文章编号 1004-924X(2000)03-0238-04

PLS 定标法在近红外光谱分析仪中的应用研究

张 玲
(中国科学院长春光学精密机械与物理研究所科技总公司，吉林 长春 130021)

摘要：近红外光谱分析仪在农产品、水质监测和石油等许多领域有广泛的应用。这是一种在线实时分析仪器。光谱分析仪的定量分析有许多种方法。PLS(Partial Least－Squares)法是在传统的多元线性回归的基础上发展起来的一种回归方法。文中针对40种烟草样品的红外光谱数据分别采用 PLS 回归法和 MLR(Multiple Linear Regression)回归法进行分析。两种回归方法的结果与样品的化学值对比，表明 PLS 法是一种比较好的多变量定标方法，比较适合于实时在线定标分析。

关 键 词：近红外光谱分析仪；PLS 方法；定量光谱分析
中图分类号：O433.4 **文献标识码**：A

1 引 言

在土壤研究、饲料加工、粮食加工、酿酒、水质监测和石油化工等许多行业中，对样品的多种化学成份含量的测定，以往多采用化学方法，例如测定蛋白质含量的典型化学方法是凯氏定氮法。利用化学方法分析的结果准确度高，但是时间长，在实验中需要配备一定数量的化学分析试剂，并要经过一系列复杂冗长的化学反应过程才能实现。这很难适应大量样品短时间内获取样品成份含量的工作需要。而且在采样和获取分析结果之间有较多的时间拖延，会严重影响数据分析结果。

近红外光谱分析技术[1]是用光谱方法分析物质的成份和结构，具有样品制备简单、费用低、速度快和不破坏样品化学性质的特点。正好能满足大量样品的成份分析工作。由于计算机技术的迅速发展，计算机在近红外光谱仪中的应用不但提高了仪器的自动控制及操作能力，而且使得大量光谱数据的实时处理更加快速准确。近红外光谱仪已经成为重要的现场实时在线分析技术设备。

近红外光谱仪的光谱分析技术原则上就是利用物质的光谱参数经过分析，得出物质不同成份的含量。具体可以分为定标和预测两个过程。定标过程是利用一定数量已知成份含量的样品组应用回归方法求出光谱参数与样品成份浓度之间的关系式。预测过程就是将未知样品成份浓度的光谱参数代入定标过程求得的公式，得出未知样品的成份含量。由于仪器的分析精度既取决于仪器本身的性能，又与定标好坏密切相关。因此回归分析的精度直接影响光谱仪器的精度。

近红外光谱分析仪器的回归定标方法有好多种，如多元线性回归法 MLR (Multiple Linear Regression)、主元素分析法 PCA (Principal Component Analysis)、主元素回归法 PCR (Principal component regression)等。偏最小二乘法 PLS (Partial least-squares)[2~5]是在以上几种回归运算的基础上发展起来的一种多元非线性迭代回归方法。在这几种回归运算方法中，MLR 法是最常用的算法，以其物理意义明确、公式简单为大多数人接受。PLS 法是矩阵法，以其对谱峰严重重叠的体系进行多元校正见长。它的特点是速度快，准确可靠，预测能力强，具有一定的消除非线性的能力，比较适用于同时测定成批样品的多种成份。由于近红外光谱的谱峰重叠的厉害，如果只选择几个测量点的数据进行定标和预测，预测结果的相关系数不会很高，而且使得其它测量点的数据信息被浪费。本文应用 MLR 和 PLS 两种定标方法，对40种烟草样品成份含量进行定标和预测。预测结果与化学值作比较。

2 原 理

首先假设有 n 种样品，m 个测量波长点，l 是样品组份数。$A_{n\times m}$ 是吸光度矩阵，$C_{n\times l}$ 是浓度矩阵。

收稿日期：1999-12-15；修订日期：2000-02-22

2.1 MLR 方法的基本原理

由朗伯比尔定律有:

$$C = AP + E_C \quad (1)$$

$P_{n\times l}$是系数矩阵,E_C 是浓度残差矩阵,确定矩阵 P 的过程就是定标过程。由方程(1)可得:

$$P = (A'A)^{-1}A'C \quad (2)$$

因此,只要知道待测样品的吸光度 A,就可由下式求得待测样品的不同组份的含量。

$$C = AP \quad (3)$$

2.2 PLS 方法的基本原理

PLS 法的基本思想是利用非线性迭代方法对吸光度矩阵 A 和浓度矩阵 C 进行分解:

$$A = t_1p_1 + t_2p_2 + \ldots + t_hp_h + E = TP + E \quad (4)$$

$$C = u_1q_1 + u_2q_2 + \ldots + u_hq_h + F = UQ + F \quad (5)$$

T 是 $n\times h$ 矩阵,t_1、t_2...t_h 是 T 矩阵的列向量,P 是 $h\times m$ 矩阵,p_1、p_2...p_h 是 P 矩阵的行向量。U 是 $n\times h$ 矩阵,u_1、u_2...u_h 是 U 矩阵的列向量,Q 是 $h\times l$ 矩阵,q_1、q_2...q_h 是 Q 矩阵的行向量。E 和 F 分别是吸光度矩阵和浓度矩阵的残差矩阵,h 是迭代次数。迭代过程中,应用新变量 w 代替 p,迭代结束后再计算 p,w 是与 p 长度相同的行向量。"'"表示矩阵或向量的转置。在 A、C 进行分解的过程中以 U 和 T 的每个向量建立 A 与 C 之间的内部关系:

$$U = BT \quad (6)$$

B 称为 *PLS* 法的回归系数矩阵。

表1. 30个烟草样品应用 *PLS* 法和 *MLR* 法定标的计算结果。

Table 1　The calibration of 30 tobacco samples using PLS and MLR method.

No	Sugar					Nicotine					Protein				
	Chem	PLS	Bias	MLR	Bias	Chem	PLS	Bias	MLR	Bias	Chem	PLS	Bias	MLR	Bias
1	13.51	13.47	0.04	13.50	0.01	0.86	0.88	−0.02	0.89	−0.03	10.18	9.58	0.60	9.68	0.50
2	11.54	11.61	−0.07	11.67	−0.13	0.73	0.78	−0.05	0.79	−0.06	9.61	10.05	−0.44	9.98	−0.36
3	14.26	14.00	0.26	13.96	0.30	1.45	1.53	−0.08	1.55	−0.10	13.68	13.53	0.15	13.46	0.22
4	24.29	24.62	−0.33	24.65	−0.34	0.76	0.65	0.11	0.63	0.13	7.80	7.95	−0.15	8.11	−0.31
5	16.48	17.51	−1.03	17.67	−1.19	1.06	0.80	0.26	0.80	0.26	9.85	10.30	−0.45	10.63	−0.78
6	23.93	24.10	−0.17	24.05	−0.12	0.71	0.59	0.12	0.61	0.10	7.04	6.96	0.08	6.89	0.15
7	24.83	24.92	−0.09	24.91	−0.08	1.70	1.59	0.11	1.60	0.10	9.85	10.19	−0.34	10.15	−0.30
8	21.57	20.55	1.02	20.55	1.02	0.81	1.02	−0.21	1.01	−0.20	8.25	8.35	−0.10	8.29	−0.04
9	14.08	13.22	0.86	13.15	0.93	0.81	1.01	−0.20	1.03	−0.22	10.06	9.62	0.44	9.39	0.67
10	11.64	11.87	−0.23	11.91	−0.27	1.18	1.25	−0.07	1.27	−0.09	11.10	11.16	−0.06	10.99	0.11
11	15.79	15.88	−0.09	15.93	−0.14	0.96	1.08	−0.12	1.07	−0.11	10.34	10.29	0.05	10.14	0.22
12	14.63	14.62	0.01	14.42	0.21	1.00	1.01	−0.01	1.05	−0.05	10.30	10.33	−0.03	10.39	−0.09
13	12.11	11.71	0.40	11.69	0.42	1.25	1.31	−0.06	1.31	−0.06	11.96	12.19	−0.23	12.27	−0.31
14	11.84	12.07	−0.23	12.05	−0.21	1.89	1.73	0.16	1.72	0.17	12.90	12.89	0.01	13.00	−0.10
15	9.84	10.49	−0.65	10.69	−0.85	1.12	0.76	0.36	0.71	0.41	12.11	12.72	−0.61	12.86	−0.75
16	17.20	17.06	0.14	17.05	0.15	0.96	1.25	−0.28	1.27	−0.30	10.64	10.53	0.11	10.39	0.25
17	16.91	17.35	−0.44	17.37	−0.46	2.43	2.20	0.23	2.21	0.22	8.75	9.34	−0.59	9.22	−0.47
18	18.31	18.83	−0.52	18.86	−0.55	1.95	1.60	0.35	1.60	0.35	8.46	8.91	−0.46	8.83	−0.37
19	17.44	17.79	−0.35	17.77	−0.33	1.34	1.26	0.08	1.26	0.08	10.61	10.72	−0.11	10.71	−0.10
20	13.09	12.92	0.17	12.89	0.20	1.67	1.89	−0.22	1.88	−0.21	16.57	15.70	0.83	15.67	0.90
21	19.41	19.29	0.12	19.44	0.03	1.07	1.09	−0.02	1.07	−0.00	11.16	10.13	1.03	10.05	1.11
22	13.38	13.41	−0.03	13.36	−0.02	0.80	0.67	0.13	0.68	0.12	13.20	13.28	−0.08	13.16	0.04
23	21.02	20.11	0.91	20.09	0.93	1.26	1.40	−0.14	1.39	−0.13	9.95	10.06	−0.11	10.15	−0.20
24	16.67	16.84	−0.17	16.88	−0.21	1.96	1.95	0.01	1.95	0.01	12.88	12.91	−0.03	13.00	−0.12
25	13.63	13.24	0.39	13.20	0.43	0.81	1.03	−0.22	1.03	−0.22	13.00	12.86	0.14	13.10	0.10
26	10.42	10.53	−0.11	10.46	−0.04	2.23	2.28	−0.05	2.30	−0.07	13.47	12.93	0.54	12.87	0.60
27	22.11	22.06	0.05	22.14	−0.03	0.84	1.24	−0.40	1.27	−0.43	9.91	9.41	0.50	9.35	0.56
28	19.68	19.24	0.44	19.19	0.49	0.68	0.68	0.00	0.70	−0.02	9.58	9.38	0.20	9.29	0.29
29	16.25	16.59	−0.34	16.62	−0.37	1.15	1.04	0.11	1.03	0.12	10.26	10.48	−0.22	10.64	−0.38
30	12.72	12.57	0.15	12.59	0.13	2.10	1.91	0.19	1.89	0.21	14.36	15.21	−0.85	15.23	−0.87
R*		0.9945		0.9868			0.9328		0.9261			0.9794		0.9736	
SEP**		0.4391		0.4773			0.1797		0.1881			0.4219		0.4768	

