

smund Rinnan, Frans van den Berg, Søren Balling Engelsen

摘要：预解决在近红外(NIR)光谱数据解决化学计量学建模中已经成为不可分割一某些。预解决目的是消除光谱中物理现象在为了提高后续多元回归、分类模型或摸索性分析。最广泛使用预解决技术可以分为两类：散射校正办法和光谱间隔办法。综述和比较了算法基本理论和当前预解决办法以及定性和定量后果应用程序。其目的是提供更好 NIR 最后模型建立，在此咱们通过对光谱预解决基本知识进行梳理。

核心词：乘法散射校正；近红外光谱法；原则化；诺里斯威廉姆斯推导；预解决；Savitzky-Golay 平滑；散射校正；光谱导数；原则正态变量；综述

1. 引言

当前为止，没有可以优化数据来进行代替，但是通过恰当数据收集和解决将会起到优化效果，对光谱数据进行预解决是最重要一步(例如最优化之前叠层建模)，惯用办法有主成分分析(PCA)和偏最小二乘法(PLS)。在大量文献中，多变量光谱应用食品、饲料和医药分析，比较不同预解决成果研究模型预测成果是不可分割构成某些。近红外反射/透射率(NIR / NIT)光谱光谱技术，到当前为止最多被使用和最大多样性在预解决技术，重要是由于入非线性光散射光谱可以引起明显影响。由于类似规模波长电磁辐射和粒子大小近红外光谱在生物样品，近红外光谱技术是一种不被广泛使用是由于存在散射效应(涉及基线转变和非线性)，这将会影响样品光谱成果记录。然而，通过应用适当预解决，可以很大限度上消除这些影响。

在应用研究中，比较了几乎完全不同定标模型(定量描述符和相应关系)。几乎没有浮现评估差别和相似性报道。代替技术即修正含义(例如，谱描述符数据)在研究中很少被讨论。本文旨在讨论建立了预解决办法对近红外光谱和模型之间关系，更详细地说，这些技术都是相应独立响应变量，因此咱们只讨论办法，不需要一种响应值。咱们同步关注预解决工艺理论方面和实际效果，这种办法合用于近红外光谱/ NIT 光谱。

对固体样品，干扰系统差别重要是由于光散射不同和有效途径长度不同。这些不受欢迎变化经常构成了样本集总变异重要某些，可以观测到得

转变基线(乘法效应)和其他现象称为非线性。普通来说,近红外光谱反射率测量一种示例将测量普及性反映和镜面反射辐射(镜面反射)。镜面反射普通由仪表设计和几何采样最小化,由于它们不含任何化学信息。这个 diffusively 反射光,这反映在广泛方向,是信息重要来源在近红外光谱。然而, diffusively 反射光将包括信息化学成分不但示例(吸取)并且构造(散射)。重要形式光散射(不涉及能量转移与样品)瑞利和洛伦兹米氏。两者都是过程中电磁辐射是分散(例如,通过小粒子,泡沫,表面粗糙度,水滴,晶体缺陷, microorganelles、细胞、纤维和密度波动)。

当粒子尺寸不不大于波长,由于普通状况下, NIR 光谱,是重要洛伦兹米氏散射。相比之下,瑞利散射,是各向异性,洛伦兹米氏散射依赖形状散射粒子和不强烈波长依赖性。

对生物样品,散射特性是过于复杂,因此软或自适应补偿,光谱预解决技术,正如咱们近红外光谱在本文中进行讨论,规定删除散射从纯粹、抱负吸取光谱。

显然,预解决不能纠正镜面反射率(直接散射),自谱不包括任何精细构造。光谱重要由镜面反射率应当总是被移除之前为离群值多元数据分析,由于她们仍将是局外人,甚至在预解决。图 1 显示了一组 13 好蔗糖和样品不同粒径加一坏蔗糖例子展示如何(极端)镜面反射率体现比正常光谱。

图 1 还演示了总体布局大多数数据在本文中。上部图,一种条形图显示了主成分得分值第一主成分(PC)后样本集数据意味着定心[1]。下面某些显示预解决效果数据集(或者,在这种状况下,任何预解决)。有关系数 r 平方值之间酒吧和一种选定参照变量包括(在本例中,已知平均粒径 13 蔗糖样本)。蔗糖数据集,这种关系应当是低,例如,当假设散射是一种阻碍粒子始发;尽量小信息粒度应当保持在对的预解决。

