小波变换在分析化学解析中的应用

汪　琳　唐晓菲（安徽省生态环境监测中心，安徽　合肥　230071）

摘　要：连续小波变换（CWT）已被认为在多刻度的信号处理技术中具有很高效率。然而，被CWT控制的信号（附带一个具体小波）只解释一部分信息。为了有效利用更多充足的分析信号，一种方法，被称为偏最小二乘法的提出，在这种方法中，已测得的数据库第一次被CWT用不同的小波拓展，偏最小二乘法被应用于去开发在拓展的数据库和目标价值之间的定量模型。为了选择代表性的小波，主要成分分析用于调查信号的分配，用不同的小波通过CWT获得相应的信号。这种方法的表现由血液和烟草粉末检验与用PLS方法获得的结果相比，这种WPLS与信号处理技术相结合被证明是提高复杂样品的近红外光谱的理想工具。

关键词：小波变换；偏最小二乘法

0 引言

随着现代分光计的快速发展，对复杂样品的定量分析的实际需要正在增加。多方位调整方法，特别是偏最小二乘回归技术，已经被广泛应用于分析多组分光谱数据，如近红外光谱数据，然而，NIR的信号不仅包含分析组分的信息还有噪声，背景，对浓度无关的系统变换，为了提高定量分析的准确度，基于算法和数据方法，一些方法被应用于去减少过多变量，关注察觉异常值。此外预处理操作已被运用到减少光谱数据的背景与噪音，如SNN，MSC，OSC，小波变换（WT）。

WT已被证明是分析信号处理的一种有效的工具，WT的主要特点是它运用测量信号的不同信息把信号分解成不同频率的组分，一般来说，高频率的组分被认为是噪音信息，低频率的组分被认为是背景信息。建议的方法是使用信号信息的已选的组分去重建或者直接在回归中使用。考虑到光谱的任何一种组分有定量模型的有用信息，利用所有频率组分去开发一种合并模型。在这些研究中DWT一般用来分解信号，和DWT相比，CWT有更好的空间分辨率，也相对更容易操作。通过使用CWT，当信号信息因回归保留时，背景和噪音被有效的去除了，在大部分的研究中，CWT被特定的小波和比例参数执行，被CWT处理的使用具体设备的信号，也许只能解释测量信号的一部分信息。

主要组成分析和PLS都一般被用于多方位调整，由向量组成的数据模型被称为第一组数据，每一个矢量代表分析信号，如样品光谱或色谱图，随着分析仪器的发展，提出了为第二组数据三矢量甚至或是高顺序的数据方法。在NIR光谱分析中，N-PLS方法因此被提出。另一种方法是建立一个提高矢量数据，这些是为了一个样品而被保留的组或列，然后应用第一顺序算法比如PLS，由于需要这样的一个方法释放三线性，应用会更广泛。

在本文中，一种方法被称作WPLS，被应用与NIR光谱定量分析，利用CWT，测定的数据库第一次用不同的小波扩展，然后运用PLS去建立在扩展数据库和目标价值之见的定量模型。所以，在NIR光谱中包含的充足的信息，可在模型中有效的使用。为了选择有代表性的小波，PCA被应用与调查信号的分配（由CWT用不同小波的获得）。血液和烟草粉末样品被用于调查这种方法，结果表明这种方法可能是复杂的NIR光谱定量分析的较高水平方法。

现代近红外光谱分析是将光谱测量技术、计算机技术、化学计量学技术与基础测试技术的有机结合，是将近红外光谱所反映的样品基团、组成或物态信息与用标准或认可的参比方法测得的组成或性质数据采用化学计量学技术建立的校正模型，然后通过对未知样品光谱的测定和建立的校正模型来快速预测其组成或性质的一种分析方法。

与常规的分析技术不同，近红外光谱是一种间接分析技术，必须通过建立校正模型（标定模型）来实现对未知样品的定性或定量分析。具体的分析过程主要包括以下几个步骤: 一是选择有代表性的样品并测量出其近红外光谱。二是采用标准或认可的参考方法测定所关心的组分或性质的数据。三是将测量的光谱和基础数据，用适当的化学计量方法建立校正模型。四是未知样品组分或性质的测定。由近红外光谱分析技术包括了近红外光谱仪，化学计量学软件和应用模型三部分。三者的有机结合才能满足快速分析的技术要求，是缺一不可的。

与传统的分析技术相比，近红外光谱分析技术具有诸多优点，它能在几分钟内，仅通过对被测样品完成一次近红外光谱的采集测量，即可完成其多项性能指标的测定（最多可达到十余项指标）。光谱测量时，不需要对分析样品进行前处理，分析过程中不消耗其他材料或破坏样品，分析重现性好，成本低。对于经常的质量监控是十分经济且快速的，但对于偶然做一两次的分析或分散性样品的分析则不太适用。因为建立近红外光谱方法之前必须投入一定的人力、物力、财力才能得到一个准确的校正模型。

