

## 目标识别因子分析解析完全未知混合体系的研究

何锡文\* 陈 鼎

王永泰

(南开大学化学系 天津 300071)

(南开大学中心实验室 天津 300071)

**摘要** 本文以目标识别因子分析为基础, 将因子分析与聚类分析相结合, 给出了从完全未知混合体系中提取纯物种光谱的新方法. 所得纯物种光谱经光谱检索作定性判别, 然后利用外标法对混合物中的各组份进行标定. 该方法用于计算机模拟体系及三组份实际混合体系的红外光谱解析, 结果令人满意.

**关键词** 目标识别因子分析, 最小支撑树, 胆甾型液晶

完全未知混合体系指的是组份数、物种及含量均属未知的体系. 由于对体系的组成知之甚少, 因而对其定性定量研究都相当困难. 通常采用的色谱与其它波谱仪器联用技术, 可以部分解决这个难题. 但是寻求分离完全的色谱柱是很繁杂的工作, 有时甚至根本无法得到分离完全的色谱峰. 在这种情况下, 利用数学方法解析完全未知混合体系, 就成了具有实际意义的课题. 因子分析以其解析多元数据的强大功能受到重视. 如文献介绍的自模型曲线法<sup>[1]</sup>及许多提取纯谱的其它方法<sup>[2,3]</sup>皆以因子分析为基础. 国内有关工作已经展开<sup>[4,5]</sup>. 如梁逸曾等<sup>[6]</sup>采用迭代目标转换算法, 解析了实际混合体系的紫外可见与荧光光谱. 但是至今对完全未知混合体系提取纯谱、光谱检索以及标定含量的工作尚未见报道. 本文以目标识别因子分析法 (TRFA)<sup>[7]</sup>为基础, 从完全未知混合物光谱中直接提取纯物种光谱, 利用光谱检索作定性判别, 再经外标法获得定量结果. 该法是将因子分析与聚类分析相结合, 从观测数据本身引出目标向量, 克服了目标转换因子分析法经验性强的弱点, 同时避免了自模型曲线法预先确定可行解域的困难. 在解析过程中未对光谱形状作附加要求. 本文先以计算机模拟三组份混合物体系, 进行解析验证方法的可行性; 进而对胆甾型液晶的三组份混合物实际体系进行解析, 结果令人满意.

### 1 理论与方法

根据 Lambert-Beer 定律, 多组份混合物的光谱应满足下式:

$$d_{ij} = \sum_{k=1}^n \varepsilon_{ik} c_{kj} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, p) \quad (1)$$

式中  $d_{ij}$  为第  $j$  个混合物在第  $i$  个波数点上的吸收值,  $m$  为波数的测点数量,  $p$  为混合物个数,  $n$  为组份数,  $\varepsilon_{ik}$  为组份  $k$  在波数  $i$  处的摩尔吸收系数,  $c_{kj}$  为样品  $j$  中组份  $k$  的浓度.

将式 (1) 写成矩阵形式:

$$[D]_{m \cdot p} = [E]_{m \cdot n} [C]_{n \cdot p} \quad (2)$$

其中 $[D]$ 为混合物光谱阵,  $[E]$ 为纯物种标准光谱阵,  $[C]$ 为浓度阵. 对混合物光谱阵 $[D]$ 进行主成份分析<sup>[8]</sup>, 可得存在于混合物体系中纯组份数 $n$ , 同时将 $[D]$ 分解为抽象光谱阵 $[A]$ 和抽象浓度阵 $[F]$ 的乘积. 即:

$$[D]_{m \cdot p} = [A]_{m \cdot n} [F]_{n \cdot p} \quad (3)$$

$[A]$ 和 $[F]$ 是没有物理意义的, 需经进一步变换, 将抽象解转化为实际解:

$$[E]_{m \cdot n} = [A]_{m \cdot n} [T]_{n \cdot n} \quad (4)$$

$$[C]_{n \cdot p} = [T]_{n \cdot n}^{-1} [F]_{n \cdot p} \quad (5)$$

其中 $[T]$ 为转换矩阵.  $[T]$ 中的列向量 $t$ 可按最小二乘法求得:

$$t = ([A]^T [A])^{-1} [A]^T b \quad (6)$$

$$\bar{b} = [A]t \quad (7)$$

式中 $b$ 为 $m$ 维的初始目标向量, 它是对 $[E]$ 的某一列所虚拟的假设值.  $\bar{b}$ 是由 $b$ 得到 $[E]$ 的某一列的预测值. 对于完全未知混合体系, 构造出合理的初始目标向量 $b$ 是非常困难的.

根据目标识别因子分析法的原理, 以“单一向量”作为 $b$  (所谓单一向量相当于单位坐标向量, 即一个分量为1, 其余为零的向量), 得到 $\bar{b}$ 后将 $\bar{b}$ 作为新的 $b$ 迭代直至收敛. 所得最后的 $\bar{b}$ 与 $[E]$ 阵的某一列成比例或近似成比例. 改变单一向量中1所在的位置, 得到的 $\bar{b}$ 可能与 $[E]$ 阵的另一列成比例. 单一向量随1所在位置的不同可以有 $m$ 个. 利用图论聚类的方法, 对 $m$ 个收敛的 $\bar{b}$ 进行分类. 收敛于同一因子的 $\bar{b}$ 彼此相似, 以此作为聚类的基础. 对于任意两迭代结果 $\bar{b}_\alpha$ 和 $\bar{b}_\beta$ , 定义其距离为 $y_{\alpha\beta}$ :

$$y_{\alpha\beta} = \sqrt{1 - \gamma_{\alpha\beta}^2} \quad (8)$$

式中 $\gamma_{\alpha\beta}$ 为两向量 $\bar{b}_\alpha$ 和 $\bar{b}_\beta$ 间的相关系数.

$$\gamma_{\alpha\beta} = \frac{\sum_{i=1}^m (\bar{b}_{i\alpha} - \bar{\bar{b}}_\alpha)(\bar{b}_{i\beta} - \bar{\bar{b}}_\beta)}{\left\{ \left[ \sum_{i=1}^m (\bar{b}_{i\alpha} - \bar{\bar{b}}_\alpha)^2 \right] \left[ \sum_{i=1}^m (\bar{b}_{i\beta} - \bar{\bar{b}}_\beta)^2 \right] \right\}^{1/2}} \quad (9)$$

$$\text{上式的 } \bar{\bar{b}}_\alpha = \frac{1}{m} \sum_{i=1}^m \bar{b}_{i\alpha}, \quad \bar{\bar{b}}_\beta = \frac{1}{m} \sum_{i=1}^m \bar{b}_{i\beta}$$

定义了距离 $y_{\alpha\beta}$ 后, 即可求最小支撑树. 根据选择标识向量的原则<sup>[7]</sup>, 从 $m$ 个单一向量中选出 $n$ 个, 分别作为相应因子的标识向量, 与它们对应的 $n$ 个列向量 $t$ 构成完整的变换矩阵 $[T]$ , 则可得:

$$[\bar{E}] = [A][T] \quad (10)$$

$$[\bar{C}] = [T]^{-1} [F] \quad (11)$$

$[\bar{E}]$ 与实际 $[E]$ 的相应列成比例,  $[\bar{C}]$ 与实际 $[C]$ 的相应行成比例. 设它们之间存在一组比例因子 $S_k$  ( $k = 1, 2, \dots, n$ ), 则有:

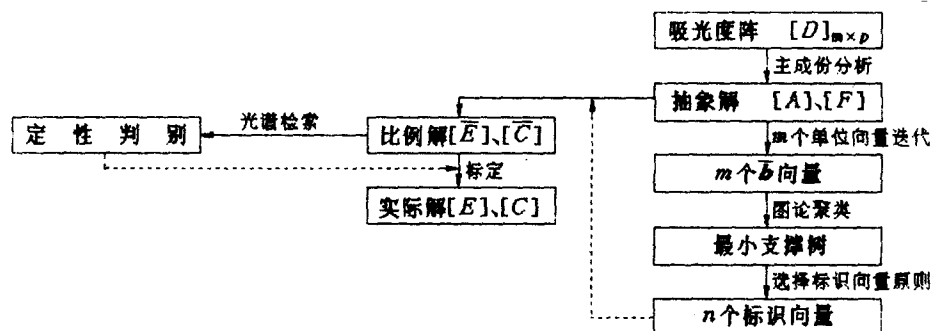
$$d_{ij} = \sum_{k=1}^n \bar{e}_{ik} \bar{c}_{kj} = \sum_{k=1}^n \varepsilon_{ik} c_{kj} \quad (12)$$

$$\text{式中, } \varepsilon_{ik} = \bar{e}_{ik} / S_k \quad (13)$$

$$c_{kj} = S_k \bar{c}_{kj} \quad (14)$$

求得的 $[\bar{E}]$ 可利用光谱检索的方法, 进行定性判别, 然后采用外标法求得比例系数 $S_k$ , 将 $[\bar{C}]$ 换算成实际浓度 $[C]$ .

目标识别因子分析解析完全未知混合体系的操作框图如下:



为验证方法的可行性, 对邻、间、对氯甲苯三组份体系进行了计算机模拟实验. 以操作框图步骤进行运算. 以 50 个单一向量 $(b_1, b_2, \dots, b_{50})$ 为初始向量, 迭代至收敛, 得到 50 个收敛的向量 $(\bar{b}_1, \bar{b}_2, \dots, \bar{b}_{50})$ . 计算任意 $\bar{b}_\alpha$ 和 $\bar{b}_\beta$ 之间的距离 $y_{\alpha\beta}$ , 得最小支撑树如图 1 所示. 图中圆代表 $\bar{b}$ , 如 $\bar{b}_3$ , 数字 3 表示该 $\bar{b}$ 是由 $(0,0,1,0,\dots,0)^T$ 这个单一向量迭代产生的. 图中三个高密度的“树梢”代表三个因子. 从每类中选择处于中心位置、与相邻点距离最小的点, 可得三个标识向量 (50 维): 19, 35, 5, 分别作为三个组份的标识向量, 与之对应的三个列向量 $t$ 构成变换矩阵 $[T]$ , 将主成份分析过程中获得的抽象解转化为比例解.

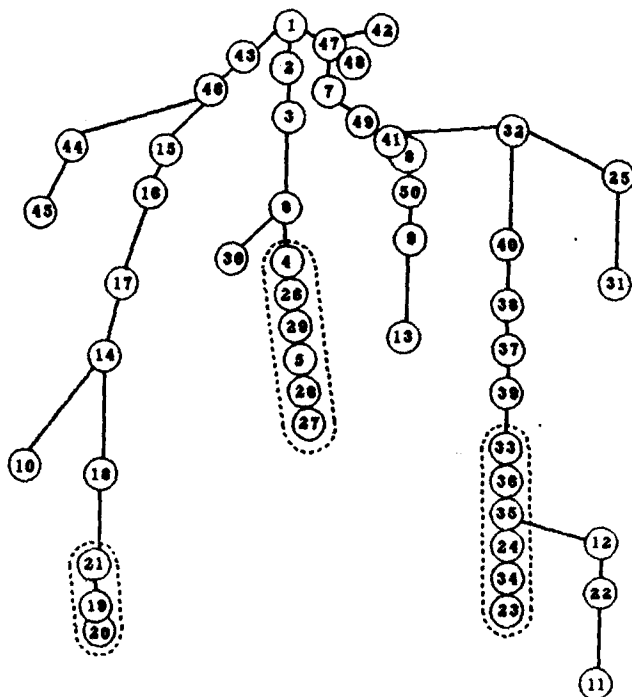


图 1 邻、间、对氯甲苯三组份模拟体系的最小支撑树

图 2 给出了该模拟体系纯谱解析与标准光谱的比较, 两者相差比例系数 $S_k$ . 将求得的纯谱数据输入至 620 红外光谱数据库工作站, 进行光谱检索. 检索结果按相似程度给出 5 种可