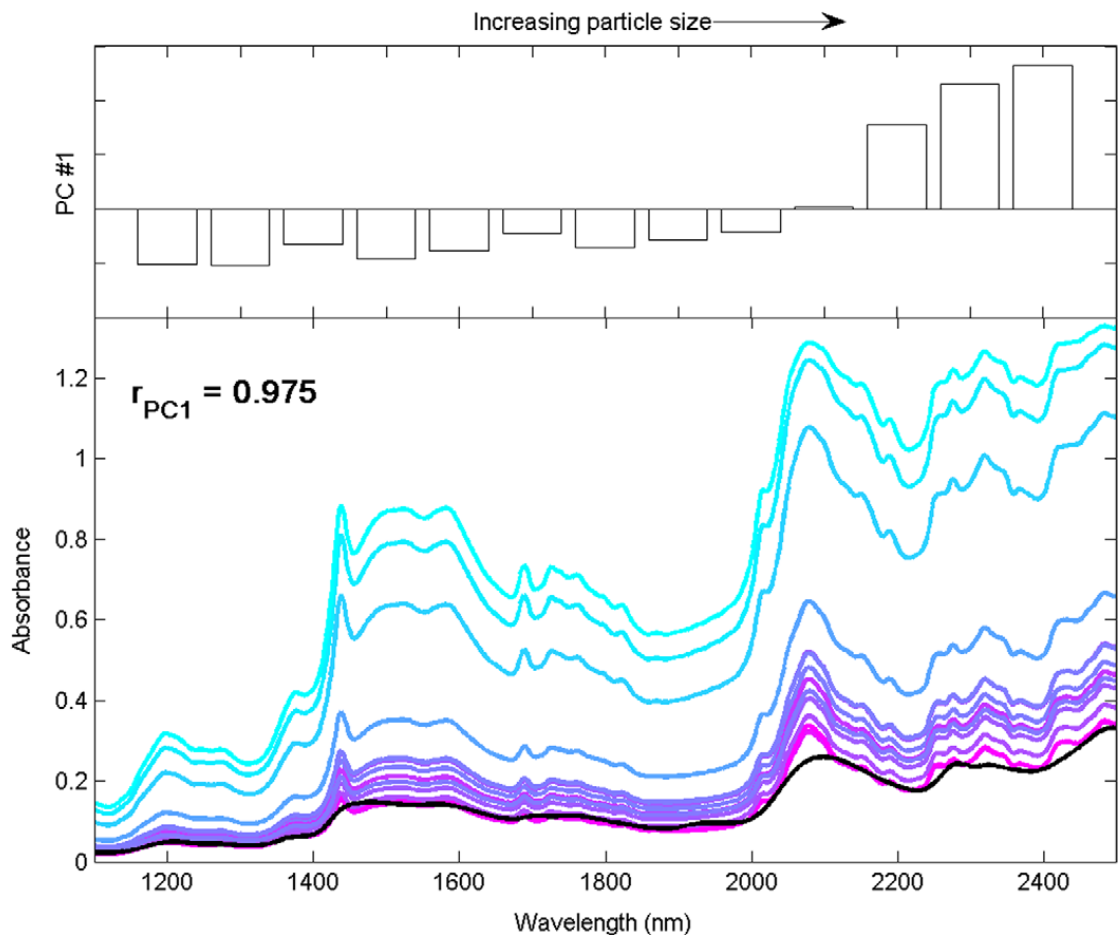


图 1 近红外光谱 13 蔗糖和样品不同粒径(最小粒子在底部, 最大顶部; 粒子尺寸范畴在 20 - 540 μm 。黑色光谱显示了一种镜面反射率蔗糖样本。酒吧是分数值第一主成分 13 个蔗糖样品主成分分析模型在完整光谱。

一种示例数据预解决蔗糖中可以看到图 2, 其中也包括一种其实一种示例数据预解决蔗糖中可以看到图 2, 其中也包括一种其实

从当前起, 在这篇文章中, 咱们将演示效果不同预解决技术在小果胶数据集只包括 7 个样品有不同程度酯化(%德; 范畴在 0 - 93%) [2]。这些样品测定近红外光谱反射率模式在光谱范畴 1100 - 2500 海里(收集每 2 海里区间; 图 3)。咱们提供相应第一因素 PCA 样本得分后作为一种条形图意味着定心, 连同集中吸光度值在波长 2244 纳米。咱们选取这个峰值, 由于它在理论上应当描述% DE 完美。对于本文, 咱们假设信息在光谱相联系果胶粒子大小和形状应当被预解决技术, 条形图应当显示一种线性行为与%德。

为了阐明预解决影响在量化, 咱们使用数据取自克里斯腾森 et

al. [3]。她们研究了一组 32 杏仁蛋白软糖混合物，基于不同食谱，九个不同数据都在互联网上可用 (www.模型生活 ku dk)。所有杏仁蛋白软糖样品测定了六种不同近红外光谱仪器和化学参照分析了水分和糖含量。在构建一种定量回归模型，重要是要打扫预测数据从非系统性散射变化，由于她们可以产生重大影响预测模型性能和模型复杂性或吝啬。在本文中，咱们使用请预测这种定量响应信息[4]。

