第31卷，第12期
2011年12月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 31，No. 12，pp3228-3231
December，2011

正交信号校正应用于多元线性回归建模的研究

张 娴，袁洪福*，郭 峥，宋春风，李效玉，谢锦春

北京化工大学材料科学与工程学院，北京 100029

摘 要 通过近红外光谱技术建立二元、三元调和食用油中花生油含量模型以及二甲亚砜水溶液浓度模型，比较了分别采用原始光谱和正交信号校正(OSC)处理后光谱进行MLR建模的结果，并对所建的正交信号校正后光谱MLR模型与原始光谱PLS模型进行预测结果比较。比较过程中使用交互验证参数(包括决定系数R_C，标准偏差SEC，预测值和实际值线性拟合方程的斜率a和截距b)以及外部预测统计参数(包括R_V，标准偏差SEP，预测值和实际值线性拟合方程的斜率a和截距b)来评价模型能力。研究结果表明：与原始光谱相比，使用OSC处理后的光谱进行MLR建模(采用相同波长)，得到的模型交互验证结果以及外部预测结果均有变好的趋势。通过优选建模波长组合(无共线性影响)，OSC与MLR联用建模可得到优于PLS模型的结果。

关键词 正交信号校正；近红外；多元线性回归；PLS

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2011)12-3228-04

引 言

近红外光谱是近年来分析化学领域迅猛发展的一种新技术。近红外分析基于被测样本分子振动倍频和组合频吸收光谱相对变化信息，与光谱绝对强度比较，这种相对变化强度是微弱的(10^{-3}～10^{-5} A)，而且样品中不同基团近红外谱带之间重叠严重，谱图解析困难，必须依靠化学计量学多元校正技术进行近红外定性和定量分析，因此，需对光谱进行降噪处理减少噪声对模型的影响，其中，正交信号校正(orthogonal signal correction，OSC)是一种通过剔除与被测目标无关信息的有效光谱预处理方法[1，2]。

OSC，一种可有效剔除与被测性质无关的信号，提高校正后光谱与性质之间线性关系的校正方法。前期研究结果表明，OSC用于偏最小二乘(PLS1)建模并不能有效改善模型预测能力，反而会导致过拟合现象使预测结果变差。虽然OSC在剔除与被测性质无关信息方面效果是显著的，但至今在改善模型预测能力方面并没有找到合适的用途，以发挥其应有的重要作用。为此，本文以近红外光谱法测定三类不同的样本为研究对象，提出OSC应用于多元线性回归(MLR)模型的新方法，并与PLS1建模进行对比分析。发现新方法明显改善MLR模型预测能力，并且避免了与PLS1联用产生的过拟合现象。OSC和MLR联用建模，获得其模型预测能力优于PLS1模型。

1 实验部分

1.1 样品制备

二甲亚砜水溶液是一种用于碳纤维制备的凝浴液，其样品制备：二甲亚砜与纯水，用电子天平按比例称量之后放入锥形瓶静置24 h后进行测量。调合食用油样品制备：采用从市场购买的三种国家标准一级食用油，使用移液管分别配制含量均匀分布的二组分调合油(大豆油、花生油)、三组分调合油(葵花籽油、大豆油、花生油)，放入锥形瓶中摇匀后进行测量。样品含量数据如表1所示。

Table 1 Concentration statistics of all samples

名称	二组分花生油含量/v%	三组分花生油含量/v%	二甲亚砜含量/W%
最小值	0	0	4.973
最大值	100.00	100.00	77.923
平均值	50.00	33.04	41.511
样品数	61	78	60
精度	0.05	0.05	0.001

收稿日期：2010-12-06，**修订日期**：2011-03-10

基金项目：国家(863计划)项目(2009AA04Z129)资助

作者简介：张 娴，女，1986年生，北京化工大学材料科学与工程学院硕士研究生 e-mail：zhangxian1986@163.com

*通讯联系人 e-mail：hfyuan@mail.buct.edu.cn

1.2　数据采集

使用聚光科技公司 SupNIR-5500 型近红外光谱仪采集样品光谱，分辨率 6 nm，扫描范围 1 000～1 800 nm，扫描次数 64 次，2 mm 石英比色皿，以空气为参比，恒温 60 ℃条件下对每个样品重复测量 10 张光谱取平均值作为最终光谱。