近红外光谱主要是反映C-H，O-H，N-H，S-H等化学键的信息，因此分析范围几乎可以覆盖所有的有机化合物和混合物。加之其独有的诸多优点，决定了它应用的领域广阔，使其在国民经济发展的许多行业中都能发挥积极作用，并逐渐扮演着不可或缺的角色。主要的应用领域包括：石油和石油化工，基本有机化工，精细化工，冶金，生命科学，制药，医学临床，农业，食品，饮料，烟草，纺织，环境保护，高校及科研院所等。在石化领域可测定的油品的辛烷值，族组成，十六烷含量等。

与传统化学分析法相比，近红外光谱分析技术有鲜明的技术特点。①分析速度快，扫描速度快，可在数十秒内获得一个样品的全光谱图，通过数学模型即可快速计算出样品的浓度；②多种组分同时分析；③无污染分析。样品

不需特别的预处理，不使用有毒有害的试剂；④实时分析和远距离测定。实时在线分析特别适合工业生产上的应用；⑤操作简单，分析成本低。

近红外光谱分析技术也有其相应的弱点。该项技术是一种间接的分析技术，它必须依赖常规的化学分析方法，测定出特定背景范围内多个标准样品成分的化学值，利用化学计量学的方法建立数学模型，并通过数学模型计算出待测样品的成分含量。数学模型预测的准确性与常规的化学分析的准确性，建模样品的代表性，模型的使用的合理性有很大关系。另外，近红外光谱分析的测试灵敏度较低，待测样品的成分含量一般不少于0.1%。

1 理论和计算

1.1 连续小波变换（CWT）

小波被定义为来源于扩张和翻译的一系列的函数。CWT被认为是在小波中投影的信号，用CWT，在信号中查找包含不同频率的信息。因此，作为一个高水平的预处理工具CWT已经被成功应用到多方位调整，被发现消失瞬间和规模参数只是影响CWT表现的两个重要因素。由于小波过滤器有合适的消失瞬间，背景可以被成功移除而没有重大信息丢失。此外，一个合适的规模参数不仅能保留有用的信息也可以压制噪音。然而，被CWT处理的信号（用特定的小波）也许只能解释这一数据的一部分信息。由于不同的小波，CWT掌握的组分互相有所不同，从小波族视角，小波过滤器比如db，coif，sym有不同过滤长度和对称性，使他们投影互相不同，另一方面，CWT可被视为区分过程，在这里小波充当区分处理器。它已经被证明可以作为平稳函数，因此N阶倒数表示一个具有N个消失矩的小波，一个单一信息的各个部分可以有不同消失的瞬间，因此结合不同组件，通过小波回归，产生更丰富的信息模型。

1.2 WPLS

利用CWT，WPLS是为了NIR光谱分析，它使用含不同小波的CWT，扩展测定光谱，然后建立在扩展数据和目标价值的定量模型，如上述，由CWT用不同小波获得组分代表不同频率的信息或是不同衍生物顺序，这些组分的组合可能会提高定量模型的表现。

当在模型中用光谱的不同组分，另一维，也就是小波被引入，假设X是调整样品的测定光谱数据，第二组顺序的数据矢量将在每个光谱中被不同的小波处理来获得，然而这样第二组顺序数据不是三线性的数据，因为在小波中没有确定的关系，这使得第二顺序调整方法不可行。因此，建立一个提高矢量串联由不同小波获得的组分是建立数据定量模型的有效方法，扩展的数据可以通过演示CWT到特定的光谱获得，如果X是N样品和P波长的光谱模型，因为Xwi是已经转变的n*p模型的光谱，如果K个不同的小波被使用，扩展的数据是K*（N*P）模型，因为Xwi是通过不同的小波来获得的，每块编码光谱信息的一特定方面，扩展数据模型X，因此可以更有效的描述样本比实测光谱的PLS模型，扩展数据可以预期WPLS模型比传统更有预测性，很明显，多样性的Xwi需要改善WPLS模型，为了选择获取各种信息Xwi小波，PCA用不同的小波来调查光谱信息，散落在PC空间的小波被选择，此外规模参数在RMSECV标准被决定。应该被指出调整系统在模型中用语参数的决定，独立的有效系统用于检测座钟模型的表现，此外MCCV和ostens f标准被用于决定模型中潜在的变量。计算是用MATLAB软件演示的，源密码包括在计算中，如PLS，CWT，MSC等，这些可以从计算机资源找到，结合的项目从作者可以得到。

