

- 318, (1997).
- 39 J. M. Jones, T. D. Davies, S. R. Dorling, Water, Air, Soil Pollut. 85 1569-1574 (1995).
- 40 J. T. W. E. Vogels, L. Terwel, A. C. Tas, F. van der Berg, F. Dukel, J. van der Greef, J. Agric. Food Chem., 44 175-180 (1996).
- 41 K. Iizuka, T. Aishima, J. Food Sci., 62 101-104(1997).
- 42 F. Angerosa, L. Di Giacinto, R. Vito, S. Cumitini, J. Sci. Food Agric., 72 323-328 (1996).
- 43 D. A. Winkler, D. J. Madellena, Ser. Math. Biol. Med., 5 126-163(1995).
- 44 H. Schmidli, Chemom. Intell. Lab. Syst., 37 125-134(1997).
- 45 N. Ghoshal, B. Achari, T. K. Ghoshal, Pol. J. Pharmacol., 48(4)359-377(1996).
- 46 P. E. Flecher, W. T. Welch, S. Albin, J. B. Cooper, Spectrochim. Acta, Part A, 53A (2) 199-206(1997).
- 47 J. B. Cooper, K. L. Wise, J. Groves, W. T. Welch, Anal. Chem., 67 1766-1771 (1995).
- 48 C. J. de Bakker, P. M. Fredericks, Appl. Spectrosc., 49 1766-1771(1995).
- 49 G. P. Firmstone, M. P. Smith, A. J. Stipanovic, Automat. Eng., [Spec. Publ.] SP-1995, SP-1116, 201-208.
- 50 B. M. Wise, N. B. Gallagher, J. Process Control, 6 329-348(1996).
- 51 P. I. Zufria, Comput-aided Chem. Eng., 6 385-408(1995).
- 52 S. J. Qin, T. McAvoy, J. Comput. Chem. Eng., 20 147-159(1996).
- 53 NAmICS Newsletter 17 April 1998, p. 11 1998 Gordon Reseach Conference.

# 现代光谱分析中常用的化学计量学方法

## Popular Chemometrics Used in Spectrometric Analysis

袁洪福 陆婉珍

(石油化工科学研究院 北京 100083)

**【摘要】** 详细介绍了现代光谱分析中常用的化学计量学分析方法(主成分回归、偏最小二乘法和马氏距离聚类分析)的原理、算法和特点。使用这些算法编写了化学计量学分析软件。通过对含 MTBE 汽油的近红外光谱分析介绍其应用。

**【Abstract】** The principles and algorithms of popular chemometrics methods in modern spectrometric analysis are introduced. The software written for multivariate linear regression(MLR), principle component regression(PCR), partial least square (PLS) and cluster analysis methods are compared in Near Infrared analysis of MTBE Content of gasoline.

**关键词:** 化学计量学 光谱分析 汽油

**Key Words:** Chemometrics Spectrometric analysis gasoline

近几年由于化学计量学的发展和应  
用,使得光谱分析进入了一个新时代,光  
谱和化学计量学的结合已成为一种快速  
和高效的分析技术,尤其是近红外光谱,  
其进展情况已在本刊登载<sup>[1]</sup>,但有关现  
代光谱分析中常用化学计量学方法的原  
理、算法和特点并未提及。本文着重对  
此作介绍。

经典定量分析方法不适合复杂样品  
的分析,即由于多组分的相互干扰,应用  
比尔定律受到限制。这是由于实际测得  
的光谱数据不仅包括了被测样品的组成  
和结构信息,而且还包括了噪音,即测量  
误差与不同组分之间的干扰等。化学计  
量学方法可有效地剔除这些噪音。如化  
学计量学采用了因子分析方法,引入了  
“主成分”和“得分”的概念。

主成分相当于一个度量单位,得分  
相当于权重。以人民币举例说明,其面  
值的度量单位有 100 元,50 元,10 元,1  
元等。假设有 287 元,用主成分和得分  
的概念去表征:  $100 \times 2 + 50 \times 1 + 10 \times 3$

