化学模式识别——岭回归分光光度测定混合酚*

李惕川 黄 敏 李新云 张宝泉 张 昱
（北京工业大学化学与环境工程学系，100022）
程维虎
（北京工业大学应用数学系，100022）

摘 要 将苯酚、间甲酚、对氯酚、邻氯酚和间氯酚等5种挥发酚与MBTH反应显色后，用分光光度法测定，能获极难分辨的混合物的吸收光谱．本文采用聚类分析对波长-吸光度数据进行特征删选后，再以因子分析-岭回归法解析光谱，成功的确定了混合体系中的物种数、种类和含量．将该法用于印染废水中5种酚的同时定性与定量测定，结果令人满意．

关键词 聚类分析，岭回归，分光光度法，工业废水，酚
分类号 X832，O635

0 引言

由于不同酚的环境行为、生态效应、毒性及致毒机制有较大差别，混合酚类的同时分别测定代替总量测定，有利于推进实施有毒化学物的分级、风险评价、管理与控制，具有较大的环境化学和生态学意义．计算光度法同时测定多种酚已有报道[1~3]，该法快速、准确、简便的特点受到分析化学工作者的重视．实践证明，对这类化学与光谱性质十分相近，且吸光度加和性又欠佳的病态分析体系，用最小二乘回归(OLS)、卡尔曼滤波(KF)等无偏校正法，难以获得准确结果，而用有偏校正法较为适宜[1]．有人成功地利用偏最小二乘(PLS)光度法同时测定了河水中的4种酚[2]，也有用岭回归分析(RRA)处理人工模拟样品中的3、4和5种混合酚体系法，也取得满意结果[3]．本文用聚类分析——岭回归方法对更为复杂的工业废水中5种挥发酚进行了同时定性、定量测定，并以印染废水为例介绍了方法的基本原理及应用．

1 基本原理

为克服各组分光谱严重重叠等干扰引起的数据共线性关系，以保证定性、定量结果的

收稿日期：1996—10—20
* 北京市自然科学基金资助项目

准确性，拟用2种补救方法：①用聚类分析优化波长集合，保证准确定性；②用岭回归有偏估计实现精确定量.

1.1 系统聚类法

样本、测量值构成原始数据矩阵 X_{ij}，按下式规范化，以消除变量变化幅度的影响. 将 n 个测量值用距离和相关性为度量相似性尺度来进行分类.

$$Y_{ij}=\frac{X_{ij}-\min X_{ij}}{\max X_{ij}-\min X_{ij}} \tag{1}$$

设任意值 k 和 l 其间距离可由勾股定理求得，即

$$D_{kl}^2=(0k)^2+(0l)^2=(X_{k1}-X_{l1})^2+(X_{k2}-X_{l2})^2 \tag{2}$$

推广至 n 维的简写式:

$$D_{kl}=\sqrt{\sum_{j=1}^{n}(X_{kj}-X_{lj})^2} \tag{3}$$

用协方差矩阵考查相关性，矩阵元素由下式给出:

$$C_{kl}=1/(m-1)\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)(x_{il}-\bar{x}_l) \tag{4}$$

由(3)、(4)式计算并组成相似矩阵 D_0，寻找 D_0 的非对角最小元素 d_{pq}，$j=p$ 和 $j=q$ 最相似则合并为新类 p' 组，删去 p，q 列及行，以 p' 代 p 形成的新矩阵记为 D_1，同样步骤得 D_2，D_3……，直至所有对象合并为一类为止. 可将上述聚类动态过程绘制成直观的谱系图，方便确定要选择的类属.

1.2 岭回归估计

最小二乘估计的正规方程为:

$$\hat{C}=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{5}$$

式中 $\hat{C}$ 为组分浓度估计矢量，X 为 $n\times m$ 阶校正矩阵，Y 为 m 个组分在 n 个波长处测得的吸光度组成的矢量. 估计值的准确度由均方差给出:

$$\mathrm{MSE}(\hat{C})=E\{\|\hat{C}-C\|^2\}=\sigma^2\sum_{i=1}^{m}(L/\lambda_i) \tag{6}$$