2. 预解决办法

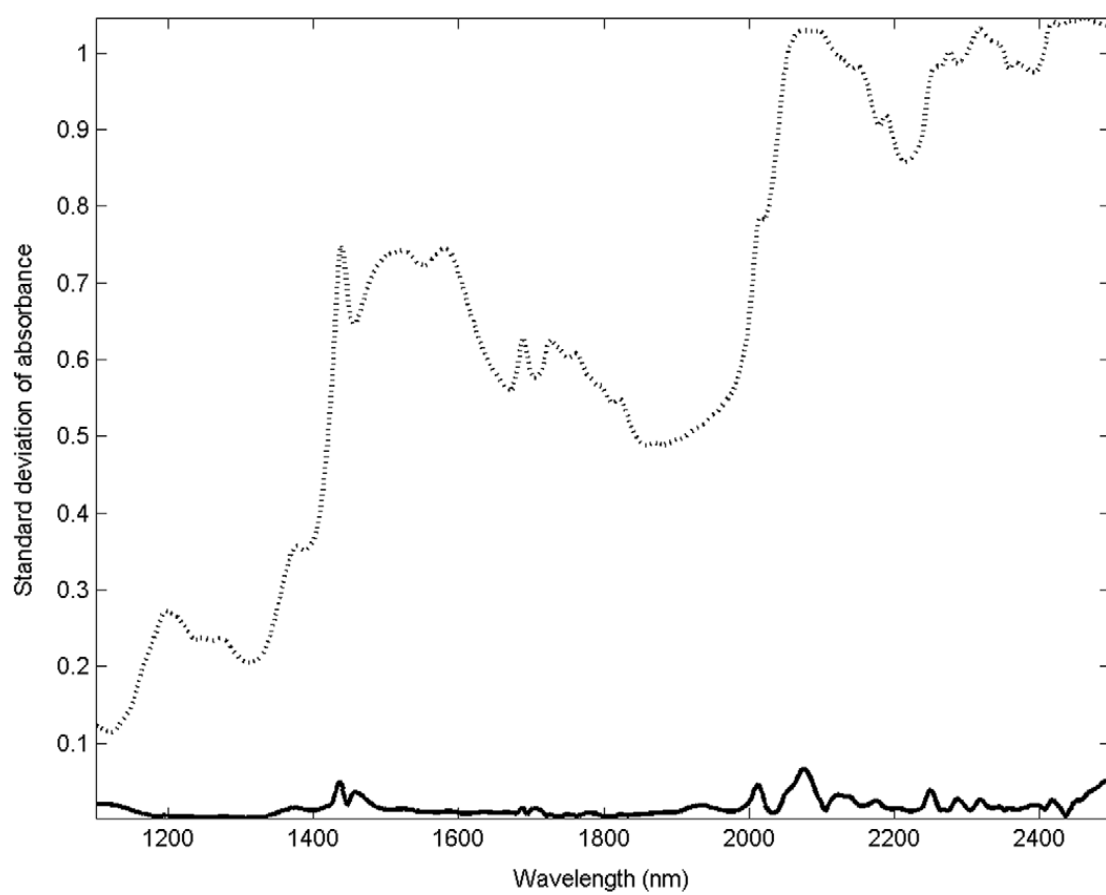
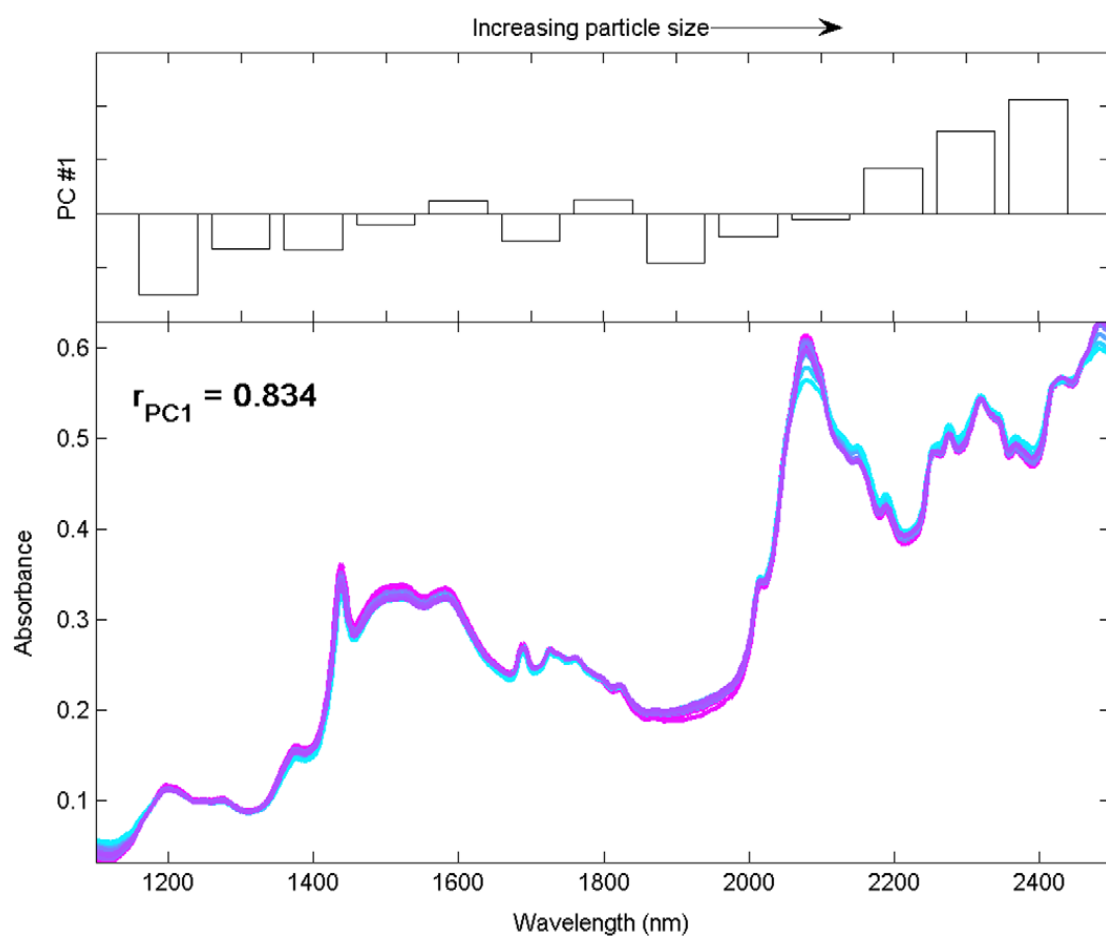