1.3　OSC 处理

在众多 OSC 算法中，O-PLS(orthogonal projections to latent structures)处理效率和正交保持性能较其他算法优秀[3]，因此，本文中所选用了 O-PLS 算法。使用 OSC 最佳主成分数确定方法(即将在该刊登出)确定三个样本集预处理的最佳 OSC 成分数依次为 3，8，4。根据样本光谱及含量数据分别建立二元调合食用油，三元调合食用油和二甲亚砜水溶液的样本集，进行校正之前将样本集按比例随机分为校正集和验证集，其中校正集∶验证集＝80∶20。使用 MATLAB7.0 软件对样品集进行 OSC 校正处理，首先使用 OSC 对校正集处理得到 OSC 权重(W)和主成分光谱(P)，结合 W 和 P 对验证集进行 OSC 校正，得到 OSC 验证样本集。

1.4　模型建立

分别使用原始光谱样品集和 OSC 校正后得到的样本集，通过 CM2000(聚光科技公司)化学计量学软件建立定量校正模型。使用原始光谱建立的样品集，将其随机分为校正集(80%)和验证集(20%)，进行均值中心化处理之后，采用 CARS(competitive adaptive reweighted sampling 竞争适应性加权抽样法)方法[4]选择的波长(表 2 和表 3)组合分别建立 MLR 和 PLS1 模型。对 OSC 校正后的样本集分为校正集和验证集(与原始光谱样本集相同)，进行均值中心化处理之后，采用表 2 所示波长组合分别建立 MLR 模型，使用 OSC 验证集对其进行外部验证。

Table 2　Wavelengths used in MLR modeling

性质	波长/nm
花生油(二元调合油)	1 016,1 145,1 152,1 155,1 232,1 324,1 409,1 507,1 524, 1,564,1 566,1 739,1 743,1 748,1 753,1 760,1 786
花生油(三元调合油)	1 034,1 087,1 110,1 288,1 373,1 450,1 605,1,695,1 732,1 789
二甲亚砜	1 115,1 135,1 161,1 173,1 318,1 376,1 393,1 428,1 439,1 453,1 457,1 459, 1 464,1 473,1 480,1 484,1 502,1 505,1 542,1 585,1 698,1 729,1 792

Table 3　Wavelengths used in PLS1 modeling

性质	波长范围/nm
花生油(二元调合油)	1 203～1 218,1 686～1 783
花生油(三元调合油)	1 281,1 321,1 322,1 324,1 447,1 449,1 452,1 453,1 454,1 697,1 708,1 711,1 726
二甲亚砜	1 010,1 015,1 016,1 029,1 042,1 053,1 137,1 144,1 168,1 188,1 227,1 238, 1 282,1 299,1 327,1 407,1 420,1 488,1 520,1 545,1 566,1 709,1 776

1.5　模型评价指标

模型优劣的判定依据：交互验证决定系数 R_C 和外部验证的决定系数 R_V，RC 和 R_V 应最接近于 1；预测标准误差(SECV，SEP)应最接近于参考方法的误差；样本预测含量值(y)与真实含量值(x)之间的线性拟合方程 $y=ax+b$ 中的斜率 a、截距 b，理想模型中，a 和 b 应分别为 1 和 0。

2　结果与讨论

前期工作中发现，OSC 应用于偏最小二乘(PLS1)建模时只能改善其交互验证结果，而作为判别模型优劣的重要指标外部验证结果却得不到改善，其原因是 OSC 与 PLS1 重复剔除与被测性质无关的信息会误删部分有效信息，从而导致过拟合效应使预测结果变差。因此本文提出采用 MLR(可避免重复误删除的算法)结合 OSC 建立模型，为 OSC 的合理应用另辟途径。

使用上述三种样本的原始光谱样本集以及 OSC 校正后的样本集建立 MLR 模型，对模型的交互验证结果与外部预测数据，如表 4 和表 5 所示，进行分析和讨论。

2.1　建模波长的选择

CARS 基于蒙特卡洛采样原理，避免陷入局部最优化，借助计算机可以在上千波长中选择出稳定的、无共线性影响的关键波长组合，是一种优秀的波长选择方法。由于波长采样具有随机性，重复使用 CARS 方法获得的波长组合结果并不是唯一的。为了使选择的波长具有代表性，实验中采取多次重复运行 CARS 程序选择最优化波长组合。每一个模型选择关键波长时，首先重复运行 CARS 200 次得到 200 组波长组合，其次在 200 组波长组合中按照最小 RMSECV 值的原则选出 10 组波长组合，将 10 组波长应用到实际模型中，最终选择得到 SEC 及 SEP 最小的模型所选波长为最优化波长组合。