$+ 1 \times 7$ , 即 100 元,50 元,10 元和 1 元为  
主成分,它们的得分分别为 2, 1, 3 和 7,  
在这个例子中,按照面值的大小分类,  
100 元为第一主成分,50 元为第二主成  
分,余类推。使用主成分和得分概念去  
表征光谱,即任何一个原始光谱图都可  
以使用主成分光谱和其得分的线性组合  
来重建。化学计量学方法通过数学方法  
对原始光谱处理,得到光谱的主成分和  
得分,根据一定的规则选取一定数目的  
主成分光谱重建光谱,该重建光谱最大  
限度地反映了被测样品的组成和结构信  
息,而最小限度地包含噪音。克服了经  
典方法的缺点,并保留了其优点。常用  
的化学计量学定量分析方法包括多元线  
性回归、主成分回归和偏最小二乘法。  
定性分析包括主成分分析和聚类分析。

## 1 光谱定量分析

经典光谱定量分析的基础是比耳定律:

$$A_{\lambda} = \epsilon_{\lambda} l C \quad (1)$$

其中,  $A_{\lambda}$  为吸收度,  $\epsilon_{\lambda}$  为消光系数,  $l$   
为光程,  $C$  被测组分浓度。

不同物质具有不同的  $\epsilon_{\lambda}$ 。对一定的  
物质,如果已知光程  $l$ , 只要求出  $\epsilon_{\lambda}$ , 就可  
建立起  $A$  和  $C$  之间简单的线性关系。  
建立该线性关系的过程称之为校正。化  
学计量学光谱分析校正方法是在以下经  
典方法基础上发展起来的。

### 1.1 经典校正方法

#### 1.1.1 最小二乘回归模型

通常,光程  $l$  是常数,可将  $l$  和  $\epsilon_{\lambda}$  乘  
积表示为常数  $k_{\lambda}$ , 因此,比尔定律可写  
为  $A_{\lambda} = k_{\lambda} C$ 。只要知道浓度  $C$  和吸收度  
 $A_{\lambda}$ , 就可以确定校正方程。原则上,使用  
一个样品就可以完成校正,但是,为提高  
校正的准确度,通常使用一系列不同浓  
度的多个样品进行校正,也就是靠平均  
效应提高模型的稳健度。其算法用矩阵  
形式表示如下:

$$1. \text{校正模型: } A_{\lambda(n \times 1)} = k_{\lambda} C_{(n \times 1)} \quad (2)$$

其中,  $A_{\lambda(n \times 1)}$  为  $n$  个样品在波长  $\lambda$  处的吸收值组成的向量,  $C_{(n \times 1)}$  为  $n$  个样品的被测组分的浓度值组成的向量;  $k_{\lambda}$  为  $A_{\lambda(n \times 1)}$  和  $C_{(n \times 1)}$  之间的关联系数。  $C'$  为  $C$  的转置矩阵。

$$2. \text{最小二乘解: } k_{\lambda} = (C'C)^{-1}C'A \quad (3)$$

$$3. \text{预测方程 } c = k_{\lambda} a_{\lambda} \quad (4)$$

其中:  $a_{\lambda}$  为未知样品在波长  $\lambda$  处的吸收值,  $c$  为未知样品被测组分的浓度值。

方法原理简单, 容易计算和理解。对于多组分样品, 必须确定被测组分所对应谱带, 在该谱带处比尔定律成立, 此谱带不能与其它组分对应的谱带重叠。应用仅仅限于纯样品或简单体系分析, 而不适合复杂体系的分析。