图 2 上图：数据被一种二阶蔗糖乘法散射校正；底：相应原则差每个波长，虚线是原始/未加工数据(见图 1)，固体是预解决数据。

最广泛使用预解决技术在近红外光谱法(在两个反射和透射模式)可以分为两类：
scattercorrection 办法和光谱衍生品。

第一群散射校正预解决办法涉及乘法散射校正(MSC)，逆 MSC，扩展 MSC(EMSC)，扩展逆 MSC，de 趋势，原则正态变量(SNV)和原则化。

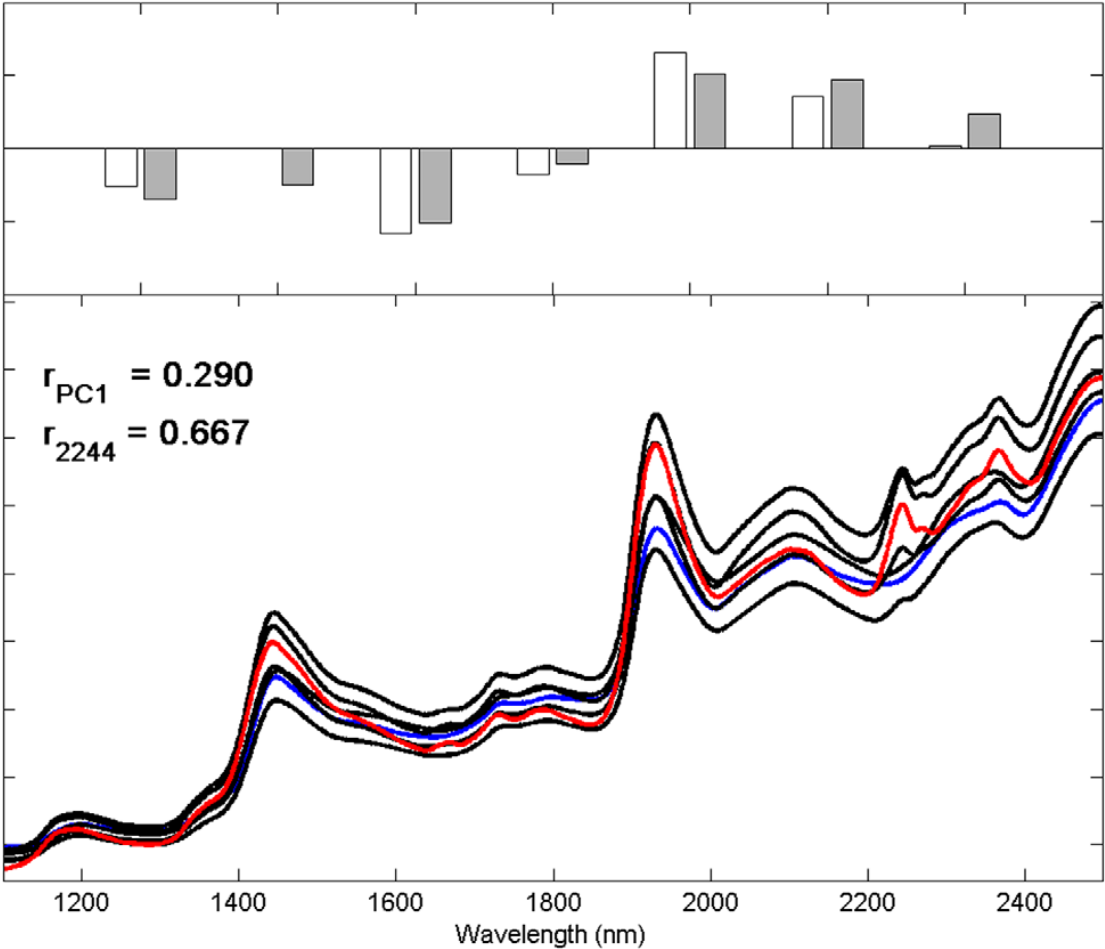


图 3 生/未加工光谱，7 果胶样本。蓝线是一种示例有 0%限度果胶酯化(DE)，红色线是一种样本 93%德。打开条批示主成分分析(PCA)评分值在第一种 PC 为完整光谱，意味着定心，关闭了酒吧后光谱值在波长 2244 纳米。

光谱推导集团是为代表两种技术在本文中：诺里斯威廉姆斯(NW)衍生品和 Savitzky-Golay(SG)多项式导数过滤器。这两种办法都使用一种平滑光谱导数计算之前，以减少有害影响，老式信噪比有限差分衍生品会。

预解决环节的目可以三种：

- 1) 提高后续摸索性分析
- 2) 提高后续叠层校正模型(力数据服从兰伯特啤酒法律);或者,
- 3) 提高后续分类模型。

朗伯比尔定律(方程(1))是经验对近红外光谱/ NIT 和显示成线性关系, 吸取光谱和浓度(s)成分(s):

$$A_{\lambda} = -\log_{10}(T) = \varepsilon_{\lambda} \cdot l \cdot c \quad (1)$$

其中 A_{λ} 作为吸光度, ε_{λ} 依赖波长摩尔吸光系数, l 代表光通过容器有效途径长度, 是 c 成分浓度。朗伯比尔定律严格有效只有纯透射系统没有散射。在反射率测量, 方程(1)是重新定义在类比透射率测量为:

$$A_{\lambda} = -\log_{10}(R) \cong \varepsilon_{\lambda} \cdot l \cdot c$$

其中, R 是反射率。

选取适当预解决应当总是被以为是相对于持续建模阶段。例如, 如果数据集兴趣不遵守法律, 额外因素兰伯特啤酒或组件在请回归普通可以弥补这一抱负行为光谱预测[5]。缺陷涉及额外因素是提高模型复杂性, 反过来, 最有也许减少对将来预测模型鲁棒性。所有预解决技术目的减少变化减少数据以增强特性谋求在光谱, 经常一种线性(简朴)关系现象(例如, 一种成分)兴趣。通过使用一种适当预解决技术, 这可以实现, 但总是有危险应用了错误类型或应用太严重预解决, 将移除有价值信息。对的选取预解决是很难评估模型验证之前, 但是, 普通来说, 执行几种预解决环节是不可取, 作为最低规定, 预解决应保持或减少复杂性有效模型。