1.3 数据库

血液数据库包含着NIR辐射和反射光谱和血红蛋白，葡萄糖和231种血液的胆固醇，网站可以下载。所有的光谱在波长范围1100~2498纳米有2纳米间隔700种变量，在研究中调查了反射光谱和血红蛋白内容，得到样品的反射光谱。由烟草公司提供的烟草粉末数据组，包含有312张NIR谱。每一张扩散反射NIR谱都是用MPA FT-NIR（布鲁克公司，德国）测量的，是由从3999.8~11995.6cm^{-1}每4cm^{-1}一个变量，共2074变量记录组成的。被测样品浓度沿用工业标准方法。用图形的方式可以显示烟草粉末的NIR谱。在图中，波长2500~833.6nm用来和血液数据组保持一致。计算中，网页描述的计算组和验证组被使用在血液数据组中，它们的取样数分别是173和58。在烟草粉末数据组中，2/3的样本（208）被选为Kennard-Stone方法的计算组剩下的104组样本作为验证组。

2 结果和讨论

2.1 典型波段的选取

为了获得不同谱图信息的波段，用PCA方式获得的信号是CWT的31组波段，包括db2~db20，sym2~sym-8，和coif1~coif5。在计算中，使用了计算组的平均谱，尺度常数从5~50每个5做一次计算。每一个尺度常数下得到一个31XP的矩阵做PCA计算。用图形来展示在血液数据组中当尺度常数是10，20，30，40下，这些波段在PC1~PC2空间的分布。可以得出：一些波段形成一簇是具有相似性的结果。当尺度常数增加的时候，这些波段的趋势是形成一个环形分布，这表明他们见的差异变小了。考虑到模型必须建立在复杂信息的基础上，波段必须布满整个空间。因此，从图形中选取八个波段。进一步降低计算复杂度，只有六个标记为绿色的波段被使用，另外两个因为接近其余六个被舍去。备选的波段分别是coif1，db2，db4，sym3，sym5，sym7。可以注意到，计算烟草数据也会得到相似的结果。

2.2 决定尺度常数

尺度常数是关系到CWY结果的另一个重要参数。采用六组选定波段扩展数据的MCCV。在MCCV计算中，RMSECV使用了80%随机选择的样本来做计算。不同尺度常数下的RMSECV结果显示，对血液数据，RMSECV很快从10变到了15，在15到30之间形成一个平台，后续上升不明显。所以合适的值应该在15~30之间。对烟草数据，RMSECV值随着常数的增加而下降，但是差异越来越小。同时考虑两组数据，最后选取了30作为尺度常数。

2.3 WPLS组元的贡献

通过绘制六个波段下尺度常数30的血液平均CWT谱。

可以看到，它有两组强峰，分别在 1400nm 和 1900nm，和原始 NIR 谱的两个明显的吸收峰相对应。然而，它的形状和峰位置在不同的波段下在转换后的谱图中不明显。另外，和原始谱相比，小峰被显著的增强了。这种结果显示，NIR 谱的信息扩展到了不同的组元。

观察血液数据每个区块的 WPLS 模型的回归数据。通观全局分布，这个值在每一个区块都差不多，这显示了模型中每个区块拥有相似的贡献。进一步的比较可以发现，那些对模型有很大贡献的变量也是相似的，分布在 1100~1200nm，1500~1800nm 和 2200~2350nm 的区间内。仔细检查后看到，对模型有重大扰动的变量在不同的区块是显著不同的。这显示了 WPLS 模型相比传统的单区块模型具有很大优势。

比较了用单区块模型的构建的 WPLS 模型的回归系数。为方便比较，所有系数被归一到区间 [–1.1]。虽然不是很明显，但是还是很容易找到，尤其在 1300~1600nm 和 1800~2000nm 范围内，这些变量的贡献值在 WPLS 模型中相比单区块模型被显著的弱化了。因此上，结合 CWT 的不同组元的 WPLS 模型可能会提供一种方法来提取 NIR 谱所包含的丰富信息，并且改进 NIR 谱分析的定量模型。

2.4 预测结果的比较

为了测试几种方法的预测能力，罗列了它们预测的均方根和 100 计算的标准差。为方便比较，PLS 原始谱，CWT 谱，MSC 和 CWT 结合 MSC 都一一列出。在 100 次重复计算中，计算组中的样本是被随机抽取建立模型的，剩下的样本就作为预测的有效性组。CWT 前处理使用 Haar 波段过滤器，尺度常数选取 20。

