第21卷，第4期
2001年8月

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

Vol. 21, No. 4, pp459-463
August, 2001

# 近红外光谱定量校正模型适用性研究*

徐广通　　袁洪福　　陆婉珍

石油化工科学研究院，100083　北京

摘　要　近红外光谱作为一种依靠模型进行分析的技术，对测定的样品进行模型适用性判断是得到可靠分析结果的前提。对于通过校正集样本近红外光谱测量和标准方法测定的基础数据依靠因子分析技术建立的多元校正模型，提出将因子分析与Mahalanobis距离相结合判断定量模型适用性的方法，以近红外光谱测定柴油十六烷值为例，对影响模型适用性判断的一些因素进行了讨论。

主题词　近红外光谱，　因子分析，　Mahalanobis距离，　模型适用性，　十六烷值

现代近红外光谱以其分析速度快、重现性好、成本低、不消耗样品、易于实现在线分析等鲜明的特点正得到越来越多的应用[1~6]，并成为90年代以来发展最快的分析测试技术之一。现代近红外光谱技术不是通过观察待分析样品谱图特征或量测谱图参数直接进行定性或定量分析，而是先通过校正集样品的光谱和组成或性质数据(组成或性质数据须通过其它认可的标准方法测定)的量测，采用合适的化学计量学方法建立校正模型，然后再通过建立的校正模型和测定的未知样品光谱实现定性或定量分析。显然，对所建立的校正模型是否适合未知样品作出判断，是实现准确分析的前提。

主成分分析(PCA)和Mahalanobis距离都常用于光谱的判别分析。PCA是通过光谱主成分得分构筑的主成分空间进行样品的簇分布分析[7,8]，该方法虽可将复杂的多维空间信息压缩到低维空间进行分析，但对离群点判别的定量阀值不易界定。Mahalanobis[9]距离是研究多维空间矢量相似性的有效方法之一，在光谱的定性、离群点判别分析中也得到使用[10~12]。Whitfield等[13]曾将其用于近红外光谱定量校正模型的适用性判断，在Mahalanobis距离计算时，采用几个波长下的光谱数据(如吸光度)进行。这种Mahalanobis距离的计算存在两个问题，一是如果将光谱数据减少，波长的选择是一个重要又比较困难的问题，因为波长选择不合适可能会丢失样品的信息；二是如果采用全谱运算，计算工作量极大，且可能由于共线性的存在导致矩阵运算不稳定。因子分析是从多维矢量空间中抽提有用信息的有效方法，它可以用较低维数的矢量来表示原来的多维矢量空间，在保存有效信息的同时，去除噪音。在此提出将PCA与Mahalanobis距离相结合解决校正模型的适用性判断。首先用PCA对校正集样本的原始光谱或预处理后的光谱进行处理，然后再用各光谱得到的因子得分计算Mahalanobis距离。这样既利用了PCA对光谱降维处理不丢失信息的特点，又发挥了Mahalanobis距离对离群值有效识别的优点。

## 1　方法原理

### 1.1　主成分分析[14]

校正集样品的主成分分析：通过校正集样品的光谱矩阵$A[m\times n]$($m$为样本个数，$n$为光谱数据点数)，利用非线性迭代偏最小二乘算法(non-linear iterative partial least squares，简称NIPALS算法)，计算指定光谱主因子$f$下的载荷矩阵$P[f\times n]$和校正集样品的光谱得分矩阵$T[m\times f]$。

### 1.2　Mahalanobis距离计算

利用校正集样品的得分矩阵$T$，计算校正集样品的Mahalanobis距离，步骤如下：

$$T=\frac{\sum_{i=1}^{m} t_i}{m} \tag{1}$$

$$T_{cen}=T-T' \tag{2}$$

$$M=\frac{T_{cen}T_{cen}}{m-1} \tag{3}$$

$$MD_i=[(t_i-T)\cdot M^{-1}\cdot(t_i-T)']^{1/2} \tag{4}$$

式中$t_i$为校正集样本$i$的光谱得分，$T$为校正集$m$个样本的平均得分矩阵；$T_{cen}$为$T$的均值中心化矩阵；$M$为校正集样品的Mahalanobis矩阵；$MD_i$为校正集样本$i$的Mahalanobis距离。根据定量校正允许的误差和对应的Mahalanobis距离，确定离群点Mahalanobis距离阀值限$MD_L$。

