第28卷,第4期 光 谱 学 与 光 谱 分 析 Vol. 28,No. 4,pp825-828
2008年4月 Spectroscopy and Spectral Analysis April, 2008

小波多尺度正交校正在近红外牛奶成分测量中的应用

彭 丹,徐可欣*,李晨曦

天津大学精密测试技术及仪器国家重点实验室,天津 300072

摘 要 光谱分析中,干扰信号的存在直接影响所建分析模型的质量。基于信号和干扰的不同特性,提出了一种扣除背景和噪声干扰的新方法——小波多尺度正交校正(WMOSC)法。首先将原始光谱进行小波变换(DWT),消除噪声及背景信息,然后采用正交信号校正(OSC)滤除与待测组分浓度无关的全部信息。与单纯的小波变换及正交信号校正相比,WMOSC能有效地扣除背景和噪声干扰,使模型具有更强的抗干扰能力,提高了模型的预测精度。利用该方法对牛奶样品的近红外光谱进行处理,采用偏最小二乘法建立校正模型,其牛奶中脂肪、蛋白质和乳糖的预测均方根误差(RMSEP)分别为0.101 6%,0.087 1%和0.110 7%。实验结果表明该方法能有效地去除干扰,保留有用信息。

关键词 小波多尺度正交校正;干扰;扣除;近红外光谱

中图分类号:O657.3 **文献标识码**:A **文章编号**:1000-0593(2008)04-0825-04

引 言

近红外光谱技术作为一种快速、简便、非破坏性的定性和定量分析方法,已广泛应用于食品、石油、化工、农业、医药等领域[1]。它不仅可应用于实验室分析,而且适用于现场快速检测和实时在线分析。

近红外光谱除样品的自身信息外,还包含了其他无关信息和噪声,如电噪声、样品背景和杂散光等。因此,在建立校正模型时,消除光谱数据无关信息和噪声的预处理方法变得十分关键和必要。特别是背景和噪声干扰的扣除,对建立预测精度高、稳健性好的分析模型至关重要,有时甚至起决定作用[2]。但是利用已有的处理方法,预测结果并不理想。为此,一些光谱工作者采用小波变换(WT)[3-6]、正交信号校正(OSC)[7-10]和净分析信号(NAS)[11]等新的方法来解决这一问题。小波变换是一种强有力的信号处理方法,具有灵活的多尺度特征,能将重叠混合信号分解为一系列不同频率的基元信号,实现对信号时频域的局部化分析。这种方法最终将不同频段的小波系数重建光谱,利用PLS等校正方法建立定量分析模型。因此,小波系数的准确与否将直接影响模型的校正效果,进而影响到预测效果。要想提高校正模型的预测能力和稳健性,就必须对小波系数重构的信号进行校正,从而得到最佳的重建光谱。本研究将离散小波变换(DWT)和正交信号校正法相结合,提出了一种应用于近红外光谱分析中扣除背景和噪声干扰的新方法。实验结果表明,与已有的DWT-PLS,OSC-PLS和PLS方法相比,本方法处理后建立的PLS校正模型具有更好的预测精度和抗干扰能力。

1 原理与方法

1.1 基础理论

小波变换[12]的实质是对原始信号的滤波过程,小波函数选取的不同,分解结果也不同。但无论小波函数如何选取,每一分解尺度所用的滤波器中心频率和带宽成固定的比例,即具有所谓的"恒Q"特性。因此,各尺度空间内的尺度信号和细节信号能提供原始信号的时频局域信息。若使不同尺度上的小波系数转换成与原始光谱维数相同的信号,则可将这些信号形成特征向量供信息提取使用,这就是基于小波变换提取多尺度空间有用信息的基本原理。具体的小波变换过程如图1所示。

其中c和d分别为逼近系数和细节系数;H和G分别为

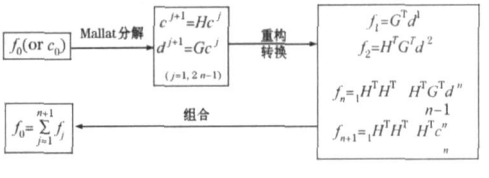

Fig. 1 Diagram of wavelet decomposition and reconstruction

收稿日期:2007-06-16,**修订日期**:2007-09-28

基金项目:"十一五"国家科技支撑计划项目(2006BAI03A03);国家自然科学基金项目(30700168)和天津市自然科学重点基金项目(023800411)资助