### 1.1.2 经典最小二乘回归模型 (CLS)<sup>[2-5]</sup>

CLS 方法基于比尔定律的加和性质。为获得更准确的校正模型, CLS 方法一般使用较多的样品进行校正。只要选用与组分数一样多或更多数目的波长, 就可能同时对所有组分进行校正。实际上, CLS 校正常常使用全谱, 其方程的矩阵形式为:

$$A_{n \times p} = C_{n \times m} K_{m \times p} \quad (5)$$

其中,  $A$  为光谱矩阵,  $K$  为消光系数矩阵,  $C$  为组分浓度矩阵,  $E$  为光谱残差矩阵, 其维数与  $A$  相同。下标  $n$  为样品数,  $p$  为用于校正的波长数,  $m$  为样品的组分数。由式(5)求矩阵  $K$ :

$$K = C^{-1}A \quad (6)$$

其中,  $C^{-1}$  为浓度矩阵的逆。通常使用矩阵广义逆代替矩阵逆:

$$K = C'(CC')^{-1}A \quad (7)$$

使用全谱数据校正, 具有平均效应, 提高了抗噪音能力。当使用全谱时,  $K$  矩阵恰为被分析组分的吸收度系数光谱, 并且与组分纯光谱很相近, 因此, CLS 方法也称  $K$  矩阵法。CLS 的缺点是必须对样品中的每个组分进行校正, 比如, 在波长  $\lambda$  处的吸收为  $A_{i,j} = C_{i,1}K_{i,1} + C_{i,2}K_{i,2} + \dots + E_{i,j}$ , 如果忽略了样品的任何一个组分的浓度, 其预测吸收度都是不准确的。因此, 校正时, 要求了解样品的任何一个组分, 并知道其浓度。在预测时, 未知样品的组分数和校正样品的组分数相同, 不能有污染, 组分之间不能有相互作用, 否则, 预测吸收度都将是不准确的。实际情况较难满足这些限制条件。

### 1.1.3 逆最小二乘法模型 (ILS)<sup>[6,7]</sup>

将比尔定律形式改写为:

$$A_{n \times p} = T_{n \times f} P_{f \times p} + E_A \quad (12)$$

$$C = \frac{A_{\lambda}}{\epsilon_{\lambda} b} \quad (8)$$

将  $\epsilon_{\lambda}$  和  $b$  合并为一个常数  $P$ , 比尔定律可表示为

$$C = PA_{\lambda} + E \quad (9)$$

其中  $C$  为组分浓度,  $A_{\lambda}$  为波长  $\lambda$  处的吸收度,  $E$  为浓度拟合线与浓度实际值间的残差。这种形式称之为逆最小二乘法, 也称多元线性回归 (MLR), 或  $P$  矩阵法。浓度  $C$  仅仅是吸收度的函数, 只需知道被分析组分的浓度和在  $\lambda$  处的吸收度, 就可以求出校正系数  $P$ , 克服了 CLS 方法的缺点。由式(9)求校正系数矩阵:

$$P = CA^{-1} \quad (10)$$

如果矩阵  $A$  不是方阵, 用广义逆代替:

$$P = CA'(AA')^{-1} \quad (11)$$

ILS 方法选择的波长必须对被分析组分浓度是物理相关的, 可以根据化学背景知识, 或数学方法相关比较进行选择。由于矩阵方程的维数限制, 所选择的波长数不能超过校正样品的数目。理论上, 增加校正样品数目, 就可以使用更多的波长数目, 但是, 随着使用波长数目增加, 不同波长之间吸收存在“共线”的可能性增加, 使方程解不稳定。增加波长数目的另一个问题就是产生过度拟合现象, 降低模型的预测能力。ILS 具有不必知道化合物中所有组分就可以校正的优点, 缺点是使用波长数目受到限制。

### 1.2 常用化学计量学方法

上述分析表明, 经典定量分析方法不适合复杂样品的分析, 即由于多组分的相互干扰, 应用比尔定律受到限制。常用化学计量学方法引入了主成分和得分概念, 克服了经典方法的缺点, 并保留了其优点。光谱分析中常用的化学计量学定量分析方法有主成分回归和偏最小二乘法。