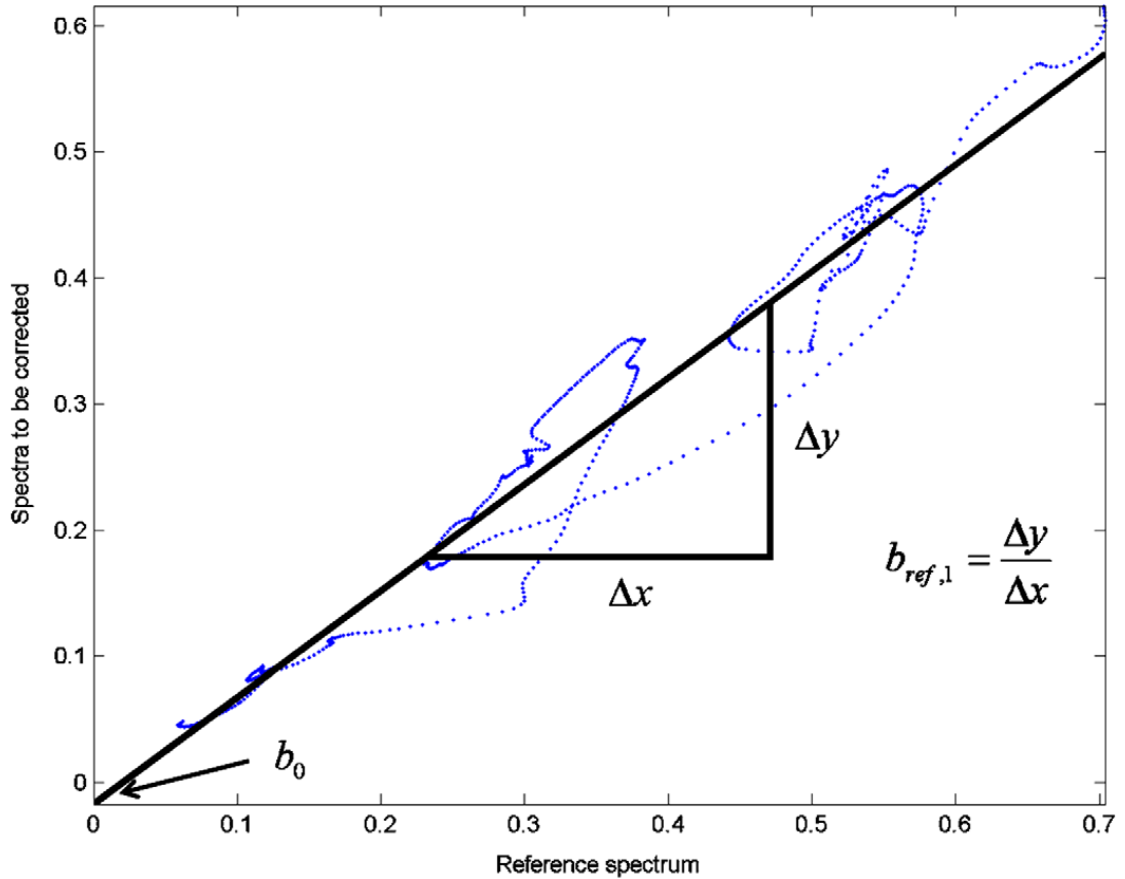


图 4 样品光谱(蓝色点)密谋反抗一种选定参照光谱。标量修正条款发现随着拦截和边坡黑色线条，就是发现从最小二乘回归适合通过所有点。

3. 散射校正

依照散射校正办法，咱们考虑三个预解决概念：研究生，SNV 和原则化。这些技术是为了减少(物理)可变性样本之间由于散射。所有三个也调节基线样本之间变化。

3.1 MSC

乘法散射(或者，在普通状况下，信号)校正(MSC)也许是最广泛使用预解决技术对近红外光谱(紧随其后 SNV 和派生)。MSC 其基本形式是由 Martens 初次引入等人 1983 年[6]，并进一步阐述了通过 Geladi 等人 1985 年[7]。MSC 背后概念是，工件或缺陷(如，不良散射效应)将被删除数据矩阵之前，数据建模。MSC 包括两个环节：

1) 预计校正系数(加法和乘法贡献。

$$X_{org} = b_0 + b_{ref,1} \cdot X_{ref} + e \quad (2)$$

2) 校正记录光谱

$$X_{corr} = \frac{X_{org} - b_0}{b_{ref,1}} = X_{ref} + \frac{e}{b_{ref,1}} \quad (3)$$

其中： X_{org} 代表近红外光谱仪器对一种原样品进行光谱测量， X_{ref} 代表参照光谱用于预解决整个数据集， e 是减少一某些 X_{org} ， X_{corr} 代表了修正后光谱，其中 b_0 和 $b_{ref,1}$ 都是标量参数，这在每个样品中具备不同代表含义，这个在图 4 中进行阐明，对每个标量参数进行理解释。

在大多数应用中，平均频谱校准设立用作参照光谱。然而，一种通用参照光谱也可以应用。在最初论文 Martens et al. [6]，这是建议只使用那些某些光谱轴，不涉及有关信息(基线)。而这使得好光谱意义上说，很难拟定这些地区在实践中，特别是在近红外光谱测量，信号从不同化学成分是强烈重叠和有关，很少或主线没有真正基线是发现。这是为什么，在大多数状况下，整个频谱用于查找标量校正参数在 MSC。图 5 演示了应用程序原则 MSC 对果胶数据。光谱特性果胶粉是守恒，而背景偏移和斜坡基本上是删除(与图 3)。线性关系光谱和%德是不错，但它并不完美。

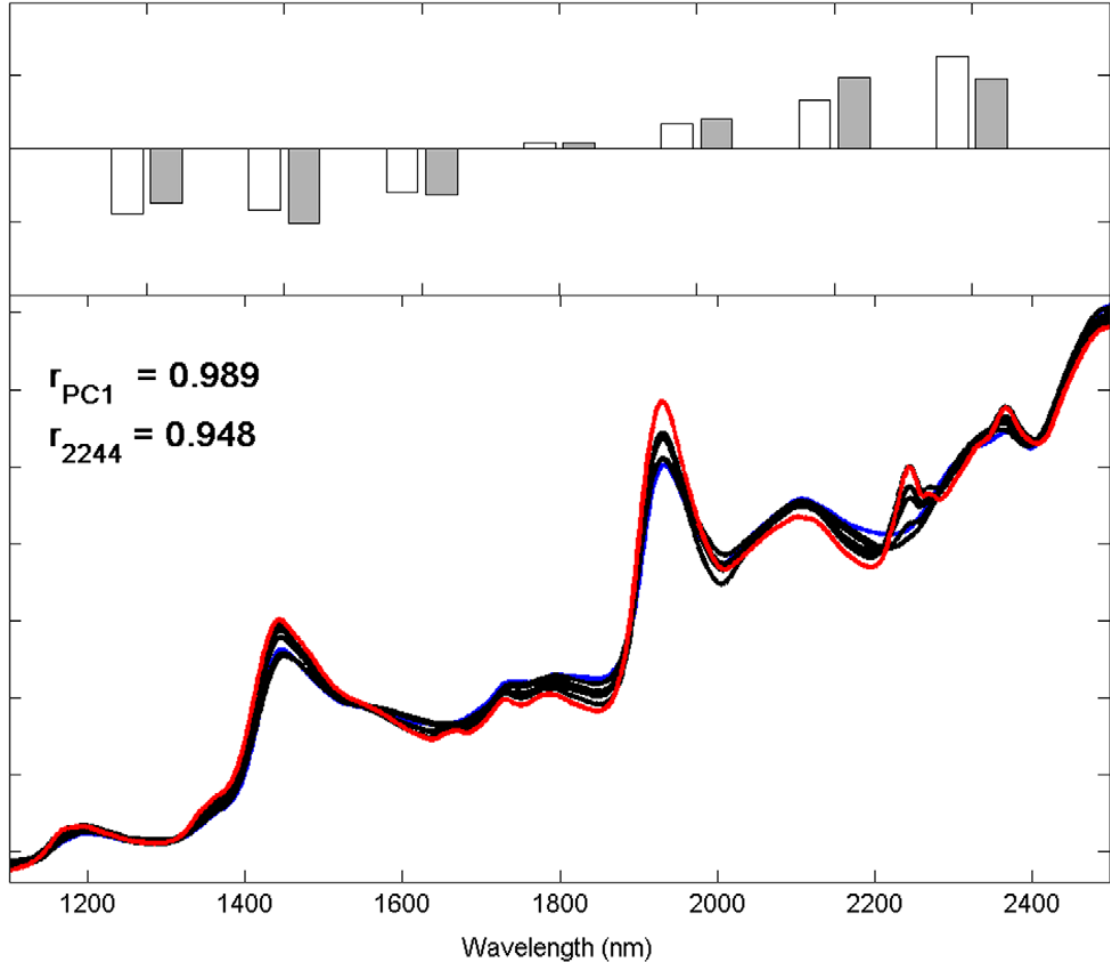


图 5 数据预解决乘法散射校正用一阶校正对平均频谱。

基本形式 MSC 已经扩展成更复杂扩展 (8 - 12) 普通被称为 EMSC。这个扩展涉及二阶多项式拟合参照光谱, 拟合一种基线波长轴, 和运用先验知识从光谱兴趣或光谱干扰物。在本文中, 所有这些代替品被称为 MSC 为简朴起见, 由于她们可以总结为一种单一方程:

$$X_{org} = [1, X_{ref}, X_{ref}^2, \lambda, \lambda^2, X_{known,1}, X_{known,2} \dots] b + e \quad (4)$$