σ^2 为测量误差的方差，λ_i 为 $S=X^{\mathrm{T}}X$ 的第 i 个特征值. 当组分谱重叠十分严重，即 S 接近退化(X 的列矢量线性相关)时，$\min\lambda_i$ 很小，$\|\hat{C}-C\|^2$ 很大，则最小二乘很难准确估计出 $\hat{C}$. 此时岭回归估计可弥补这一缺憾，基本方程如下:

$$\hat{C}(k)=(X^{\mathrm{T}}X+kL)^{-1}X^{\mathrm{T}}Y \tag{7}$$

式中 k 为岭常数，L 为 $m\times m$ 阶单位矩阵. 用 $S+kL$ 代替 S 使最小特征值 $\min\lambda_i$ 变为 $\min\lambda_i+k$，可望找到某个 $k>0$，使 $\hat{C}(k)$ 的均方差较 $\hat{C}$ 的要小:

$$\mathrm{MSE}[\hat{C}(k)]=\sigma^2\sum_{i=1}^{m}L/(k+\lambda_i) \tag{8}$$

确定岭常数的方法很多，本文选择计算量较小的岭迹图法，由下式算出岭迹[4].

$$\hat{C}_{j-1}(k) = -\left[\sum_{i=1}^{m+1} a_{i1}a_{ij}/(\lambda_i + k)\right] / \sum_{i=1}^{m+1}\left[a_{i1}^2/(\lambda_i + k)\right] \quad (9)$$

$$(j = 2, \cdots, m+1)$$

式中 a_{il}, a_{ij} 是正交矩阵 A 中的元素，λ_i 为矩阵

$$U = \begin{vmatrix} Y^{\mathrm{T}}Y & Y^{\mathrm{T}}X \\ X^{\mathrm{T}}Y & X^{\mathrm{T}}X \end{vmatrix}$$

的特征值.

2 实验

2.1 试剂与仪器

所用苯酚、对氯酚、间甲酚、邻氯酚、间氯酚及硫酸铈铵、MBTH、氢氧化钠、硫酸铜、EDTA、硼酸等均为分析纯试剂；样品在 500 mL 全玻璃蒸馏器处理；测量用岛津 UV-256 分光光度计；数据在联想 486 微机上处理.

2.2 测定步骤

2.2.1 模拟样品

1) 在 25 mL 比色管中加入 2.5 mL 浓度为 10 mg/L 的苯酚、间甲酚、对氯酚、邻氯酚、间氯酚及 2.5 mL MBTH (0.5 g/L)，摇匀，5 min 后加入质量分数为 0.4% 硫酸铈铵 2.5 mL，5 min 后再加 5 mL 缓冲液 (8 g NaOH + 2 g EDTA + 8 g 硼酸 + 蒸馏水至 250 mL，再与等体积乙醇混合)，混匀静置 15 min 后，在 450～600 nm 范围，读取 150 个波长点吸光度值，计算各波长吸光系数作为目标检测矢量.

2) 用聚类分析对 150 个波长点进行分类、归并，经加合性检验后确定优化集合.

3) 对 5 种酚不同配比的 15 个混合溶液，按步骤(1)显色后，测定在优化波长处的吸光度值，得到吸光度数据矩阵 [A]，在 486 计算机上运算.

2.2.2 实际水样测定

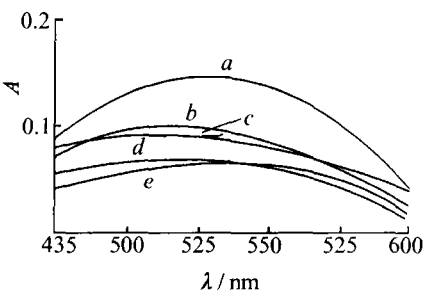
1) 由于印染、炼油、焦化等含酚废水样品的颜色较深、浑浊、含油及其他干扰物，测定前应进行简单前处理. 首先于分液漏斗中分离出浮油后，加粒状氢氧化钠使 pH 为 12.0～12.5，用四氯化碳萃取 2 次后，用磷酸调水样至 pH=4.0，然后用 500 mL 蒸馏瓶蒸馏，收集 250 mL 蒸馏液备用.

2) 在 25 mL 比色管中加入 2.5 mL 处理后水样，另一管中为加标后水样，均按 2.2.1 中步骤进行测定和计算.