#### 1.2.1 主成分回归 (PCR)<sup>[6-9]</sup>

包括主成分分析 (PCA) 和主成分回归分析 (PCR) 两步。

##### 1.2.1.1 主成分分析 (PCA)

选取具有代表性的校正样品集。对光谱数据进行中心化, 即求出所有光谱的平均光谱, 从每个光谱中减去平均光谱得到的就是中心化光谱, 中心化具有突出光谱之间的微细差别作用。因为 PCA 是采用特征值方法, 特征值是以光谱吸收度值的变化为基础得到的, 而不是根据吸收度的绝对值。通过 PCA 代数把光谱矩阵  $A$  分解为两个维数较小的矩阵的乘积:

其中  $A_{n \times p}$  为光谱矩阵,  $T_{n \times f}$  为所有光谱的得分矩阵,  $P_{f \times p}$  为特征值或主成分矩阵,  $E_A$  为预测光谱残差, 维数与  $A_{n \times p}$  相同。  $n$  为样品数目,  $p$  为波长数目,  $f$  为 PCA 主成分数目。  $E$  为光谱残差矩阵。

PCA 和 CLS 比较, 校正模型的形式相类似。但, PCA 中, 光谱数据不是由浓度和吸收度系数光谱构成, 而是由主成分光谱和得分构成的。与 CLS 相似之处为, 波长数目使用不受限制, 可以使用全谱数据, 保留了平均效应。PCA 中, 浓度  $C$  并没有起作用, 因此, 单独使用 PCA 并不能预测被分析组分浓度。但是, PCA 的  $P$  矩阵在功能上类似于 CLS 的  $K$  矩阵, 可理解为单位光谱矩阵。而 PCA 的得分  $T$  矩阵对每个校正集光谱是唯一的, 与其组分浓度可能非常相关, 因此, PCR 的第二步就是将浓度矩阵  $C$  对得分矩阵  $T$  进行回归。

##### 1.2.1.2 主成分回归分析 (PCR)

回归方程:

$$C_{n \times m} = T_{n \times f} B_{f \times m} + E_c \quad (13)$$

其中,  $B$  为校正系数矩阵,  $m$  为组分数, 其它与上述相同。

这恰好类似于 ILS, 不需要知道校正样品中的所有组分浓度, 就可以建立被测组分的校正模型。  $B$  系数矩阵的解为:

$$B_{f \times m} = (T'_{n \times f} T_{n \times f})^{-1} T'_{n \times f} C_{n \times m} \quad (14)$$

与 ILS 比较, 很容易看出, PCR 方法问题出在得分与浓度回归的一步, 因为, ILS 的波长选择根据波长与浓度物理相关的原则, 而, PCA 计算主成分和得分时, 并没有考虑组分的浓度, 得分可能与组分浓度变化非常相关, 但不能保证任何情况下这种可能性都存在。事实上, PCA 的主成分中某些主成分与组分浓度并无任何关系, 如何选择有用的主成分去建立被分析组分的模型并非容易。另外, 如果化合物中组分浓度之间存在“共线”现象, 也会影响 PCR 模型的预测能力。

##### 1.2.2 偏最小二乘法 (PLS)<sup>[10-13]</sup>

###### 1.2.2.1 PLS 基本原理和特点

在 PCR 中, 只对光谱矩阵作了分解, 消除光谱矩阵中无用的信息; 同样, 浓度矩阵中也包含了无用信息, 也应作同样的处理, PLS 就是基于这种思想的回归方法。

PLS 第一步, 作矩阵分解, 其模型为:

$$A_{n \times p} = T_{n \times r} P_{r \times p} + E_A \quad (15)$$

$$C_{n \times m} = U_{n \times r} Q_{r \times m} + E_C \quad (16)$$

其中,  $T$  和  $U$  分别为光谱矩阵和浓度矩阵的得分矩阵,  $P$  和  $Q$  分别为光谱矩阵和浓度矩阵的载荷(即主成分)矩阵,  $E_A$  和  $E_C$  分别为用 PLS 模型拟合  $A$  和  $C$  时所引进的误差矩阵。

