DOI:10.16866/j.com.app.chem2011.07.037

汽油辛烷值NIR数据处理与建模仿真

王瑾，蒋书波

(南京工业大学自动化与电气工程学院，江苏，南京，210009)

摘要：随着当今计算机与各类程序软件的开发使用，化学计量学不断发展，人们可以在近红外光谱区内采集大量的数据，并使用各种有效的统计方法，把近红外光谱技术应用于定性与定量。近红外光谱为分子振动光谱的倍频和组合频谱带，主要是对含氢基团的吸收，包含有绝大多数类型有机物的组成与分子结构的丰富信息。原理是基于不同的基团或同一基团在不同化学环境之中吸收波长的差异。该技术具有快速、简便、样品无需处理、适合在线分析等特点。近红外光谱技术主要应用于农产品、医学、制药等行业，研究的热点方向是仪器的相关改进和新的化学计量学算法。对近红外光谱数据进行处理的常用化学计量学方法为多元校正方法，主要有逐步多元线性回归、主成分回归、偏最小二乘法、拓扑学方法以及人工神经网络法等。本文介绍了近红外光谱分析技术以及在相关领域中的应用。通过对汽油近红外光谱数据的考察，建立汽油辛烷值的定量校正模型。分别采用偏最小二乘回归与主成分回归建立校正模型并实现分析。实验结果表明，应用仿真工具可以对近红外光谱数据进行有效并可靠的处理与分析。

关键词：近红外光谱分析；汽油辛烷值；偏最小二乘；主成分分析

中图分类号：TP273; TP319:TQ02; TP391.9　　**文献标识码**：A　　**文章编号**：1001-4160(2011)07-947-950

1 引言

近红外光谱分析技术(near infrared，NIR)是近年来发展起来的一种快速检测技术，具有无污染、无破坏、分析速度快、效率高、成本低及可以实现在线分析等特点，在食品、医药、化工、石油等领域获得了空前的发展，并且其涉及的领域越来越广[1]。

近红外光谱(NIR)分析技术以其快速、简便及无损等特点，在复杂样品化学成分的测定中占有日趋重要的地位。近红外光谱产生于分子振动，吸收偏弱，吸收峰重叠严重，并且从多组分复杂样品得到的近红外光谱往往不是各组分光谱的简单叠加，必须借助于化学计量学方法才能进行定性与定量分析。因此，化学计量学方法已成为近红外光谱分析中的研究热点。各种多元校正技术，如多元线性回归(multiple linear regression，MLR)、主成分回归(principal component analysis，PCR)、偏最小二乘回归(partial least squares regression，PLSR)和人工神经元网络(artificial neural network，ANN)等方法在近红外光谱分析中已得到了广泛的应用[2]。

2 辛烷值

辛烷值是表征汽油抗爆性的重要指标，其标准方法是ASTM-CFR辛烷值机台架方法，按照国标分别测定汽油的研究法辛烷值(RON)和马达法辛烷值(MON)，测定时样品需用量较大，传统辛烷值机分析时间较长，且测试费用较高。可以看出，以上分析方法不能满足实际生产中快速分析的要求。而成品汽油的生产、出厂要严格控制，严把质量关，同时油品的生产还要从源头抓起，从半成品抓起，搞好汽油调和就需要建立一种快速、准确的分析方法，解决分析滞后的问题，及时指导生产并提高经济效益。

为了建立一种快速、简便、适于在线分析的方法，Kelly 等首先采用近红外光谱对测定汽油的辛烷值及组成进行了尝试，并取得较好的结果。这些技术被很快用于炼油工业，并实现了过程的在线分析。我国近红外光谱技术在石油化工领域中的应用虽起步较晚，但已取得较大的进展：石油化工科学研究院对近红外光谱在油品分析中的应用作了较系统的研究，为了填补我国近红外仪器的空白，还自行研制开发了多通道近红外光谱仪，并对该仪器测定汽油的性质及组成进行了较系统的研究，取得了较好的结果[3]。

3 模型的理论分析

3.1 主成分回归

主成分回归法(PCR)用于待测样品的光谱 x，首先由主成分分析得到载荷矩阵，求取其得分向量：t=x*P。然后，通过主成分回归模型 b 得到最终的结果：y=t*b。在主成分回归中，确定参与回归的最佳主成分数最为重要。

收稿日期：2011-05-16；**修回日期**：2011-06-26
基金项目：南京工业大学学科建设基金重点项目资助(项目编号：39710002)
作者简介：王瑾(1986—)，男，吉林人，硕士研究生.
联系人：王瑾，E-mail:wangjinww@gmail.com.