作者简介:彭 丹,女,1979年生,天津大学精密仪器与光电子工程学院博士研究生 e-mail:pengdantju@gmail.com *通讯联系人

低通滤波器和高通滤波器；f_0 (or c^0) 为原始信号；f_1，f_2，…，f_n 为不同尺度下重构后的细节信号；f_{n+1} 为分解 n 次后重构的逼近信号。重构后的逼近信号反映了原始信号的“骨架”信息，而细节信号反映了局部的细微信息，它们与原始信号具有相同的维数。

1.2 小波多尺度正交校正法

小波多尺度正交校正(wavelet multi-scale orthogonal signal correction，WMOSC)是一种基于小波变换多尺度空间提取有用信息的方法，它的基本思想是先对近红外光谱进行小波变换，得到一系列尺度下阈值量化后的细节信号及逼近信号，将其与浓度矩阵 y 正交，滤除与组分浓度无关的信号，然后对新的信号(逼近和细节信号)进行重新组合。

有关 OSC 算法可参考文献[13]，一般光谱信号是离散的，因此采用与之匹配的离散小波变换进行分解与重构。WMOSC 方法的校正过程如图 2 所示，其具体步骤为：(1) 首先对光谱数据进行小波分解，采用阈值法[12]去除噪声和背景信息，然后对分解得到的低频系数和阈值量化处理后的高频系数进行重构，转换成与原始光谱维数相同的信号 f_j，即逼近信号和细节信号；(2) 将不同频率区域的信号 f_j 分别与浓度矩阵 y 进行 OSC 校正，得到相应的得分矩阵 w_j 和载荷矩阵 p_j，即 $\mathrm{OSC}(f_j, y) \rightarrow f_j^{\mathrm{OSC}}$，$w_j$，$p_j$；(3) 对 OSC 处理后的信号 f_j^{OSC} 进行组合，得到新的光谱数据 X^{WMOSC}，即 $X^{\mathrm{WMOSC}} = \sum_{j=1}^{n+1} f_{\mathrm{OSC}\,j}$；(4) 校正预测集 X_{test}，先对 X_{test} 进行小波变换得到不同尺度下的信号 $f_{j,\mathrm{test}}$，再利用步骤(2)中保存的 w_j，p_j 与 $f_{j,\mathrm{test}}$ 进行 OSC 正交，即 $f_{j,\mathrm{test}}^{\mathrm{OSC}} = f_{j,\mathrm{test}} - f_{j,\mathrm{test}} w_j (p_j^{\mathrm{T}} w_j)^{-1} p_j^{\mathrm{T}}$，然后将各个尺度下校正后的信号 $f_{j,\mathrm{test}}^{\mathrm{OSC}}$ 进行求和，得到预测集 X_{test} 的 WMOSC 校正光谱，即 $X_{\mathrm{test}}^{\mathrm{WMOSC}} = \sum_{j=1}^{n+1} f_{\mathrm{OSC}\,j,\mathrm{test}}$。