PLS 第二步, 将  $T$  和  $U$  作线性回归:

$$U_{n \times r} = T_{n \times r} B_{r \times r} \quad (17)$$

其中,  $B$  为关联系数矩阵

在预测时, 由样品的光谱矩阵  $A_{未知}$  和校正得到的  $P_{校正}$  求出样品光谱的  $T_{未知}$ , 再按式(18)求未知样品的浓度:

$$C_{未知} = T_{未知} B Q \quad (18)$$

实际上, PLS 计算并非如此, PLS 把数据矩阵分解和回归并为一步, 即光谱数据和浓度数据的分解是同时进行的, 并且, 将浓度信息引入到光谱数据分解过程中, 在每计算一个新主成分之前, 将光谱得分和浓度得分进行交换, 使得到光谱主成分光谱直接与被分析组分浓度关联。

PLS 又分为 PLS1 和 PLS2。虽然方法差别不大, 但所得到的结果有差异。所谓 PLS2, 和 PCR 相似, 在校正过程中, 光谱矩阵分解只给出一个  $T$  矩阵和一个  $P$  矩阵, 显然, 这样得到的  $T$  和  $P$  对个别组分并不是最优化的, 在预测时, 对于复杂体系, 会降低结果精度; 在 PLS1 中, 校正得到的  $T$  和  $P$  是对个别组分进行优化的, 换句话说, 对应于不同的组分, 其  $T$  和  $P$  矩阵不同。当校正样品中不同组分在浓度变化上相差很大时, 比如, 一个组分浓度范围为 60~70%, 另一个组分的为: 0.2~0.8%, PLS1 预测结果普遍优于其它方法。缺点就是在校正时, 对每个被分析组分都要计算一套主成分矩阵和得分矩阵, 需要校正计算时间长, 但随着计算机计算速度的发展, 这一问题将不再明显。

### 1.2.2.2 PLS 算法

#### 1.2.2.2.1 模型建立

在化学计量学中对光谱矩阵分解的数学方法主要有两种, 一种是非线性迭代最小二乘法<sup>[14]</sup>, 另一种是线性代数中常用的奇异值分解法<sup>[15]</sup>。一般是采用前一种。与 PCR 相同, 在进行 PLS 计算前, 光谱数据和浓度数据都经过中心化处理。

(1) 取浓度矩阵  $C$  的一列记为  $c$ , 作为分解的初始迭代矢量赋值给矩阵  $U$  的分解矢量:  $u = c$

(2) 以  $u$  代替  $T$  矩阵的分解矢量  $t$  来计算  $v$  ( $P$  矩阵的分解矢量)

$$v' = uA(u'u)^{-1}$$

(上述两步使用浓度阵  $C$  的一列作为初始迭代矢量来进行  $A$  矩阵的正交分解, 即分解  $A$  矩阵时引入了浓度的信息)

(3) 标准化处理:  $v' = v' |v'|^{-1}$

(4) 由  $v'$  计算矢量  $t$ :  $t = Av(v'v)^{-1}$

(5) 以  $t$  替代  $U$  矩阵的分解矢量  $u$  计算  $Q$  矩阵的分解矢量  $q$

$$q' = t'C(t't)^{-1}$$

(这一步相当于使用  $A$  矩阵分解矢量  $t$  对  $C$  矩阵作分解, 即在  $C$  矩阵分解时引入了光谱信息)

(6) 标准化处理:  $q' = q' |q'|^{-1}$

(7) 由  $q'$  计算矢量  $u$  ( $U$  矩阵的分解矢量):  $u = Cq(q'q)^{-1}$

(8) 判断  $s$  是否收敛, 即  $\|t_{前一轮} - t_{后一轮}\|$  是否小于某一常数  $\theta$ 。如果收敛则到(9)步, 否则, 返回到(2)步, 继续进行分解迭代。注意, 此时返回(2)步应使用第(7)步计算的  $u$ 。