：相关系数， *：标准偏差。

3 结果与讨论

文中以40个烟草样品为例来说明PLS法的应用，同时比较了PLS法和MLR法的预测结果和精度。烟草样品有三种组份：蛋白质、糖和尼古丁。组份的浓度应用化学方法测定。吸光度参数由近红外光谱仪获得。光谱仪有19个波长测量点，这19个测量点的波长值分别是：1680nm，1722nm，1734nm，1759nm，1778nm，1790nm，1818nm，1840nm，1940nm，2100nm，2130nm，2139nm，2180nm，2230nm，2250nm，2270nm，2310nm，2336nm，2348nm。

PLS法的具体迭代过程如下：

对于样品每一种组份，设 $u=c_j$，即等于浓度矩阵的某一列，

① 设 $A=uw+E$，则：$w=u'A/u'u$ (7)

$w=w/|w|$

② 设 $A=tw+E$，则：$t=Aw'/w'w$ (8)

③ 设 $C=tq+F$，则 $q=t'C/t't$ (9)

$q=q/|q|$

④ 设 $C=uq+F$，则：$u=Cq'/q'q$ (10)

⑤ 第②步中的 t 与以前的 t 比较，如果相同，则进行第⑥步运算，否则返回第一步。

⑥ 设 $A=tp+E$，则：$p=t'A/t't$ (11)

$p=p/|p|$

$t=t|p|$

$w=w|p|$

⑦计算回归系数：$b=u't/t't$ (12)

⑧计算残差：$E_i=E_{i-1}-t_ip_i$ (13)

$F_i=F_{i-1}-b_it_iq_i$ (14)

其中 $E_0=A$，$F_0=C$。

⑨返回第①步，确定 A、C 的下一个分量，直到 $i=h$。

PLS 法由以下步骤确定待测样品的浓度，A 是待测样品的吸光度，设 $E_0=A$，$F_0=0$则：

①$t_i=E_{i-1}w_i'$ (15)

②$E_i=E_{i-1}-t_ip_i$ (16)

③$F_i=F_{i-1}+b_it_iq_i$ (17)

④返回第①步，进行下一次迭代。迭代 h 次后，$C=F_h$，C 就是待测样品的预测浓度矩阵。

迭代次数 h 的确定是PLS法中关键问题之一。原则上是采用交叉校验的方法确定的，即从全部样品中选择一定数量的样品建立模型，用模型预测剩余样品的浓度，计算预测浓度与实际浓度的残差平方和；每次用不同的样品建立模型，保证每个样品都曾在建立模型时出现过一次。计算总的残差平方和。比较不同迭代次数时总的残差平方和，最小残差平方和所对应的迭代次数就是所要求得的PLS回归模型的迭代次数。