能物质的编号、名称. 图2中各光谱检索结果如下:

编 号	名 称
图 2a	619 邻氯甲苯
5128 间乙基苯甲酸甲酯	
556 9,10-二甲基蒽	
3564 2-苄基咪唑啉	
1576 邻氯苯酚	
图 2b	643 间氯甲苯
8512 间甲苯硫酚	
3140 3-碘-苯胺	
2240 1,4-萘醌	
8516 1,3-苯二硫酚	
图 2c	667 对氯甲苯
794 4,4'-对二碘联苯	
8546 对甲苯硫醚	
6785 4-溴盐酸吡啶	
481 1,2,3,4-四甲基苯	

由上述检索结果可知, 每张图检索得到的第一种可能物质即是所求的组份, 故利用该法求得的纯谱能够正确检索出被测物质的名称, 表明方法可行.

## 2 实验

### 2.1 仪器

Nicolet 170SX FTIR 光谱仪, 620 工作站, 长城 386 微机. 运算程序用 Pascal 语言编写. 检索用红外光谱库为 ALDRICH 凝聚态光谱库, 容量为 10607 张光谱.

### 2.2 试剂

邻、间、对氯甲苯(分析纯)为北京化工厂产品. 胆甾型液晶: 胆甾烯氯、苯甲酸胆固醇酯、壬酸胆固醇酯均为分析纯, 天津试剂二厂产品.

### 2.3 实验方法

2.3.1 计算机模拟体系 以固定厚度 KBr 液池分别测得邻、间、对氯甲苯三种样品的红外光谱. 通过线性加合, 构成 6 个三组份混合样品. 在  $980\sim 600\text{cm}^{-1}$ , 以  $8\text{cm}^{-1}$  的间隔读取数据, 形成光谱数据矩阵.

2.3.2 实际混合体系 用经研磨烘干的 KBr 固体配制浓度为 2% 的三种胆甾型液晶储备样. 按不同比例混合成 6 个实验样品. 用 KBr 压片法测定各样品的红外光谱. 在  $1520\sim 1140\text{cm}^{-1}$ , 以  $8\text{cm}^{-1}$  的间隔读数据, 形成光谱数据矩阵. 红外光谱数据均经基线法校正.

## 3 结果与讨论

胆甾型液晶的三种组份(胆甾烯氯、苯甲酸胆固醇酯、壬酸胆固醇酯)的红外光谱重叠严重. 本文选择特征性较强的指纹区用于光谱计算. 首先对 6 个实际样品的光谱数据阵进行