(注意, 这样计算出的  $t$  已失去了正交性质, 但对多元校正问题无关紧要。上述计算得到的  $t, v, u, q$  均应保留, 供预测和模型校验使用。)

(9) 计算  $t$  与  $u$  之间的内部关系, 进行回归运算:  $b = u't(t't)^{-1}$

上述各分解矢量均对应于第一主成分, 记为  $t_1, v_1, u_1, q_1, b_1$ 。下一步计算残差阵:

$$E_A = A - t_1 v_1'$$

$$E_C = C - u_1 q_1' = C - b_1 t_1 q_1'$$

下一步使用  $E_A$  和  $E_C$  分别代替矩阵  $A$  和矩阵  $C$ , 返回上述主成分迭代计算过程, 计算下一主成分, 得到对应的矢量  $t_2, v_2, u_2, q_2, b_2$ , 依次类推, 直到计算出  $A$  和  $C$  的全部主成分为止。

#### 1.2.2.2.2 预测计算

对未知样品, 光谱为  $a_{未知}$  或  $A_{未知}$ , 根据关系  $A = TP'$  和校正时储存的  $P'$ , 计算出  $t_{未知}$ , 再根据校正时储存的  $B$  向量和矩阵  $Q$ , 就可以计算出  $C_{未知}$  或  $c_{未知}$ 。

#### 1.2.2.2.3 最佳主成分数的确定方法

上述使用因子分析方法建立模型, 第一主成分的信息量贡献最大, 第二主成分次之, 余类推。用主成分和得分来拟合重建的光谱, 理想化是重建的光谱包含样品的全部光谱信息而不含任何噪音。取主成分数过少, 对重建光谱产生拟合不够, 丢掉了样品光谱信息, 取主成分数过多, 对重建光谱产生过度拟合, 引

入噪音。问题是模型应该需要多少个主成分? 其方法较多, 其中, 性能最好, 也是最常用的是交互验证 (Cross Validation) 方法。首先求残差平方和 (PRESS), 它是评价模型预测能力的一个有效指标, 其定义为:

$$PRESS = \sum_{i=1}^n \sum_{j=1}^m (C_{p,j} - C_{i,j})^2$$

矩阵说明: 样品集实际浓度矩阵  $C$ , 使用校正模型对校正样品集预测的浓度矩阵  $C_p$ 。

维数说明: 校正样品数  $n$ , 被分析组分分数  $m$ 。

最小 PRESS 对应的主成分数  $f^*$  并不是最佳主成分数, 往往是造成过度拟合。通常用  $F$  统计方法或人工分析来确定最佳主成分数。  $F$  统计方法如下:

$$F(f) = PRESS(f) / PRESS(f^*)$$

最佳的  $f$  比  $f^*$  小, 尽可能小且满足  $F(f) < F_{\alpha, \beta}$  ( $\alpha = 0.25, \beta$  为自由度) 条件。

### 1.2.3 化学计量学定量方法特点

光谱定量分析是间接测量方法, 影响模型预测能力因素很多, 主要是校正集浓度数据的准确度和精度及变化范围、光谱测量误差等。因此, 要求校正集具有代表性, 换句话说, 校正集应在化学成分上接近被分析的未知样品, 校正集浓度变化范围应覆盖未知样品的组分浓度变化。在模型建立过程中, 如果发现浓度残差较大的样品, 应分析其原因, 是来自于光谱测量还是浓度测量, 还是样品来源等? 对样品进行必要的筛选, 以保证所建立的模型具有较好的适应能力。原理上这种光谱分析方法是依靠标准测量方法提供基础数据, 它的准确度不会超过标准方法。然而, 光谱分析具有很好的重现性和精密度, 因此, 它往往比标准测量方法是具有更好的重现性和精密度。