3 结果与讨论

3.1 优化波长集合

5 种挥发酚的标准可见光谱如图 1 所示，但各吸收曲线重叠严重，无法直观选择分辨较好



a.苯酚 *b*.对氯酚 *c*.间甲酚 *d*.邻氯酚 *e*.间氯酚

图1 5种挥发酚的吸收光谱

波长点，可按1.1方法进行波长点选择．聚类后对各类波长点数据按符合 $|A-A_0/A_0|<e=0.05$（A为实测吸光度，A_0为理论吸光度，e为保留下限）者保留，获得不同类中的23个波长点构成优化波长集合（546，544，542，539，537，533，530，524，521，515，513，510，508，506，503，502，500，498，497，495，493，491，490 nm）．

3.2　模拟样品测定结果

3.2.1 因子分析定性

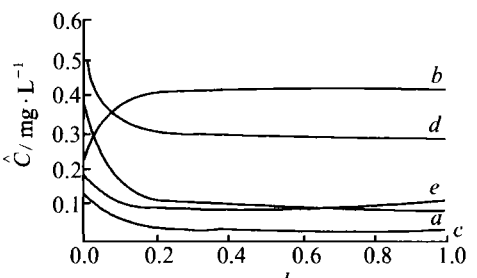
5种酚不同配比的15个模拟样品的抽象因子分析计算结果见表1，可见当 $n=5$ 时，RE（真实误差）和REV（约化特征值）的数值基本稳定，ER（特征值比）在 $n=5$ 时出现折点，*IND*（指示函数）值在 $n=4\sim6$ 之间出现最小值，综合考虑上述判据，可推断模拟样品中有5种组分，与事实相符．

表1　优化波长集合的因子分析结果

特征值 $\lambda\times10^3$	RE$\times10^3$	IE$\times10^3$	IND$\times10^3$	ER	REV$\times10^3$
411 801	38.204	95.509	1.698	19.282	1 119.02
21 356.9	6.187	21.875	0.316	100.856	64.718
211.76	4.994	21.623	0.295	2.076	0.720
102.00	4.303	21.514	0.299	1.965	0.592
89.32	3.934	20.992	0.257	1.420	0.228
36.56	3.656	22.391	0.366	1.135	0.185
32.21	3.358	22.210	0.415	1.179	0.189
27.32	3.045	21.529	0.476	1.213	0.190
22.52	2.716	20.371	0.554	1.168	0.188
19.28	2.322	18.360	0.645	1.662	0.197
11.60	2.038	16.895	0.815	1.221	0.149
9.50	1.677	14.530	1.049	2.794	0.158
3.40	1.619	14.590	1.798	1.069	0.077
3.18	1.530	14.309	3.824	1.325	0.106
0.23					

3.2.2 岭回归估计定量结果

由(9)式计算岭迹，获得岭迹图（图2）．可见，当 k 从0开始变大时，各组分质量浓度估计值 $\hat{C}$ 变化十分明显，这是由于校正矩阵中存在共线性关系对最小二乘解产生显著影响所致，随 k 的增大 $\hat{C}$ 趋平稳，选用适当 k 就能得到较稳定可靠的解．取岭常数 $k=0.1$，计算15个模拟水样的组分质量浓度的结果见表2．



a.苯酚 *b*.对氯酚 *c*.间甲酚 *d*.邻氯酚 *e*.间氯酚

图2　5种组分酚体系的岭迹图

表2 模拟水样岭回归计算质量浓度结果 mg/L

样品	苯酚		对氯酚		间甲酚		间氯酚		邻氯酚	
	实验值	计算值	实验值	计算值	实验值	计算值	实验值	计算值	实验值	计算值
1	0.240	0.236	0.270	0.255	0.180	0.176	0.300	0.282	0.200	0.212
2	0.050	0.047	0.640	0.627	0.150	0.137	0.480	0.479	0.240	0.235
3	0.760	0.747	0.080	0.075	0.100	0.110	0.260	0.265	0.020	0.021
4	0.200	0.225	0.400	0.413	0.050	0.049	0.300	0.291	0.500	0.490
5	0.100	0.107	0.240	0.234	0.065	0.068	0.720	0.689	0.300	0.319
6	0.360	0.366	0.200	0.101	0.540	0.529	0.150	0.140	0.028	0.029
7	0.100	0.094	0.260	0.249	0.200	0.213	0.240	0.248	0.700	0.677
8	0.270	0.254	0.260	0.247	0.100	0.093	0.360	0.374	0.250	0.232
9	0.320	0.323	0.180	0.170	0.420	0.410	0.230	0.228	0.140	0.136
10	0.560	0.548	0.250	0.227	0.100	0.096	0.100	0.097	0.270	0.257
11	0.020	0.027	0.520	0.513	0.180	0.172	0.280	0.265	0.536	0.523
12	0.240	0.234	0.720	0.716	0.606	0.564	0.200	0.206	0.048	0.050
13	0.360	0.343	0.230	0.235	0.350	0.333	0.210	0.197	0.140	0.131
14	0.080	0.073	0.320	0.325	0.100	0.108	0.260	0.264	0.720	0.737
15	0.580	0.566	0.240	0.243	0.150	0.143	0.060	0.054	0.246	0.233
平均相对偏差%	2.0		2.2		0.9		2.0		1.2	
总平均相对偏差%	1.7									

从表2可见，15个模拟样品的总平均相对偏差为1.7%，说明此法对测定上述混合酚体系是可行的.