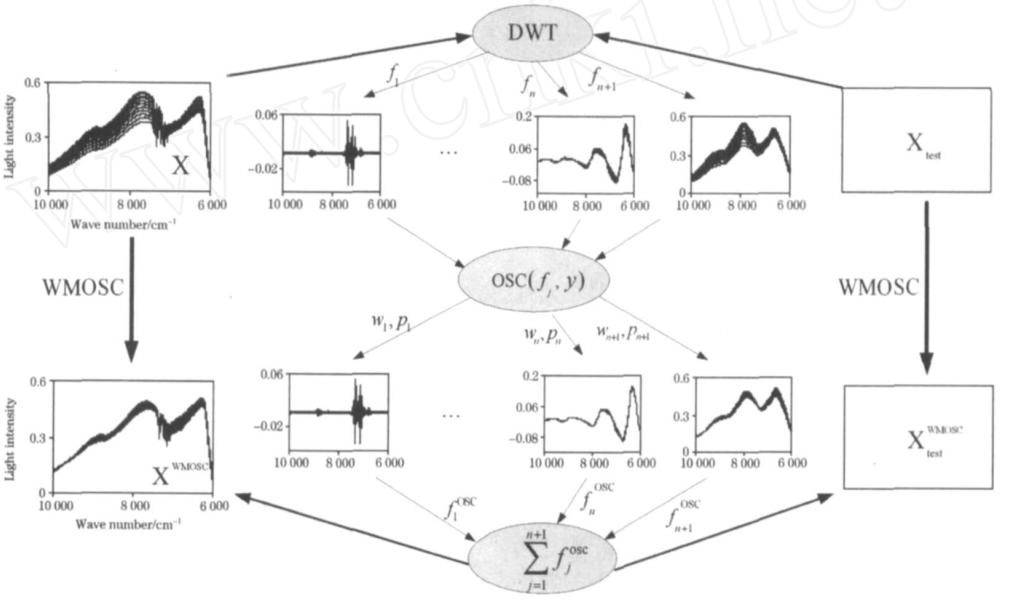

Fig. 2 Diagram of wavelet multi-scale orthogonal signal correction algorithm

2 实验部分

2.1 仪器及光谱采集

Bruker 公司的 VECTOR22 型 FTIR 光谱仪，利用积分球漫反射系统采集光谱。光谱范围为 6 000～10 000 cm^{-1}，扫描分辨率为 8 cm^{-1}，扫描次数为 64 次，PbS 检测器。测量前样品经搅拌后置于样品杯中。

2.2 样品制备

样品由北京牛奶中心提供，共 100 个样品，牛奶中脂肪、蛋白质和乳糖的参考值由丹麦 FOSS 公司生产的乳品综合成份指标分析仪测定。

2.3 数据分析

利用 CAMO 公司的 Unscrambler 7.8 软件建立 PLS 校正模型。光谱经预处理后，分别建立脂肪、蛋白质和乳糖含量的定量模型，对模型进行完全交互验证，引入校正均方根误差(RMSEC)和预测均方根误差(RMSEP)作为判断标准。

DWT 和 OSC 算法程序，均由 Matlab2006 语言编制。

3 结果与讨论

3.1 牛奶的近红外光谱

图 2(左 X)为牛奶的近红外漫反射光谱，谱带较宽且重叠严重，很难精确谱带的归属，为提取有用信息，建立定量模型，需要借助于信号处理与多元统计方法。在建模前对光谱进行处理，提高光谱的信噪比。图 2(左 X^{WMOSC})为采用小波正交校正(WMOSC)处理后的牛奶近红外漫反射光谱，与原始光谱 X 相比可以看出，经 WMOSC 处理后的光谱发生了较大变化，这是 WMOSC 将原始光谱中与组分不相关的所有信息去除的结果。在 6 000～10 000 cm^{-1} 谱区内，牛奶近红外光谱主要反映脂肪、蛋白质和乳糖中含氢基团一级和二级倍频的吸收信息，与近红外其他区域相比，这一波段的吸

收相对较强，为开展牛奶成分含量的定量分析提供相对丰富的信息。此外，不同物质的信息位置不同，最佳建模谱区也不相同。因此，需要建立每种物质的独立校正模型。

3.2　主成分数对预测结果的影响

在建立校正模型时，主成分数对模型的质量起着决定性作用。如果建模时使用的主成分数过少，则不能充分反映未知样品被测组分产生的量测数据变化，其模型预测准确度就会降低；反之，使用过多的主成分数建模，则会将一些代表噪声的主成分加到模型中，使模型的预测能力下降。本实验考察了 RMSEP 随主成分数变化情况，结果如图 3 所示。从图 3 可以看出，对于 WMOSC-PLS 模型下脂肪、蛋白质和乳糖的预测指标 RMSEP 的变化规律基本相同，即随主成分数的增加，先快速下降后趋于平坦，但仍有微小变动。考虑到样品的复杂性、模型的通用性以及模型的预测准确性，对于脂肪、蛋白质和乳糖的 WMOSC-PLS 建模分别选择 6，8 和 7 作为主成分数。