### 1.3　未知样品模型适用性判断及最佳校正模型的选择

对测定的未知样本光谱矢量$A_{un}$，通过校正集样品求得的光谱载荷$P$，计算其光谱得分$t_{un}$，由$t_{un}$根据式(4)再计算其Mahalanobis距离$MD_{un}$。将未知样本与符合误差要求的Mahalanobis距离阀值$MD_L$进行比较，如$MD_{un}<MD_L$，说明未

2000-03-10收，2000-09-16接受；*本课题得到中石化总公司科研开发项目资助(合同号596001)

徐广通，1964年生，博士学位，副教授，石油化工科学研究院博士后

知样品与校正集样品属于一类，由建立的校正模型预测出的组成或性质结果误差应在模型建立时设定的误差范围之内（误差允许范围的大小通常依据测定基础数据标准方法的再现性确定）。反之，未知样品即为离群样本，不适合用已建立的校正模型预测其组成或性质，否则预测出的结果与标准方法测定结果的误差可能超过允许的误差范围。如有多个校正模型供未知样品选择（假设为 $c$ 个），则可计算未知样品相对于每个校正集样本的 Mahalanobis 距离 $MD_j(j=1,2,\cdots c)$，从中选择最小的 $MD_{\min}$ 值，与 $MD_{\min}$ 相对应的校正模型即为最准确的预测模型，但该模型能否满足测定准确性的要求，再通过比较 $MD_{\min}$ 与 $MD_L$ 确定。

## 2 实验部分

### 2.1 仪器

Bomem MB-160 型傅里叶变换近红外光谱仪，光谱采集范围 1 000～2 000 nm，InGaAs 检测器，5 mm 石英样品池。PC 计算机（主频 233 MHz），计算程序由 MATLAB 语言编写。

### 2.2 样品来源及十六烷值测定

样品分别取自石油化工科学研究院的催化裂化柴油、齐鲁石化胜利炼油厂的成品柴油、洛阳国家柴油检测中心抽检的成品柴油。十六烷值基础数据由台架试验按 GB/T386 方法测定。

### 2.3 样本光谱的测量

以空样品池作空白，样品放入光路 1 min 后采集光谱，每个样品扫描 10 次，得到的光谱图见图 1a～图 1c。

### 2.4 定量校正方法

定量校正采用偏小二乘法（PLS）方法，计算程序由 MATLAB 语言编写。光谱经均值中心化处理，与十六烷值间通过 PLS 回归，采用交互检验法预测残差平方和（PRESS）确定最佳主因子，并建立校正模型。