把原始谱中用通常的 PLS 的得到的结果作为比较标准，可以发现，对血液数据，CWT 和 MSC 都对改进结果有很大影响，结合 CWT 和 MSC 能够进一步的增强预测性能。WPLS 的结果不 CWT–PLS 要好一点点，但是 MSC–WPLS 给出了最好的预测，得益于 MSC 和 WPLS 的优势。对于烟草数据库，CWT 似乎没有成效，MSC 只能稍微改善结果。然而 WPLS 和 MSC 有相似的作用。MSC–WPLS 可以有显著的提高，这些结果很清楚表明更多的在扩展数据中被编码的信息。和 CWT–PLS 比较，WPLS 会产生对于这两个数据库更好的效果，结果可能是结论的进一步证据，另一方面，MSC 似乎是改善结果的有效工具。这也许依靠在光谱中测量的分散效应，幸运的是 MSC–WPLS 可以从这两个方面都受益。

3 结论

在这篇文章提出一种方法 WPLS，应用于 NIR 的定量分析，第一次被 CWT 用不同性质的小波扩展到已测的数据库，PLS 被应用于开发扩展数据库和目标价值之间的定量模型，不同小波中包含的组分编码用来分析信号的不同细节信息，结合的模型比传统的 PLS 模型具有更精确的准确度。

结合血液和烟草粉末样品，这种方法可以合理的提取和利用光谱中不同的信息，因此能够提高定量效果，偏最小二乘法是复杂样品 NIR 光谱分析中重要的研究方法。

参考文献：

[1] 秦侠，沈兰荪 . 小波分析及其在光谱分析中的应用 [J]. 光谱学与光谱分析 ,2000,20(6):892-897.

[2] 李肃义，嵇艳鞠，刘伟宇，等 . 小波变换与神经网络融合法在油页岩近红外光谱分析中的应用 [J]. 光谱学与光谱分析 ,2013(04):968-971.

[3]Wheeler O H.The girard reagents[J].Chem Educ,1968,45(12):435-442.

[4] 高建波，胡东成 . 小波变换和神经网络用于红外光谱定量分析 [J]. 清华大学学报 (自然科学版),2001(3):121-124.

[5] 金钦汉 . 从 2000 年匹兹堡会议看分析化学和分析仪器发展的一些新动向 [J]. 现代科学仪器 ,2000,3(4):14-16.

[6] 严国光，严衍禄 . 仪器分析原理及其在农业中的应用 [M]. 北京 : 科学出版社 ,1982.

[7] 李庆春，张玉良 . 近红外漫反射光谱分析法 (NIR—DRSA) 在作物品质育种中的应用 [J]. 作物学报 ,1992,18(3):235-240.

[8] 孙通，徐惠荣，应义斌 . 近红外光谱分析技术在农产品 / 食品品质在线无损检测中的应用研究进展 [J]. 光谱学与光谱分析 ,2009(09).

[9] 严衍禄，吉海彦 . 傅里叶变换近红外光谱技术及应用 [M]. 北京 : 科学技术文献出版社 ,1994.

[10] 吉海彦，严衍禄 . 在国产近红外光谱仪实验样机上用偏小二乘法定量分析大麦成分 [J]. 分析化学 ,1998,26(5):607-611.

[11]M.Blanco,I.Villarroya.NIR spectroscopy: a rapid-response analytical tool[J].Trends in analytical chemistry,2002,21(4):240-250.

[12] 严衍禄，赵龙莲，韩东海等 . 近红外光谱分析基础与应用 [M]. 北京 : 中国轻工业出版社 ,2005.

[13] 张叔良 . 红外光谱分析与新技术 [M]. 北京 : 中国医药科技出版社 ,1993.

[14] 吴瑾光 . 近代傅立叶变换红外光谱技术及应用 [M]. 北京 : 科学技术文献出版社 ,1994.

[15] 陆婉珍，袁洪福，徐广通 . 现代近红外光谱分析技术 [M]. 北京 : 中国石化出版社 ,2000.

[16] 胡守仁 . 神经网络导论 [M]. 北京 : 国防科技大学出版社，1993.

[17] 陈明 . 神经网络模型 [M]. 大连 : 大连理工大学出版社，1993.

[18] 高隽 . 人工神经网络原理及仿真实例 [M]. 北京 : 机械工业出版社 ,2003.

[19] 乔晓艳，王艳景，李刚 . 偏最小二乘法荧光光谱预测啶虫脒农药残留 [J]. 光学精密工程 ,2010,18(11):2369-2374.

[20] 王动民，金尚忠，陈华才，陈星旦 . 棉—涤混纺面料中棉含量的近红外光谱分析 [J]. 光学精密工程 ,2008,16(11):2051-2054.

作者简介：

汪琳（1989- ），女，汉族，硕士学历，安徽省生态环境监测中心助工，研究方向：重金属检测。