2.2　有、无 OSC 校正的 MLR 模型交互验证结果对照

由表 4 可以看出所有样本集建立的定量模型相关系数 R_C 均接近理想化状态，并且交互验证得到的拟合方程的斜率 a 与截距 b 分别趋于 1 和 0，说明被测光谱与性质有着非常好的线性关系，预测值和实际值之间没有系统偏差。经过 OSC 处理之后，光谱与性质之间的线性相关型进一步得到改善，a 和 b 以及 SECV 均向理想化状态靠近，说明 OSC 有效地剔除了光谱中的无用信息，达到改善交互验证结果的目的。考察一个模型的优劣，最主要是要考察其预测能力，所以还需要进一步研究模型的外部预测结果。

Table 4 Cross validation results of MLR models and MLR+OSC models

建模方法及结果	二元调合油(花生油)		三元调合油(花生油)		二甲亚砜水溶液	
采用方法	MLR	MLR+OSC	MLR	MLR+OSC	MLR	MLR+OSC
剔除校正样本数*	0	0	4	4	3	3
决定系数 R_C	0.999 93	0.999 95	0.999 7	0.999 9	0.999 99	1.000
交互验证 a	0.999 8	0.999 9	0.999 4	0.999 8	0.999 98	1.000
交互验证 b	0.007	0.005	0.023	0.005	4.5E-4	−3.7E-5
SECV	0.39	0.33	0.92	0.44	0.112	0.021

* 对花生油含量(三元调合油)有、无 OSC 处理 MLR 建模剔除的样本均为相同的 4 个样本，均为性质残差过高。* 二甲基亚砜有、无 OSC 处理 MLR 建模剔除样本均为相同的三个样本，理由为性质残差过高

Table 5 Predication results of MLR models and MLR+OSC models

建模方法及结果	二元调合油(花生油)		三元调合油(花生油)		二甲亚砜水溶液	
采用方法	MLR	MLR+OSC	MLR	MLR+OSC	MLR	MLR+OSC
决定系数 R_V	0.999 76	0.999 77	0.999 7	0.999 8	0.999 95	1.000
外部验证 a	1.009	1.002	1.002	0.993	1.001	0.997
外部验证 b	−0.339	−0.073	0.481	0.392	−0.131	0.123
SEP	0.62	0.60	0.69	0.59	0.227	0.094

2.3 有、无 OSC 校正对 MLR 模型外部验证结果比较

使用三个样本集中随机分出来的验证集(与性质变化无关)对模型预测能力进行考察，结果表明(如表 5 所示)，对于三种样本应用，与原始光谱建立模型的预测结果相比，用 OSC 校正光谱集所建模型预测结果的决定系数 R_V 均有所改善。

经 OSC 校正后，外部预测拟结果中 a 与 b 变化并无明显规律，可能与 OSC 剔除的光谱信息有关。值得注意的是，外部验证预测标准偏差 SEP 毫无例外地全部明显变小，这说明，OSC 与 MLR 联用，可以得到更好的预测结果，但是，比其他算法建立的模型是否更为优越还需做进一步比较。

2.4 OSC 和 MLR 联用模型与 PLS1 模型的外部预测结果对照

根据表 5 和表 6 不难发现，原始光谱建立的 MLR 模型与 PLS1 模型相比，PLS1 得到更小的 SEP，这就体现出 PLS1 比 MLR 优势。但是经过 OSC 处理后，MLR 模型的 SEP 大大减小，比原始光谱建立的 PLS1 模型 SEP 还要小(如表 6 所示)，并且建立 MLR 模型所使用的波长数也大大减少。三种模型都得到了同样的规律，说明 OSC 能够改善 MLR 建模结果并非偶然现象。