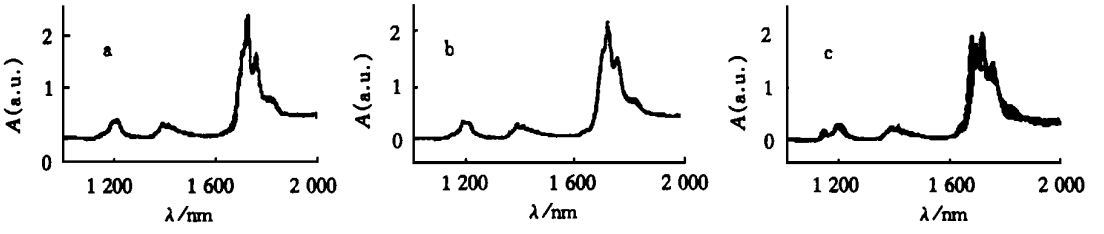


**Fig 1 Spectra of different diesel fuels**

a-Luoyang, b-Shengli, c-FCC

## 3 结果与讨论

### 3.1 校正集模型的建立

用前述的方法对洛阳和催化裂化柴油分别建立校正集，计算其 Mahalanobis 距离，由偏最小二乘（PLS）建立定量校正模型。两个柴油校正集对应的实测值与预测值的相关曲线、Mahalanobis 距离及残差图分别见图 2 和图 3。从图 2c 和图 3c 结果可以看出，洛阳油模型的预测误差较催化裂化柴油预测误差偏大，主要原因是洛阳柴油来源复杂，一部分为不同原油的直馏油，一部分为直馏与催化裂化柴油加氢后的调和油，当预测十六烷值的误差小于 3 时，光谱 Mahalanobis 距离的最大阀值不能超过 3。催化裂化柴油模型尽管重油来源不同，但加工工艺变化不大，得到较正模型的精度较高。由图 3 可以看出，当光谱 Mahalanobis 距离的阀值限定为 2.5 时，其十六烷值的预测误差不大于 2。

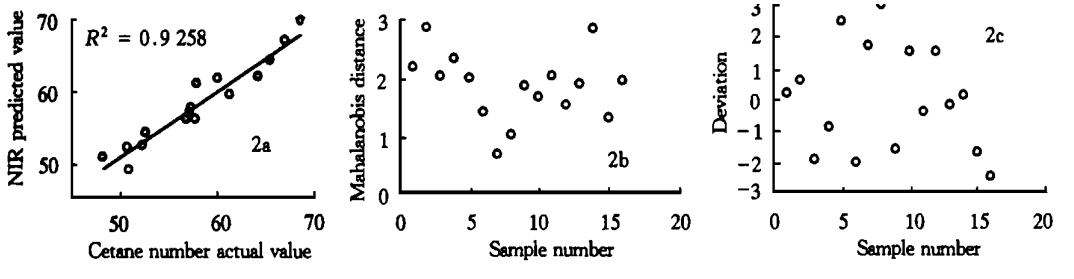


**Fig 2 Luoyang diesel fuels calibration set (Factor=4)**

a-correlation curve; b-Mahalanobis distance; c-derivation distribution

### 3.2 定量校正模型适用性分析

用上述建立的两个模型，分别对未知样品光谱的 Mahalanobis 距离进行计算，并用 PLS 定量模型预测其十六烷值结果见表 1。可以看出，当未知样品光谱的 Mahalanobis 距离小于校正集样品为达到确定误差限设定的 Mahalanobis 距离阀值时，所预测结果的误差均小于建模时所允许的误差限。由图 4 可以看出，随未知样品光谱的 Mahalanobis 距离增大，所预测结果的误差呈明显增长趋势。关于模型选择的判断，表 1 中 SL01、SL02、SL03、LY11、CC13、CC14 分别由两个校正集得到的 $MD$ 可以看出，$MD$ 较小者预测结果的误差也较小，同厂生产

的SL01、SL02、SL03 三个批次的柴油, SL01 和 SL02 用洛阳模型预测 *MD* 较小, 预测结果与实测结果偏差也小, 而SL03 从 *MD* 看则更接近催化裂化柴油模型, 实验结果也证明, 用催化裂化模型预测的结果较洛阳模型预测的结果准确。显然, 通过 PCA 计算未知样品光谱的得分, 再计算其 Mahalanobis 距离方法, 来选择定量校正模型并判断模型对未知样品的适用性是可行的。