3.3 实际样品测定结果

经预处理后的北京印染二厂总排放口水样及6个加标水样，按与模拟样品相同的测定方法得到的实验数据进行因子分析，结果见表3.

表3 实际水样的因子分析结果

样品	特征值λ	RE/10^3 真实误差	RE/10^2 嵌入误差	IND/10^3 指示函数	ER 特征值比	REV/10^3 约化特征值
1	153.596	8.210	3.103	2.281	471.033	954.012
2	0.326	3.959	2.116	1.584	7.842	2.470
3	0.041 6	3.033	1.985	1.895	2.370	0.396
4	0.017 5	2.532	1.914	2.814	2.036	0.219
5	0.008 6	2.305	1.948	5.762	1.369	0.061 1
6	0.006 3	2.081	1.946	2.080	1.454	0.017 5
7	0.004 3					

从表3可以看出，当$n = 5$时，IE和REV值基本趋于稳定，IND出现极值，而ER有一个折点，综合判定废水样中酚的组分数为5．对5组分在各波长处的吸光系数构成该组分的目标，进行目标检验，所得结果列入表4．

表4　目标检验结果

目标参数	AET 误差的均方根	RET 真实误差	REP 目标向量均方根	SPOIL 损坏函数
苯　酚	425.21	105.950	411.80	0.257
对氯酚	494.86	118.924	480.36	0.247
间甲酚	229.50	127.340	190.93	0.667
间氯酚	359.43	132.990	333.92	0.389
邻氯酚	521.41	129.552	505.06	0.256

由表4可见，苯酚、对氯酚、间甲酚、间氯酚和邻氯酚等5种组分的目标的SPOIL函数值都小于3.0，可判断5个目标都是真实目标，即该印染废水中同时含有上述5种酚．用岭回归估计法进行数据处理，得到实际水样各组分的质量浓度值，结果列于表5．

表5　实际水样中各组分质量浓度(ρ)计算结果

组　分	$\rho_{标}$	$\rho_{计}$	$\rho_{测}$	R/%
苯　酚	0	0.23	1.44	104
	0.5	0.74		
对氯酚	0	0.19	1.19	105
	0.2	0.40		
间甲酚	0	0.42	2.63	94.4
	0.18	0.59		
邻氯酚	0	0.10	0.63	97.8
	0.9	0.98		
间氯酚	0	0.09	0.56	103
	0.34	0.44		

由表中结果可以看出，5种酚的加标回收率R在94.4%～105%之间，岭回归估计法处理印染为废水混合酚的定量结果令人满意．

参　考　文　献

1　方国桢，郭忠先．6种化学计量学法同时分光光度测定五组分的比较研究及食品分析．分析化学，1994，22(3)：265～271

2　范华均，张薇，晏蓉，等．偏最小二乘光度法同时测定多种酚的研究及应用．高等学校化学学报，1994，15(9)：1305

3　梁逸曾，卢志强，俞汝勤，等．多元光度测定病态系统的岭回归估计．高等学校化学学报，1989，10(7)：704～708

4　陈希孺，王松桂，编著．近代实用回归分析．南宁：广西人民出版社，1984，155

Ridge Regression Spectrophotometry to the Determination of Multicomponent Phenols in Waste Water

Li Tichuan Huang Min Li Xinyun Zhang, Baoquan Zhang Yu
(Department of Chemistry and Environmental Engineering,
Beijing Polytechnic University, 100022)
Cheng Weihu
(Department of Applied Mathematics, Beijing Polytechnic University, 100022)

Abstract A method of spectrophotometry of combining ridge regression with cluste-factor analysis in simultaneous determination of multicomponents was proposed. The basic principle and computing of this method were discussed. This method was applied to simultaneous qualitative and quantitative of five volatile phenols in waste water. Satisfactory results were obtained.

Keywords cluster analysis, ridge regression, spectrophotometry, waste water, phenols