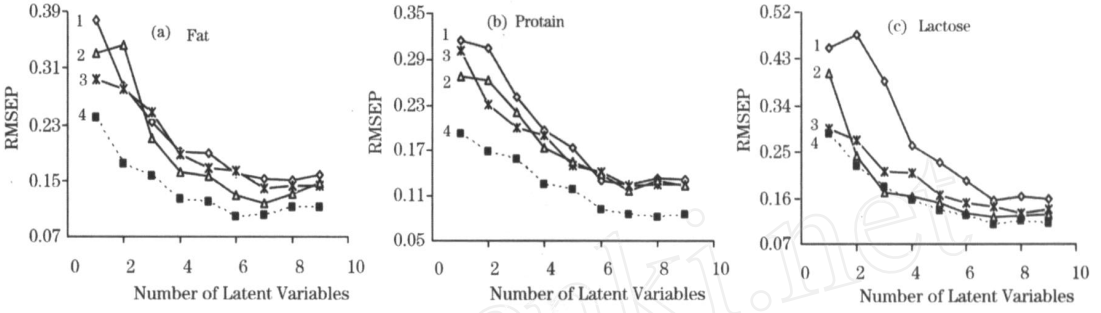

Fig. 3　Changes of RMSEP curves of milk content with number of latent variables

1: PLS; 2: DWT-PLS; 3: OSC-PLS; 4: WMOSC-PLS

3.3　扣除干扰信息的方法

背景和噪声干扰是光谱分析中经常遇到的问题，在近红外光谱应用中，往往存在着有用信号较弱，背景信号及噪声较强的情况。但传统的基于单一尺度的数据处理方法，只能通过校正使待测组分与背景组分峰尽量分开以减小背景干扰。然而，实测光谱大多是多尺度的，即在不同尺度下有用信息和干扰信息的数据分布并不完全相同。在频域中，光谱信息(包括有用信息和背景信息)主要集中在低频部分，噪声主要分布于高频部分，因此可采用小波变换滤除光谱中的高频噪声，从而提高光谱信号的信噪比。

通常不同频段上的光谱信号又可分为相关信息与不相关信息两类。对于牛奶的主要成分含量测量而言，相关信息主要是指组分信息，不相关信息主要是指背景信息和高频噪声，尤其是组分与背景信息，它们无论在时域或是频域都复杂地重叠在一起。因此，仅通过小波变换方法不能完全消除样品背景的影响。从图 3 可以看出，与 PLS 模型的 RMSEP 结果相比，DWT-PLS 模型对脂肪和乳糖的预测精度有明显提高，而对蛋白质的预测精度几乎没有影响，这是由于小波变换只处理光谱数据本身，忽略了浓度信息的影响，从而导致预测结果的较大偏差。这说明仅采用 DWT 扣除干扰信息难以得到最佳预测结果，有可能损失部分的建模有用信息或者对无关信息扣除得不完全，进而影响到建模的质量[14]。Wold 等提出了正交信号校正(OSC)方法，其目的就是去除原光谱矩阵中包含的与浓度阵不相关的变异信息，以提高模型的预测能力[15]。本研究将 DWT、OSC 和 PLS 相结合建立牛奶近红外光谱的定量校正模型(WMOSC-PLS 模型)，选择 db4(Daubechies 小波，$N = 4$)分解原始光谱，分解水平为 6。将阈值量化后的小波系数分别进行重构，得到与原始光谱信号维数相同的一系列信号 f_1 ：f_7，然后分别对其进行 OSC 校正，过程如图 2 所示。

3.4　模型的比较

图 3 为不同主成分数时采用 PLS，OSC-PLS，DWT-PLS 和 WMOSC-PLS 建模方法的预测结果，可以看出采用 WMOSC-PLS 建模方法的稳健性明显好于其它方法。进一步比较，如表 1 所示，WMOSC 法处理后建立的分析模型使预测精度得到了显著的提高。显然，采用频域展开方式和信息正交化提取方法可以最大程度地扣除背景和噪声干扰，有利于提高被测物质吸收光谱的信噪比，为建立更准确可靠的分析模型提供了有效方法。

4　结　论

背景和噪声干扰一直是光谱分析上的难题，人们一直在为寻找客观有效的方法而不断努力，并取得了一定的成果。本文以牛奶样品为具体研究对象，利用 WMOSC 法消除不同尺度下背景和噪声干扰，最大程度地提取光谱中的有用信

Table 1　Results of the calibration with different pretreatment methods

方法	脂肪		蛋白质		乳糖	
	RMSEC/ %	RMSEP/ %	RMSEC/ %	RMSEP/ %	RMSEC/ %	RMSEP/ %
PLS	0.143 7	0.152 2	0.125 9	0.134 7	0.157 3	0.159 1
OSC-PLS	0.137 1	0.140 3	0.123 6	0.124 4	0.130 5	0.133 7
DWT-PLS	0.114 3	0.118 8	0.111 0	0.112 4	0.122 3	0.123 8
WMOSC-PLS	0.099 6	0.101 6	0.081 3	0.087 1	0.108 2	0.110 7

息。实验结果表明，DWT 多尺度分析能够消除背景和噪声干扰，而 OSC 是对 DWT 算法的一种校正，避免 DWT 消除干扰的缺陷。因此，两种方法的结合能有效扣除测量中不相关信息的干扰，提高测量结果的准确性，具有简单、速度快、校正性能好、分析精度高等优点，适用于近红外光谱分析中干扰信号的有效扣除，尤其是适合于食品、药物、农产品等领域的在线检测。

参 考 文 献

[1] YAN Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄, 赵龙莲, 韩东海, 等). Principle and Application of Near Infrared Spectroscopy(近红外光谱分析基怠与应用). Beijing: Light Industry Press of China(北京: 中国轻工业出版社), 2005.