其中, λ 代表依赖波长轴修正向量, $X_{known,1}$ 包括先验知识涉及了想要/不必要光谱信息 (例如一种已知光谱干扰物种)。方程 (4) 可以容易地扩大到涉及任何其他任何具备恰当意义修正。

b 代表了一组标量 (校正系数) 给出了方程 (5)。

$$b = [b_0, b_{ref,1}, b_{ref,2}, b_{\lambda,1}, b_{\lambda,2}, b_{known,1}, b_{known,2} \dots] \quad (5)$$

其中: b_0 偏差校正

$b_{ref,1}$ 是依照 i^{th} 修正

$b_{\lambda,1}$ 改正对我

$b_{known,1}$ 是对 i^{th} 校正订单波长轴依赖, $b_{known,1}$ 是纠正第 i 已知信息吗

相比喻程 (2), 它可以观测到方程 (4) 只是一种高阶扩张之一想法。在本文中 $X_{known,1}$ 将不再做进一步讨论, 由于, 在诸多实际状况, 参照光谱对想要和不必要成分并没有现成可用。参照校正是最惯用办法只有一种一阶多项式。虽然没有数学限制扩大到高阶增长, 有普通没有光谱参数这样做(除了也许如果重要瑞利散射是出当前短波长区域。

图 6 显示了成果一种二阶多项式校正果胶数据。修正条款用于二阶多项式参照校正只是发现了拟合二阶(二次)多项式点在图 4。只有边际改进获得了比一阶修正在图 5。

波长轴依赖是最常涉及作为一种二阶多项式拟合波长轴光谱。当没有参照校正涉及在内, 这个简朴波长配件也名义光谱 de 趋势[13], 它可以被视为一种基线校正。重要是要注意, 涉及波长依赖性在完整校正方程 (4) 而不是让它作为一种单独环节会导致一种较小矫正效果。这是由于一种矩阵求逆操作同步执行所有校正参数在 MSC, 不同修正会互相影响最小二乘法拟合准则。当一种波长依赖是独立决定只有波长轴(而不是参照光谱)影响校正, 这将导致一种趋平加工谱。这个效果可以看出通过比较无花果。7 和 8。

正如前面提到, 更加复杂修正(例如, 高阶多项式或其他转换波长依赖性)可以很容易地纳入 MSC。Thennadi I 和马丁[12]建议使用对数值波长, 由于这是更多声音来判断光谱。然而, 区别使用对数变换波长与使用一种一阶多项式校正是最小, 使这两个办法相似所有实用目。

彼得森指出, et al。[9], 这是一种相称简朴过程, 应用逆版本 MSC, 称为逆信号校正 (ISC) [14]。预计校正参数, b 系数, 发现以类似方式来定期 MSC:

$$X_{ref} = [1, X_{org}, X_{org}^2, \lambda, \lambda^2, X_{known,1}, X_{known,2} \dots] b + e \quad (6)$$

请注意, x_{org} 和 x_{ref} 已经互换了地方比喻程 (4)。一种运用(扩展) ISC (EISC) 是简朴修正方程:

$$X_{corr} = [1, X_{org}, X_{org}^2, \lambda, \lambda^2, X_{known,1}, X_{known,2} \dots] b \quad (7)$$