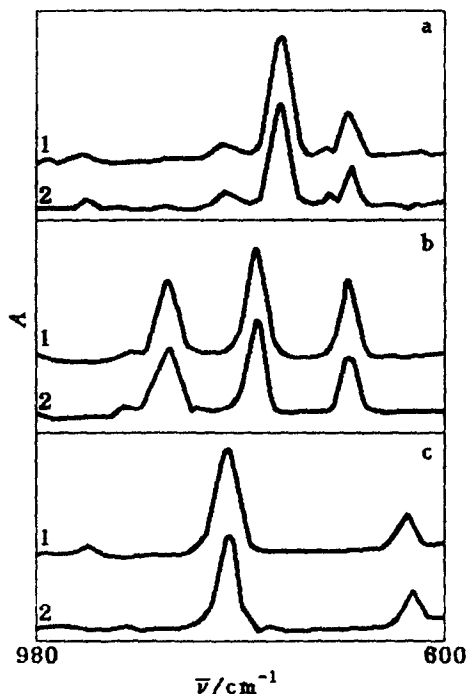


图2 邻、间、对氯甲苯三组份模拟体系的纯物种光谱分辨结果

a. 邻氯甲苯; b. 间氯甲苯; c. 对氯甲苯

1—标准光谱; 2—计算纯谱

主成分分析, 继以 50 个单一向量迭代至收敛, 计算任意两收敛向量间的距离, 得最小支撑树如图 3 所示。

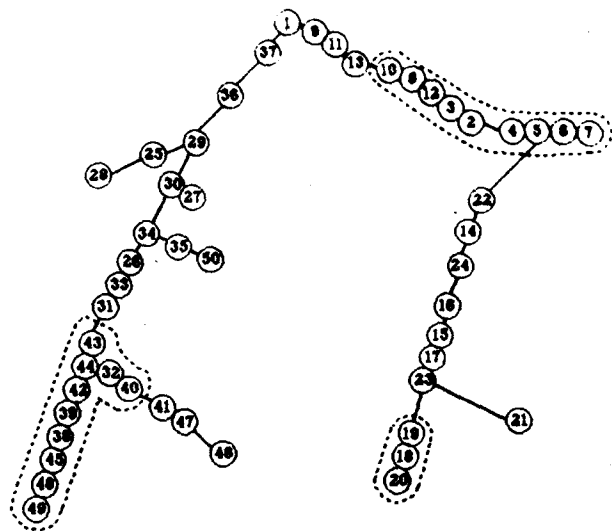


图 3 胆甾型液晶三组份混合体系的最小支撑树  
构编号为 6611, 名称是胆甾烯氯, 其它 4 种结构从略; 图 4b, 按相似程度也给出 5 种可能的结构, 第 1 种结构编号为 6665, 名称是苯甲酸胆固醇酯, 其它 4 种结构略; 图 4c, 相应给出最相似的结构编号为 6660, 名称是壬酸胆固醇酯。检索结果表明, 该法求得的纯谱正确反映了各组份的光谱特征, 可作为定性判别的依据。检索结果与实际相吻合。

表 1 胆甾型液晶混合体系浓度解析结果

样品 编号	$c_1$ (%)			$c_2$ (%)			$c_3$ (%)		
	计算 值	实 际 值	相对 误差 %	计算 值	实 际 值	相对 误差 %	计算 值	实 际 值	相对 误差 %
1	0.113	0.115	-1.74	0.302	0.304	-0.66	0.935	0.970	-3.61
2	0.200	0.187	6.95	0.414	0.423	-2.13	1.012	1.140	-11.2
3	0.287	0.280	2.50	0.491	0.513	-4.29	0.209	0.190	10.0
4	0.385	0.381	1.05	0.577	0.608	-5.10	0.406	0.387	4.91
5	0.478	0.498	-4.02	0.130	0.115	13.0	0.536	0.543	-1.29
6	0.537	0.556	-3.42	0.228	0.220	3.64	0.555	0.521	6.52

经光谱检索判定各物种后, 可采用外标法作定量分析。具体步骤是以纯物种光谱作为标准光谱, 按式 (13) 求得比例系数  $S_k$ , 再用式 (14) 将浓度比例解转化为实际浓度。胆甾型液晶三组份实际混合体系的浓度计算, 结果如表 1 所示。表中  $c_1$ 、 $c_2$ 、 $c_3$  分别代表苯甲酸胆固醇酯、壬酸胆固醇酯和胆甾烯氯的重量百分比浓度。

按照选择标识向量的原则, 图 3 中各点被低密度区和长边分割为三类, 选择高密度的“树梢”, 从中选取三个标识向量 (50 维): 39、20 和 10, 分别作为三种液晶组份的标识向量。与之对应的三个列向量  $t$  构成转换矩阵  $[T]$ , 将抽象解转化为比例解。

目标识别因子分析法解析胆甾型液晶混合体系所得的纯物种光谱与标准光谱的比较如图 4 所示。将所得的纯谱输入 620 工作站, 进行光谱检索。检索结果如下:

图 4a, 按相似程度, 给出 5 种可能的结构, 相似程度最接近的第 1 种结

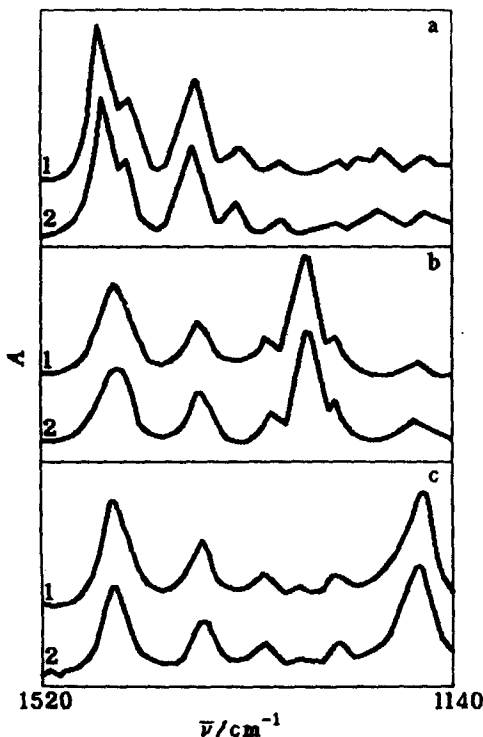


图 4 胆甾型液晶三组份混合体系的纯物种光谱分辨结果

a. 胆甾烯氯; b. 苯甲酸胆固醇酯; c. 壬酸胆固醇酯  
1—标准光谱; 2—计算纯谱

## 4 结论

本文应用目标识别因子分析成功地解析了计算机模拟体系及三组份实际混合体系, 从混合物光谱中提取了纯谱, 并经光谱检索和外标法, 达到了定性定量分析的目的. 这对完全未知混合体系的解析具有实际意义. 由于该法对光谱形状未加任何限制, 且不需确定可行解域, 故它的解析功能有很强的能力.

## 参考文献

- 1 W.H. Lawton, E.A. Sylwester, *Technometrics*, **1971**, *13*, 617.
- 2 P.C. Gillette, J.B. Lando, J.L. Koenig, *Anal. Chem.*, **1983**, *55*, 630.
- 3 H.B. Friedrich, Yu, Jung-Pin., *Appl. Spectrosc.*, **1987**, *41*(2), 227.
- 4 胡鑫尧, 汪国柄, 谭泽光, 朱蓉芳, 宋烈侠, 科学通报, **1985**, *30*, 398.
- 5 李 科, 陶 亢, 王宗明, 光谱学与光谱分析, **1986**, *6*(2), 48.
- 6 梁逸曾, 谢玉珑, 俞汝勤, 化学学报, **1991**, *49*, 394.
- 7 戴树桂, 曾幼生, 环境科学学报, **1986**, *6*(2), 1.
- 8 E.R. Malinowski, D.G. Howery, "Factor Analysis in Chemistry", John Wiley & Sons, New York, **1980**, p. 38.

## Studies on Target Recognition Factor Analysis Resolving Unknown Multicomponent Mixture

HE Xi-Wen\* CHEN Ding

(Department of Chemistry, Nankai University, Tianjin, 300071)

WANG Yong-Tai

(Central Laboratory, Nankai University, Tianjin, 300071)

**Abstract** In this paper, target recognition factor analysis (TRFA) is applied to estimating the component spectra from unknown multicomponent mixture system. With the use of pattern recognition techniques, TRFA derived target vector from the original data. A set of procedures which include extracting pure component spectra, spectra searching and standardizing are developed. This method made it possible to complete qualitative and quantitative analysis of an unknown mixture system in the meantime. The computer simulation technique is first used to verify the algorithm, and then the practical data collected from FTIR are also analyzed by this method. Satisfactory results are obtained.