#### 1.2.4 应用实例

使用上述算法 MLR、PCR 和 PLS 编写的软件, 已用于近红外及紫外光谱测定油品烃族组成及性质, 如辛烷值、密度、馏程、蒸汽压、MTBE 含量、冰点等。该软件可以计算相关光谱, 用于波长的选择, 图 1 是 40 个含 MTBE 汽油样品的近红外光谱和 MTBE 的相关近红外光谱图。

相关光谱值无论接近 1 或 -1, 都说明 MTBE 浓度与光谱强度之间具有很好的相关性。根据图 1, 可以选择合适的波长 (如波数为 4600、5500 和 7400  $\text{cm}^{-1}$ ) 对应的光谱数据参加 MLR

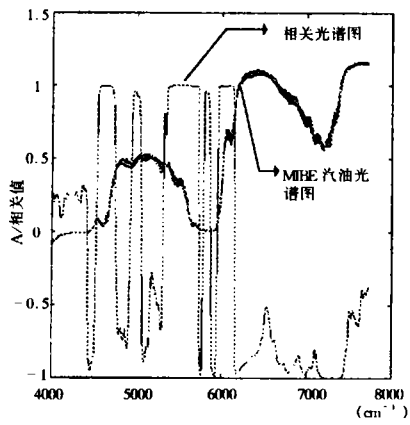


图1 相关光谱图

计算。而 PCR 和 PLS 都是全谱方法,即全谱数据都参加运算。图 2 是分别使用 MLR、PCR 和 PLS 方法对上述 40 个样品的光谱计算得到的 MTBE 含量结果与实际值对照情况,可以看出,3 种结果都与实际值相吻合,PCR 和 PLS 结果与实际结果更接近,MLR 方法对样品 28 #、32 #、38 # 和 39 # 的计算结果与实际值之间有较大的误差,其原因是这 4 个样品的光谱测量有较大的误差。这些结果表明,与 MLR 方法相比,PCR 和 PLS 方法使用全谱数据,具有平均效应的优点,能够有效地剔除噪音,得到更接近实际值的结果。

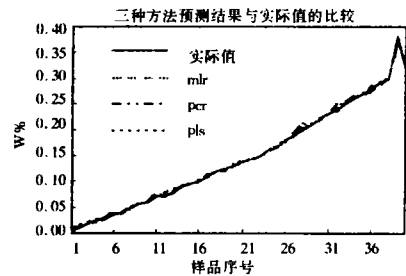


图2 MLR、PCR 和 PLS 结果与实际值对照

## 2 化学计量学光谱定性分析

光谱定性也有广泛的应用。化学计量学有很多定性的方法,也就是所谓的聚类分析,目前国际流行的方法是依靠 PCA 分析结合马氏距离判据 (Mahalanobis Distance) 方法<sup>[16,17]</sup>。

### 2.1 马氏距离

矩阵说明:校正集光谱数据矩阵  $A_{n \times p}$ , 平均光谱数据矩阵  $\bar{A}_{1 \times p}$  中心化 (即光谱数据减去平均光谱) 光谱矩阵  $A_{cen}$ , 未知样品光谱  $A_{un1 \times p}$ , 马氏矩阵  $M_{p \times p}$ , 马氏距离  $D$ 。

维数说明:校正集样品数  $n$ , 所选参加模型建立的波长数目  $p$ , 波长序号  $j$ 。

$$\bar{A}_j = \frac{(\sum_{i=1}^n A_{i,j})}{n}$$

$$Acen_j = A_j - \bar{A}_j$$

$$M = \left( \frac{Acen' Acen}{(n-1)} \right)$$

$$D^2 = (A_{un} - \bar{A}) M^{-1} (A_{un} - \bar{A})'$$

这种方法的局限性与 ILS 相同,遇到波长选择的困难。

### 2.2 使用主成分计算马氏距离

正象前边所述,使用 PCA 的得分代替 ILS 中波长选择而产生了 PCR 一样,可以避免选择过多波长造成过度拟合,同时能保留全谱的平均效应。以下为使用得分向量代替选择波长的光谱矩阵计算马氏距离:

矩阵说明:马氏矩阵  $M_{f \times f}$ , 校正集 PCA 得分矩阵  $T_{n \times f}$ , 未知样品得分向量  $T_{un}$ , 马氏距离  $D$ 。

维数说明:校正集样品数  $n$ , PCA 主成分数  $f$ 。

$$M = \left( \frac{T'T}{n-1} \right)$$

$$D^2 = (T_{un}) M^{-1} (T_{un})'$$

### 2.3 应用实例

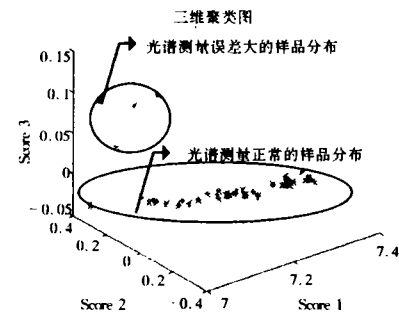


图3 簇分析图

使用上述算法编写 PCA 分析结合马氏距离的软件。它在建立校正模型过程中起着重要作用,可为判断光谱测量误差提供了依据。例如在建立测量汽油中甲基叔丁基醚 (MTBE) 含量校正模型中,对 40 个不同 MTBE 含量的汽油样品近红外光谱进行 PCA 分析。使用各样品的第一、二、三主成分的得分作三维图 (如图 3), 发现 4 个样品偏离了主群。从图 4 可以看出,如果将马氏判据的阈值设为 2.5, 则样品 28 #、32 #、38 # 和 39 # 均位于界外。通过分析,偏离原因是光谱测量误差造成的,是由光池放置位置差别引起的。使用这个方法可用不同物质的鉴定,如在环保中使用近红外光

谱对废塑料进行分类。该方法也可用于产品的纯度检验,以纯品建立模型,设定合适的马氏距离判据,就可以使用近红外进行检验产品纯度是否合格。

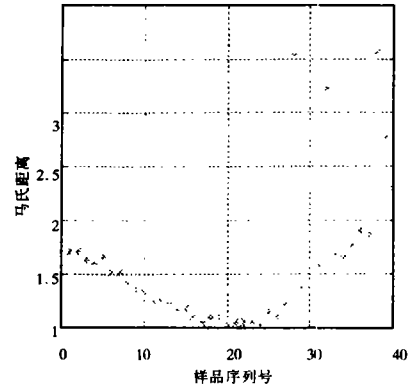


图4 马氏建立分布图

## 3 参考文献

- 徐广通, 袁洪福, 陆婉珍, 现代科学仪器 1997, 3, 9
- Haaland D M, et al. Appl. Spec., 1980, 34: 539
- Haaland D M, et al. Appl. Spec., 1982, 36: 665
- Haaland D M, et al. Appl. Spec., 1985, 39: 73
- Brown C W, et al. Anal. Chem., 1982, 54: 1472
- Mark H. Anal. Chem., 1986, 58, 2814
- Malinowski E H, J. Chemometrics, 1987, 1: 33
- Miyashita Y et al. J. Chemometrics, 1990, 4: 97
- Fredericks P M, et al. Appl. Spec., 1985, 39: 303
- Lindberg W, et al., Anal. Chem., 1983, 55: 643
- Haaland D M, et al. Anal. Chem., 1988, 60: 1202
- Kowalski B R, et al. Appl. Spec., 1990, 44: 1464
- 俞汝勤. 化学计量学导论, 湖南教育出版社, 1991
- Brown. C. W. et al., Anal. Chem., 1982, 54: 1472
- Brown, C. W. et al., Anal. Chem., 1983, 55: 1703
- Mahalanobis P C. Proc. Natl. Inst. of Science of India, 1986, 2: 49
- Mark H L. Anal. Chem. 1987, 59: 790