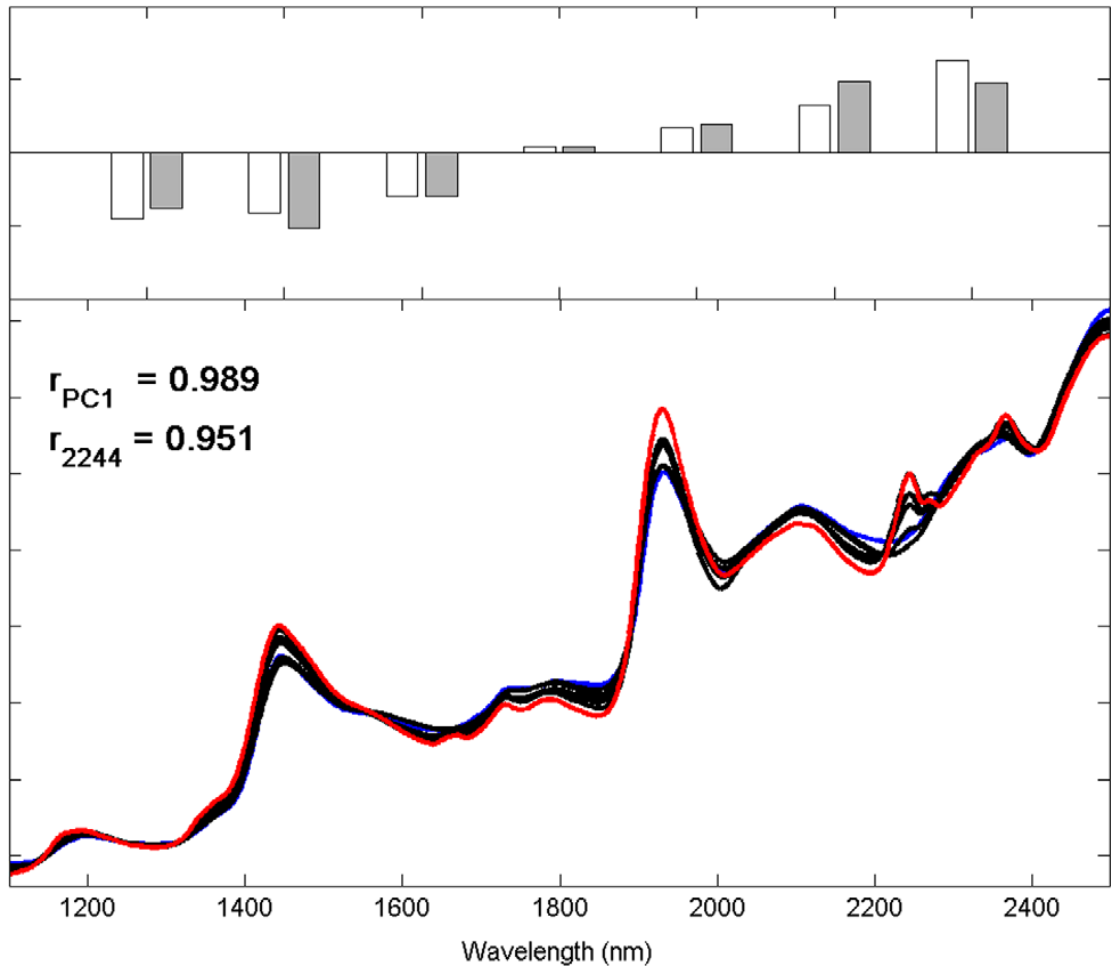


图 6 乘法散射校正光谱使用二阶参照校正对平均频谱。

在 ISC 和 EISC, 无论是预计校正系数和校正自身中执行可以被描述为一种迈进方式, 使它能以便涉及附加条款和/或参照信号[9]。前面提到矩阵求逆操作所需参数预计在 MSC 可以很容易地成为数字坏脾气如果它涉及高阶多项式参照改正。这是一种论点支持 ISC。然而, 假设, ISC 最小二乘法拟合, 误差在记录光谱(纠正)不大于误差为参照光谱。在大多数实际应用, 参照是平均谱计算样本数据集从 n (例如, 校准设立)。预期噪音水平数量级这个引用是不大于个人光谱(忽视了偏见由于散射差别集)。这是一种反对 ISC, 由于一种小错误在光谱预解决将会影响到更大限度比最初 MSC。

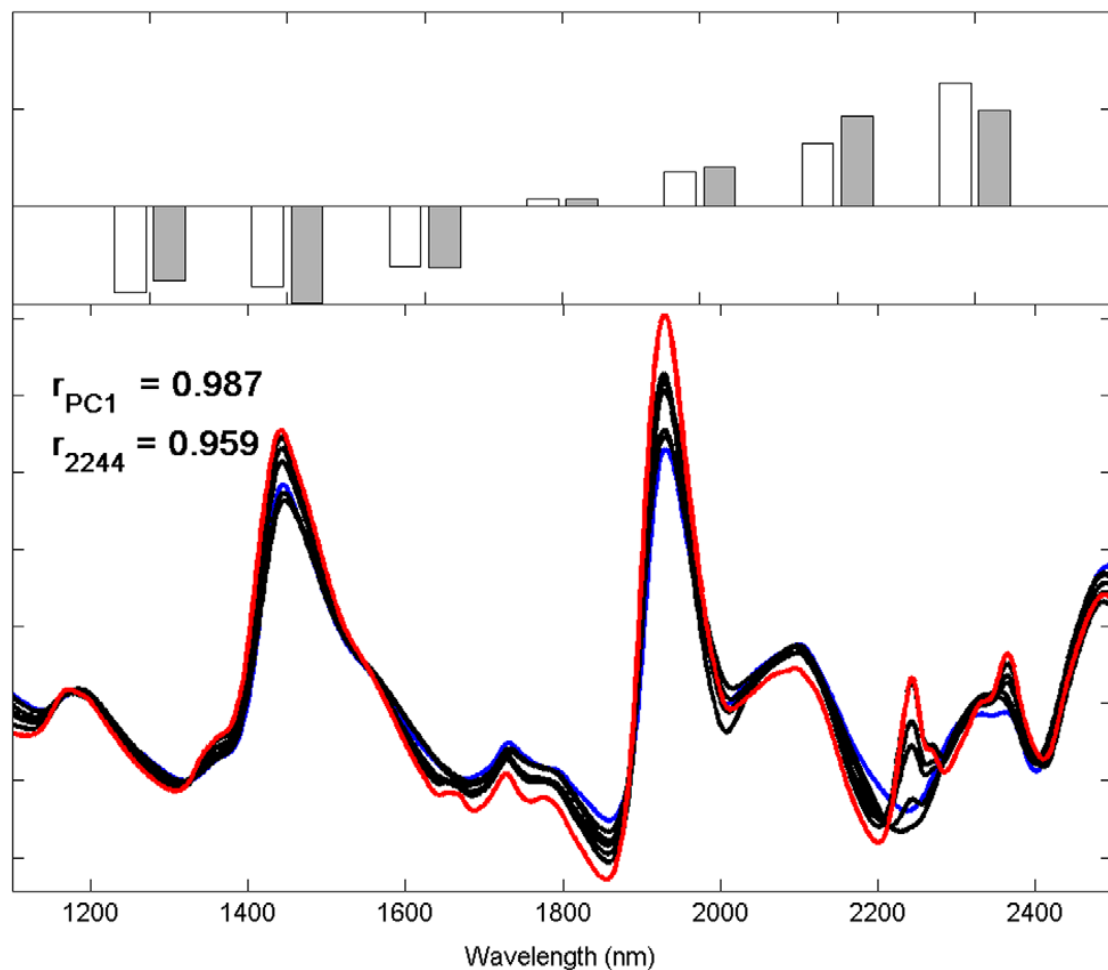


图 8 乘法散射校正(MSC)与一种一阶多项式参照校正对平均频谱，另一方面是一种单独 MSC 与二阶多项式波长校正(去趋势)。

重要挑战是定义一种适当 MSC 参照光谱。正如前面提到过，这是最经常设立为普通校准光谱。加拉格尔等人^[15]提供了一种天然变化来 MSC 通过包括一种加权方案在预解决环节。提出了两个选取：

- 1) 使用一种预定义权重向量波长轴向
- 2) 迭代搜索最优权重向量

迭代解决方案是通过给低体重发现变量或波长与高残留差别原始数据和修正解决方案。权重计算将继续，直到区别两个后续迭代修正光谱不大于假定数据中噪声水平。不幸是，这个相称简朴直接办法并不总是适合近红外光谱数据，自蔓延在更高波长范畴普通表白更分散，应当改正而不是权重小。图 9 显示了权重。用于最后修正给强调到 shortwavelength 地区，而长波长区域并不有助于校正所有。

另一种建议寻找参照校正 MSC 都暗示了 Windig et al.。——所谓糊涂 MSC[16]。这种办法发现平均频谱从 msc 修正数据集。接下来，MSC 是多次重复更新参照光谱平均数据集纠正正在每次迭代中环节。