如果选取的主因子过少，将会丢失原始光谱较多的有用信息，拟合不充分；如果选取的主因子过多，会将测量中的噪声过多的包含进来，出现过拟合现象，所建模型的预测误差也会明显增大。因此，合理确定主成分数是必要的工作。

对于多元线性回归问题，当变量线性相关及矩阵病态时，无法获得精确解。主成分回归的应用能够很好的解决线性回归所遇到的变量线性相关、矩阵病态或变量过多所带来的相关问题。

设有化学计量模型：

$$Y_{n\times p} = X_{n\times m}B_{m\times p} + E_{n\times p} \quad (1)$$

首先对 X 阵进行主成分分析：

$$T = XP \quad (2)$$

T 阵的维数可以与 X 阵相同，如果使用整个 T 阵参加回归，这样得到的结果与多元线性回归没有很大的差别。因为主成分是原变量的线性组合。前面的 k 个主成分包含了 X 矩阵的绝大部分有用信息，而后面的主成分则往往与噪声和干扰因素有关。因此参与回归的是少数主成分组成的矩阵。在维数上远小于 X。

将 T 与 Y 阵进行多元线性回归：

$$Y = TB + E \quad (3)$$

$$B = (T^TT)^{-1}T^TY \quad (4)$$

对于未知的样本有：

$$Y_{un} = T_{un}B = X_{un}PB \quad (5)$$

由此可以看出，主成分回归通过对参与回归的主成分的合理选择，可以去掉噪音。主成分之间相互正交，解决了多元线性回归中的共线性问题。主成分回归能够充分利用数据信息，有效的提高了模型的抗干扰能力。但是，主成分回归运算速度缓慢，建立的数学模型不容易理解[4]。

在 PCR 中，只对光谱阵 X 进行分解，消除无用的噪声信息。同时，浓度阵 Y 也包含有无用信息，应对其作同样的处理，且在分解光谱阵 X 时应考虑浓度阵 Y 的影响。偏最小二乘法(PLS)就是基于以上思想提出的多元回归方法[5]。

3.2 偏最小二乘回归

偏最小二乘法(PLS)是把模型式的方法和认识性的方法有机的结合起来。在一个算法下，可以同时实现回归建模(多元线性回归)、数据结构简化(主成分分析)以及2组变量之间的相关性分析(典型相关分析)。它的提出是多元统计数据分析中的一个飞跃。偏最小二乘法是一种多因变量对多自变量的回归建模方法，可以较好的解决许多以往用普通多元回归无法解决的问题。

偏最小二乘回归的建模方法：设有 q 个因变量和 p 个自变量。为了研究因变量与自变量的统计关系，观测了 n 个样本点，由此构成了自变量与因变量的数据表 X 和 Y。偏最小二乘回归分别在 X 与 Y 中提取出 t 和 u，要求：(1) t 和 u 应尽可能大地携带它们各自数据表中的变异信息；(2) t 和 u 的相关程度可以达到最大。在第一个成分被提取后，应用偏最小二乘回归分别实施 X 对 t 的回归以及 Y 对 t 的回归。如果回归方程已经达到满意的精度，则算法终止；否则，将利用 X 被 t 解释后的残余信息以及 Y 被 t 解释后的残余信息进行第二轮的成分提取。如此往复，直到能达到一个较满意的精度为止。若最终对 X 共提取了多个成分，偏最小二乘回归将通过施行 y_k 对 X 的这些成分的回归，然后再表达成 y_k 关于原自变量的回归方程。

同传统的多元线性回归模型相比，偏最小二乘法的特点有：能够在自变量存在多重相关性的条件下完成回归建模；允许在样本点个数少于变量个数的条件下完成回归建模；偏最小二乘回归法在最终模型中将包含原有的所有自变量；偏最小二乘回归模型更便于识别系统信息与噪声；在偏最小二乘回归模型中，每一个自变量的回归系数将更易于解释。本文是在 MATLAB 环境下进行数据处理与建模过程。

MATLAB 语言比较易于学习，因为其只有一种数据类型，一种标准的输入输出语句，不用“指针”，不需编译，比其他语言少了很多内容[6]。

在对产物的测量光谱所得的数据处理方面，应用MATLAB进行建模与仿真已经是一种十分方便与可靠的方法：翁欣欣等使用一种改进的BP算法Levenberg-Marquardt方法用于橄榄油掺杂的NIR分析，使初榨橄榄油的识别率达到了 100%[7]。方利民等使用基于独立分量分析-神经网络回归(independent component analysis-neural network regression，ICA-NNR)用于分析柴油的近红外光谱，测定值与预测值的相关性及相对误差均优于现行常用的PLS、PCR方法[8]。阎宇等使用实数编码遗传算法 (real-coded genetic algorithm，RNCGA)优化的BP神经网络建立了快速、准确测定石脑油馏程的分析仿真模型[9]。Roman M. Balabin等在用NIR对汽油分类的研究中对多种光谱数据处理技术做出了详细的对比及分析，其中概率神经网络(probabilistic neural network，PNN)及支持向量机(support vector machine，SVM)的结果要优于传统的PLS，线性判别分析(linear discriminant analysis，LDA)等方法，但需考虑人工神经网络的训练时 间[10]。综上分析，本文选用了PCR与PLS算法进行谱图数据建模仿真与对比。