表1中列出了30个烟草样品的化学值、PLS法估计值、MLR法估计值、估计值与化学值的相对误差、平均标准偏差和定标相关系数。从表1中可以看出，PLS法的定标相关系数和平均标准偏差优于MLR法。表2列出了应用两种方法各自定标公式预测10种样品成份的估计结果。表2中的PLS法的平均标准偏差较MLR法低。从以上的运算讨论可以看出MLR法公式简单，物理意义明确，而PLS法公式复杂。MLR法的缺点是要求 $L=A'A$ 有逆矩阵，即要求

$$|L| = |A'A| \neq 0$$

因此求解(1)式中的系数矩阵 P 比较困难。当测量波长点少时，MLR法还是相当有效的，但当测量波长点多时，MLR法相当复杂。对于线性度好的系统，MLR法的相关系数较高，可是对于光谱重叠严重的体系，MLR法也是捉襟见肘。PLS法虽然迭代步骤多，但当应用计算机运算时，迭代次数多已不是缺点，而且对吸光度矩阵没有特殊要求。测量波长点越多，越显现出PLS法的优越性。对于光谱重叠严重的体系，PLS法在多元回归的定量光谱分析中是十分有效的手段。加之PLS法还具有一定的消除非线性的能力。因此，近红外光谱仪采用PLS定标回归方法更适宜。

表2. 应用PLS法和MLR法预测10个烟草样品的计算结果.

Table2. The calculation of 10 tobacco samples using PLS and MLR method

No	Sugar					Bicotine					Protein				
	Chem	PLS	Bias	MLR	Bias	chem	PLS	Bias	MLR	Bias	chem	PLS	Bias	MLR	Bias
1	14.38	14.15	0.23	14.43	−0.35	1.47	0.89	0.59	0.97	0.50	12.16	13.61	−1.45	14.66	−2.50
2	12.81	11.90	0.91	11.82	0.99	1.08	1.04	0.05	1.33	−0.25	12.39	13.02	−0.63	14.46	−2.07
3	10.85	12.99	−2.13	13.71	−2.85	1.11	0.54	0.57	0.45	0.66	12.18	8.72	3.46	12.61	−0.43
4	18.69	18.90	−0.21	18.85	−0.16	0.87	0.92	−0.04	1.14	−0.24	9.08	10.74	−1.66	9.70	−0.62
5	19.89	18.78	1.11	18.60	1.30	0.89	1.36	−0.47	1.77	−0.88	9.70	9.68	0.02	10.93	−1.23
6	19.02	17.26	1.76	21.45	−2.43	1.11	1.45	−0.34	1.86	−0.75	9.51	10.81	−1.30	9.73	−0.22
7	11.27	10.79	0.48	10.05	1.22	0.77	1.35	−0.57	1.62	−0.84	10.83	9.62	1.21	12.63	−1.80
8	13.88	15.02	−1.14	15.03	−1.15	1.06	1.02	0.04	1.16	0.10	10.12	9.25	0.87	11.62	−1.50
9	10.59	11.19	−0.60	11.28	−0.70	0.91	1.03	−0.12	1.28	−0.37	9.81	10.81	−0.99	13.54	−3.73
10	16.03	16.41	0.38	16.71	0.68	1.17	1.24	−0.06	1.69	−0.52	9.57	11.03	−1.46	11.18	−1.61
SEP**		1.0875		1.4351			0.3685		0.5723			1.5565		1.8606	

**：标准偏差。

参考文献：

[1] Norris K H, Williams P C. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. influence of particle size[J]. Cereal Chemistry, 1984, 61(2): 158−165.

[2] Geladi Paul, Kowalski Bruce R. Partial least-squares regression: a tutorial[J]. Analytica Chimica Acta, 1986, 185: 1−17.

[3] Geladi Paul, Kowalski Bruce R. An example of 2-block predictive partial least-squares regression with simulated data[J]. Analytica Chimica Acta, 1986, 185: 19−32.

[4] Haaland David M. Thomas Edward V. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information[J]. Anal. Chem., 1988, 60: 1193−1202.

[5] Mcshane Michael J, Cote Gerard L, Spiegelman Clifford H. Assessment of partial least-squares calibration and wavelength selection for complex near-infrared spectra[J]. Applied Spectroscopy, 1998, 52(6): 878−884.

Partial least-squares method for calibration with NIR spectrophotometer

ZHANG Ling

(*Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun* 130021, *China*)

Abstract: Near Infrared Reflectance (NIR) spectrophotometer is widely used in analyzing grains, water and oil. The spectrophotometer is able to analyze without destroy the chemical property of sample. There are many methods in its quantitative spectral analyses. Partial Least Squares (PLS) method is a good alternative to the more classical multiple linear regression. 40 tobacco sample data are described to illustrate the spectrophotometer calibration on PLS regression and Multiple Linear Regression (MLR). The result is related to calibration in chemical analysis. The values of components from PLS method is more precise than that from MLR method. It is proved that the PLS regression is a powerful multivariate calibration method and more suitable to the on-line or real-time analysis.

Key words: NIR spectrophotometer; partial least-squares method; quantitative calibration methods

作者简介:张玲(1964−),女,江苏杨州市人,1987年毕业于长春光机学院,学士,毕业后一直在光机所从事光谱仪器的研究和开发工作。现任长春光机所医疗仪器公司副总经理。