[2] Seasholtz M B, Kowalski B R. Anal. Chim. Acta, 1993, 277: 165.

[3] TIAN Gao-you, YUAN Hong-fu, CHU Xiao-li, et al(田高友, 袁洪福, 褚小立, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(4): 516.

[4] Tan H W, Steven D. Brown. J. Chemometrics, 2002, 16: 228.

[5] Mittermary C R, Tan H W, Brown S D. Appl. Spectrosc., 2001, 55: 827.

[6] WU Rong-hui, SHAO Xue-guang(吴荣晖, 邵学广). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(4): 617.

[7] ZHAO Jie-wen, ZHANG Hai-dong, LIU Mu-hua(赵杰文, 张海东, 刘沐华). Acta Optica Sinica(光学学报), 2006, 26: 136.

[8] QU Hai-bin, OU Dan-lin, CHENG Yi-yu. J. Zhejiang Univ. (B Life Science), 2005, 6B(8): 838.

[9] Woody N A, Feudale R N, Myles A J, et al. Anal. Chem. 2004, 76: 2595.

[10] Blanco M, Coello J, Montoliu I. Anal. Chim. Acta, 2001, 434: 125.

[11] LIU Rong, LÜLi-na, CHEN Wen-liang, et al(刘 蓉, 吕丽娜, 陈文亮, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(9): 1042.

[12] HAO Yong, CHEN Bin, ZHU Rui(郝 勇, 陈 斌, 朱 锐). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(10): 1838.

[13] Wise B M, Gallagher N B. www.eigenvector.com/MATLAB/OSC.html.

[14] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Progress in Chemistry(化学进展), 2004, 16: 528.

[15] Wold S, Antti H, Lindgren F, et al. Chemometrics and Intelligent Laboratory Systems, 1998, 44: 175.

Application of Wavelet Multi-Scale Orthogonal Signal Correction in Milk Components Measurement Using Near-Infrared Spectroscopy

PENG Dan, XU Ke-xin*, LI Chen-xi

State Key Laboratory of Precision Measuring and Instruments, Tianjin University, Tianjin 300072, China

Abstract Spectral interferences can have a significant impact on the spectral variation and as a consequence can adversely affect the results of calibration model in spectra analysis. Wavelet transform (WT) and orthogonal signal correction (OSC) were both the popular preprocessing algorithms. It was known that the former can effectively eliminate the background and noise and the latter can effectively filter out the interference information irrelevant to analyte concentration during the preprocessing of spectra. According to the different characteristics of analyte information and interference information in near-infrared (NIR) spectra, a new hybrid algorithm (WMOSC) that was the combination of discrete wavelet transform (DWT) and OSC was proposed to eliminate the spectral interferences including background, noise and systemic spectral variation irrelevant to the concentration. First, DWT was used to split the spectral signal into different frequency components, which keep the same data points as the original spectra data, to remove noise and background information by threshold method. Then OSC was applied to each frequency components to remove the information uncorrelated to the concentration independently. Finally, the spectra preprocessed by WMOSC were achieved through the summation of all frequency components. WMOSC was successfully applied to preprocess the NIR spectra data of milk. After elimination of the interference in the NIR spectra data by WMOSC, the partial least squares (PLS) regression was used to develop the calibration models for estimating the contents of main constituents in milk. The prediction ability and robustness of models obtained in subsequent PLS calibration using WMOSC were superior to those obtained using either DWT or OSC alone. The root mean square errors of prediction (RMSEP) of the models for fat, protein and lactose were 0.101 6%, 0.087 1% and 0.110 7%, respectively. The experimental results show that WMOSC is an effective method for eliminating the interferences information in NIR spectra.

Keywords Wavelet multi-scale orthogonal signal correction; Interference; Removal; Near-infrared spectroscopy

*Corresponding author

(Received Jun. 16, 2007; accepted Sep. 28, 2007)