4 仿真验证

本文考察的近红外数据来自参考文献[11]。其中数据集 1 包括小麦样品的近红外数据及水分和蛋白质的含量。数据集 2 包括 60 个汽油样本的近红外光谱和辛烷值，(900~1700) nm 每 2 nm 进行 1 次采样。本文从数据集 2 中选取特定的校正采样值(GASc)与验证采样值(GASv)。

在 MATLAB 中有专门的求 PLSR 的函数[XL，YL，XS，YS，BETA，PCTVAR，MSE，stats] = PLSREGRESS(X，Y，ncomp，...)与求 PCR 的函数[COEFF，SCORE，latent，tsquare] = princomp(X)；在图 1 中可以比较 PLSR 与 PCR 对光谱数据进行 2 个主成分分析的效果，可以看出：PLSR 对输出响应的预测已经很好，呈良好的线性，但 PCR 对 2 个主成分的回归相比 PLSR 要差一些。

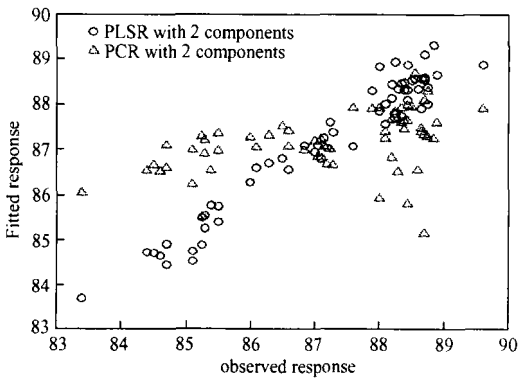

Fig.1 Observed response of PLSR and PCR with 2 components.
图 1 PLSR 和 PCR 在 2 个主成分下的观察响应

由图 2 可以得知：在主成分数为 2 时，PCR 相比 PLSR 对 X 阵的信息利用反而要好些，但在对 Y 值拟合时，PCR 忽略了有关的重要信息。

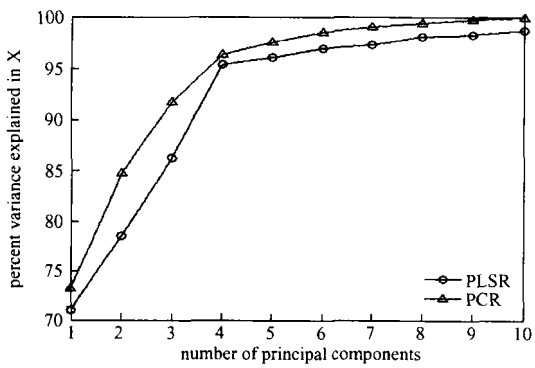

Fig.2 PLSR and PCR in the different principal components regression under the X array.
图 2 PLSR 与 PCR 在不同主成分数下 X 阵回归

最后在图 3 中显示的是不同主成分数下 PLSR 与 PCR 的期望误差，可以看出，在 2 个主成分数时，PLSR 的误差已经很小，而 PCR 的误差还没有得到控制，在主成分数为 4 时，两者的拟合精度都得到保证。

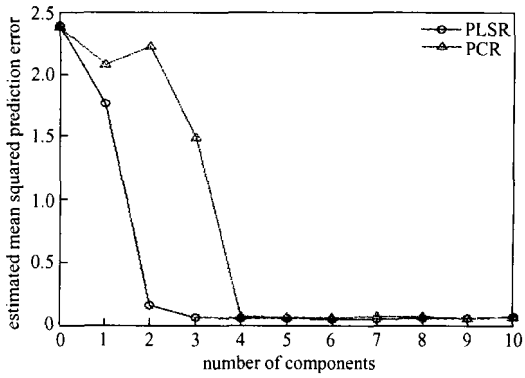

Fig.3 PLSR and PCR in the different expectations of the principal component scores of errors.
图 3 PLSR 与 PCR 在不同主成分数下的期望误差

