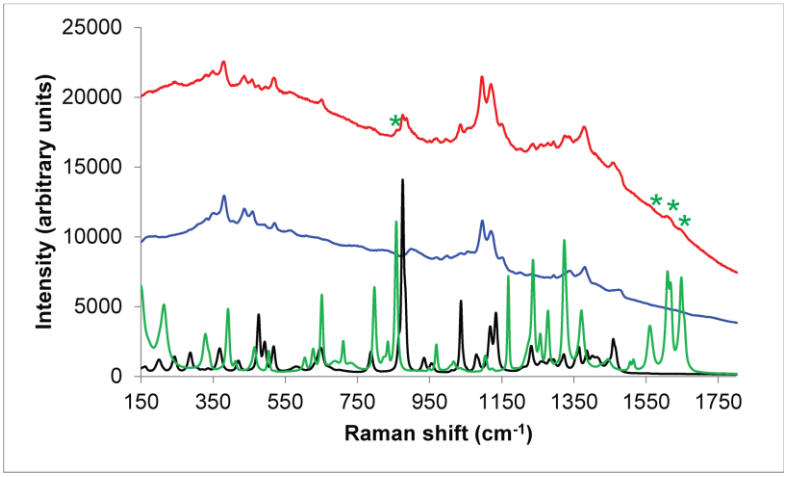

图 7. 比较 1.5% APAP 药片样品（红色）、纤维素（蓝色）、纯对乙酰氨基酚（绿色）、甘露醇（黑色）的光谱

预处理步骤

片剂中的辅料在 785nm 的激光激发下会产生较高的荧光背景。在这种情况下不建议用基线校正来消除荧光背景，因为一元多项式拟合不可能适合每个光谱。相反，Savitzky-Golay 求导法是一个更简单的方法。对数据使用 Savitzky-Golay 一阶求导（立方阶，w=25）可消除荧光背景。

当检查光谱时，我们可以看到在指纹区有明显的拉曼信号，而在 1800cm^{-1} 以上则没有明显的信号。为了在模型中排除 1800-2800 cm^{-1} 这段不重要的光谱区域，我们可以进行手动区域选择，设置模型只选择指纹区域（约 200-1800 cm^{-1}）。当在

BWIQ 软件中执行手动区域选择步骤时，软件将总是会回到归一化步骤之前。**图 8** 显示了经过 Savitzky-Golay 一阶求导、手动区域选择和 SNV 处理的数据。在~860 cm^{-1}和~1500 cm^{-1}处的信号显示出了明显的强度变化，这与对乙酰氨基酚浓度的增加相对应。

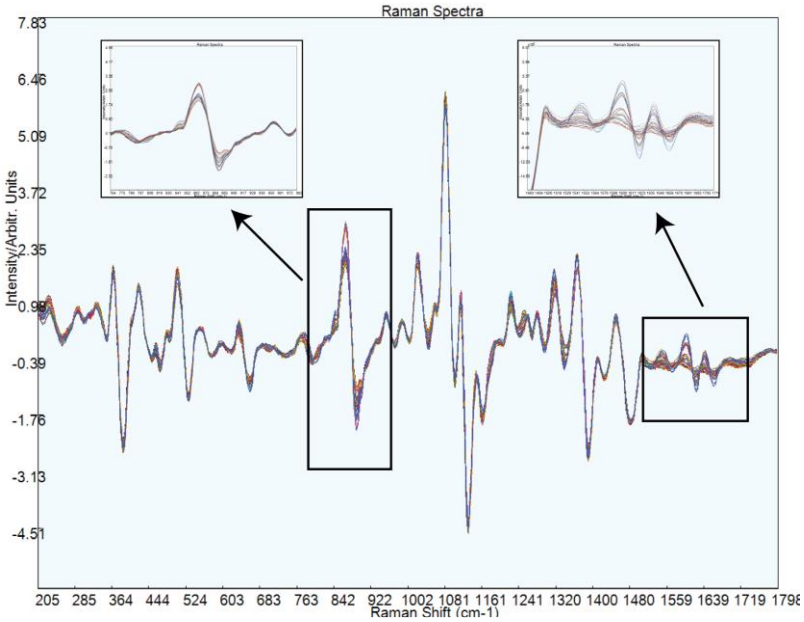

图 8. 用 S-G 一阶求导、手动区域选择和 SNV 处理的光谱。光谱显示，随着对乙酰氨基酚浓度的增加，位置~860 cm^{-1}和 1500-1650 cm^{-1}的信号显示出了明显的强度变化，这信号的位置与对乙酰氨基酚的拉曼特征峰一致。