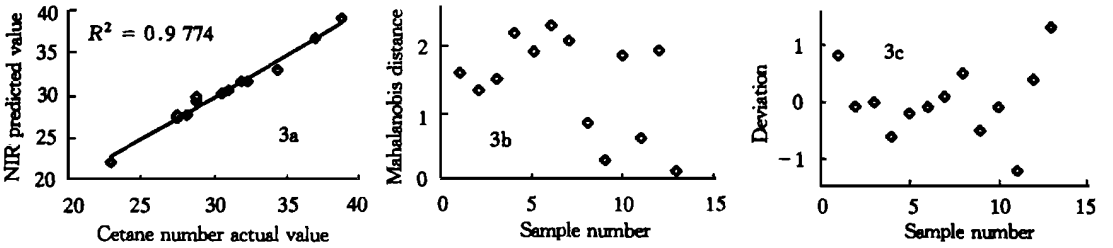


**Fig 3　FCC diesel fuels calibration set (Factor=3)**

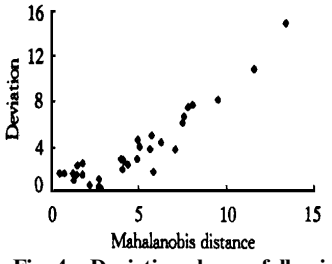


**Fig 4　Deviation change following Mahalanobis distance change (Luoyang calibration set)**

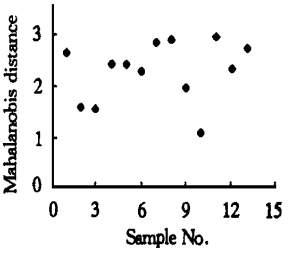


**Fig 5　Mahalanobis distance of FCC diesel fuels calibration set (Factor= 5)**

**Tab. 1　Validation of suitability of quantitative calibration model①**

| Luoyang model | | | | | FCC model | | | | |
|---|---|---|---|---|---|---|---|---|---|
| No. | $MD_{un}$ | GB386 | NIR | Dev. | No. | $MD_{un}$ | GB386 | NIR | Dev. |
| LY06 | 1.85 | 51 | 49.4 | 1.6 | **CC13** | **1.37** | **30.2** | **30.1** | **0.1** |
| **LY11** | **2.25** | **65.2** | **64.5** | **0.7** | **CC14** | **1.60** | **30.6** | **30.6** | **0** |
| LY13 | 1.50 | 53.3 | 54.9 | 1.6 | CC22 | 2.18 | 22.9 | 22.6 | 0.3 |
| LY17 | 1.86 | 44.2 | 46.8 | 2.6 | CC28 | 1.91 | 28.5 | 28.0 | 0.5 |
| LY20 | 1.34 | 61.8 | 60.7 | 1.1 | CC29 | 2.29 | 30.9 | 30.8 | 0.1 |
| LY22 | 0.80 | 55.8 | 57.5 | 1.7 | CC31 | 0.83 | 27.4 | 28.0 | 0.6 |
| LY23 | 0.54 | 56.3 | 58.0 | 1.7 | CC32 | 0.87 | 27.4 | 27.7 | 0.3 |
| LY24 | 1.30 | 57.8 | 56.1 | 1.7 | CC33 | 1.10 | 35.2 | 34.4 | 0.8 |
| LY25 | 1.56 | 64.4 | 62.0 | 2.4 | CC34 | 0.94 | 32.8 | 31.2 | 1.6 |
| **SL01** | **2.89** | **46.6** | **47.0** | **0.4** | CC38 | 1.93 | 37.7 | 38.9 | 1.2 |
| **SL02** | **2.77** | **45.3** | **45.8** | **0.5** | CC39 | 0.15 | 28.5 | 30.0 | 1.5 |
| SL13 | 2.80 | 46.4 | 47.6 | 1.2 | **SL01** | **3.77** | **46.6** | **44.3** | **2.3** |
| **SL03** | **4.02** | **42** | **39.0** | **3.0** | **SL02** | **3.39** | **44.6** | **42.7** | **1.9** |
| SL05 | 4.11 | 41.9 | 39.8 | 2.1 | **SL03** | **2.77** | **42** | **39.7** | **2.3** |
| SL12 | 4.17 | 41.3 | 38.4 | 2.9 | SL04 | 3.13 | 43.7 | 41.6 | 2.1 |
| SL35 | 4.95 | 42.4 | 38.4 | 3.0 | SL05 | 2.80 | 41.9 | 40.0 | 1.9 |
| SL36 | 5.08 | 42.4 | 38.3 | 4.1 | SL16 | 3.25 | 44 | 42.3 | 1.7 |
| SL60 | 5.76 | 41.6 | 36.5 | 5.1 | SL17 | 3.18 | 41.6 | 41.1 | 0.5 |
| SL55 | 5.68 | 41.8 | 37.9 | 3.9 | SL18 | 4.07 | 46.4 | 44.7 | 1.7 |
| SL39 | 5.31 | 42.1 | 37.6 | 4.5 | SL19 | 3.76 | 45 | 43.5 | 1.5 |
| SL43 | 4.41 | 40.7 | 38.2 | 2.5 | LY01 | 4.84 | 57.3 | 49.4 | 7.9 |
| **CC13** | **8.10** | **30.2** | **22.4** | **7.8** | LY02 | 5.99 | 68.9 | 54.1 | 14.7 |
| **CC14** | **7.53** | **30.4** | **24.2** | **6.2** | LY05 | 4.56 | 56.5 | 49.7 | 6.8 |
| CC18 | 7.60 | 31.7 | 24.9 | 6.8 | LY08 | 4.92 | 61.4 | 50.3 | 11.1 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| CC22 | 13.4 | 22.9 | 7.8 | 15.1 | **LY11** | **5.39** | **65.2** | **52** | **13.2** |
| CC27 | 11.6 | 28 | 17.0 | 11 | LY12 | 5.45 | 65.7 | 52.3 | 13.4 |
| CC37 | 7.83 | 34.2 | 26.6 | 7.6 | LY15 | 5.75 | 67.2 | 54.6 | 12.6 |
| CC39 | 9.56 | 28.7 | 20.4 | 8.3 | LY25 | 5.03 | 64.4 | 53.3 | 11.1 |