图 10 显示了成果糊涂 MSC 应用到果胶数据集——在这种状况下体现非常类似于呆头呆脑 MSC 单纯 MSC。在呆头呆脑 MSC，可以遵循增长模型记录信息，然后停止在收敛(两个迭代环节普通是足够)。叠加在图 10 变化参照光谱从原始光谱平均值。

3.2 原则正态变量 (SNV)

SNV 预解决也许是第二个最散射校正应用办法 NIR / NIT 数据[13]。在本文中，规范化(也称为对象明智原则化)光谱将检查在同一种小节由于明显相似性两个原则。SNV 基本格式和正常化修正与老式 MSC:

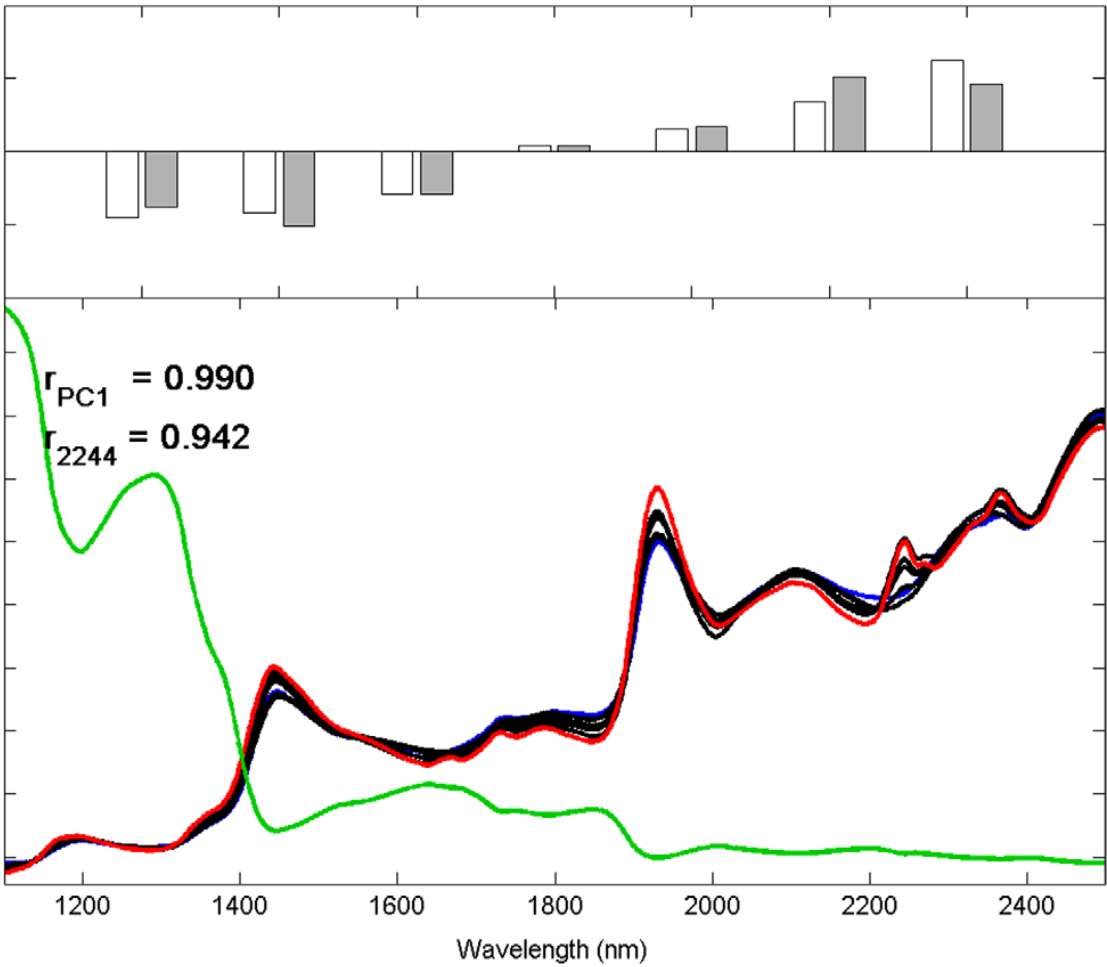


图 9 加权乘法散射校正，基于迭代重量决心。绿线显示了最后权向量(任意规模贡献)。

$$X_{corr} = \frac{X_{org} - a_0}{a_1} \quad (8)$$

对于 SNV, a_0 是平均值样品光谱需要修正同步, 为原则化, a_0 设立等于零。对于 SNV, a_1 是原则偏差样品光谱。

图 11 演示了 SNV 校正果胶数据集。对于正常化, 不同向量规范可以用于比例因子 a_1 , 最常用是总金额绝对值元素向量(都市街区或出租车规范)或平方根之和平方元素(欧几里得规范)。其他选项, 有时也用正最大吸光度变量和正常化向着单一选定波长。这两个最后选项应当小心使用, 由于她们可以有不良影响后续分析在例嘈杂数据。

图 12 显示了效果。欧几里得正常化, 迄今为止最惯用正常化, 果胶数据集。

这个信号校正背后概念 SNV 和原则化是相似 MSC 除了常用参照信号不是必须。相反, 每个观测自己解决, 孤立于剩余设立。缺少需要一种共同参照也许是一种实际优势。

由于 SNV 和归一化不涉及一种最小二乘方拟合在她们参数预计, 她们可以敏感嘈杂条目谱。相反, 使用平均值和原则偏差作为校正参数, 有人也许会考虑使用更健壮等价物这些记录时刻。郭 et al. [17] 建议使用中档或意味着内部四分位范畴和原则偏差内部四分位数预计为 a_0 , a_1 , 分别命名为“办法”健壮正常变量”。这将是特别适当状况下光谱很模糊(如。在超快在线近红外光谱应用程序, 在强劲办法将会更受散射噪声(例如, 波长选取性反射颗粒液体流)]。鲁棒预计影响她们演示郭 et al. [17] 对于模仿和实际数据。

正如已经讨论了 Dhanoa et al. [18], 有一种明显相似性 SNV 和 MSC。这种关系可以通过如下简朴近似提出:

$$X_{msc} \approx X_{SNV} \cdot \overline{s_X} + \overline{\overline{X}} \quad (9)$$

其中: $\overline{s_X}$ 普通原则偏差光谱, $\overline{\overline{X}}$ 大意味着在所有光谱, 都发现了从原始/未修正光谱(参见图 13)。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。

如要下载或阅读全文，请访问：

<https://d.book118.com/076025051124010120>