Table 6 Predication results of MLR models and PLS1 models

建模方法及结果	花生油(二元调合油)		花生油(三元调合油)		二甲亚砜水溶液	
采用方法	PLS	MLR+OSC	PLS	MLR+OSC	PLS	MLR+OSC
决定系数 R_V	0.999 74	0.999 77	0.999 6	0.999 8	0.999 992	0.999 996
外部验证 a	1.006	1.002	0.999	0.993	0.998	0.997
外部验证 b	−0.311	−0.073	0.509	0.392	0.103	0.123
SEP	0.61	0.60	0.66	0.60	0.109	0.094

2.5 OSC 和 MLR 联用模型改善预测能力的理论分析

无论 MLR 还是 PLS 建模，都是依靠回归方法建立定量关系，其方法之间的很大区别在于在剔除噪声干扰方面有着明显不同。PLS 校正算是通过对 $\boldsymbol{X}$ 矩阵和 $\boldsymbol{Y}$ 矩阵分别进行主成分分析，通过选取能够反映分析体系数据最大方差变化的一系列主成分(通常是最初的几个主成分)建立模型，由于不同主成分是相互正交的，因此，PLS 建模可有效地克服了 MLR 中遇到“共线性”局限。为解决单纯 PCA 得到的主成分并非与被测性质信息相关，PLS1 在矩阵分解迭代过程中引入了 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 交叉步骤，旨在这样得到的光谱主成分与被测性质信息相关。与 MLR 相比，PLS1 建模可以使用更多的谱带信息或全谱信息，通过抛弃后边对有用信息贡献小的主成分(主要反映噪声信息)来剔除噪音，保证了建模变量之间的正交性，因此，显著地改善了建模效果，并成为近红外光谱分析建模的主流方法。OSC 是通过类似 PLS 迭代算法，得到与被测性质无关的信息部分，从原始光谱中扣除这些噪音。OSC 与 PLS 二者在剔除噪声目的上是相同的，但采用的方法路线不同。但是，对于被测性质在剔除噪声方面，OSC 比 PLS 更具有针对性。因此，如果在选取足够好的波长(即能够充分反映建模有关的化学信息，并保证不存在“共线性”现象)前提下，发挥 OSC 剔除与被测性质无关信息的优势，MLR 和 OSC 联用建模的预测能力优于 PLS 建模是可能的。

3 结 论

OSC 校正能有效地剔除光谱中与被测性质无关的信号，

但 OSC 用于 PLS1 建立模型会出现过拟合的风险，这也是 OSC 至今没有得到广泛应用的原因。OSC 与无迭代算法如多元线性回归联用建模可避免该风险，不仅能明显改善 MLR 模型预测能力，并且得到了比 PLS1 建模更小的 SEP。OSC 和 MLR 联用建模模型可以充分发挥其剔除无关信息改善光谱与被测性质之间线性关系的优点，获得优于 PLS1 模型的预测能力。

References

[1] Wold S, Antti H, Lindgren R, et al. Chelnometrics and Intelligent Laboratory Systems, 1998, 44: 175.

[2] Bertran E, Iturriaga H, Maspoch S, et al. Analytica Chimica Acta, 2001, 431: 303.

[3] YU Hao, CHENG Yi-yu, QU Hai-bin(余 浩，程翼宇，瞿海斌). Pretreating Near-infrared Spectra by Orthogonal Signal Correction(基于正交信号校正算法的近红外光谱预处理). Hangzhou: Zhejiang University(杭州：浙江大学), 2004.

[4] LI Hongdong, Liang Yingzeng, Xu Qingsong, et al. Analytica Chimica Acta, 2009, 648: 77.

Study on Building MLR Model Using Orthogonal Signal Correction

ZHANG Xian, YUAN Hong-fu*, GUO Zheng, SONG Chun-feng, LI Xiao-yu, XIE Jin-chun

Beijing University of Chemical Technology, Beijing 100029, China

Abstract MLR and PLS models with or without OSC were studied through establishing quantitative calibration models for the peanut oil content in blending edible oils, and for the dimethylsulfoxide concentration in water solution. The cross validation results and the predication results of MLR models, were compared to evaluate the effectiveness of OSC for improving the performance of MLR model. The results show that the SEC or SEP of MLR models using OSC gets smaller. Selecting appropriate wavelengths combination by CARS method, prediction capacity of MLR model using OSC is better than PLS1 model using raw spectrum.

Keywords Orthogonal signal correction; Near infrared spectroscopy; MLR; PLS

(Received Dec. 6, 2010; accepted Mar. 10, 2011)

* Corresponding author