①: $MD_L$ is 3 when deviation of cetane number is litter than 3 in Luoyang model

$MD_L$ is 2.5 when deviation of cetane number is litter than 2 in FCC model

**3.3** 主因子数的选取对 **Mahalanobis** 距离阀值的影响

由于 Mahalanobis 距离是通过主成分分析的光谱得分求得, 而光谱的得分直接与主成分分析时所选择的因子数有关, 如图 5 为催化裂化柴油校正集在因子数取 5 时计算的 Mahalanobis 距离分布图。图 5 与图 3b 比较可以看出, 随因子数的增加, 校正集样本 Mahalanobis 距离值的差值也增加。因此, 当因子数变化时, 根据允许的误差限, Mahalanobis 距离的阀值也应作出相应的调整。图 6 为以洛阳柴油为校正集, 在不同因子数下, 对不同类型柴油的识别情况。由图可以看出, 当因子数为 2 时(见图 6a), 光谱信息没得到充分应用, 洛阳柴油(1~20 号)与胜利柴油谱图(21~40 号)没识别开。当因子数为 4 时(建立定量校正模型时的最佳主因子数), 洛阳油, 胜利油和催化裂化油(41~60 号)都明显识别开, 催化裂化油 Mahalanobis 距离差别较大, 是由于用了 3 种不同的渣油原料, 在催化剂和工艺参数上也有所差别。当因子数取 6 时, 尽管洛阳油与胜利油间的识别较好, 但胜利油与催化裂化油间的识别又变差。这主要是因为随因子数的增加, 光谱噪音也参与计算, 使谱图的识别性变差。因此, PCA 与 Mahalanobis 距离结合判断离群点或样品分类时, 因子数的选择要适当。由于 PLS 主因子数的选择是根据交互验证的残差平方和确定, 求 Mahalanobis 距离时因子数的选择可与用 PLS 建立定量校正模型时选择的最佳主因子数相同。