在 BWIQ 软件中，中心化处理是作为一个单独的步骤使用的（在 Vision 软件中是自动完成的）。当中心化处理被使用时，光谱会以零线为中心。图 9 显示了使用了包含中心化处理在内的所有预处理算法的数据集。经过处理的数据现在就适合用来搭建一个稳健的 PLS 模型。

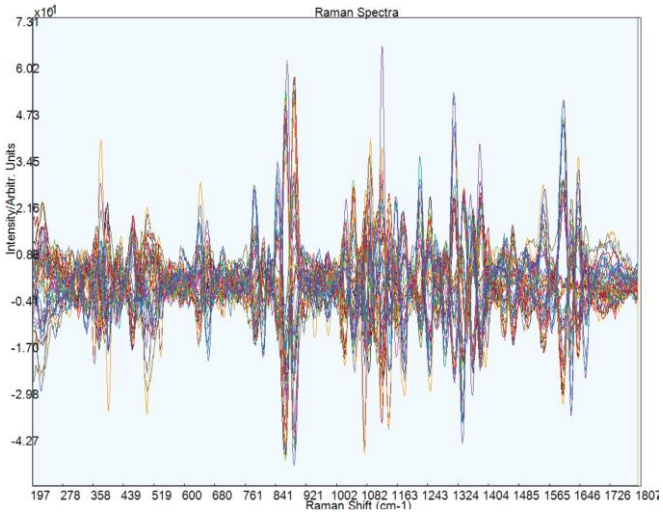

图 9. 预处理后的数据集，包含了中心化处理

表 1 显示了模型中会使用的预处理步骤，以及它们对应的目的。通过执行这些预处理步骤，即使是拉曼光谱学的初学者和没有敏锐的化学计量学意识的人，通常也能搭建一个稳健的模型。

表 1。预处理步骤用于准备示例模型和每个步骤的目的。

使用的预处理算法	目的
Savitzky-Golay 一阶求导（三阶，w=25）	去除荧光背景
手动选择区域	将感兴趣的光谱区域隔离并去除无信息的信号
标准正态变量变换	根据强度波动对光谱归一化处理
中心化处理	从数据集中去掉平均值

参考文献

1. J. Huang, S. Romero-Torres and M. Moshgbar. American Pharmaceutical Review. 13, 116-127 (2010).

2. J.M. Shaver. Chemometrics for Raman Spectroscopy. In Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line; I.R. Lewis, H.G.M. Edwards, Eds.; Marcel Dekker, Inc.: New York, 2001; Vol. 28, pp 275-306.

3. M.J. Pelletier. Appl. Spectroscopy. 57, 20A-42A. (2003) https://doi.org/10.1366/000370203321165133

4. Vision software user manual

5. BWIQ software user manual

6. QTRam for Content Uniformity Analysis- A Simple Demonstration. Internal B&W Tek reference document 400000352-B

7. B&W Tek, LLC (2019). QTRam® for Content Uniformity Analysis of Low-Dose Pharmaceutical Tablets (Application Note 410000046), https://bwtek.com/appnotes/qtram-for-content-uniformity-analysis-of-low-dose-pharmaceutical-tablets/

Analytes:	Carbon
	Pesticides
	Biopharmaceuticals
Matrix:	Tablets, capsules,
	pharmaceutical powders
Method:	Raman Spectroscopy
Industry:	Chemical
	Food & Beverage
	Pharmaceuticals

B&W TEK OEM Application Note 101, Version 1 – 2023-05