5 结果与讨论

在现代对 NIR 光谱模型建立多元定量校正法主要分为 2 类：线性的有多元线性回归(MLR)，主成分回归(PCR)和偏最小二乘法(PLS)等；非线性的有人工神经网络(ANN)和支持向量回归(support vector regression，SVR)等。MLR、PCR 和 PLS 是一脉相通的：PCR 采用 PCA 对光谱阵 X 进行分解，通过得分向量进行 MLR 回归，提高了模型的预测能力；PLS 则对光谱阵 X 与浓度阵 Y 同时分解，考虑了两者间的关系。可以说，PLS 是多元线性回归、典型相关分析和主成分分析的完美结合，在 NIR 光谱分析中，PLS 的应用最为广泛。

PLS 和 PCA 法均是从传统多元回归方法种发展而来的线性回归方法，由于采用了主成分分析的方法，提取化学测量中的有用信息，去除了无用的信息和噪音，在复杂体系分析中显现出其特有的优越性。本文应用 PLS 和 PCA 方法在 MATLAB 中进行了建模仿真验证，2 种方法的预测结果均较为理想。实验表明，在主成分数较小的情况下应用 PLS 法进行近红外光谱分析的准确度相对更高。

References:

1 Lu Wanzhen, Yuan Hongfu, Xu Guangtong. Modern Near Infrared Spectroscopy. Beijing:China Petrochemical Press, 2000.
2 Li Yankun, Shao Xueguang, Cai Wensheng. Based on multi-model consensus partial least squares quantitative analysis of near infrared spectroscopy. Chemical Journal of Chinese Universities, 2007, 28(2):246-249.
3 Lv Qiuling. Near infrared spectroscopy in gasoline of rapid determination of gasoline properties. Shandong University, 2008.
4 Shi Yonggang, Feng Xinlu, Li Zicun. Chemometrics. Beijing: Petrochemical Press, 2003.
5 Lu Wanzhen, Yuan Hongfu, Chu Xiaoli. Near Infrared Spectroscopy Instruments. Beijing:Chemical Industry Press, 2010.
6 Bian Chunyu. Application of MATLAB. Science and Technology Communication, 2010, (8):118-120.
7 Weng Xinxin, Lu Feng, Wang Chuanxian, Qi Yunpeng. Nearinfrared spectroscopy method for BP neural network PLS

doping of olive oil. Spectroscopy and Spectral Analysis, 2009, 29(12):3283-3286.
8 Fang Limin, Lin Min. Diesel near infrared spectra of independent component analysis. Acta Petrolei Sinica (Petroleum Processing Section), 2008, 24(6):726-732.
9 Yan Yu, Cheng Mingxiao, Lin Jinguo, Li Junhua, Wang Jing gang. GABP Raman spectroscopy combined with neural network determination of naphtha boiling range. Automation and Instrumentation, 2010, 25(4):10-15.
10 Balabin R M, Safieva R Z, Lomakina E I. Gasoline classification using near infrared (NIR) spectroscopy data: Comparisonof multivariate techniques. Analytica Chimica Acta, 2010, 671(1):27-35.
11 ftp://ftp.clarkson.edu/pub/hopkepk/Chemdata/Kalivas/

NIR gasoline octane data processing and modeling

Wang Jin and Jiang Shubo

(School of Automation & Electrical Engineering, Nanjing University of Technology, Nanjing, 210009, Jiangsu, China)

Abstract: With the development of today's computer, various types of softwares and chemical metrology, people can collect large amounts of data in the near infrared spectral region, using a variety of effective statistical methods used in the NIR qualitative and quantitative analysis. Near infrared spectroscopy for molecular vibrational spectra of the frequency and combination of spectral bands, mainly on the absorption of hydrogen groups, containing most types of organic matter composition and molecular structure of the wealth of information. Principle is based on the same groups or different groups in different chemical environment differences in absorption wavelength. The technology is fast, simple, sample without treatment, suitable for online analysis and so on. Near infrared spectroscopy is mainly used in agricultural, medical, pharmaceutical and other industries, a hot research direction is related to improved equipment and new chemometric algorithms. The near infrared spectral data processing methods commonly used chemometric multivariate calibration methods, there are stepwise multiple linear regression, principal component regression, partial least squares, topological methods and artificial neural network method. Near infrared spectroscopy technology and applications are introduced in related fields in this article. A quantitative calibration model is established by near-infrared spectral data on gasoline investigation. PLS and PCR are used to establish calibration models and to achieve analysis. Experimental results demonstrate that for the NIR data, MATLAB can be effective and reliable to process and analysis.

Keywords: Near infrared spectrum, gasoline octane, PLS, PCR

(Received: 2011-05-16; Revised: 2011-06-26)