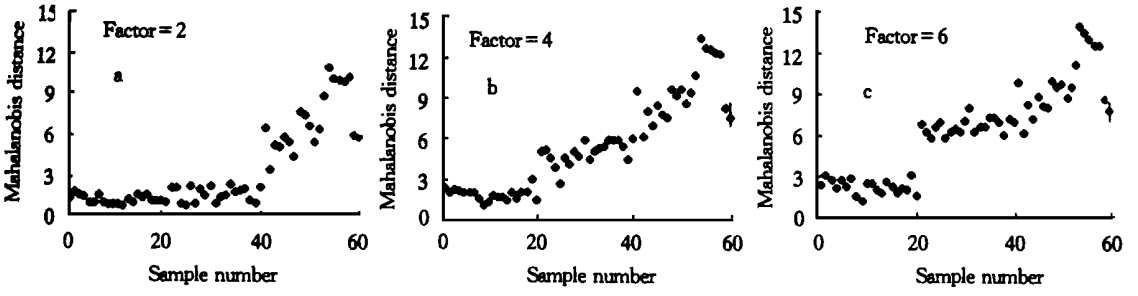


**Fig. 6 Effect of factor number for sample discrimination through Mahalanobis distance**

## 4 结 论

鉴于近红外光谱是通过校正集样本的光谱信息、基础测量方法测定的化学组成或理化性质结合化学计量学方法建立分析模型进行测量的技术, 在近红外光谱分析中, 用已建立的定量校正模型对未知样品的组成或性质进行测量时, 对分析模型的适用性作出判断是非常重要的, 是判断测量的结果是否与基础方法的测量结果一致的重要前提。将主成分分析(PCA)与 Mahalanobis 距离相结合解决模型的适用性判断, 可以充分利用 PCA 对大量的光谱数据进行降维不丢失光谱信息的优点, 将 PCA 计算得到的光谱得分用于 Mahalanobis 距离计算, 较好地解决了 Mahalanobis 距离计算时波长点的选择问题, 也可以避免大量的光谱数据直接进行 Mahalanobis 距离计算出现的共线性或计算量大的问题, 同时也避免采用 PCA 自身进行判断界限不易量化的问题。通过柴油十六烷值的测定, 对提出的方法进行了检验, 并将建立的方法用于柴油和汽油组成和性质等指标的分析, 取得良好的结果。提出的方法同样也适用于其他类似的光谱分析方法。

## 参 考 文 献

1 W F McClure. *Anal. Chem.*, 1994, **66**(1): 43A

2 J D Kirsch and J K Drennen. *Appl. Spectrosc. Rev.*, 1995, **30**(3): 139

3 J C Sreven et al. *Anal. Chem.*, 1996, **68**(20): 3525

4 M L Lysaght et al. *Fuel*, 1993, **72**(5): 623

5 G A Lang et al. *Hydrocarbon Processing*, 1994, **2**: 69

6 Guangtong XU, Hongfu YUAN and Wanzhen LU(徐广通, 袁洪福, 陆婉珍). *Spectroscopy and Spectral Analysis*(光谱学与光谱分析), 2000, **20**(2): 134

7 W R Gemperline, L D Webber and F O Cox. *Anal. Chem.*, 1987, **59**: 138

8 J S Shenk and M O Westerhaus. *Crop Sci.*, 1991, **31**: 460

9 P C Mahalanobis. *Proc. Natl. Inst. of Science of India*, 1936, **2**: 49

10 H Mark and D Tunnell. *Anal. Chem.*, 1985, **57**: 1449
11 H Mark. *Anal. Chem.*, 1986, **58**: 379
12 H Mark. *Anal. Chem.*, 1987, **59**: 790
13 R G Whitfield, M K Gerger and R L Sharp. *Appl. Spectrosc.*, 1987, **41**(7): 1204
14 俞汝勤. 化学计量学导论, 湖南教育出版社, 1991

# Study of Quantitative Calibration Model Suitability in Near-infrared Spectroscopy Analysis

Guangtong XU, Hongfu YUAN and Wanzhen LU
*Research Institute of Petroleum Processing*, 100083 *Beijing*

**Abstract** Near-infrared spectroscopy is a fast, high efficiency and low cost analytical technique, that depended on analytical models which are built through near-infrared spectra, the primary method results of calibration set and chemometrics methods. The reliability of analytical results mostly depends on the suitability of analytical model to unknown samples. To judge the suitability of analytical model, a new method is presented that combine principle factor analysis (PCA) and Mahalanobis distance. It is different with traditional method, that the Mahalanobis distance is calculated through PCA factor scores. It can avoid the wavelengths selection problem in traditional method of spectral Mahalanobis distance. Through the determination of cetane number of diesel fuel, some factors that affect model suitability judgement are discussed.

**Keywords** Near-infrared spectroscopy, Model suitability, Factor analysis, Mahalanobis distance, Cetane number

(Received March 10, 2000; accepted Sep. 16, 2000)